Forest&Wood Products Australia
*Knowledge for a sustainable Australia*

SUSTAINABILITY & RESOURCES

PROJECT NUMBER: PNC052-0708 JUNE 2010

# Accelerated breeding for high pulp yield in E. nitens using DNA markers identified in 100 cell wall genes: The Hottest 100

# Accelerated breeding for high pulp yield in *E. nitens* using DNA markers identified in 100 cell wall genes: The Hottest 100

Prepared for

**Forest & Wood Products Australia**

by

**Bala Thumma, Colleen MacMillan, Simon Southerton, D. Williams, K. Joyce and I. Ravenwood.**

# Publication: Accelerated breeding for high pulp yield in E. nitens using DNA markers identified in 100 cell wall genes: The Hottest 100

**Project No: PNC052-0708**



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

ISBN: 978-1-921763-05-2

**Researcher/s:**
B. Thumma, C. MacMillan and S. Southerton
CSIRO Plant Industry, GPO Box 1600 Canberra ACT

D. Williams
Forestry Tasmania
79 Melville St, Hobart, Tasmania 7001

K. Joyce and I. Ravenwood
Gunns Limited
de Boer Drive, Ridgley, Tasmania 7321

**Final report received by FWPA in June, 2010**

**Forest & Wood Products Australia Limited**
Level 4, 10-16 Queen St, Melbourne, Victoria, 3000
T +61 3 9927 3200 F +61 3 9927 3288
E info@fwpa.com.au
W www.fwpa.com.au

## Table of Contents


Executive Summary ..... iii
Introduction ..... 1
  Pedigree markers ..... 1
  Population markers ..... 1
Methodology ..... 3
  Experimental populations ..... 3
    Association population ..... 3
    Validation populations ..... 3
  Wood phenotyping ..... 3
    Meunna population ..... 3
    Tarraleah population ..... 4
    Loudwater population ..... 5
  Candidate genes ..... 6
  SNP discovery ..... 6
  SNP genotyping ..... 6
    Meunna population ..... 6
    Validation populations ..... 7
  Association analyses ..... 7
Results: ..... 8
  Cellulose and KPY predictions ..... 8
  Gene walking of candidate genes ..... 8
  SNP Discovery ..... 9
  SNP genotyping in Meunna population ..... 10
  SNP-trait associations in Meunna ..... 11
    Cellulose and KPY ..... 11
    Lignin traits ..... 11
    Physical traits ..... 11
  SNP validation in the Loudwater and Tarraleah populations ..... 12
    Genotyping ..... 12
  SNP validation in the Loudwater and Tarraleah populations…cont'd ..... 14
    SNP trait associations ..... 14
  Meta analysis ..... 17
  Marker-assisted selection for KPY ..... 17
Discussion ..... 20
Conclusion ..... 21
Recommendations ..... 21
References ..... 22

## Executive Summary

The hottest 100 project examined naturally occurring allelic variation in genes influencing pulp yield in *Eucalyptus nitens* that can be captured in breeding programs using marker-assisted selection (MAS). Three industry populations derived from Central Victorian provenances were phenotyped for kraft pulp yield (KPY), cellulose, growth and other traits. This included an association population at Meunna TAS (420 unrelated families) and validation populations located at Loudwater TAS (420 trees from 140 second generation families) and Tarraleah TAS (520 trees from 150 first generation families). About 100 cell wall genes were selected for study based on previous research and literature searches. Full length gene sequences of these candidate genes were obtained by gene walking and single nucleotide polymorphisms (SNPs) identified by 454 high throughput sequencing in unrelated individuals. Over 500 SNPs were genotyped across the Meunna association population. And association analyses revealed 80 SNPs significantly associated with several wood quality traits including 17 SNPs associated with cellulose content and KPY. Significant SNPs were genotyped across both validation populations. Eleven out of the seventeen SNPs (65%) showed significant effects in at least one of the validation populations. The direction of the SNP effect was the same at six SNPs, however, for about half the SNPs the allele effect is reversed (flips) between population(s). This suggests that environment is moderating the impact of specific alleles. Meta analysis, performed by combining the data from the three populations revealed 11 (stable) SNPs significantly associated with cellulose and KPY.

The major benefits for the industry from this project are (1) comprehensive phenotypic data on pulp-related traits such as cellulose, lignin and KPY and solid wood traits such as density and microfibril angle and (2) SNP markers significantly associated with pulp yield and growth traits in the three populations in this study which can now be used for MAS. For example, selecting trees homozygous for six superior KPY alleles is predicted to shift the KPY by 2.2% to 51.2% from the population mean of 50.1% with a concurrent improvement in growth of 10%. This improvement in KPY is predicted to improve the profitability of a standard eucalypt pulp mill operation by over $40M per year. Further validation of these markers is recommended in order to be more confidant about the magnitude of the gains likely to be achieved by MAS. This could be achieved by genotyping the KPY SNPs across additional populations growing in different environments and comparing trait values of trees with favourable, unfavourable and average genotype rankings. These experiments would effectively mimic genetic gain trials. Markers thus validated should be robust for MAS of superior trees across all breeding populations of *E. nitens*.

It is recommended that the markers be used by industry to screen existing clonal seed orchards to identify inferior genotypes for culling and to screen future genotypes entering the orchards. As more markers become available in future research, wider screening in provenance trials is recommended to identify elite individuals and parents for controlled crosses to pyramid desirable alleles into future generations. An expanded study with many more genes should identify a large proportion of the alleles affecting wood quality traits.

## Introduction

The goal of tree breeding is to produce improved progeny of greater economic value. This is achieved by selection during breeding of individuals with more desirable combinations of genes. Traditionally selection has been performed on phenotype (traits) alone, but the development of methods for identifying allelic differences in genes (DNA markers) controlling trait variation has sparked considerable research activity among forestry scientists keen to increase accuracy by selection directly on genotype.

The relatively long generation times and poor juvenile-mature trait correlations in forest trees makes the application of marker-asissted selection (MAS) particularly attractive. Use of markers has the potential to deliver the following benefits to tree breeding:

- early selection in seedlings for traits expressed in mature trees.
- reduced phenotyping costs
- screening of large numbers of trees for desirable genotypes
- selection of favourable parental genotypes for crosses

The efficacy of MAS relies on identifying DNA markers which explain a significant proportion of additive variation in phenotypic traits. Most phenotypic traits of interest for tree breeding are characterised by continuous (quantitative) variation. Such traits are usually influenced by a number of genes of small effect, interacting with other genes and the environment.

## Pedigree markers

Initially, MAS research in trees employed genome-wide linkage analysis of progeny arrays (LANDER and BOTSTEIN 1994), an approach used widely in crop and animal breeding. By identifying patterns of co-segregation of polymorphic markers with complex trait and these studies revealed causative regions of the chromosome that were implicated in controlling phenotypic variation (quantitative trait loci or QTLs). The QTL approach is mainly used for within family MAS and a number of studies carried out in forest tree species including eucalypts have revealed QTL markers associated with important commercial traits (Butcher and Southerton 2007).

While QTL studies improved our understanding of the genetic control of complex traits, none of the markers identified 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. 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. Due to their outcrossing habit, large randomly mating populations, and long evolutionary histories, the genomes of tree are highly diverse. Additionally, there is a frequent tendency in tree genomes for SNPs within a gene to be weakly correlated with adjacent SNPs in unrelated individuals. The rapid breakdown in non-random association of SNPs or alleles, known as linkage disequilibrium (LD), is a major impediment to transfer markers between different pedigrees.

## Population markers

New marker discovery approaches that are population-based, rather than pedigree-based, are now feasible, largely due to the development of high throughput genomics (SNP discovery)

technology. Association mapping or linkage disequilibrium (LD) mapping is a population based approach which exploits the low LD in trees to detect marker:trait associations. In forest trees where LD typically breaks down within the length of a gene, candidate gene-based association studies are particularly attractive for detecting markers tightly linked to traits. No specialised populations are required for association studies as most of the current breeding populations are ideally suited. As the markers affecting a trait are detected using breeding populations, the results from association mapping are directly applicable across populations. While association studies are common in humans, there are a few published studies in plants. In the first association mapping study reported in forest trees, Thumma et al. (2005) identified polymorphisms in a lignin biosynthetic pathway gene that affect MFA. Recently in another association study, a number of SNP markers from different cell wall genes were found to be associated with a range of wood quality traits in *Pinus taeda* (GONZALEZ-MARTINEZ *et al.* 2007).

