Resources

Eucalypt MAS: Implementation of marker-assisted selection in Australia's major plantation eucalypts

Project number: PNC378-1516 February 2018

Level 11, 10-16 Queen Street
Melbourne VIC 3000, Australia
T +61 (0)3 9927 3200 E info@fwpa.com.au
W www.fwpa.com.au

Forest & Wood Products Australia

Eucalypt MAS: Implementation of marker-assisted selection in Australia's major plantation eucalypts

Prepared for

Forest & Wood Products Australia

By

Saravanan Thavamanikumar, Simon Southerton, Robert Southerton, Jeremy Brawner and Bala Thumma

Publication: Eucalypt MAS: Implementation of marker-assisted selection in Australia's major plantation eucalypts

Project No: PNC378-1516

IMPORTANT NOTICE

This work is supported by funding provided to FWPA by the Australian Government Department of Agriculture, Fisheries and Forestry (DAFF).

ISBN: 978-1-925213-79-9

Researcher/s:

Saravanan Thavamanikumar, Simon Southerton, Robert Southerton, Jeremy Brawner and Bala Thumma

Gondwana Genomics Pty Ltd
1 Wilf Crane Crescent
YARRALUMLA ACT 2600

Forest & Wood Products Australia Limited
Level 11, 10-16 Queen St, Melbourne, Victoria, 3000
T +61 3 9614 7544 F +61 3 9614 6822
E info@fwpa.com.au
W www.fwpa.com.au

Kelsey Joyce

Jim Wilson

Andrew Jacobs

Forico Pty Ltd
De Boer Drive
RIDGLEY TAS 7321

Ben Bradshaw

Australian Bluegum Plantations
3/191 Chester Pass Road
ALBANY WA 6330

Christopher Barnes

RMS Timberlands Australia Pty Ltd
22 Cameron Street
LAUNCESTON TAS 7250

Dean Williams

Sustainable Timber Tasmania
Level 1, 99 Bathurst St
HOBART TAS 7000

Stephen Elms

HVP Plantations
50 Northways Road
CHURCHILL VIC 3842

Final report received by FWPA in October 2017

Executive Summary

Using largely *in-house* methodology, large numbers of markers associated with density were identified in both *E. nitens and E. globulus.* These along with pulp yield and growth markers were genotyped across a large number of trees in existing trials. Several validation tests were performed in both species to demonstrate the predicting power of markers using data from field trials owned by partnering companies. The accuracy of the marker predicted traits (molecular breeding values – MBVs) was benchmarked against estimated breeding values (EBVs) which are routinely used for selecting superior trees in traditional tree breeding. High accuracies were observed for MBVs of different traits. In the second phase of the project, more than 11,000 seedlings from partnering companies were screened with the markers. MBVs of these seedlings are currently being used by companies to select superior seedlings and to set up genetic gains trials. This represents one of the first ever applications of marker-assisted selection (MAS) in tree breeding.

High accuracies from the validation tests indicate that the markers discovered in this project can be used for selecting superior lines while they are still seedlings. This can reduce the normal breeding cycle which takes about 10-15 years to as little as 5 to 7 years. In addition to accelerating the breeding cycle, markers can be used for screening thousands of seedlings, which will result in larger genetic gains. Application of MAS in seedlings results in 2 to 3 times more genetic gain per annum compared to traditional breeding. Results from financial modeling indicate a return on investment (ROI) in the order of $8 for every $1 invested. While markers can be used for correcting pedigree errors and pedigree reconstruction improving the breeding value estimation, the main application of markers is to screen the seedlings derived from selected parents. In addition, markers can be used for selecting superior, diverse parents for crossing. Companies can capture significant gains if markers are incorporated into routine breeding and selection programs.

Table of Contents

Introduction

Australian plantation forestry is dominated by the temperate eucalypts blue gum (*E. globulus*) and shining gum (*E. nitens*) which account for about 80% of hardwood plantations. Blue gum is the most widely planted hardwood species with a total of about 400,000 ha in plantations. Most of these plantations are being harvested for export wood chips. Since the early 1990s, the Southern Tree Breeding Association (STBA) has conducted two cycles of selection, which achieved modest genetic gains (Dutkowski et al., 2000). The main breeding objectives for blue gum were increased volume growth, wood density, and pulp yield. Shining gum is the second most widely planted (~150,000 ha) eucalypt species in Australia. The first large-scale *E. nitens* progeny trials were established in the 1970s and up to two cycles of breeding have been completed (Hamilton et al., 2008). Forico and Sustainable Timber Tasmania (ST Tas) have the most active *E. nitens* breeding programs. Significant areas of *E. nitens* plantations are also harvested for veneer, appearance and structural products but most plantations are harvested for export wood chips.

Genetics of eucalypt wood traits

The goal of forest tree breeding programs is to increase the economic value of products derived from plantation forests. Genetic studies have shown that eucalypt wood quality traits have moderate-to-high heritability compared to growth traits (Costa e Silva et al., 2009). High heritability (0.76) was estimated for NIR predicted cellulose (Schimleck et al., 2005), while moderate heritability (0.50) was estimated for density (Raymond, 2002). This suggests that genes control a sizeable proportion of phenotypic variation in eucalypts.

Until now, selections made by tree breeders have always been based on phenotypic measurements on large numbers of trees in numerous field trials. Traditional tree breeding cycles are long (typically 10-14 years) compared to other crops, as many traits are not expressed until trees are over 7 years old. In view of these difficulties, there has been considerable interest in the development of molecular markers for identifying superior genotypes while they are seedlings. The potential of marker-assisted selection (MAS) to accelerate tree breeding has been appreciated by scientists for decades but has not reached industrial application due to a lack of reliable markers and a framework for the application of markers in practical tree breeding populations.

Dissection of the molecular basis of trait variation in forest trees began in the 1990s with the introduction of quantitative trait locus (QTL) mapping in controlled-cross pedigrees (Neale, 2007). Using base populations derived from the Australian Tree Seed Centre collections, CSIRO assisted with breeding strategies and developed genetic linkage maps and identified QTLs for growth and wood traits (Thamarus et al., 2004; Thumma et al., 2010). Marker loci controlling variation in growth and wood properties have been reported for several tree species (Butcher & Southerton, 2007) including some in *Eucalyptus* (Byrne et al., 1997; Freeman et al., 2013; Thamarus et al., 2004; Thumma et al., 2010; Verhaegen et al., 1997). This work has shown that variation in wood quality and other traits in trees is polygenic. Because wood traits are under polygenetic control (quantitative), genetic improvement will rely on the selection of multiple alleles, each of relatively small individual effect. While QTL studies improved our understanding of the genetic control of complex traits, none of the markers identified to date using this approach have proven to be useful in operational breeding programs. The major disadvantage of markers identified using pedigrees is that they

are generally not transferable from the pedigree of identification to other pedigrees in the same population (Grattapaglia et al., 2012). In addition, special populations such as full-sib families are required for detecting markers. This greatly limits their use in tree breeding programs that are based on large numbers of families, as is the case in Australian breeding programs.

Marker discovery using association genetics

First applied in humans, association studies are a powerful method for discovering markers in outcrossing populations like forest trees, which contain many unrelated families. In this method, large numbers of unrelated individuals are accurately measured for various traits and genotyped at large numbers of markers. Due to unique features of the eucalypt genome and eucalypt populations, markers discovered using association genetics occur within genes that directly influence the trait in question and frequently the marker causes the variation (Thumma et al., 2009). These markers are ideal for MAS in eucalypt breeding programs because they target the better genes that cause the improved trait. Since the markers are discovered using different populations in different environments, they predict well in unrelated populations and in different environments.

The unique features of forest trees, including eucalypts that make them ideal for association genetics include outcrossing breeding systems, a long history or recombination and a relatively short history of domestication. As a result, breeding populations of most forest trees closely resemble the wild state (Butcher & Southerton, 2007) and contain vast stores of genetic variation for tree breeding. An important consequence of these life history characteristics is the very low linkage disequilibrium (LD) observed in forest trees such as eucalypts (Southerton et al., 2010; Thavamanikumar et al., 2011; Thumma et al., 2009) and pines (Brown et al., 2004). LD is a measure of the tendency for adjacent markers (Single Nucleotide Polymorphisms or SNPs) within a gene to be correlated in unrelated individuals. Generally in eucalypts, LD breaks down within 500 to 1000bp or within the length of the gene (Southerton et al., 2010). In most crop plants, LD extends to much larger distances (Kraakman et al., 2004). The rapid breakdown in LD is a major impediment to the transfer of QTL markers between different pedigrees however, low LD makes trees ideally suited to candidate gene-based association studies (González-Martínez et al., 2006; Neale & Savolainen, 2004). This research approach seeks to find alleles that affect a phenotypic trait and remain linked to the trait across populations and over many generations. This methodology is also well suited to tree breeding programs that aim to maintain a broad genetic base (i.e., programs with a large number of families).

The first reported association study in forest tree species uncovered polymorphisms in the *cinnamoyl CoA reductase (CCR*) gene that were associated with microfibril angle (MFA) in *E. nitens* (Thumma et al., 2005). Thumma et al. (2009) also reported the discovery of an SNP in a COBRA-like gene (*Eni*COBL4) that is associated with pulp yield and cellulose content in *E. nitens*. This study was the first to verify that the associated SNP caused the trait variation and that it did this through its impact on allelic expression. In a recent study in *E. globulus*, nine markers affecting growth and wood quality traits were discovered in one population and subsequently validated in an independent population (Thavamanikumar et al., 2014). Similarly in pines, several SNP markers in different cell wall genes that are associated with wood quality traits were identified (Dillon et al., 2010; González-Martínez et al., 2007).

CSIRO scientists (now located in Gondwana Genomics) have developed novel research strategies for identifying genes and alleles that control complex traits in eucalypts using association genetics (Thumma et al., 2005, 2009 and Gondwana Genomics unpublished methods). These genomic techniques have been powerfully demonstrated for wood quality traits and growth in both *E. nitens* and *E. globulus* in methods described in previous FWPA supported research (Thumma et al., 2010, 2015). Between 60 and 70 SNP markers associated with pulp yield and about 200 markers associated with growth were identified in each eucalypt species. In addition, 96 SNPs were found to be associated with MFA in *E. nitens*.