## Pulp yield marker discovery

Identification of genes and gene variants controlling wood quality traits is an important objective in many forest tree breeding programs, as small fraction of a percentage point improvements in traits such as pulp yield, can deliver large gains for a pulp mill. Genetic studies have shown that wood quality traits have moderate-to-high heritability compared to growth traits (COSTA E SILVA et al. 2009; RAYMOND 2002). High heritability (0.76) was estimated for NIR predicted cellulose (SCHIMLECK *et al.* 2004), while moderate heritability (0.50) was estimated for density (RAYMOND 2002). This suggests that genes control a sizeable proportion of the phenotypic variation observed in wood quality traits. In the Hottest 100 project we used association mapping to identify SNP markers that contribute to the genetic control of variation in pulp yield in *E. nitens.* The main objectives of the Hottest 100 project were to (a) measure several important wood quality traits in three *E. nitens* populations (b) identify SNPs in 100 candidate genes and genotype these across these populations and (c) identify and validate SNPs significantly associated with pulp yield related traits in different populations.

# Methodology

## Experimental populations

### Association population

The Forestry Tasmania used in the present study for association analysis was located at Meunna (elev. 270m, rain. 1400mm) in north-western Tasmania. The trial was planted in 1993 and included 420 open pollinated families (5 tree line plots) in 6 reps. Prior to the commencement of the Hottest 100 project in 2002 CSIRO had obtained pulp yield, cellulose and other wood quality data from 300 unrelated *E. nitens* trees which had been used in previous studies (THUMMA *et al.* 2005). In this project leaf and wood core samples were collected from an additional 120 trees in 2006 to increase the size of the population to 420 trees.

### Validation populations

Two validation populations were used in the current study. One was a Forestry Tasmania population located at Tarraleah (elev. 600m, rain. 1700mm) located in the Central Highlands of Tasmania. The trial was planted in August 1993 on the site of a failed *P. radiata* plantation and included 420 open pollinated families (5 tree line plots) in 6 reps. The trial had been measured for about 13 traits including survival, growth, form, bark thickness, transition to adult foliage, pilodyn penetration, flowering and possum damage. Wood samples were obtained from 520 trees from 150 families from a disk cut from a felled tree at 5.6m from the base. The other validation population was a Gunns Ltd second generation population located at Loudwater Rd near Hampshire (elev. 520m, rain. 1500mm) in northern Tasmania. The progeny trial was planted in October 1997 and contains 177 open pollinated families in two tree line plots with eight replicates.

## Wood phenotyping

### Meunna population

Near infrared reflectance (NIR) spectra were acquired on each Meunna core at 4 mm spatial resolution from bark to bark. Spectra were acquired using a Bruker MPA between 1,000 and 2,500 nm at 2 nm spectral resolution. The NIR spectra can be thought of as representing the chemical “fingerprint” of each core. By analysing the spectra of individual cores it is possible to rank, or classify, them into discrete populations based solely on their NIR response and not their wood quality traits. This is done via Principle Component Analysis (PCA). Alternately the spectral response can be regressed against a set of samples with known properties (eg chemical composition, stiffness, KPY). This is done using projection to latent structures (PLS) regression. Details of PCA and PLS can be found in Martens and Næs (1989). Prior to analysis a single core average spectrum was computed by averaging the individual spectra taken every 4 mm into a single spectrum. In order for the data to be compared to the measurements undertaken 4 years previously on the 300 Meunna trees, the outer ~4±~1 years growth was kept out of the averaging process. Existing calibration models for KPY and cellulose were used to predict KPY and cellulose for each core. Similarly a PCA was performed on the core-average spectra to determine the key variance between cores. This results in classifying the individual samples according to the variance in their chemical fingerprint.

NIR spectra were also obtained from wood meal obtained by grinding the 120 wood cores. To make the data comparable to the previous data from 300 trees collected in 2002, the

outer four annual rings were removed. Wood meal samples from sixty samples covering a range of cellulose values predicted from NIR spectra were used in measuring cellulose in the lab. Spectra were measured on the wood meal in diffuse reflectance mode in a scanning spectrometer (NIR Systems Inc., Model 5000). A ceramic standard was used as the instrument reference. Spectra were collected at 2-nm intervals over the 1100–2500 nm wavelength range. Fifty scans were acquired per sample and the results were averaged. The Vision® software was used to convert the data to the second-derivative mode using a segment width of 10 nm and a gap width of 0 nm.

Cellulose content and KPY for each wood-meal sample was predicted from the NIR spectra using previously developed NIR calibrations. Sixty samples (representing a wide range of NIR predicted cellulose contents) from Meunna were selected for analysis of crude cellulose content. Chemical assays were done on these samples according to the diglyme method of WALLIS *et al.* (1997): NIR calibrations were done according to the procedures outlined by SCHIMLECK *et al*. (2004) calibrations were developed within the Vision® software (version 2.51) using partial least-squares (PLS) regression with four cross validation segments and a maximum of 10 factors (vectors) as described in SCHIMLECK et al (2004).

The degree of fit of the NIR calibration to the chemical assay data was measured by the standard error of calibration (SEC) (WORKMAN 1992). The calibration resulted in a coefficient of determination ($R^2$) of 0.89 and a SEC of 0.65 for the cellulose calibration set. A further 10 samples (not included in the calibration sets) representing a wide range of NIR predicted cellulose contents were selected for chemical assay to test the predictability of NIR spectra analysis compared to chemically assayed cellulose content. High coefficient of determination ($R^2$, 0.91) and relatively small standard error of prediction (SEP, 0.97) indicated that the NIR calibrations could be used to accurately predict cellulose content.

**Tarraleah population**

Wood properties in 520 trees from 150 families were obtained from a disk cut by David Blackburn (CRC) from each felled tree at 5.6m from the base. The disk was cut into three segments and NIR spectra collected from pith to bark on the tangential face of the disk (Fig. 1). A range of solid wood traits on this trial including shrinkage, collapse, MOE and MOR were obtained.

Figure 1. Preparation of wood discs from Tarraleah for Phazir scans

Near infrared (NIR) spectra were acquired on each tree using a Polychromix Phazir between 950 and 1,800 nm at 8 nm spectral resolution. Spectra were recorded at five (5) points on the radial face of the wedges and averaged to provide a single spectrum per wedge. Multivariate data processing was performed using *The Unscrambler* v9.7 (CAMO A/S, Norway, www.camo.no). A calibration model was previously developed using Projection to Latent Structures (PLS) regression using spectra obtained on bark windows of solid wood stems. A total of 59 samples were used in the development of the calibration model, which was based on first derivative spectra (Savitsky-Golay, 7 point window). Two samples were removed from the calibration after identification as spectral outliers.

**Loudwater population**

Wood samples were collected from three trees from the best 140 families in April and May 2007 when the trees were aged 9.5 yrs, for a total of 420 trees. A 20mm wood disk was cut from 10, 30 and 70% of the merchantable log and the discs processed at Gunns Fibre Technology Laboratory at Ridgley for KPY at Kappa18 and basic density (See Fig. 2). Standing tree and log acoustic velocity readings were also taken for another FWPA project. A 20mm disc cut from breast height (1.3m) and a plinth of wood cut from the disc were used for NIR assessment at Clayton. Growth and form assessment was done in 2006.


Figure 2. Sampling of Gunns breeding population at Loudwater showing breast height disc for NIR-KPY analysis and three discs for laboratory-KPY analysis

NIR spectra were acquired using a Bruker MPA FT-NIR between 10,000 and 4,000 $cm^{-1}$ (1,000 and 2,500 nm) at 4 $cm^{-1}$ spectral resolution. A total of 409 samples were received with unique identifiers from the 420 samples that were originally collected. Core samples were translated radially (from bark end to pith) past a fibre optic bundle designed to acquire a 1 x 3 mm area. The linear translation unit was stepped in contiguous 5 mm increments with each 5 mm increment comprising 32 scans. All 5 mm increment scans for each core were then averaged to provide one single spectrum for each core. Note that no radial weighting of the spectra was applied. The model used for prediction was that based on radial cores. Models were built based on first derivative spectra (Savitsky-Golay, 15 point window, 2nd order

polynomial), using Projection to Latent Structures (PLS) regression with spectra and associated lab-KPY values from a series of trials over the previous 3-4 years. All multivariate data processing was performed using *The Unscrambler* v9.7

## Candidate genes

CSIRO had previously identified expressed sequence tags (ESTs) for over 170 candidate cell wall genes. In addition, several other eucalypt genes and ESTs were available in the published literature. Many of the CSIRO genes were identified in comparisons of wood properties and gene expression in *E. nitens* branches, where we observed major changes in cellulose, lignin and MFA (QIU *et al.* 2008). In other experiments gene expression was compared between high and low pulp yield trees using a Gunns Ltd *E. nitens* population (Bhuiyan et.al. unpublished data). Candidate genes from CSRO research and through searches of the published literature were prioritized on the basis of their known function and their pattern of expression to identify the most promising candidates for later SNP discovery and association studies. A central goal of the project was to obtain the full open reading frame (ORF) and promoter region of approximately 100 genes. Originally it was estimated that on average each gene would contain 25 common SNPs. Subsequent work suggests that this estimate was high and that the average number of common SNPs per gene is between 10 and 15. The complete open reading frame (ORF) of each gene was obtained using the gene walking technique (Clontech, Palo Alto, CA) in both directions from the EST sequence and by obtaining the full sequence of the cDNA clone from which the EST was derived. Promoter regions were obtained using the gene walking method. A few genes were selected from database searches (e.g. cellulose synthase and MYB1) and promoter regions obtained using the gene walking method. The completed ORF was used to carry out a BLAST search of the Arabidopsis genome in order to identify the Arabidopsis homologue for the gene.

## SNP discovery

DNA was isolated from mature leaves in methods commonly employed in our laboratory. Equal amounts of DNA from each of the 420 trees in the Meunna population were combined in a DNA bulk. Gene-specific PCR primers were designed matching sequences at the 3' and 5' end of each candidate gene. PCR reactions were carried out for each gene using the gene-specific primers, and bulked DNA as template. The PCR reaction included a long range Taq polymerase capable of amplifying gene fragments up to 10kb in length. PCR products were analysed by gel electrophoresis and were purified using Millipore plates. Equi-molar amounts of each amplified gene were pooled together and the DNA bulk sent to the Australian Genome Research Facility (AGRF) for 454 sequencing. The 454 sequence reads were used to assemble complete genomic reference sequences for each gene. The genomic reference sequence was then used as a scaffold against which all of the short 454sequence reads were aligned in order to identify SNPs. Alignments were made using the CLC genomics workbench software.

## SNP genotyping

### Meunna population

Approximately 200 base pairs of gene sequence spanning 906 SNPs identified in approximately 90 genes were sent to the AGRF. The high density of SNPs in *Eucalypts* results in the frequent occurrence of SNPs in the sequences flanking SNP positions (See Figure 3). Each of these SNPs was highlighted in the flanking sequences that were sent to the AGRF.

CCTTCCCAACCACCGCCATACCATCTGCTTTAAGCATTCCGATGA
GTCCCTGATCCACCGCCTTCTCACWGAGCCTTCCCGCYCTCCCTC
TTCTCGTCCC[G/A]CTTTCTCATATAAAGAAGTGAAAGAATACGA
GGATACTCCACTTGGGTATCGCCAAGAACTCATTGGGTCGCGAG
AAGATTGGCCAACATGATGGAATCC

Fig 3. Flanking sequences of the SNP EnCesA1m-4. The SNP under investigation is indicated by the red text and the flanking SNPs indicated by the back text. Y = C/T SNP; W = A/T SNP

Based on these sequences, the AGRF designed short DNA primers in optimal positions adjacent to each SNP. These DNA primers were used in the genotyping reaction. Primers were designed for 775 SNPs without requiring any modification. For 93 of the SNPs it was necessary to insert a "neutral" DNA base (inosine) where a flanking SNP occurred in the primer site. For the remaining 38 SNPs no suitable primer sites were identified and the SNPs could not be genotyped. DNA primers for genotyping 868 SNPs were synthesised by the AGRF. Genotyping was performed by the AGRF using the Sequenom genetic analysis system.

**Validation populations**

DNA isolated from 420 trees Loudwater population and 520 trees from Tarraleah population was sent to AGRF for genotyping. A total of 80 SNPs were genotyped on both populations based on the results from association analysis in the Meunna population.

## Association analyses

Hardy-Weinberg Equilibrium (HWE) of the SNP genotypic data was tested with 'Golden Helix" software. The GLM function of 'TASSEL'(BRADBURY *et al.* 2007) software was used to identify SNP-trait associations in the Meunna association population. All wood quality and growth traits were tested for associations. Significance parameters were based on 1000 permutation tests. Experiment-wise *p*-values were calculated based on the minimum *p*-value across all tests from permuted trait data compared to the original *p*-value for each marker. Adjusted *P*-values representing the percentage of times the permuted *P*-value was lower than the original *P*-value were used in testing for significant associations.

Marker-trait associations in validation populations were tested using a regression method under an additive model with 'PLINK' software (Purcell et al. 2007). Regression under an additive model is more informative as it tests for the direction of allelic effects. To test the allelic effects, the Meunna association population was re-analysed using the regression method under an additive model. Meta analysis was performed using 'PLINK' software (http://pngu.mgh.harvard.edu/purcell/plink/).

## Results:

## Cellulose and KPY predictions

To expand the size of the Meunna population wood cores were taken from an additional 120 families. Cellulose and KPY data predicted from NIR spectra taken from intact cores were used in the initial analyses. However, in the present study NIR spectra collected from wood meal sample were used to predict cellulose and KPY values. To make the data from 120 trees comparable to the data from the original 300 trees, last four annual rings were taken out before grinding the samples. Wood meal from sixty trees was also used to measure cellulose in the lab. This data was used in the calibration model to predict cellulose. The correlation ($r^2$) between cellulose measured in the lab and cellulose predicted from NIR was 0.85. Apart from cellulose, NIR data from the extra 120 samples was also used to predict KPY and lignin traits. The predicted trait data from the extra 120 trees was normalised by adjusting mean and variance so that it matched the data for the previous 300 trees. The normalised data was used in the analysis of marker trait associations presented in this study. Summary statistics for the NIR-predicted values for cellulose composition and KPY from the 420 trees in the Meunna population are shown in Table 1.

Table 1. Summary statistics of NIR-predicted cellulose and KPY for 420 trees growing at Meunna.

| | **Cellulose (%)** | **KPY (%)** |
|---|---|---|
| **Average** | 41.11 | 49.15 |
| **Min** | 38.29 | 45.49 |
| **Max** | 43.41 | 52.00 |
| **Range** | 5.11 | 6.51 |

Cellulose and KPY predictions were obtained for 420 trees in the Gunns Ltd Loudwater population and 520 trees in the Forestry Tasmania Tarraleah populations. Data for these trees is in the possession of the respective industry partners.

## Gene walking of candidate genes

In most cases the starting sequence information for candidate genes was a short expressed sequence tag (EST) obtained in prior research. Genome walking was used to obtain full genomic sequence data for each gene including in most cases between 500 and 1000 base pairs from the upstream promoter region. A full list of the candidate genes and a summary of the genomic sequences obtained for each gene in the project are presented in Appendix 1. The full DNA sequence of over 130 genes was obtained. This included the full gene and greater than 500bp of promoter and 5'UTR for 81 genes. The full gene and greater than 250 base pairs of promoter and 5'UTR were obtained for about 100 genes. The full gene sequence of a further 36 genes was also obtained. In total, DNA sequence for approximately 79,000bp of promoter and 5'UTR and 146,000bp of gene and 3' UTR were obtained. Average promoter length obtained was about 840bp and average gene length was about 1560bp.