Genomic selection

While the markers identified using association studies are useful across the population, individual marker effects are too small to be useful for making selections in breeding programs. Genomic selection (GS) is one approach in which marker effects of several to many markers are used together to make selections. Genomic selection is widely used by animal breeders and has been applied in crop breeding as well as forest tree populations (Resende et al., 2012). In this approach, large numbers (thousands) of markers randomly distributed across the genome are genotyped in advanced breeding populations derived from small numbers of parents (small effective populations). Marker prediction models are first developed using a training population, such as the parents of controlled pollination (CP) families or seed orchard parents, and used to link (model) marker genotype data with trait data. This model is then applied in a closely related test population, such as the progeny from the seed orchard parents, to predict their traits using only the marker genotype data of the progeny (test) population. Traits estimated with markers in test populations are known as genomic estimated breeding values (GEBVs) or molecular breeding values (MBVs). The accuracy of the MBVs is estimated by correlating the MBVs obtained with estimated breeding values (EBVs) obtained using traditional methods. The accuracy of GS is affected by the strength of the marker-trait associations and the genetic relationships captured by the markers. In the first study of GS in forest trees, (Resende et al., 2012) observed accuracies ranging from 0.63 to 0.75 for tree diameter and height in 800 loblolly pine clones replicated across four sites. Similarly, in the first eucalypt GS study in *E. grandis* and *E. urophylla* hybrids, accuracies ranging between 0.55 and 0.88 were observed for growth and wood traits (Resende et al., 2012). Since then several studies have been published in different tree species (Durán et al., 2017; Gamal El-Dien et al., 2015; Müller et al., 2017). In all of these studies, markers randomly distributed throughout the genome were used. The use of random markers in GS results in several problems including low prediction accuracies between different populations of the same species (Resende et al., 2012) and low accuracies in advanced generations. This is due to the breakdown of LD and recombination between marker-trait associations (Hayes et al., 2009). Moreover, a recent study has shown that predictions cannot be made across generations using large numbers of random markers (Tan et al., 2017). In view of these issues, we proposed to use associated markers rather than random markers for predicting traits in GS (Thavamanikumar et al., 2013). The main advantage of using associated markers in GS is that the markers occur in candidate genes that directly influence the trait. As a result, recombination will not affect accuracies in advanced generations. Also, since the markers are identified using populations from different environments, the accuracy of GS using associated markers is expected to be high across different populations and environments. Recently, several studies have shown that incorporating associated markers

with random markers lead to a high accuracy of predicting traits with markers (Boichard et al., 2016; Porto-Neto et al., 2015; Spindel et al., 2016; Thavamanikumar et al., 2015).

In-House marker discovery and genotyping method development

In order to build marker models based on large numbers of trait-associated markers we developed *in-house* methods for high-throughput marker discovery and genotyping. Traditional ways of marker discovery using association studies are laborious and time-consuming. To overcome these difficulties we have developed *in-house* high-throughput methods using next-generation sequencing to identify potential markers (candidate SNPs). This is done by comparing the frequency of SNPs in genomic DNA of trees from extremes of trait variation across several populations growing in different environments (unpublished data). This results in the discovery of several hundred associated markers with each trait of interest. These markers are then used for developing assays (marker panels) for genotyping with the new genotyping method developed by Gondwana Genomics (GG).

Our previous genotyping approach employed a Fluidigm machine and cost-limited us to genotyping 96 SNPs per tree. To overcome the limitation, we developed a novel proprietary method for genotyping that uses next-generation sequencing (unpublished data). The new method allows us to genotype thousands of SNPs per tree in a cost-effective manner. It can be used for genotyping large numbers of targeted (trait-enriched) markers across a large number of samples. This change to the way we genotype markers has significant advantages. The new method allows us to genotype many more SNPs per trait in a single test, which will improve prediction accuracies. It also allows us to capture family effects in our marker predictions. This can significantly increase the accuracy of our marker predictions, particularly in more advanced breeding populations with an effective population size less than 100. Finally, the new approach employs next-generation sequencing, which is increasingly being used for genotyping. As sequencing costs fall, we are confident that this will lower the cost of genotyping each tree. The new methods of identifying high throughput markers and high throughput genotyping were applied in the current project.

The main aims of the current project are 1) to identify large numbers of new markers controlling growth, KPY and the additional trait density in *E. nitens* and *E. globulus* and to incorporate them into the existing marker panels of growth and pulp yield, 2) to test and develop the new marker genotyping method developed by GG, 3) to test different models for applying these markers in GS, 4) to conduct validation tests in industry breeding populations to test the accuracy of marker predictions, and 5) to apply the markers, after validation, in operational screening of seedlings in industry breeding programs.

Materials and Methods

Trial sampling and phenotyping - *E. globulus*

In previous CSIRO research (Blue Gum Genomics, 2010-2014), four genetically diverse *E. globulus* trials had been measured for KPY and sampled for DNA. This included three trials of the Otways race located at West Ridgley, Latrobe Tas and Busselton WA and a trial of the Flinders Island race at Busselton. In this project, we expanded the number of trials to include 2 Flinders race trials, 2 Tasmania race trials and 2 Gippsland race trials; one of each race located at Latrobe Tas and Busselton WA. We also included a mixed race 2nd generation trial located at Marri Downs WA (see Table 1).

Density was measured at breast height in approximately 500 trees in each of the four Busselton trials using a pilodyn in November 2015. This was undertaken prior to the commencement of the project because the trials were about to be felled. Density had already been measured in the Latrobe trial using whole cores on between 350 and 500 trees per race. Cambial scrapes for DNA were collected from these trees in May 2016. KPY was estimated on the Busselton Gippsland and Tasmania trials and on the Latrobe Gippsland, Tasmania and Flinders trials by NIR analysis of swarf drilled from breast height.

Saravanan Thavamanikumar collecting *E. globulus* cambial scrapes for DNA at Latrobe Tas

Trial sampling and phenotyping - *E. nitens*

In the Blue Gum Genomics project five genetically diverse *E. nitens* trials (Central Victorian race) were measured for KPY and sampled for DNA. Two of these trials were growing on cold sites (Florentine and Tarraleah), two were growing on warmer sites (Meunna and Southport) and one was growing on an intermediate site (Loudwater). In the current project, we expanded the number of trials to three cold and three warm sites by sampling a trial at Blythe Road, Tas (cold site) and Hollow Tree, Tas (warm site).

Geoff Downes operating the resistograph on *E. nitens trees* growing at Blythe Rd

Wood swarf and cambial scrapes were collected from the Blythe Road and Hollow Tree trials in June 2016. The swarf samples were used for NIR analysis to predict KPY. Geoff Downes (Forest Quality) obtained resistograph (resi) data for each tree at the Blythe Road and Florentine trials from which density estimates were later obtained. Density is strongly correlated with the resistance to the probe penetrating the tree from bark to bark. We expect the resi data to give better estimates of density than the pilodyn, which only penetrates the outer wood. Whole core density had previously been estimated for the trees sampled at the Hollow Tree trial.

E. nitens growing at Blythe Rd near Hampshire Tas

Table 1. Genetic material used for marker discovery

Trial name	Owner	Race	Trees sampled	KPY	Density	Growth
E. globulus						
Busselton, WA	FPC	Otways	520	✔	✔*	✔
Busselton, WA	FPC	Flinders	550	✔	✔*	✔
Busselton, WA	FPC	Tasmania	460	✔	✔*	✔
Busselton, WA	FPC	Gippsland	520	✔	✔*	✔
Latrobe, TAS	Forico	Otways	470	✔	✔#	✔
Latrobe, TAS	Forico	Flinders	440	✔	✔#	✔
Latrobe, TAS	Forico	Tasmania	450	✔	✔#	✔
Latrobe, TAS	Forico	Gippsland	350	✔	✔#	✔
West Ridgley, TAS	Forico	Otways	470	✔		✔
Marri Downs, WA	ABP	Mixed	800	✔		✔
E. nitens						
Florentine, TAS	FT	Central Victorian	420	✔	✔♭	✔
Meunna, TAS	FT	Central Victorian	400	✔	✔§	✔
Loudwater, TAS	Forico	Central Victorian	500	✔	✔#	✔
Tarraleah, TAS	FT	Central Victorian	520	✔	✔§	✔
Southport, TAS	Ft	Central Victorian	520	✔	✔§	✔
Blyth Rd, TAS	Forico	Central Victorian	400	✔	✔♭	✔
Hollow Tree, TAS	Norske Skog	Central Victorian	375	✔	✔#	✔

* = pilodyn; # = whole core; ♭ = resistograph; § = silviscan. Blue indicates new trials sampled and new data obtained or accessed

Development of DNA capture libraries and preparation of high throughput sequencing libraries

DNA capture libraries were prepared using gene sequences of 2500 cell wall genes. Cell wall genes are expected to play a major role in the development of the three traits under investigation. Most of these genes were selected from our previous study comparing gene expression in two populations of *E. nitens* (Thavamanikumar et al., 2014). Gene expression was compared between high and low pulp yield and high and low growth trees in two *E. nitens* trials at Meunna and Florentine. Genes with consistent differences in expression between trait extremes at both sites were selected for developing the DNA capture libraries. In addition to these, other important genes such as transcription factors and genes involved in cell wall biosynthesis identified from literature searches were included in the development of DNA capture library. We re-sequenced the previously sequenced pulp yield and growth samples from the BGG project with the newly expanded DNA capture library as it contains a

comprehensive list of genes involved in growth and wood development. This will help us to discover a significant proportion of SNPs associated with growth and wood traits.

DNA was isolated from 48 trees at each extreme of trait variation in each trial using Qiagen high throughput DNA extraction kits. In total, DNA was isolated from 3,936 samples. Bulks were created by combining equimolar amounts of DNA from each tree into a pool. Forty-eight bulks were created in *E. globulus* and 36 bulks were created in *E. nitens* (Table 2). Pooled DNA samples were used for preparing whole genome DNA libraries by fragmenting DNA and adding adapter sequences for sequencing with Illumina next-generation sequencing platform. Whole genome DNA libraries were hybridised with the DNA capture libraries to enrich the whole genome DNA libraries for growth and wood quality candidate genes. The enriched DNA libraries were sequenced with the NextSeq module of Illumina sequencing.

Table 2. Number of samples selected for sequencing

	Trial	Trait	high	low
E. globulus				
	BO	DBH	48	48
	BF	DBH	48	48
	BT	DBH	48	48
	BG	DBH	48	48
	LO	DBH	48	48
	LF	DBH	48	48
	LT	DBH	48	48
	LG	DBH	48	48
	BO	KPY	48	48
	BF	KPY	48	48
	BT	KPY	48	48
	BG	KPY	48	48
	LO	KPY	48	48
	LF	KPY	48	48
	LT	KPY	48	48
	LG	KPY	48	48
	BO	density	48	48
	BF	density	48	48
	BT	density	48	48
	BG	density	48	48
	LO	density	48	48
	LF	density	48	48
	LT	density	48	48
	LG	density	48	48
	sub total		***1152***	***1152***
E. nitens				
	FL	DBH	48	48
	Meu	DBH	48	48
	Tarr	DBH	48	48
	SP	DBH	48	48
	BR	DBH	48	48

HT	DBH	48	48
FL	KPY	48	48
Meu	KPY	48	48
Tarr	KPY	48	48
SP	KPY	48	48
BR	KPY	48	48
HT	KPY	48	48
FL	density	48	48
Meu	density	48	48
Tarr	density	48	48
SP	density	48	48
BR	density	48	48
HT	density	48	48
sub total		***816***	***816***
Total		**1968**	**1968**

BO, Busselton Otways, BF, Busselton Flinders, BT, Busselton Tasmania, BG, Busselton Gippsland, LO, Latrobe Otways, LF, Latrobe Flinders, LT, Latrobe Tasmania, LG, Latrobe Gippsland, WO, West Ridgely Otways, FL, Florentine, Meu, Meunna, LW, Loud Water, Tarr, Tarraleah, SP, South Port, BR, Blythe Road, HT, Hollow Tree

High throughput sequencing and selection of associated SNPs

DNA libraries from high and low trait pools were sequenced with Illumina Next generation sequencing. Sequence reads from high throughput sequencing were mapped to the *E. grandis* reference genome. The sequence reads mapped to the candidate genes were analysed to identify single nucleotide polymorphisms (SNPs). Read counts at each SNP position were used to estimate allele frequencies. SNPs with large and consistent differences in allele frequencies between high and low trait pools across the populations were selected as candidate SNPs for genotyping.

Genotyping with the new method

In a recently completed *Teratosphaeria* leaf disease (TLD) project in *E. globulus* (Thumma et al., 2017)**,** GG developed a novel high throughput genotyping method to genotype large numbers of trait-enriched markers across large numbers of samples using next-generation sequencing technology. We used this method in the current project to genotype around 3000 trees of *E. nitens* and *E. globulus* in the first phase and more than 11,000 trees in the second phase of the project. Details of the trees used in the first phase of the study are shown in Table 3.

Table 3. Summary of the trees genotyped in the first phase

Trial	Number of samples genotyped	
	E. globulus	***E. nitens***
HVP parents	131	
HVP RES 1448	296	
HVP RES 1504	293	
Blythe Road		204
Forico Arboretum		523
Florentine		192
Hollow Tree		276
Meunna		192
Tarraleah		192
Huddlers		401
MiddleSex		271
Total	**720**	**2251**

Association analyses

Individual association studies were conducted in each population separately with single marker association analysis. Results from individual association studies were combined in a meta-analysis to identify markers that are stable across different populations. The software package 'plink' was used for these analyses.

Validation tests and commercial screening

Several validation tests were performed in the breeding populations of different companies. Training (parental) populations were used to develop prediction models which were then applied in test populations (progeny) to estimate MBVs. Different models such as 'RRBLUP', Bayesian models such as 'BayesB', 'BLR', 'BRR' and dimensionality reduction methods such as the 'partial least squares'(PLS) method were used. MBV prediction from these models was compared by correlating MBVs estimated from different models. Generally, high correlations were observed among all the models.

In addition to these models, we also used 'single step Bayesian Regression' (SSBR). SSBR gives similar results as 'single step genomic best linear unbiased prediction' (ssGBLUP) but is faster and it doesn't require inversing the dense relationship matrix which is computationally demanding (Fernando et al., 2014). In a 'single step' method, marker and trait data from genotyped samples, trait data from non-genotyped relatives, and the pedigree relationships are all used together to predict MBVs. Since all the available information is used in the models, accuracies from single step methods are expected to be higher. All the models were run using R software.

The number of samples included in operational screening is shown in Table 4. More than 11,000 samples were genotyped. For most of these leaf samples from seedlings were used for DNA isolation. However, cambium samples were used for DNA isolation for some of the RMS samples. One of the main features of our new genotyping method is that the quantity and quality of the DNA samples required are minimal. Just a single leaf was collected for all the leaf samples used for DNA isolation.

Table 4. Summary of the samples used in Phase 2 screening

Company	Number of samples
Forico	7,300
ABP	2,300
HVP	1,200
RMS	900
ST Tas	120

Results and Discussion

Marker discovery

Gene libraries containing cell wall and growth genes were used for discovering the candidate markers. Candidate markers i.e. markers potentially associated with different traits were validated using the high throughput methods developed by GG. Data from high throughput sequencing was used for comparing allele frequencies between different populations at different environments to identify robust markers that are stable across different populations. For each species, marker panels consisting of 1000 probes were developed. These marker panels are made up of 250 candidate SNPs for each of the two wood traits (KPY and density), and 500 candidate SNPs for DBH. The candidate SNPs were selected from the largest and consistent allelic differences between high and low trait pools across different populations. Marker panels developed from the candidate markers were genotyped using our new genotyping method.

Marker genotyping

Marker panels consisting of 1000 candidate SNPs were genotyped using the new method in *E. nitens* and *E. globulus*. After filtering the markers based on minor allele frequency and call rate, 2363 markers in *E. nitens* and 3200 markers in *E. globulus* were selected for association and GS analyses. Initially, to test the consistency or concordance of the genotype calls, we genotyped 20 duplicate *E. nitens* DNA samples from earlier in the project. We observed an average concordance rate of 97% (ranging from 90% to 99%) between the duplicate samples. We also compared the concordance of the new method with our previous chip-based genotyping system (Fluidigm). Three duplicate samples were observed in *E. globulus* parents previously genotyped with 96 SNPs using Fluidigm system. The same three samples were found to be duplicates with more than three thousand markers genotyped in the new method. In addition to these initial tests, we have done further tests with similar results. These results show that the genotype calls with the new method are accurate and comparable to those from the Fluidigm system. In the current project we have genotyped more than thirteen thousand trees of *E. nitens* and *E. globulus* using the new method.

Association analyses

Association analyses were conducted in *E. nitens* to validate the candidate markers. In total, five populations were used in association analyses. First, analyses were performed in each population separately to identify markers associated with each trait. Results from individual association analyses were then analysed in a meta-analysis to identify markers that are the most robust across different populations. Meta-analysis revealed a different number of

markers associated with different traits. KPY had the highest number of markers associated (13%), followed by growth (7%) and density (4%) at a significance level of $P<0.05$. The larger number of detected markers for KPY may derive from the inclusion of most of the cell wall genes used in the gene library from an RNA-Seq study comparing KPY extremes (Thavamanikumar et al., 2014).

With the previous genotyping system, we were restricted to selecting a subset of SNPs based on association studies results for use in GS analyses as only a small number of markers could be genotyped. However, with the new genotyping method the restriction on the number of markers is removed. All of the candidate SNPs from different traits were genotyped together for application in GS analyses. The rationale behind this approach is that while significant markers are useful for capturing marker-trait associations, the non-significant markers are useful for capturing genetic relationships both of which contribute to the accuracy of MBVs.

Genomic selection

The main aim of the current project is to develop markers and application methods that can be used for marker-assisted selection (MAS) in eucalypt breeding. GS is a marker-based approach that can be used for selecting superior genotypes based on marker data. In GS, marker effects from large numbers of markers are used together to predict traits based on marker genotype data alone. A prediction model is developed using marker and trait data of training populations (typically parents). This model is then applied in test populations (typically progeny) to predict traits based on their marker data alone. Traits predicted with markers are known as molecular breeding values (MBVs) or genomic estimated breeding values (GEBVs). The accuracy of marker predictions was tested by correlating MBVs with the estimated breeding values (EBVs). When raw data or measured trait is used for training the model, the correlation between MBVs and measured data gives the predictive ability of the markers. To obtain accuracy, predictive ability is divided by the square root of heritability. The accuracy of marker-based predictions (MBVs) is dependent on the strength of the marker-trait associations and the genetic relationships captured by the markers. When the training and test populations are related, genetic relationships captured by the markers contribute to the accuracy of MBVs. When the training and test populations are distantly related or unrelated, marker-trait associations captured by the markers contribute to the accuracy of the MBVs.

The main application of genomic selection is in screening the seedlings and selecting the top-ranking seedlings based on MBVs. The main advantage here is that within family selection can be made without the phenotype data. As markers capture the Mendelian segregation term i.e. markers can differentiate between the sibs within a family, within family selections can be made at seedling stage using markers. This is not possible with EBVs as all members of a family will have the same EBVs in the absence of measured phenotype data. With traditional methods used to calculate EBVs, within family selection is only possible with data obtained when progeny are 6 to 7 years old. As a consequence, within family variation is generally less exploited compared to between family variation in tree breeding. In animal breeding, which is more advanced compared to tree breeding most of the recent genetic gain is due to capturing Mendelian sampling term and exploiting within family variation (Avendaño et al., 2004).

Validation tests

Several validation tests were conducted to test the performance of markers in different populations of *E. nitens* and *E. globulus*. These tests were performed in breeding populations of different companies. All the validation tests are done 'blind', that is we did not have access to trait data of the test (progeny) samples and trait data was held by the partner companies participating in the project. MBVs of the test samples were estimated with marker and trait data of the training (parental) samples. These MBVs were sent to partner companies who estimated the accuracies by correlating MBVs with their EBVs.

While the main aim of the validation tests was to test the performance of the markers, these tests were also useful for demonstrating the potential gains possible with the markers. As it takes a long time to see the genetic gains from marker-based selections in operational tree breeding, these tests can be used to gauge the gains possible as data from existing field trials is used in the validation tests.

Validation tests in E. nitens

Validation tests in *E. nitens* were performed in existing field trials of Forico. To test the performance of the markers in *E. nitens* three validation tests were conducted.

Validation *test 1*

In the first validation test, MBVs of 523 seed orchard trees were estimated by using measured trait data from several unrelated populations. Three traits (DBH, wood density, and KPY) were used for estimating MBVs. The main aim of this test is to estimate the power of marker effects alone in predicting traits. Moderate accuracies ranging from 0.15 for tree volume to 0.37 for KPY were observed (Table 5). These results with unrelated training and test populations indicate that the marker-trait associations captured by the markers are contributing to the accuracies. Similar tests with random markers have yielded accuracies close to zero or negative (Müller et al., 2017; Resende et al., 2012).