Ninety five percent of the eucalypt genes selected for this study had significant homology to a gene in the Arabidopsis genome. Candidate gene selection was biased towards the selection of genes involved in lignin (13) and cellulose and hemi cellulose biosynthesis (19), as there is a higher probably that these genes will influence KPY. The largest category

of known genes included transcription factors and genes that activate other genes (25), as these genes are expected to have a larger impact on cell wall development. Other important classes of genes were those influencing cytoskeletal (actin and tubulin) development (12), including EgrTUB1, which was recently shown by our group to influence MFA in eucalypts (SPOKEVICIUS *et al.* 2007). Another important cell wall protein in this group is EnCOBL4A. EnCOBL4A (PY1) contains a SNP that associates with KPY (THUMMA *et al.* 2009). Interestingly, eucalypts possess two genes, EnCOBL4A and EnCOBL4B, which are strongly homologous to COBL4 in Arabidopsis. A large number of genes with an unknown function were included in the list of candidates. Five genes were included that have no homologue in other plants. Several more novel genes were included at the start of the project but as sequencing progressed several of these were found to have homologues in the Arabidopsis genome.

## SNP Discovery

DNA bulked from 300 trees from the Meunna population was used as a template to PCR amplify candidate genes. A total of 112 candidate genes were amplified using the bulked DNA (Figure 4).

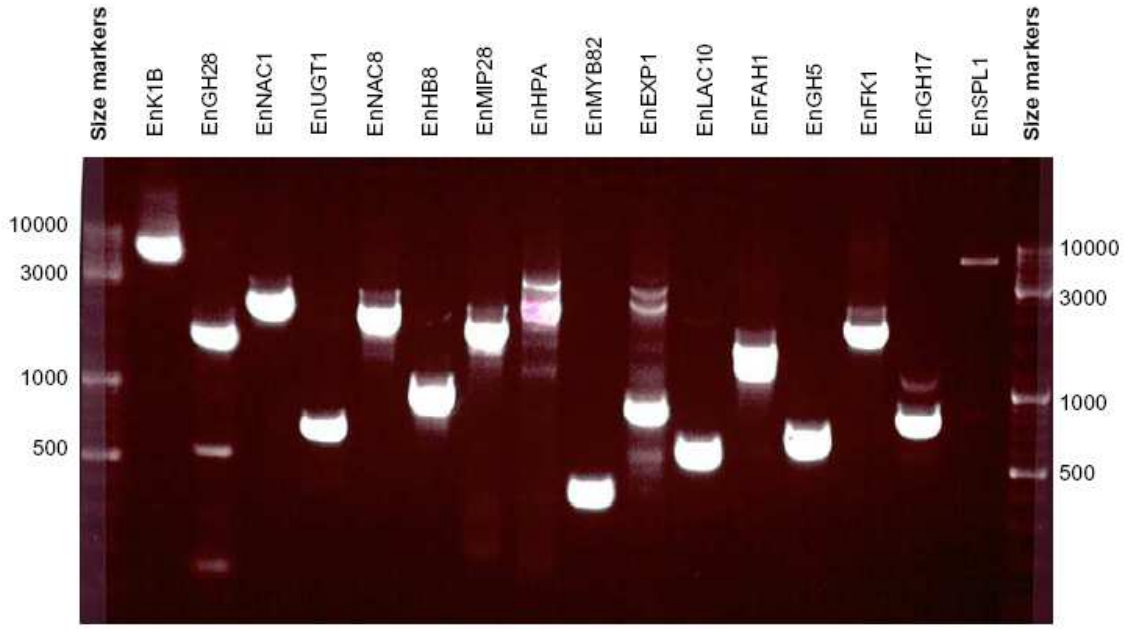


Figure 4. PCR amplification products from 16 genes separated by gel electrophoresis. The approximate size of the amplification products can be estimated by comparison with the size markers at the side of the gel (size indicated in base pairs).

Amplified DNA fragments were sequenced at the AGRF using the Life Sciences 454 FLX high throughput sequencing system. A total of 158,529 sequencing reads were generated from the pooled DNA. CLC genomics workbench software was used for analysis and mapping 454 sequence reads. The average length of each read was 223 bp. Reads were mapped using reference genomic sequence when sequence information was available. 'De novo' assembly was used for mapping the genes without prior sequence information. An example of a typical assembly using the CLC genomics workbench software is shown in Figure 5.

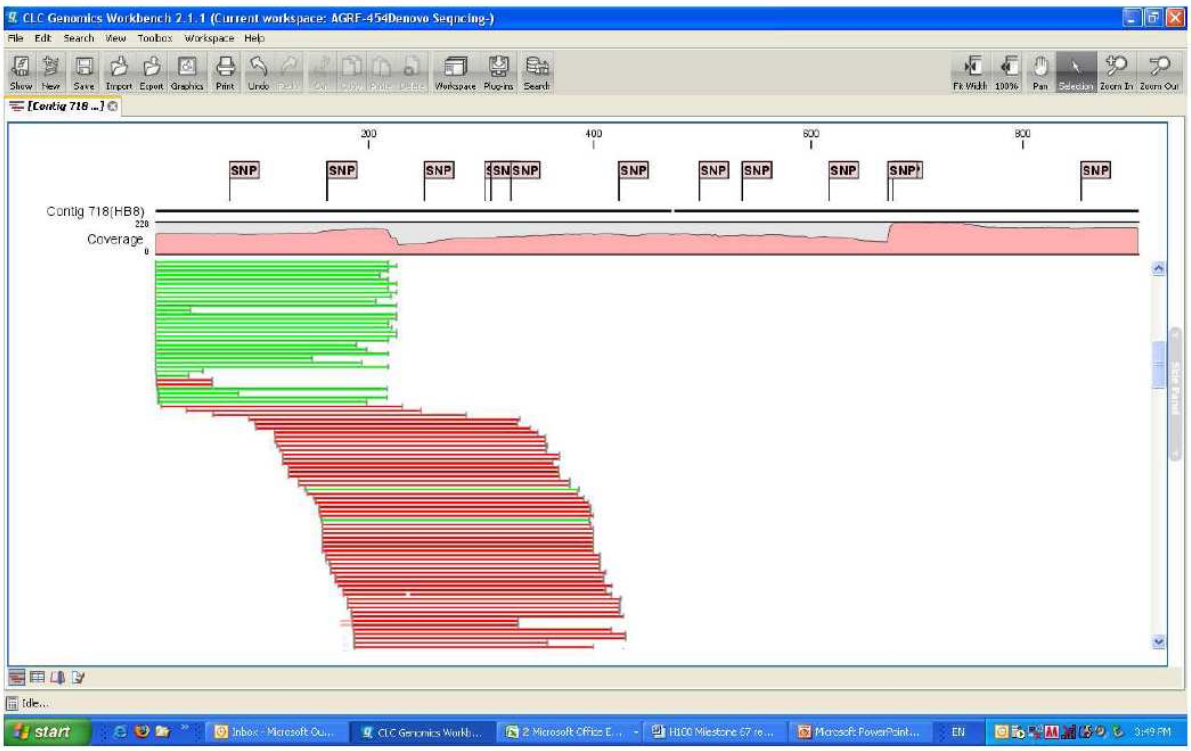


Figure 5: A screen shot of a typical 454 contig assembly in a portion of the EnHB8gene. SNP positions are indicated above the contig and 454 sequence reads appear below the contig as horizontal green or red lines. Coverage (number of sequence reads across a section of the gene) is indicated by the height of the pink region.

A total of 1116 common SNPs (frequency >10%) were identified among the 82 genes. The average number of sequence reads (coverage or depth of sequencing) used for SNP detection was 75 with a range from 11 to 260 sequences. On average 7 SNPs were identified per 1000 bp (1kb) of sequence and an average of 14 SNPs were identified in each gene. Only 2 out of 82 genes did not show any SNPs, reflecting the high level of polymorphism in eucalypt genes. We compared the SNPs identified using 454 sequencing with those previously identified using standard 'Sanger' sequencing in the *EnCCoMT2* gene. All but one SNP detected using Sanger sequencing were also identified with 454 sequencing. However, 454 sequencing detected 14 more common SNPs compared to Sanger sequencing. It is likely that the single SNP unique to Sanger sequencing was a sequencing artefact. Many more (14) low frequency (<15%) SNPs were detected with 454 sequencing compared to Sanger sequencing reflecting the much deeper coverage with 454 sequencing.