Table 5. Accuracies of MBVs of seed orchard parents using unrelated populations for training

Traits	Accuracy
KPY	0.37
Density	0.20
Volume	0.15

Validation test 2

In the second validation test, MBVs of the progeny trees (testing population) were estimated using genotype and EBVs of the parent trees (training population). Two types of MBVs were estimated in this test. First, MBVs were estimated with a model trained in parents and second, MBVs were estimated with a model trained in parents as well as progeny without the genotype data using SSBR model. Three progeny trials were used in this test. Four hundred trees from the Hudler trial (a warm site), 271 trees from the Middlesex trial (a cold site) and 204 trees from the Blythe Road (a cold site) were used for estimating MBVs. Accuracies of MBVs were estimated by correlating MBVs with EBVs. As expected, accuracies from this test were significantly higher compared to the first validation test (Table 6). Generally, accuracies were higher at the Hudler site compared to the other two sites. The accuracy of the

MBVs with parents and non-genotyped progeny are generally higher than the MBVs using only the parents in model training. The accuracy of tree volume MBVs at Middlesex was higher with EBVs estimated using normal model compared to cold site-specific model even though Middlesex is a cold site. For the Blythe Road site, however, high accuracies were observed with cold-site specific EBVs as expected.

Table 6. Validation test 2 - Accuracies of MBVs of progeny trees estimated with 523 parental data

Hudler Trial	Accuracy[1]	Accuracy[2]
Kpy_tree_13_ebv	0.77	0.78
Bd_tree_13_ebv	0.54	0.52
Vol_stand_normal_13_ebv	0.60	0.70
Vol_stand_cold_13_ebv	0.58	0.65
Middlesex Trial		
Kpy_tree_13_ebv	0.48	0.58
Bd_tree_13_ebv	0.50	0.47
Vol_stand_normal_13_ebv	0.55	0.58
Vol_stand_cold_13_ebv	0.36	0.56
Blythe Road		
Kpy_tree_13_ebv	0.64	0.59
Bd_tree_13_ebv	0.56	0.61
Vol_stand_normal_13_ebv	0.41	0.43
Vol_stand_cold_13_ebv	0.49	0.48

[1] – Accuracy of MBVs estimated with only the parental data. [2] – Accuracy of MBVs estimated with parents and non-genotyped progeny data

Validation test 3

Two trials (Hudler and Middlesex) were used in the third validation test. Parental data and half of the progeny data from each family were used for model training. This model was then used to predict MBVs of the remaining half of the progeny trees from each family. Accuracies of MBVs were assessed with two training models. In one model, parental and half of the progeny data was used while in the other model only half of the progeny data was used. Accuracies of validation test 3 (Table 7) are higher than those of validation test 2 (Table 6) indicating that the close relationship between the training and testing populations increased the accuracies of the MBVs. Accuracies from two training models were similar in the Hudler trial while the Middlesex trial accuracies from the second model using only the progeny data were slightly higher. Higher accuracies in both trials indicate that using information from the target site data in model training improves the accuracies. The high accuracy of the forward estimation of MBVs from these tests should give confidence in the MBVs of the seedlings raised from the selected parents, which is the main application of makers. In seedling screening, training model is developed with parental data as well as the available progeny data from progeny-tested field trials. The training model is then used for making early selections by screening the seedlings derived from the selected parental trees.

Table 7. Validation test 3 - Accuracies of MBVs of the remaining half of the progeny trees estimated with data from parents and 50% of progeny

Trial/Trait	Accuracy[1]	Accuracy[2]
Hudler Trial		
Kpy_tree_13_ebv	0.84	0.83
Bd_tree_13_ebv	0.67	0.66
Vol_stand_normal_13_ebv	0.70	0.74
Middlesex Trial		
Kpy_tree_13_ebv	0.70	0.76
Bd_tree_13_ebv	0.67	0.70
Vol_stand_cold_13_ebv	0.72	0.79

[1] – Accuracy of MBVs estimated with parental data and 50% of progeny data. [2] – Accuracy of MBVs estimated with only 50% of progeny data

Validation tests in E. globulus

Data from HVP and ABP trials were used for validation tests in *E. globulus*. In the HVP trial, data from the parents was used for model training while in the ABP trial data from progeny trees was used for model training.

Validation tests in HVP populations

Marker genotype data and estimated breeding values (EBVs) of the 60 parents were used to predict the traits (MBVs) of 577 progeny trees using only their genotype data. The progeny trees were from two trials at different sites. MBVs were estimated for three traits (wood density, wood yield, and tree volume). Accuracies of the MBV estimates were assessed by correlating MBVs of the progeny with their EBVs (Table 8). Accuracies for the three traits ranged from 0.83 (density) to 0.91 (wood yield).

For estimating MBVs only parental marker genotypes were used without any information linking the parents to the offspring. These high accuracies indicate that the markers have captured accurately pedigree relationships among the progeny which contributed to the high accuracies observed in this test. Accuracies of MBVs generated using both additive and non-additive effects in prediction models (Table 8, accuracy[2]) were higher than the accuracies of MBVs based on only additive effects. Higher accuracy, especially for density, indicates additive, as well as non-additive effects, are important for this trait.

Table 8. Accuracies of MBVs of progeny trees from E. globulus of HVP

Trait	Accuracy[1]	Accuracy[2]	Accuracy[3]
Density EBV	0.72	0.83	-
Tree vol_age3 EBV	0.88	0.89	0.86
Wood Yield EBV	0.91	0.91	-

[1] – MBVs estimated with additive effects only, [2] – MBVs estimated with both additive and non-additive effects, [3] – MBVs estimated with revised EBVs excluding data from the progeny when estimating parental EBVs

A possible reason for the high accuracies observed here for tree volume may be due to the way the EBVs were estimated. EBVs of the parents were estimated with data of the progeny used in the testing as well as the progeny over a large number of trials and families. This raised the possibility that the high accuracies of the MBVs may be due to the EBVs of the parents incorporating the progeny data (circular prediction). This issue only affects tree volume as the other two trait MBVs were correlated to mid-parental averages. To circumvent the circular problem, the tree volume EBVs of the parents were re-estimated without the data of the progeny trees by Jo Sasse of HVP. We used these EBVs for developing the prediction model and to re-estimate the MBVs of the progeny. The accuracy of the MBVs was estimated by correlating individual tree EBVs with mid-parental values of the progeny. With the revised MBVs we still observed high accuracies similar to the results from the previous analysis. The accuracy of the revised MBVs for tree volume ranged from 0.86 to 0.92 using individual tree EBVs and mid-parent values respectively. These results show that the markers are capturing the family relationships among the progeny, which is contributing to the high accuracy of the MBVs.

Validation tests in ABP populations

Two validation tests were performed in the ABP populations. One test was to predict within the same generation using progeny trees from the same generation and the other test was to predict across the generations using progeny data to estimate MBVs of the parents. We could not perform forward predictions i.e. using parents' data to estimate progeny MBVs as in HVP, as parent EBVs are not available.

Validation test 1 – testing within a generation

For the within generation test, close to 500 progeny trees from two sites (Towes and Sinclair) were used. Prediction models were developed using 245 trees selected from odd-numbered replicates from the two sites (training population). The prediction model was then used to estimate MBVs of 254 trees selected from even-numbered replicates from the two sites (test population). EBVs of the training population estimated with data from 23 trials were used for estimating MBVs of the test population. MBVs were estimated for three traits (tree volume, pilodyn density and pulp yield). The accuracy of the MBVs was tested by correlating MBVs with EBVs of the test population. High accuracies were observed for all three traits (Table 9). Accuracies of the three traits were similar and ranged between 0.74 (PPY) and 0.76 (VOL).

Table 9. Accuracies of progeny MBVs – same generation validation tests

Trait	Accuracy
PILO Density EBV	0.75
Tree volume EBV	0.76
Pulp yield EBV	0.74

Validation test 2 – testing across generations

For across the generation tests, EBVs of the 500 progeny trees were used for developing a prediction model that was then used to estimate MBVs of the parental generation (backward estimation). MBVs were estimated for 109 parents of which 44 are the parents of the 500 progeny trees used for model training. In general, high accuracies were observed for all three traits (Table 10). However, the highest accuracies were observed for 44 trees that are the parents of the 500 progeny trees used for model training. This further demonstrates that a close relationship between the training and test populations leads to higher accuracies.

Unlike the test with HVP samples, the tests in the ABP samples do not suffer from the circular prediction issue. In both tests, the trait data used in the training populations is independent of the test populations. EBVs of the parents were estimated by 'backward estimation' i.e. parents EBVs were estimated using the progeny data.

Table 10. Accuracies of parental MBVs – across the generation validation tests

Trait	Accuracy[1]	Accuracy[2]
PILO Density EBV	0.53	0.71
Tree volume EBV	0.64	0.74
Pulp yield EBV	0.74	0.92

[1] – accuracy for all 109 parents. [2] – accuracy for 44 parents of the progeny trees used in training.

Limitations of benchmarking MBV accuracies against EBVs

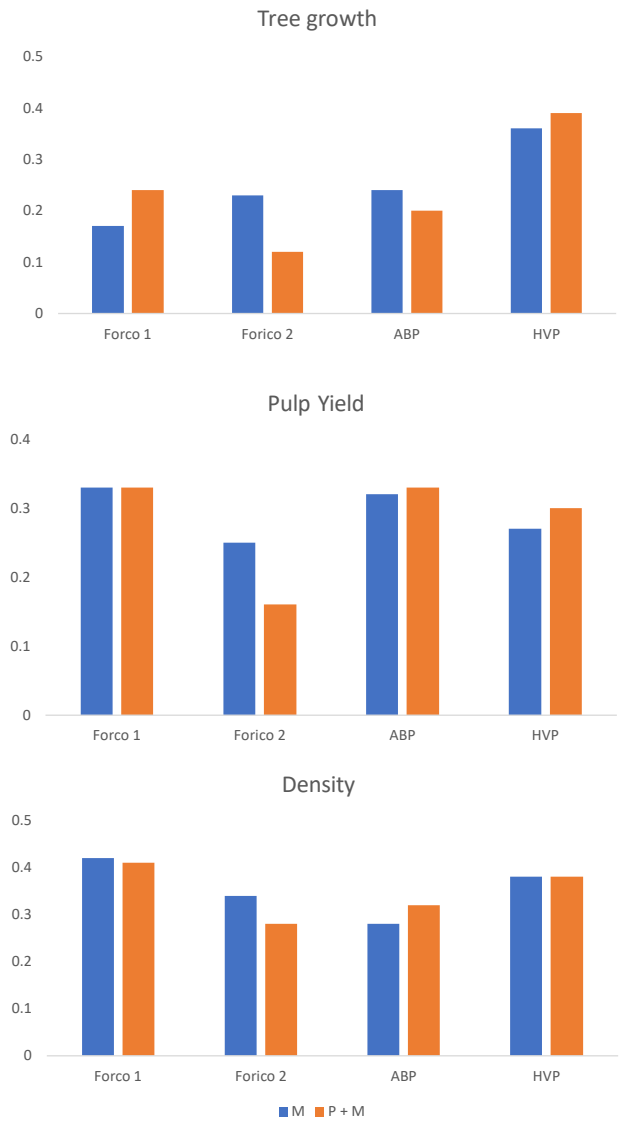
Results from all of the validation tests clearly show the high accuracy of MBVs and that a close relationship between the training and test populations is important for the high accuracy of MBVs. While the accuracy of the MBVs is benchmarked against EBVs, there is, however, a great deal of variation in the accuracies of the EBVs themselves. There are two sources of error in EBV estimates. One is pedigree errors. Most of the pedigree relationships used in traditional breeding for estimating breeding values contain errors. In this project, we have observed pedigree errors in several families. Several studies have shown the increased accuracy of genetic parameter estimation by correcting the pedigree errors with markers. Another unidentified source of error is errors in assumed pedigree structure. The progeny from an OP family were all treated as half-sibs and were given similar weights when estimating EBVs. However, several studies with markers have shown that there were unknown full-sibs within an OP family and members from different families show different levels of relationships. Similarly, in CP families, kinship coefficients among full-sibs vary around the mean value of 0.25. EBVs cannot account for these differences in relationships whereas marker derived relationships capture the full gamut of relationships among members of a family. Marker derived relationships (realised relationships) contain Mendelian sampling/segregation term (differences between the members of a family) which leads to

higher resolution of relationships compared to those with pedigree information. These factors contribute to deviations between MBVs and EBVs.

Testing the marker performance with site adjusted trait data

These tests are aimed at testing the ability of the markers to predict trait data other than EBVs. Two models are used to test the predictive ability (PA) of the markers, one is using markers alone (M) in the model and the other one is using both markers and pedigree information (P+M) in the model. Site adjusted trait data is used for these tests. Correlation between predicted traits with different models and observed trait data gives PA. The predictive ability can be converted to accuracy by dividing the PA by square root of heritability. *E. nitens* populations of Forico and *E. glob*ulus populations of HVP and ABP were used in these tests. Predictive ability of the two models was similar across all the traits and populations except for Middlesex trial (Figure 1). In Middlesex, the inclusion of pedigree with markers reduced the PA compared to M model. This suggests that there could be errors in the assumed pedigree of the Middlesex trial. This can be verified by pedigree testing with the markers. The inclusion of pedigree information (polygenic effect) in P +M model captures the genetic variation not explained by the markers. This will lead to persistence of accuracy in advanced generations and reduced bias in MBV estimates (Solberg et al., 2009).

Figure 1. Comparison of predictive abilities between two prediction models across different populations and traits.



Forico 1 – Hudler trial; Forico 2 – Middlesex trial

Implementation of MAS in industry breeding populations

One of the aims of the validation tests was to give industry partners confidence in marker-based selections using MBVs. According to the project proposal, the second phase of the project (commercial screening) would be carried out after reviewing the results of validation tests. High accuracies in the validation tests in breeding populations of partnering companies led to the continuation of the project into the second phase. The main aim of the second phase of the project was to screen seedlings in operational breeding populations and application of MAS. More than 11,000 seedlings/trees were genotyped with the marker panels developed for *E. nitens* and *E. globulus*. Details of the number of samples genotyped in each company are presented in Table 4. Below are brief descriptions of implementing MAS in breeding populations of different companies.

Implementing MAS in Forico's breeding populations

In total, 7,300 samples were genotyped with the marker panel developed for *E. nitens*. These include 6976 seedlings from the nursery. To estimate the MBVs of the seedlings, 558 seed orchard parents, 1072 progeny trees from three trials were used as training population to develop the prediction model. This model was then used to predict the MBVs of the seedlings. Marker data and EBVs of the training population were used for model training. Two models were used for estimating MBVs. The first model was based on 'RRBLUP' using the marker genotype data and EBVs of the training population. The second model was based on the 'single step Bayesian regression model' (SSBR, similar to ssGBLUP) in which marker genotype data, trait data from non-genotyped relatives of the deployment population (seedlings from the nursery) and pedigree data were all used together to estimate MBVs of the seedlings. MBVs were estimated using EBVs of stand volume, KPY, and density. For stand volume, two separate MBVs were estimated using two EBVs estimated with normal site model and cold site model.

The main advantage of the single step method is that all the available information is used when estimating MBVs, which should increase the accuracy of MBVs. However, this improvement in accuracy is dependent upon the proportion of the samples genotyped. When the proportion of the samples genotyped vs non-genotyped samples were similar, accuracies from the two model types (the model which uses data of only genotype samples such as RRBLUP and the models which use data of genotyped as well as non-genotyped samples such as SSBR) would also be similar.

To estimate the accuracy of the seedling MBVs, we correlated the MBVs with mid-parental values. This will give an indication of the potential accuracy of the MBVs when the seedlings are eventually phenotyped. In total, there were about 3000 seedlings from 243 CP families among the 6976 seedlings genotyped. EBVs were available for all the parents of these families and these were used to estimate mid-parental values. Mid-parental values of these families were used for estimating accuracies. High accuracies were estimated for all the four traits (Table 11).

Table 11. Accuracy of Forico's *E. nitens* nursery seedling MBVs

Trait	Accuracy
Volume (normal)	0.66
Volume (Cold)	0.62
KPY	0.71
BD	0.71

In traditional breeding, mid-parental values are used for making family based selections in the absence of progeny trait data. These high correlations with mid-parental values give confidence for making within family selections using MBVs. However, within-family accuracies could not be tested due to a large number of small families used in this project. Forico are currently selecting the seedlings based on the MBVs provided by GG. Selected seedlings will be established in different trials at different sites. Some of these trials will be genetic gains trials in which the performance of top-ranking seedlings selected based on MBVs will be compared with low ranking seedlings. To the best of our knowledge, this

(selection based on MBVs) represents one of the **first implementations of MAS in tree breeding** anywhere in the world.

Implementing MAS in ABP's breeding populations

In total, 2300 samples were genotyped with the *E. globulus* marker panel developed by GG. This includes 1500 seedlings established in 2015. Marker data and EBVs of 500 progeny trees used in the validation tests and 109 parents were used for model training. This model was used to estimate MBVs of seedlings established in 2015 (deployment population). Not all the parents of the seedlings had EBVs. In addition to 109 parents, we used MBVs of 153 parents for model training. Three traits, tree volume, pilodyn density and pulp yield were used for estimating MBVs of the seedlings. MBVs for tree volume were estimated using the SSBR method, as data from non-genotyped relatives was available. For the other two traits, MBVs were estimated with data from genotyped samples only. EBVs for tree volume were available from both the parents for 596 seedlings from 50 families. For these seedlings, MBV accuracy was estimated using mid-parental values. Similar to the results from Forico, a high correlation (0.66) was observed between MBVs and mid-parental values. When the EBVs of all the parents of the seedlings are available, these MBVs can be updated using the parental data for model training. This should improve the accuracy of seedling MBVs.

Screening of HVP trial to establish genetic gains trials

Seven hundred seedlings in a recently established trial were genotyped with the *E. globulus* marker panel. The main aims of this screen were to assess the accuracy of within-family MBVs, to identify top-ranking and bottom-ranking seedlings based on MBVs and to establish a genetic gains trial with the selected seedlings at different sites. These seedlings were derived from 8 families with an average of 88 seedlings per family which is ideal for assessing the accuracy of within-family MBVs.

We estimated MBVs of the seedlings with two models. In the first model, only the parental EBVs were used and in the second model, EBVs of the parents as well as the progeny data of the two trials used in validation tests (half-sibs of seedlings) were used for estimating MBVs of the seedlings. There is a high correlation (0.87) between MBVs using the two models across all the 700 seedlings (across the families). However, within family correlations between the two MBVs were lower and ranged from 0.63 to 0.75 in different families with an average of 0.67. This suggests that both models are similar in capturing across family variation but they differ in capturing within family variation. This may also suggest that the model with parent and progeny data captures within family variation better than the model using only parental data. However, this needs to be tested with EBVs estimated with measured trait data in the seedlings.

A high correlation of 0.76 was observed between the tree volume MBVs (estimated with model 2) and mid-parental values after correcting for pedigree errors in parents. MBVs of the seedlings were sent to HVP who will select the seedlings using the MBVs for establishing the genetic gains trials and forward selection of elite individuals.

MBV estimation in RMS' breeding populations

Samples from three progeny trials including their parents were genotyped with the *E. nitens* marker panel. In total, 900 samples were genotyped. For two trials, data from the progeny trees were used for estimating the MBVs of the parents. For one trial, no data was available

for either the parents or the progeny. For this trial, marker and trait data from unrelated populations from validation tests was used for developing the prediction model. The model was used to estimate the MBVs of the progeny as well as their parents. This is not ideal as the model is developed with unrelated samples. However, once the trees are measured, trait data can be used for updating the MBVs.

MBV estimation in Sustainable Timber Tasmania breeding population

In total, 108 trees from ST Tas were genotyped with the *E. nitens* marker panel. This included 26 seed orchard (SO) parents and 82 progeny trees derived from 10 of the 26 parents. All the families including the 10 families of the 82 progeny trees were OP families. Two types of EBVs were available for the 26 SO parents. One for measurement age at 8-15 years and another one for harvest age at 22 years. EBVs for several traits were available including MAI at harvest age for 26 parents. Progeny trees had DBH measured at 4 years of age. MBVs of the progeny were estimated with the marker and EBV data of the 26 parents. To test the performance of the markers, we correlated the MBVs with the DBH of the progeny trees. A correlation of 0.30 was observed between harvest age MAI MBV and DBH at 4 years for the progeny trees. We also estimated MBVs for parents using DBH data of the progeny trees (backward estimation). A high correlation (0.47) was observed between DBH MBVs and MAI EBVs of the parents (Table 12). These results reveal the accuracy of the progeny MBVs. It is interesting to note the high correlation between harvest age MAI and DBH at 4 years of age.