## SNP genotyping in Meunna population

A total 850 SNPs were selected from 80 genes sequenced using high throughput 454 sequencing. Sequence information from these SNPs was sent to the AGRF for genotyping using Sequenom's high throughput genotyping system. A total of 500 SNPs were successfully genotyped in 420 trees from Meunna population. It was expected that about 20-25% of the SNPs would fail due to competitive binding of the genotyping primers to other DNA sequences in the eucalypt genome. The failure rate of 40% was considerably higher than expected. A relatively high proportion of the failed SNPs returned a homozygous genotype for every tree (monomorphic SNP). This suggested that SNP selection from low coverage regions of 454 sequencing reads can yield false SNPs due to errors in 454 sequencing. However, there were several common SNPs (MAF>0.10) selected from regions with good coverage that yielded monomorphic genotypes.

## SNP-trait associations in Meunna

Phenotype information from seven wood quality traits was used to identify SNP-trait associations. Physical traits such as density and microfibril angle (MFA) were measured in 300 trees from the Meunna population using SilviScan. Chemical traits such as KPY, cellulose, total and klason lignin and extractives were measured in 420 trees from the Meunna population by NIR spectroscopy. Results from the single marker analysis using general linear model (GLM) are presented in Table 2. SNP-trait associations were corrected for multiple testing by permutation tests. A total of 42 SNPs from 26 genes were associated with one or more of seven wood quality traits.

The percentage of phenotypic variance explained by the SNPs ranged from 2.0 to 5.7%. Overall the number of SNP associations and the percentage of phenotypic variance explained by the SNPs are consistent with results from other similar large scale studies (ECKERT *et al.* 2009; GONZALEZ-MARTINEZ *et al.* 2007). Even though the percentage of variance explained by each SNP is small, the cumulative effect of all the significant SNPs associated with a trait is large. For example, ten SNPs from ten different genes explain about 32% of variance in cellulose content. In several cases, the same gene was associated with different traits but the associated SNPs were different. However, in some cases the same SNP was associated with different traits suggesting common bio-chemical pathways underlying the traits.

### Cellulose and KPY

*Cellulose*: Of all the traits tested, cellulose had the highest number of SNP associations. Twelve SNPs from ten genes were associated with cellulose. The percentage of phenotypic variance explained by the associated SNPs has ranged from 2.5 to 4.3%. Cumulative percentage of variance explained by ten independent SNPs is 32%.

*KPY*: Four SNPs were associated with KPY. The percentage of phenotypic variance explained by the associated SNPs ranged from 2.5 to 4.2%. The cumulative effect of all the significant SNPs is 13.5%. Two SNPs (EXGT1-SNP1 and Myb83-SNP6) are common between cellulose and KPY. However, most of the SNPs associated with cellulose were also associated with KPY at the nominal significance level of $P < 0.05$.

### Lignin traits

*Klason lignin*: Six SNPs from five genes were associated with klason lignin. EXGT1-SNP10 is common to klason lignin, cellulose and KPY. SAM1-SNP1 is associated with klason lignin and cellulose content. The cumulative percentage of variance in klason lignin explained by the SNPs is 16%.

*Total lignin*: Three SNPs from a LIM transcription factor are associated with total lignin. LIM-SNP33 is associated with both klason and total lignin.

*Extractives*: Seven SNPs are associated with extractives. The cumulative percentage of variance explained by these SNPs is 22.0%.

### Physical traits

*Microfibril angle (MFA)*: Seven SNPs are associated with MFA. The cumulative percentage of variance explained by the eight SNPs is 30%.

*Density*: Six SNPs from five genes were associated with density. One SNP CcoAMT2-SNP28 is associated with both density and KPY. The cumulative percentage of variance explained by the six independent SNPs is 24%.

Eighty SNPs, including all the significant SNPs (Table 2; Appendix 2), several SNPs associated with KPY and cellulose at the nominal level of significance and some significant SNPs identified in previous CSIRO research were genotyped in the Loudwater and Tarraleah populations in order to validate the SNP-trait association observed in the Meunna association population.

## SNP validation in the Loudwater and Tarraleah populations

### Genotyping

The 42 strongly associated SNPs (Table 2) and an additional 38 associated SNPs (Appendix 2) were genotyped across 420 trees (3 x 140 families) in the Gunn's Ltd second generation breeding population at Loudwater and 520 trees (3-5 x 150 families) in the Tarraleah population. Since all the SNPs had previously been genotyped successfully on the Meunna population, the success rate for genotyping on the 940 trees was expected to be high, and this was the case. The overall pass rate for genotyping was about 96%. About one third of the failures were due to poor quality or lost DNA in 13 trees. Patterns in the failures suggested that robotic handling errors most likely contributed to a large proportion of the remaining failures. The worst failure rate for a SNP was 17%. Almost all other SNPs had pass rates greater than 95%. In order to ensure DNA samples were in correct order, DNA was re-extracted from 2 trees from each of the 10 genotyping plates, and genotyped on the last plate. All 20 genotypes of the repeated trees matched the genotypes for the same trees on the other plates. Overall, the quality of the genotyping data from the two populations was very high.

**Table 2. SNP associations with wood quality traits in *E. nitens***

| Cellulose | | | |
|---|---|---|---|
| **Gene** | **SNP** | ***P*- Adjusted[a]** | **$R^2$(%)[b]** |
| ACT1 | SNP-4 | 0.0250 | 2.78 |
| CNX1 | SNP-1 | 0.024 | 2.89 |
| EXGT1 | SNP-10 | 0.0025 | 3.06 |
| PIP2 | SNP-5 | 0.05 | 2.50 |
| PIP2 | SNP-14 | 0.05 | 2.90 |
| RAC7 | SNP-142 | 0.025 | 3.14 |
| TUB1 | Indel-2 | 0.025 | 2.80 |
| UP3 | SNP-20 | 0.006 | 3.32 |
| UP3 | SNP-21 | 0.0005 | 4.27 |
| GH17 | SNP-2 | 0.003 | 2.96 |
| MYB83 | SNP6 | 0.0015 | 3.77 |
| SAM1 | SNP-1 | 0.0005 | 4.18 |
| **KPY** | | | |
| EXGT1 | SNP-10 | 0.018 | 2.88 |
| PEL3 | SNP-5 | 0.03 | 2.49 |
| CCoAMT2 | SNP-28 | 0.0005 | 3.86 |
| MYB83 | SNP-6 | 0.0005 | 4.22 |
| **Extractives** | | | |
| 4CL | SNP-7 | 0.0095 | 3.12 |
| NOV11 | SNP-3 | 0.0025 | 3.37 |
| GH17 | SNP-1 | 0.0035 | 3.32 |
| CCoAMT2 | SNP-23 | 0.0005 | 4.02 |
| CIP7A | SNP-11 | 0.012 | 2.94 |
| Dehydrin | SNP-5 | 0.03 | 2.71 |
| EXGT1 | SNP-7 | 0.02 | 2.81 |
| MYB1 | SNP-4 | 0.005 | 3.16 |
| UP3 | SNP-4 | 0.015 | 3.31 |
| **klason-lignin** | | | |
| EXGT1 | SNP-10 | 0.04 | 2.55 |
| LIM1 | SNP-18 | 0.01 | 2.76 |
| LIM1 | SNP-33 | 0.0009 | 3.68 |
| SAM1 | SNP-1 | 0.02 | 2.61 |
| SUS2 | SNP-6 | 0.02 | 2.00 |
| UP8 | SNP-20 | 0.03 | 2.00 |
| **Total lignin** | | | |
| LIM1 | SNP-20 | 0.05 | 2.25 |
| LIM1 | SNP-21 | 0.005 | 3.05 |
| LIM1 | SNP-33 | 0.009 | 2.94 |
| **Micro fibril angle (MFA)** | | | |
| CIP7A | SNP-8 | 0.04 | 3.50 |
| GH28A | SNP-2 | 0.007 | 4.30 |
| LIM1 | SNP-34 | 0.05 | 3.70 |
| MYB83 | SNP-3 | 0.0004 | 5.17 |
| RABA1 | SNP-6 | 0.03 | 3.53 |
| RAC7 | SNP-3 | 0.006 | 4.69 |
| ZIP | SNP-16 | 0.002 | 4.86 |
| **Density** | | | |
| EXGT1 | SNP-22 | 0.05 | 3.24 |
| HB1 | SNP-3 | 0.04 | 3.69 |
| TUB1 | SNP-2(Pr) | 0.007 | 4.30 |
| XLYT1 | SNP-35 | 0.02 | 4.10 |
| XLYT1 | SNP-47 | 0.03 | 3.80 |
| CCoAMT2 | SNP-28 | 0.0005 | 5.72 |

[a]*P*-values are adjusted for multiple testing. [b]$R^2$ is the percentage of variance explained by the SNP. SNPs associated with multiple traits are underlined

## SNP validation in the Loudwater and Tarraleah populations…cont'd

### SNP trait associations

Single marker association analysis using a regression method under an additive model was carried out with the genotype data from the Loudwater and Tarraleah populations. To make the results from the validation population comparable to the Meunna association population, data from the Meunna association population was re-analysed using a regression method under an additive mode. Of the 17 SNPs associated with cellulose in the Meunna population, 11 (65%) were also significantly associated with cellulose content in one or both of the validation populations at $p<0.05$ (Table 3). We estimated that 2 SNPs will show significant effects by chance when 17 randomly selected SNPs were tested. This strongly suggests that the majority of the associations we detected in all populations are real. This is further supported by the detection of 13 associations with cellulose among the 63 other non-cellulose SNPs when we predicted chance alone would yield 6 significant SNPs (Table3). In other words the rate of cellulose SNPs validating was 2.5 times higher than the rate of non-cellulose SNPs.

Table 3. Validation of Meunna cellulose SNPs in Loudwater and/or Tarraleahpopulations

| | **Meunna** | **Loudwater and/or Tarraleah** | |
|---|---|---|---|
| **SNPs*** | **Actual** | **Null prediction#** | **Actual** |
| Cellulose | 17 | 2 | 11 |
| Non-cellulose | 63 | 6 | 13 |

* Significant at $p < 0.05$

# Predicted number of significant associations at either extreme due to chance

Of the 7 SNPs associated with KPY in the Meunna population, 5 (71%) were also significantly associated with KPY in one or both of the validation populations at $p<0.05$ (Table 6). We estimated that 7 randomly selected SNPs would by chance give associations in 1 SNP in Loudwater or Tarraleah. There were 13 associations with KPY among the 73 other non-KPY SNPs when we predicted chance alone would yield 7 significant SNPs (Table 4). The rate of KPY SNPs validating was also 2.5 times higher than the rate of non-KPY SNPs.

Table 4. Validation of Meunna KPY SNPs in Loudwater and/or Tarraleahpopulations

| | **Meunna** | **Loudwater and/or Tarraleah** | |
|---|---|---|---|
| **SNPs*** | **Actual** | **Null prediction#** | **Actual** |
| KPY | 7 | 1 | 5 |
| Non-KPY | 73 | 7 | 13 |

* Significant at $p < 0.05$

# Predicted number of significant associations at either extreme due to chance

SNP associations with cellulose and KPY are presented in tables 5 and 6. There were several SNPs affecting both cellulose and KPY, reflecting the high phenotypic correlation between the traits. A SNP from RAC7 (RAC7-1) was significantly associated with cellulose and KPY (in Tarraleah significance of this SNP with KPY was <0.1) with allelic effects in the

same direction in all three populations. For cellulose, allelic effects from five genes were in the same direction and for KPY allelic effects from four genes were in same direction in at least two populations. For example the minor allele of CNX-11 was associated with low trait values of cellulose in both Tarraleah and Loudwater populations while the minor allele of EXP-10 was associated with high trait values in Meunna and Loudwater populations. However, allelic effects from eight genes associated with cellulose and four genes associated with KPY were opposite i.e., one allele was associated with higher trait value in one population while the other allele was associated with lower trait values in other population. For example, the minor allele of the EnCAD-2 SNP has a negative effect on cellulose in Meunna and Tarraleah, but a positive effect on the trait in Loudwater. Similarly, the minor allele of theEnCOBL4-7 SNP has a negative effect on cellulose in the Meunna population but a positive effect on the trait in the other two populations.

Table 5. Validation of SNPs associated with cellulose across all three populations. The T value indicates the direction of the allelic effect (with reference to minor allele) on the trait. *P*-values were estimated under an additive model.

| | | Meunna | | | Tarraleah | | | Loudwater | |
|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **SNP** | **T** | ***P* (add)** | | **T** | ***P* (add)** | | **T** | ***P* (add)** |
| RAC7 | SNP-1 | -3.001 | 0.002 | | -2.098 | 0.036 | | -3.002 | 0.003 |
| CCoAMT2 | SNP-23 | -2.698 | 0.007 | | 0.233 | 0.815 | | -1.899 | 0.053 |
| CNX1 | SNP-01 | 1.64 | 0.110 | | 2.086 | 0.038 | | 0.325 | 0.745 |
| EXP | SNP-10 | 0.12 | 0.010 | | -1.166 | 0.244 | | 2.935 | 0.004 |
| SAM1 | SNP-01 | -4.046 | 0.000 | | -1.261 | 0.060 | | -1.12 | 0.264 |
| EnCOBL4 | SNP-07 | -0.11 | 0.030 | | 2.003 | 0.046 | | 2.1 | 0.036 |
| EnCAD | SNP-02 | -0.11 | 0.040 | | -1.184 | 0.030 | | 2.183 | 0.029 |
| EnDHY | SNP-05 | -3.067 | 0.002 | | 0.909 | 0.364 | | 3.105 | 0.002 |
| UP3 | SNP-19 | -2.705 | 0.006 | | 2.376 | 0.018 | | 0.9399 | 0.347 |
| UP3 | SNP-20 | -3.347 | 0.001 | | 2.425 | 0.016 | | 1.217 | 0.224 |
| UP3 | SNP-21 | -3.787 | 0.000 | | 2.327 | 0.020 | | 1.506 | 0.133 |
| EnEXGT1 | SNP-01 | 2.505 | 0.013 | | -2.625 | 0.009 | | 0.9033 | 0.366 |
| EnCNX1 | SNP-11 | 0.4367 | 0.660 | | -1.552 | 0.016 | | -5.02 | 0.000 |
| EnCOBL4 | SNP-02 | 0.16 | 0.100 | | -3.474 | 0.001 | | -0.5653 | 0.572 |
| EXGT1 | SNP-10 | -0.022 | 0.650 | | 2.059 | 0.040 | | -0.9799 | 0.327 |
| EnCSLA9 | SNP-01 | 0.05 | 0.330 | | -2.531 | 0.012 | | 3.101 | 0.002 |
| MYB83 | SNP-06 | -2.531 | 0.012 | | 1.277 | 0.202 | | 1.56 | 0.119 |
| **Additional associations** | | | | | | | | | |
| | **SNP** | **T** | ***P* (add)** | **SNP** | **T** | ***P* (add)** | **SNP** | **T** | ***P* (add)** |
| | ACT1-04 | 3.085 | 0.003 | CAD07 | 2.651 | 0.008 | EXGT1-26 | -2.385 | 0.017 |
| | PIP2-05 | -2.606 | 0.010 | EXGT1-11 | 2.163 | 0.031 | EXGT1-14 | -2.1 | 0.036 |
| | PIP2-15 | 2.684 | 0.008 | EXGT1-22 | -2.118 | 0.034 | TUB1-12 | 2.006 | 0.045 |
| | TUB1 Indel-02 | -2.14 | 0.030 | EXGT1-07 | 2.397 | 0.016 | HB1-03 | -2.122 | 0.034 |
| | GH17-02 | 2.558 | 0.010 | TUB1Pr02 | 2.746 | 0.006 | UP10-01 | -2.133 | 0.033 |
| | | | | | | | GH28A-02 | -3.896 | 0.000 |

[a]Allelic effects in the same direction in all three populations are shown in red. Allelic effects in the same direction in at least two populations are shown in green. Allelic effects in the opposite direction are shown in blue. Significant SNPs in a single population are shown in brown and non significant SNPs are shown in black colour.