Table 12. Correlation between DBH and MAI MBVs in parents and progeny trees of ST Tas population

Generation	Trait 1	Trait 2	COR
progeny	DBH.4	MAI_MBV	0.30
parents	DBH.4_MBV	MAI_EBV	0.47

Comparison of genetic gain from MAS and Phenotypic Selection

Genetic gain or response to selection was estimated as the difference between the average of the selected group and average of all the samples used in selection (population mean) *i.e*. the selection differential. Percentage of gain was estimated by dividing selection differential with the population mean. The following formula was used for estimating the percentage of the genetic gain.

$$Gain = \frac{Avg(s) - Avg(p)}{Avg(p)} \times 100$$

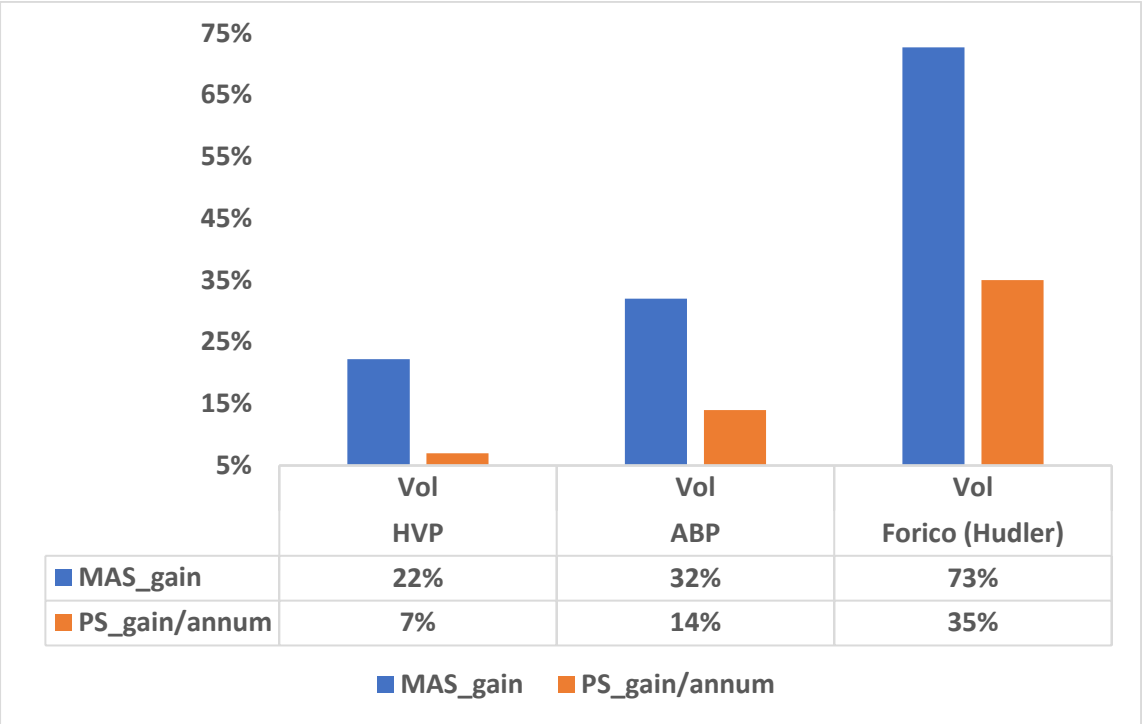
Where *Avg(s)* is the average EBV of the selected group, *Avg(p)* is the average EBV of the all samples used in selection.

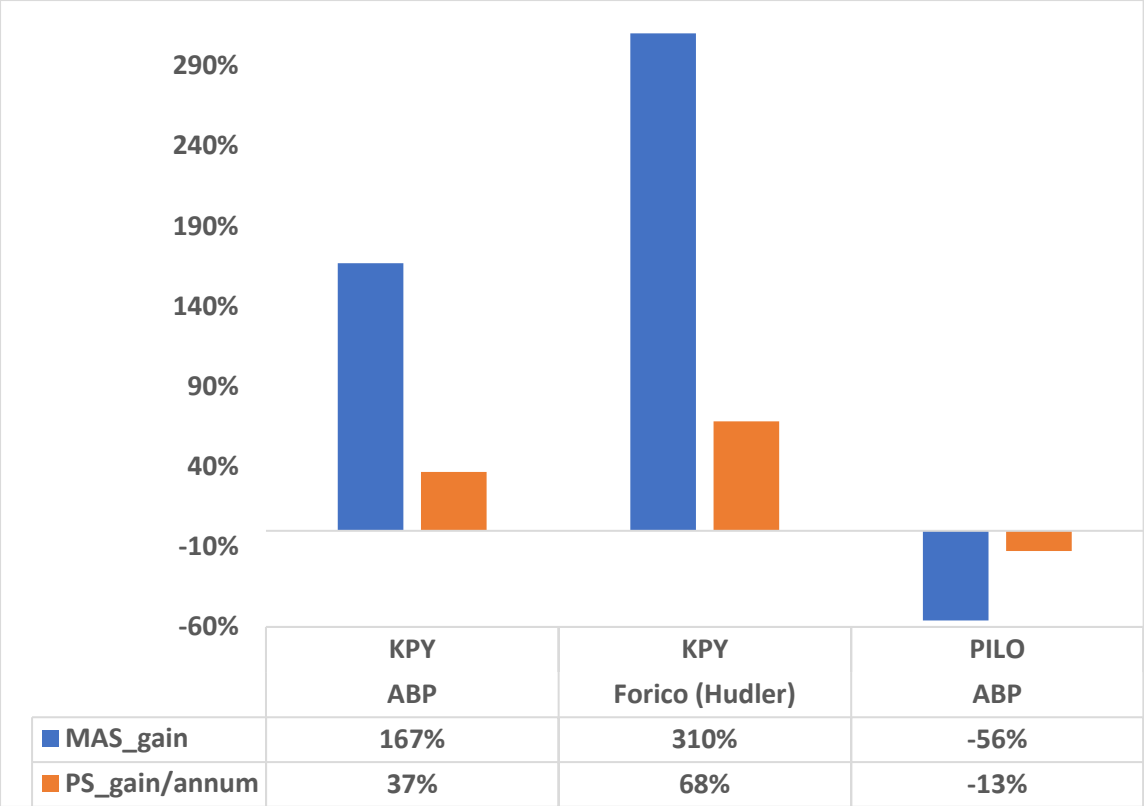
For calculating genetic gain, we used a selection intensity of 10% *i.e*. top 10% trees were selected from the population of samples used in selection. All the tests were performed in the validation populations which were used for estimating MBVs. Selections were based on EBVs. For estimating genetic gain from MAS, MBVs of the validation population were used to rank the trees. Average of the top 10% trees was calculated using the EBVs of the selected

trees. For estimating genetic gain from phenotypic selection (PS) or traditional breeding, trees from the validation population were ranked based on EBVs, and average EBV of the top 10% selected trees was estimated.

For selecting the superior lines/genotypes from the selected families, MAS is applied at the seedling stage while PS is performed at least at the age of 3 years for growth and 6 years for wood traits. To compare the gain from PS with MAS, PS gain was converted to per year by dividing the genetic gain by 3 for growth and 6 for wood traits. These tests were done in validation populations of several industry partners. To reflect the practical breeding, MBVs of the progeny estimated with the model developed in parents were used in these tests wherever possible (HVP and Hudler). Results from these tests are shown in Figure 2. Genetic gain from MAS is 2 (100%) to 3 times (200%) more than the PS for all the traits across all the populations. These are conservative estimates as growth and wood trait EBVs of some of the populations may have been estimated at later ages compared to the age used in these calculations. Genetic gain from MAS could further be increased with high selection intensity. Selection intensity can be increased by increasing the sample size of the test population.

Fig 2. Comparison of genetic gain from MAS (marker-assisted selection) and PS (phenotypic selection). Selection is for lower values for pilodyn density.





Simulation of gains from marker-assisted selection

A multivariate selection tool was developed to evaluate the impacts of incorporating MAS in a sample tree improvement program for a chip-export breeding objective. The sample program, using *E. nitens*, shows that an ROI of $7.81 can be achieved using conservative estimates of the gains.

This selection tool was developed using the standard breeder's equation:

$$\boldsymbol{GENETIC\ GAIN} = \frac{\boldsymbol{Phenotypic\ VAR} \times \boldsymbol{Heritability} \times \boldsymbol{Selection\ Intensity}}{\boldsymbol{Breeding\ Cycle}}$$

The scenario program commences with screening 3 existing progeny trials derived from 100 parent trees and the assumptions are as outlined in Table 13. Initially, we make 100 forward selections by screening 30 progeny from each of 100 open-pollinated families. We assume that the progeny trees are already producing seed. Three thousand trees (30 trees/family, 100 families) will be genotyped with *E. nitens* marker panel and 100 trees will be selected as parents (forward selections) based on EBVs and marker genotype data. Marker genotype data will be used for improving existing prediction models for *E. nitens*. We estimate improvements from this step to deliver an increase of 4.1% to growth and 1.1% to density, for an overall productivity improvement of 5.3%. Although the information derived from the screening can support this work, we do not attribute the gains from this step to the return on investment (ROI) in the program.

From these 100 forward selection parents, we screen 10,000 seedlings (100 seedlings/family) each year for 3 years, selecting the top 50 seedlings each year to be fast-tracked into developing a Clonal Seed Orchard (CSO). Truncation selection is used to identify 1 in 20 trees for growth and 1 in 10 trees for density. This is a total of 30,000 seedlings screened in the program and 150 selected to build the CSO, this is an overall selection intensity of 1 in 200. The CSO is open pollinated and after 7 years can be used for early deployment, initially 5% of the estate increasing to 100% over 4 years.

The calculations do not take into account numerous other outputs from an implementation of MAS which can support the other breeding work i.e. fingerprinting, pedigree reconstruction, reduction of inbreeding and genetic diversity. It does not take into account the change in rotation opportunity available through increased growth as trees could be harvested up to 2 years earlier.

Table 13. Scenario Assumptions

	Program Assumptions
Estate Size	90,000 ha
Crop	*E. nitens*
Time to flower	7 years
Rotation	15 years
Annual Harvest	6,000 ha (estate ÷ rotation)
Current MAI	17 m^3/ha/yr

The literature was reviewed to search for genetic parameters that were useful for estimating genetic gain from selection for a chip-export breeding objective. The pulp productivity study

of (Greaves et al., 1997) was selected as this provided a comprehensive set of genetic parameters that directly relate to profitability. We have applied the accuracies of 0.65 for volume and density which are conservative given the observed accuracies found in the blind validations in this project.

Table 14. Genetic assumptions (harvest values)

	Genetic Assumptions
Volume at harvest	255 m^3/ha (MAI × rotation)
Volume (Phenotype SD)	70
Volume (Additive SD)	30
Density	450 Kg/m^3
Density (Phenotype SD)	30.8
Density (Additive SD)	20
Marker Accuracies phenotype	0.50
Marker Accuracies MAS	0.65
Correlation Vol:Den	0.11

Table 15. Financial assumptions

	Financial Assumptions
Chip price	$150 USD/tonne
Discount rate	8% p.a.
MAS cost p/sample	$40
MAS overheads yr	$150k
Final year ROI	2054 (after 36 years)

An estate size of 90,000 hectares with a 15-year rotation was assumed, with gains calculated at the end of a single rotation in 2054. That is 15 years from the first deployment seed becoming available in 2040. Costs have been calculated using target pricing of $40 per sample with $150,000 annually allocated to program overheads, it's assumed that the cost of managing the existing tree improvement programs remain the same as these must be maintained. Revenue from current chip price is adjusted to AUD$195 by a factor of 1.3 reflecting the current USD/AUD exchange rate.