**Table 6.** Validation of SNPs associated with KPY across all three populations. The T value (+/-) indicates the direction of the allelic effect (with reference to the minor allele) on the trait.

| | Meunna | | | Tarraleah | | | Loudwater | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **SNP** | **T** | ***P* (add)** | | **T** | ***P* (add)** | | **T** | ***P* (add)** |
| C7 | SNP-01 | -2.644 | 0.008 | | -1.779 | 0.076 | | -2.203 | 0.028 |
| EXGT1 | SNP-10 | -0.931 | 0.352 | | 2.228 | 0.026 | | -1.092 | 0.276 |
| EXGT | SNP-26 | -0.502 | 0.615 | | 0.437 | 0.663 | | -2.532 | 0.012 |
| CNX | SNP-01 | 1.926 | 0.050 | | 2.050 | 0.041 | | 1.294 | 0.197 |
| CCoAMT2 | SNP-23 | -2.311 | 0.021 | | -0.034 | 0.973 | | -1.759 | 0.079 |
| EXP | SNP-10 | 0.120 | 0.020 | | -2.084 | 0.038 | | 2.436 | 0.015 |
| EnCOBL4 | SNP-02 | 0.160 | 0.002 | | -2.243 | 0.025 | | -0.128 | 0.898 |
| MYB83 | SNP-02 | -0.100 | 0.080 | | -0.764 | 0.010 | | -2.108 | 0.036 |
| CSLA9 | SNP-01 | 0.090 | 0.090 | | -2.603 | 0.010 | | 2.871 | 0.004 |
| EXGT1 | SNP-07 | -0.254 | 0.800 | | 2.083 | 0.038 | | -1.977 | 0.049 |
| EnCAD | SNP-07 | 0.050 | 0.350 | | 2.084 | 0.038 | | 1.745 | 0.082 |
| EnCAD | SNP-02 | -0.010 | 0.850 | | -1.874 | 0.062 | | 1.950 | 0.052 |
| CNX | SNP-11 | -0.853 | 0.394 | | -1.248 | 0.213 | | -4.147 | 0.000 |
| **Additional associations** | | | | | | | | | |
| | **SNP** | **T** | ***P* (add)** | **SNP** | **T** | ***P* (add)** | **SNP** | **T** | ***P* (add)** |
| | PEL-05 | -3.140 | 0.002 | EXGT1-11 | 2.154 | 0.032 | GH28A-02 | -3.138 | 0.002 |
| | MYB83-06 | -1.992 | 0.040 | EXGT1-01 | -3.556 | 0.000 | Expa1-10 | 2.449 | 0.015 |
| | | | | SUS2-06 | -2.090 | 0.037 | NOV11-03 | 2.369 | 0.018 |
| | | | | UXS5-03 | -2.223 | 0.027 | UP10-01 | -2.283 | 0.023 |
| | | | | | | | EXGT1-14 | -2.131 | 0.034 |
| | | | | | | | DHY-05 | 2.102 | 0.036 |
| | | | | | | | HB1-03 | -1.979 | 0.048 |

[a]Allelic effects in same direction in all three populations are shown in red colour. Allelic effects in same direction in at least two populations are shown in green colour. Allelic effects in opposite direction are shown in blue colour. Significant SNPs in one population are shown in brown and non significant SNP are shown in black colour.

## Meta analysis

Results presented in the tables 5 and 6 were based on analysing each population separately. Single population analysis may lack the power to detect associations due to the smaller size of the populations. Combining the data from the three populations (meta analysis) significantly increases the power to detect SNP-trait associations. Meta analysis of several studies therefore provides stronger support for the effect of a SNP variant compared to a highly significant result from a single study (MUNAFÒ and FLINT 2004). Two types of tests, fixed effect and random effect are performed in meta analysis. In the fixed effect test the true effect of an allele is tested by combining the data from all populations. In the random effect test heterogeneity between the studies is considered when testing the effect of an allele. The random effect test is therefore more conservative compared to the fixed effect test.

Meta analysis of the 80 SNPs genotyped across the three populations detected several significant associations for cellulose and KPY (Table 7). Fourteen SNPs showed a significant effect on cellulose while twelve SNPs had a significant effect on KPY. For each trait, eight of the associations were significant at $P< 0.01$. Eleven out of twelve SNPs associated with KPY were also associated with cellulose. These SNPs showed an effect on the trait in the same direction in all three populations and were usually close to the $p<0.05$ level of significance in individual populations. This demonstrates the increased power of meta analysis to detect significant associations compared to single study analysis. All of the SNPs except for two were significantly associated with both cellulose and KPY under fixed and random effect models indicating the robustness of the associations.

## Marker-assisted selection for KPY

Using the KPY markers identified in the meta-analysis, we used the SNP genotype information to identify trees containing the desirable alleles. We identified 16 trees that were homozygous for 6 of the top 9 KPY SNPs (bold SNPs in Table 7). The predicted improvement in KPY obtained by selecting favourable alleles from these six SNPs are presented in Table 8. Mean KPY of the 16 selected trees was 51.2% while the mean KPY of the combined population was 50.1%. Student's t-test indicated that the mean KPY of the sixteen selected trees was significantly higher than the population mean. There was also a significant improvement in the DBH of the selected trees compared to the total population. The mean DBH of the marker selected trees was 25.6 cm while the mean DBH of the population was 23.3 cm. This result strongly suggests that the marker effects are largely additive.

Table 7. Meta analysis of the SNPs associated with cellulose and KPY.

| Cellulose | SNP | N | P | P(R) | Q | I |
|---|---|---|---|---|---|---|
| | **RAC7-01** | 3 | 0.00 | 0.00 | 0.93 | 0.0 |
| | **SAM1-01** | 3 | 0.00 | 0.01 | 0.10 | 57.4 |
| | **CNX1-11** | 3 | 0.00 | 0.16 | 0.02 | 75.7 |
| | ACT1-05 | 3 | 0.00 | 0.00 | 0.30 | 16.1 |
| | **CCoAMT2-23** | 3 | 0.00 | 0.04 | 0.20 | 38.2 |
| | CAD-07 | 3 | 0.01 | 0.01 | 0.39 | 0.0 |
| | CIP7A-04 | 3 | 0.01 | 0.01 | 0.36 | 1.5 |
| | **TUB1-12** | 3 | 0.01 | 0.01 | 0.72 | 0.0 |
| | SUS2-06 | 3 | 0.03 | 0.03 | 0.45 | 0.0 |
| | ZIP-16 | 3 | 0.04 | 0.04 | 0.89 | 0.0 |
| | UXS5-03 | 3 | 0.04 | 0.04 | 0.64 | 0.0 |
| | **MYB83-02** | 3 | 0.04 | 0.04 | 0.89 | 0.0 |
| | CNX1-01 | 3 | 0.04 | 0.08 | 0.22 | 33.7 |
| | Expa1-10 | 3 | 0.04 | 0.04 | 0.42 | 0.0 |
| **KPY** | | | | | | |
| | **CNX1-11** | 3 | 0.00 | 0.00 | 0.22 | 34.7 |
| | **RAC7-01** | 3 | 0.00 | 0.00 | 0.87 | 0.0 |
| | CNX1-01 | 3 | 0.00 | 0.00 | 0.59 | 0.0 |
| | **MYB83-02** | 3 | 0.00 | 0.00 | 0.97 | 0.0 |
| | SUS2-06 | 3 | 0.00 | 0.00 | 0.66 | 0.0 |
| | **SAM1-01** | 3 | 0.01 | 0.01 | 0.66 | 0.0 |
| | CAD-07 | 3 | 0.01 | 0.01 | 0.71 | 0.0 |
| | **CCoAMT2-23** | 3 | 0.01 | 0.03 | 0.30 | 16.1 |
| | **TUB1-12** | 3 | 0.02 | 0.02 | 0.79 | 0.0 |
| | Expa1-10 | 3 | 0.02 | 0.02 | 0.44 | 0.0 |
| | Pel3-05 | 3 | 0.03 | 0.28 | 0.03 | 72.4 |
| | ACT1-5 | 3 | 0.03 | 0.03 | 0.50 | 0.0 |

N – Number of individual populations
P – Fixed effect p-value (Test for true effect size)
P(R) – Random effect p-value (test of range and distribution of effect sizes in all populations)
Q – Chi-squared test for heterogeneity between studies
I – Between study heterogeneity index (0- no heterogeneity; >0.50 significant heterogeneity).