Table 16. Gains from MAS

	Volume (m^3/ha)	**Density (kg/m^3)**	**Dollars**
per hectare increase	22.3	16.7	$2,915.34
per hectare increase (%)	8.4%	3.7%	12.4%

Improvements attributable to MAS are estimated at 8.4% for growth and 3.7% for density, for an overall improvement in tonnes per hectare of 12.4%. This delivers an increase in the value of $2,915 per ha.

Table 17. Return on Investment from a MAS project

Investment in MAS	$1,782 k
Project profit NPV	$13,910 k
ROI per $1	$7.81
IRR	14.3%

The program commences in 2018 with the CSO established over 3 years and initially producing deployment seed in 2025, by 2040 the first MAS improved trees are harvested. Final ROI is calculated on the returns at the year 2054 however, returns for industry partners can be realised much earlier through the increase in estate valuation. Estate valuation increase can be determined by measurement of the gains observed from CSO seedlings in the estate, or by recognising the value of the MAS program as soon as it is commenced, subject to audit verification.

Using the conservative scenario and only counting the strict MAS gains, the program estimates NET returns of $222m in the year 2054, but this is discounted by 8% pa to Net Present Value (NPV) of $13.9m in today's terms. By adjusting the discount rate so that ROI is $1:$1 we can determine the Internal Rate of Return (IRR) from the project, which is 14.3% p.a for 37 years.

Additional benefits of using markers in tree breeding

In addition to marker-based selection with MBVs, markers can also be used to improve the accuracy of genetic parameter estimation in traditional tree breeding which improves the genetic gain. The genetic relationship matrix (GRM) generated from markers can be used for confirming or identifying parent-pedigree relationships, confirming or identifying full-sib and half-sib relationships within and between the families, identifying unknown relationships among the families and individuals of different families. A recent study has shown that pedigree corrected with markers improved the accuracy of heritability estimates and breeding values (Munoz et al., 2014). It was suggested that pedigree errors need to be first corrected by markers before implementing genomic selection models. This will improve the accuracy of MBVs as more accurate EBVs with minimum pedigree errors are used for model training and for correlating with the MBVs. In addition to correcting pedigree errors, another advantage of using markers is estimating inbreeding rates which can be used for maintaining genetic diversity and reducing inbreeding depression. Markers can also be used for fingerprinting to identify clonal and labeling errors.

Pedigree reconstruction and inbreeding estimates using markers

To demonstrate the ability of our markers for pedigree reconstruction and estimating inbreeding, we used 20 *E. nitens* controlled pollinated (CP) families of Forico. The identity of mother trees of all 20 families was revealed to us and we aimed to identify the paternal parent

of individuals in these 20 families. Ten to twenty siblings were genotyped from each family. We also genotyped 525 potential parents of these families. GRM estimated with the marker data was used to identify the parents and to confirm the identity of the full-sibs within each family. While the identity of mother trees was revealed to us, GRM can also be used to confirm the identity of the known parents.

GRM is not only useful for identifying parents, it can also be used for estimating inbreeding. Diagonals of GRM show individual inbreeding values while off-diagonals show the pair-wise relationships. Inbreeding of a family may be due to inbred parents used for crossing or due to the failure of pollination resulting in self-pollination. When an inbred parent is used for crossing, all the individuals of the family show higher than the expected full-sib relationships. But the individuals within a family themselves may not have high inbreeding coefficients (Table 18). However, when controlled pollination fails resulting in self-pollination, all the individuals of the family show elevated relationships and some of the sibs may show high rates inbreeding coefficients (Table 19). These two types of inbreeding can be identified with GRM. Examples of the two types of inbreeding are shown below.

Table 18. Inbreeding due to inbred parent used in controlled crossing

	P1	P2	FS1	FS2	FS3	FS4	FS5
P1	1.33	0.24	0.73	0.76	0.73	0.69	0.71
P2	0.24	1.02	0.53	0.56	0.60	0.60	0.61
FS1	0.73	0.53	1.05	0.66	0.64	0.61	0.56
FS2	0.76	0.56	0.66	1.11	0.72	0.71	0.64
FS3	0.73	0.60	0.64	0.72	1.08	0.75	0.67
FS4	0.69	0.60	0.61	0.71	0.75	1.03	0.69
FS5	0.71	0.61	0.56	0.64	0.67	0.69	1.02

P- parents, FS-full-sibs, Off diagonals are pair-wise relationships among full-sibs and parents. Diagonals are individual inbreeding coefficients.

Table 19. Inbreeding due to failure of controlled pollination resulting in self-pollination

	P1	FS1	FS2	FS3	FS4	FS5	FS6
P1	0.90	0.76	0.72	0.73	0.74	0.71	0.73
FS1	0.76	1.25	0.82	0.84	0.80	0.77	0.76
FS2	0.72	0.82	1.26	0.68	0.68	0.60	0.63
FS3	0.73	0.84	0.68	1.23	0.84	0.84	0.69
FS4	0.74	0.80	0.68	0.84	1.18	0.82	0.71
FS5	0.71	0.77	0.60	0.84	0.82	1.28	0.81
FS6	0.73	0.76	0.63	0.69	0.71	0.81	1.40

P- parents, FS-full-sibs, Off diagonals are pair-wise relationships among full-sibs and parents. Diagonals are individual inbreeding coefficients.

Pair-wise relationships among the full-sibs range between 0.25 and 0.60. Elevated pair-wise relationships indicate inbreeding. The expected parent-offspring relationships are similar to those of full-sibs. Inbreeding coefficients of individuals above 1.25 are regarded as significant. Parent (P1) in table 14 is inbred, consequently, the pair-wise relationships among the full-sibs were higher than the expected full-sib relationships.

Results of the pedigree reconstruction

Except for two, all the full-sibs of the 20 families were confirmed by marker analysis. These two errors may indicate field labeling errors. For one family, the maternal parent could not be confirmed due to low genotype call rates. We also identified the families that were inbred. These results demonstrate that the markers were not only useful for estimating MBVs but also for pedigree reconstruction, estimating inbreeding and identifying clonal errors.

Conclusions and Recommendations

Large numbers of associated markers were identified in *E. nitens* and *E. globulus* for the three most important commercial traits. These markers were then applied using GS to build models for predicting MBVs in various industry trials. These tests were performed in different trials of different companies in two commercially important eucalypt species, *E. nitens,* and *E. globulus*. Through a series of validation tests, we have demonstrated that high accuracies can be achieved with the MBVs. The ultimate test of marker performance is when the selected trees are assessed in the field, which may take several years. However, these validation tests are an excellent surrogate, given that in most tests parental information was used to predict the performance of mature progeny. Key to the success of these tests has been the high throughput methods for identification of associated markers and our genotyping methodology. Another feature of the GG technology is the high throughput DNA isolation method and the minimal quantity and quality requirements of DNA samples, which is crucial in routine screening for commercial implementation. These technological improvements allowed rapid turnaround times. For example, close to 7,000 seedlings were genotyped, analysed and the MBVs were provided to Forico months in advance to the actual delivery date. Forico have used the MBVs to select the seedlings and set up field trials with the selected seedlings.

One of the benefits of markers in tree breeding is improving the accuracy of genetic parameter estimation with traditional methods. Markers can be used for identifying and correcting pedigree errors. The marker corrected pedigree file can be used in BLUP analyses for improving the accuracy of heritability and breeding value estimation. Markers are also useful for converting OP families to CP families. Full-sibs from OP families (generated naturally) and the paternal parents of OP families can be identified using markers. This is particularly useful in field-tested OP families. Once the top-ranking OP families are identified from field testing, they can be screened with markers to identify full-sibs and their paternal parents. This information can be used for making further deliberate crosses between the parents knowing that their progeny will be superior based on the field tests. In addition, markers can be used for estimating inbreeding rates and identifying labeling and clonal errors.

However, the main application of markers in tree breeding is in screening seedlings to make early selections. The major benefits of making selection in seedlings include a drastic reduction in the length of the breeding cycle, increasing selection intensity which contributes to high genetic gains and the ability to make within family selections. Within family variation is poorly explored in traditional breeding as data from field-grown trees is required for making within family selections. However, the accuracy of predicting traits within a family could not be tested due to the small size of the families used in this study. Another important application of markers is in selecting parents for controlled crossing. Parents for controlled

crossing can be selected based on MBV estimates. Genetic diversity and inbreeding rates among the top-ranking trees selected based on MBVs can be used for selecting complementary parents to produce superior progeny.

As shown in the validation tests, the close relationship between training and test populations leads to a high accuracy of MBVs. Therefore, for accurate estimation of seedling MBVs, data from parents (and field-tested progeny that are related to the seedlings if available) should be used in developing the prediction model. To take advantage of the marker technology selection cycles should be accelerated. With traditional breeding it takes 5-6 years for making one selection whereas with markers selections can be done annually. While it takes time for generating controlled crosses, seedlings from top-ranking families from progeny testing can be screened with the markers. Results from genetic gains calculations indicate 2 to 3 times more gain per annum compared to the traditional breeding. This results as shown in the financial modeling, in a substantial return on investment (ROI) by implementing MAS as suggested in the operational plans. In this project, we have improved the existing marker technology, developed methods for application of markers in breeding programs and demonstrated MAS in existing breeding populations of several industry partners. While the tests performed in this project revealed the high accuracy of predicting traits (MBVs) across families, within family accuracies, however, could not be assessed due to the small size of the families used in this study. One of the main applications of markers in tree breeding is making within family selections in seedlings. Within-family accuracies need to be tested in further studies using a large number of individuals per family.

References

Avendaño, S., Woolliams, J. A., & Villanueva, B. (2004). Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. *Genetical Research*, *83*(1), 55–64. https://doi.org/10.1017/S0016672303006566

Boichard, D., Ducrocq, V., Croiseau, P., & Fritz, S. (2016). Genomic selection in domestic animals: Principles, applications and perspectives. *Comptes Rendus Biologies*, *339*(7–8), 274–277. https://doi.org/10.1016/j.crvi.2016.04.007

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., & Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences*, *101*(42), 15255–15260. https://doi.org/10.1073/pnas.0404231101

Butcher, P., & Southerton, S. (2007). Marker-assisted selection in forestry species. *Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish*, 46.

Byrne, M., Murrell, J. C., Owen, J. V., Kriedemann, P., Williams, E. R., & Moran, G. F. (1997). Identification and mode of action of quantitative trait loci affecting seedling height and leaf area in Eucalyptus nitens. *Theoretical and Applied Genetics*, *94*(5), 674–681. https://doi.org/10.1007/s001220050465

Costa e Silva, J., Borralho, N. M. G., Araújo, J. A., Vaillancourt, R. E., & Potts, B. M. (2009). Genetic parameters for growth, wood density and pulp yield in Eucalyptus globulus. *Tree Genetics & Genomes*, *5*(2), 291–305. https://doi.org/10.1007/s11295-008-0174-9

Dillon, S. K., Nolan, M., Li, W., Bell, C., Wu, H. X., & Southerton, S. G. (2010). Allelic variation in cell wall candidate genes affecting solid wood properties in natural populations and land races of Pinus radiata. *Genetics*, *185*(4), 1477–1487. https://doi.org/10.1534/genetics.110.116582

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C., & Valenzuela, S. (2017). Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics and Genomes*, *13*(4). https://doi.org/10.1007/s11295-017-1158-4

Dutkowski, G. W., Apiolaza, L. A., Gore, P. L., McRae, T. A., & Pilbeam, D. (2000). *The STBA Cooperative Tree Improvement Strategy for Eucalyptus globulus - Revised for the period 2000-2005. STBA Technical Report TR00-03*.

Fernando, R. L., Dekkers, J. C., & Garrick, D. J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution*, *46*(1), 50. https://doi.org/10.1186/1297-9686-46-50

Freeman, J. S., Potts, B. M., Downes, G. M., Pilbeam, D., Thavamanikumar, S., & Vaillancourt, R. E. (2013). Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in Eucalyptus globulus. *New Phytologist*, *198*(4), 1121–1134. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84877626568&partnerID=40&md5=992012a5e471006ad5ac7fceede33bb9

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., & El-Kassaby, Y. A. (2015).

Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics*, *16*(1), 370. https://doi.org/10.1186/s12864-015-1597-y

González-Martínez, S. C., Ersoz, E., Brown, G. R., Wheeler, N. C., & Neale, D. B. (2006). DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in Pinus taeda L. *Genetics*, *172*(3), 1915–1926. https://doi.org/10.1534/genetics.105.047126

González-Martínez, S. C., Wheeler, N. C., Ersoz, E., Nelson, C. D., & Neale, D. B. (2007). Association genetics in Pinus taeda L. I. wood property traits. *Genetics*, *175*(1), 399–409. https://doi.org/10.1534/genetics.106.061127

Grattapaglia, D., Vaillancourt, R. E., Shepherd, M., Thumma, B. R., Foley, W., Külheim, C., … Myburg, A. A. (2012). Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. *Tree Genetics and Genomes*, *8*(3), 463–508. https://doi.org/10.1007/s11295-012-0491-x

Greaves, B. L., Borralho, N. M. G., & Raymond, C. A. (1997). Breeding objective for plantation eucalypts grown for production of kraft pulp. *Forest Science*, *43*(4), 465–472.

Hamilton, M. G., Joyce, K., Williams, D., Dutkowski, G. W., & Potts, B. M. (2008). Achievements in forest tree improvement in Australia and New Zealand 9. Genetic improvement of Eucalyptus nitens in Australia. *Australian Forestry*, *71*(2), 82–93.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Kraakman, A. T. W., Niks, R. E., Van Den Berg, P. M. M. M., Stam, P., & Van Eeuwijk, F. A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*, *168*(1), 435–446. https://doi.org/10.1534/genetics.104.026831

Müller, B. S. F., Neves, L. G., de Almeida Filho, J. E., Resende, M. F. R., Muñoz, P. R., dos Santos, P. E. T., … Grattapaglia, D. (2017). Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics*, *18*(1), 524. https://doi.org/10.1186/s12864-017-3920-2

Munoz, P. R., Resende, M. F. R., Huber, D. A., Quesada, T., Resende, M. D. V, Neale, D. B., … Peter, G. F. (2014). Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. *Crop Science*, *54*(3), 1115–1123. https://doi.org/10.2135/cropsci2012.12.0673

Neale, D. B. (2007). Genomics to tree breeding and forest health. *Current Opinion in Genetics and Development*, *17*(6), 539–544. https://doi.org/10.1016/j.gde.2007.10.002

Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends in Plant Science*, *9*(7), 325–330.

Porto-Neto, L. R., Barendse, W., Henshall, J. M., McWilliam, S. M., Lehnert, S. A., & Reverter, A. (2015). Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. *Genetics Selection Evolution*, *47*(1), 84. https://doi.org/10.1186/s12711-015-0162-0

Raymond, C. A. (2002). Genetics of Eucalyptus wood properties. *Annals of Forest Science*,

59(5–6), 525–531. https://doi.org/10.1051/forest:2002037

Resende, M. D. V., Resende Jr, M. F. R., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., … Grattapaglia, D. (2012). Genomic selection for growth and wood quality in *Eucalyptus*: Capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist*, *194*(1), 116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x

Resende, M. F. R., Mu??oz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., … Kirst, M. (2012). Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist*, *193*(3), 617–624. https://doi.org/10.1111/j.1469-8137.2011.03895.x

Schimleck, L. R., Kube, P. D., Raymond, C. a, Michell, A. J., & French, J. (2005). Estimation of whole-tree kraft pulp yield of Eucalyptus nitens using near-infrared spectra collected from increment cores. *Canadian Journal of Forest Research*, *35*(12), 2797–2805. https://doi.org/10.1139/x05-193

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., Odegard, J., & Meuwissen, T. H. (2009). Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genetics Selection Evolution*, *41*(1), 53. https://doi.org/10.1186/1297-9686-41-53

Southerton, S. G., MacMillan, C. P., Bell, J. C., Bhuiyan, N., Dowries, G., Ravenwood, I. C., … Thumma, B. R. (2010). Association of allelic variation in xylem genes with wood properties in Eucalyptus nitens. *Australian Forestry*, *73*(4), 259–264. https://doi.org/10.1080/00049158.2010.10676337

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395–408. https://doi.org/10.1038/hdy.2015.113

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., & Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology*, *17*(1), 110. https://doi.org/10.1186/s12870-017-1059-6

Thamarus, K., Groom, K., Bradley, A., Raymond, C. A., Schimleck, L. R., Williams, E. R., & Moran, G. F. (2004). Identification of quantitative trait loci for wood and fibre properties in two full-sib pedigrees of Eucalyptus globulus. *Theoretical and Applied Genetics*, *109*(4), 856–864. https://doi.org/10.1007/s00122-004-1699-4

Thavamanikumar, S., Dolferus, R., & Thumma, B. R. (2015). Comparison of Genomic Selection Models to Predict Flowering Time and Spike Grain Number in Two Hexaploid Wheat Doubled Haploid Populations. *G3: Genes|Genomes|Genetics*, *5*(October), 1991–1998. https://doi.org/10.1534/g3.115.019745

Thavamanikumar, S., McManus, L. J., Ades, P. K., Bossinger, G., Stackpole, D. J., Kerr, R., … Tibbits, J. F. G. (2014). Association mapping for wood quality and growth traits in Eucalyptus globulus ssp. globulus Labill identifies nine stable marker-trait associations for seven traits. *Tree Genetics and Genomes*, *10*(6), 1661–1678. https://doi.org/10.1007/s11295-014-0787-0

Thavamanikumar, S., McManus, L. J., Tibbits, J. F. G., & Bossinger, G. (2011). The

significance of single nucleotide polymorphisms (SNPs) in Eucalyptus globulus breeding programs. *Australian Forestry*, *74*(1), 23–29. https://doi.org/10.1080/00049158.2011.10676342

Thavamanikumar, S., Southerton, S. G., Bossinger, G., & Thumma, B. R. (2013). Dissection of complex traits in forest trees - opportunities for marker-assisted selection. *Tree Genetics and Genomes*, *9*(3), 627–639. https://doi.org/10.1007/s11295-013-0594-z

Thavamanikumar, S., Southerton, S., & Thumma, B. (2014). RNA-Seq using two populations reveals genes and alleles controlling wood traits and growth in Eucalyptus nitens. *PLoS ONE*, *9*(6). https://doi.org/10.1371/journal.pone.0101104

Thumma, B., MacMillan, C., Southerton, S., Williams, D., Joyce, K., & Ravenwood, I. (2010). Accelerated breeding for high pulp yield in E . nitens using DNA markers identified in 100 cell wall genes : The Hottest 100. *FWPA Report : PNC052-0708*.

Thumma, B. R., Matheson, B. A., Zhang, D., Meeske, C., Meder, R., Downes, G. M., & Southerton, S. G. (2009). Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. *Genetics*, *183*(3), 1153–1164. https://doi.org/10.1534/genetics.109.106591

Thumma, B. R., Nolan, M. F., Evans, R., & Moran, G. F. (2005). Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. *Genetics*, *171*(3), 1257–1265. https://doi.org/10.1534/genetics.105.042028

Thumma, B. R., Southerton, S. G., Bell, J. C., Owen, J. V., Henery, M. L., & Moran, G. F. (2010). Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. *Tree Genetics & Genomes*, *6*(2), 305–317. https://doi.org/10.1007/s11295-009-0250-9

Thumma, B., Thavamanikumar, S., Brawner, J., & Southerton, S. (2015). Genetic Selection Tools for Enhanced Wood Properties and Plantation Productivity in Australia ' s Temperate Eucalypts (Blue Gum Genomics) Genetic Selection Tools for Enhanced Wood. *FWPA Report : PNC209-1011*.

Thumma, B., Thavamanikumar, S., & Southerton, S. (2017). Discovery and application of DNA markers for resistance to Teratosphaeria in E. globulus. *FWPA Report : PNC363-1415*.

Verhaegen, D., Plomion, C., Gion, J. M., Poitel, M., Costa, P., & Kremer, A. (1997). Quantitative trait dissection analysis in Eucalyptus using RADP markers: 1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. *Theoretical and Applied Genetics*, *95*(4), 597–608. https://doi.org/10.1007/s001220050601

Acknowledgements

We gratefully acknowledge the contribution of David Spencer, Randall Falkiner, Philip Southerton, Geoff Downes, Ross Gillies and Michael Bird for their assistance with sampling the trials. We thank Marina Trigueros, Jessica Bovill, Jaifu Tan and Tina Liu for their contribution to the molecular genetic analyses. We acknowledge the contribution of Jo Sasse and Andrew Callister in validation test analyses and Justine Morgan in project administration. We also thank Darren Davies and Bryan Hayes for their support to the project.