Table 8. Trait means of total population and selected trees with favourable alleles from six SNPs. 95% confidence intervals (CI) are shown in brackets.

| | **KPY** | **DBH** |
|---|---|---|
| Total population (1247 trees) | 50.1(±0.08) | 23.3(±0.27) |
| Selected trees (16 trees) | 51.2(±0.82) | 25.6(±2.1) |
| *P*- Student's t-test | 0.02 | 0.04 |

## Discussion

Several SNPs were identified in the study that are associated with cellulose and KPY and which were validated in at least one of two validation populations. SNP1 from RAC7 was consistently associated with cellulose and KPY in all three experimental populations with the allele effect in the same direction in all three populations. Similarly SNP23 from CCoAMT2 was significantly associated with cellulose and KPY in the Meunna and Loudwater populations with allelic effects in the same direction. There were other SNPs which were associated with either cellulose or KPY in at least two populations with allelic effects in the same direction. The power of detecting moderate to small genetic effects in individual populations is limited due to the smaller sizes of the populations. However, combining the data from individual populations in a meta analysis increases the power to detect the associations (EVANGELOU *et al.* 2007). Meta analysis with all three populations increased the number of SNPs associated with both cellulose and KPY from one to eleven demonstrating the increased power to detect significant associations. This is one of the first studies in forest trees to detect SNPs associated with wood traits using meta analysis. This was possible as there were 80 SNPs genotyped across the three populations.

An important factor contributing to the success in identifying significant associations is likely to be the quality of the trait data on each population. Accuracy in trait measurements significantly increases the power to detect associations. The fact that a high proportion of SNP associations were validated in one or both of the validation populations strongly suggests that the quality of the trait data obtained for each population was high. Wood quality traits were measured in disks sampled at 5.6 m in the Tarraleah population whereas they were measured at 1.3m in the other two populations. Despite this sampling difference, many associations were validated in the Tarraleah population, suggesting a strong correlation between measurements at both heights. However, in future studies it may be advisable to measure wood traits on samples collected from the same position on the tree.

Tests of the genetic gain derived from selecting for multiple SNPs were carried out by identifying trees homozygous for some of the most significant SNPs. Sixteen trees were identified that were homozygous for 6 of the most significant SNPs identified in the meta analysis. These trees had 2.2% higher KPY (51.2%) compared to the population mean (50.1%). This improvement in KPY was accompanied by a 10% increase in the DBH. Growth is a complex trait controlled by a number of small effect genes. Growth exhibits substantially more genotype x environment (GxE) interaction compared to wood quality traits. A number of studies have found that cellulose and pulp yield traits are positively correlated with growth (APIOLAZA *et al.* 2005; KUBE *et al.* 2001; QUANG *et al.* 2009). As growth is a complex trait subject to significant GxE interactions, selecting for cellulose and pulp yield traits should indirectly improve tree growth.

While a number of SNPs showed association with cellulose and KPY with allelic effects in the same direction, there are a number of SNPs where the allele effect is reverse in one or both of the other populations. This phenomenon has been observed in a number of studies in humans and is referred to as the "flip-flop" effect. Clarke and Cardon (2009) showed that unless a flip flop is genuine, the likelihood of observing a significant flip flop effect in different populations is negligible. Given the high proportion of significant flip flop SNPs in our study, most of the flip-flop SNPs should therefore be genuine. These flip-flop effects may represent some of the first molecular evidence of GxE. However, the mechanism behind these interactions which causes the flip-flop effects is not clear. The most obvious environmental gradient that the genotypes may be interacting with is climate at the three trial sites. The Loudwater and Tarraleah validation populations are located at elevations of 520m and 600m respectively, while the Meunna population is located at 270m, thus the validation populations are both growing on colder sites. The largest contrast in climate occurs between

Meunna and Tarraleah due to the large difference in elevation and the central inland location of the Tarraleah trial. The largest number of flip flops occurred between the Meunna and Tarraleah populations. Experiments involving allelic expression and methylation analysis are underway to explore the functional basis of these flip-flop effects.

**Conclusion**

The results of the Hottest 100 project strongly suggest that association studies can reveal molecular markers that can be used for MAS in trees. More than 900 SNPs were identified by sequencing 100 cell wall genes. Five hundred of these SNPs were successfully genotyped in the Meunna association population. Association analyses revealed 80 SNPs significantly associated with different wood quality traits. Eleven out of 17 SNPs associated with cellulose and 5 out of 7 SNPs associated with KPY were validated in at least one of the two validation populations. Meta analysis revealed eleven SNPs significantly associated with both cellulose and KPY. Tests of genetic gain based on six highly significant SNPs revealed 2.2% improvement in KPY along with a 10% improvement in growth.

## Recommendations

The study identified 11 SNPs associated with both cellulose and KPY. Association studies in a much larger number of genes is recommended as this is very likely to reveal many more SNP markers for capturing a much larger proportion of the phenotypic variation in KPY and other wood traits. Selecting for six of the most significant SNPs improves both KPY and DBH. This suggests that expanded association studies will also deliver markers that will predict trees with higher growth rates. These 11 significant SNPs could be used in breeding programs to improve the efficiency of breeding for cellulose and KPY. The fact that they are stable across the three populations, which span a considerable climatic range in the plantation estate, strongly suggests that they will deliver significant gains for industry. An expanded study with many more genes should identify a much larger proportion of the alleles affecting wood quality traits.

The study also revealed the somewhat surprising evidence of significant environmental effects on cell wall genes. This suggests that validation of SNP markers in a wider range of environments is advisable. It is recommended that the KPY SNPs should be genotyped on a large number of trees at harvest age across a range of environments. A small number of trees with favourable, unfavourable and average predicted KPY based on the markers could then be phenotyped for KPY and the mean of each group compared. This test would mimic a genetic gain trial and SNPs validated through these tests would be expected to be environmentally robust. The detection of SNPs responding to environment in our study also suggests that further significant gains in plantation productivity may be achieved by breeding for specific environments. Further research into the nature of the environmental signal the SNP genotypes are responding may reveal further opportunities to breed for particular plantation regions.

Operationally, it is recommended that the markers be used by industry to screen existing clonal seed orchards to identify inferior genotypes for culling and to screen future genotypes entering the orchards. As more markers become available in future research, wider screening in provenance trials is recommended to identify elite individuals and parents for controlled crosses to pyramid desirable alleles into future generations.

## References


APIOLAZA, L. A., C. A. RAYMOND and B. J. YEO, 2005 Genetic variation of physical and chemical wood properties of *Eucalyptus globulus*. Silvae Genet. **54:** 160-166.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAM-DOSS *et al.*, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics**:** btm308.

CLARKE, G. M., and L. R. CARDON, 2009 Aspects of observing and claiming allele flips in association studies. Genet. Epidemiol. **34:** 266-274.

COSTA E SILVA, J., N. BORRALHO, J. ARAÚJO, R. VAILLANCOURT and B. POTTS, 2009 Genetic parameters for growth, wood density and pulp yield in *Eucalyptus globulus*. Tree Genetics & Genomes **5:** 291-305.

ECKERT, A. J., A. D. BOWER, J. L. WEGRZYN, B. PANDE, K. D. JERMSTAD *et al.*, 2009 Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits. Genetics **182:** 1289-1302.

EVANGELOU, E., D. M. MARAGANORE and J. P. A. IOANNIDIS, 2007 Meta-Analysis in Genome-Wide Association Datasets: Strategies and Application in Parkinson Disease. PLoS ONE **2:** e196.

GONZALEZ-MARTINEZ, S. C., N. C. WHEELER, E. ERSOZ, C. D. NELSON and D. B. NEALE, 2007 Association Genetics in Pinus taeda L. I. Wood Property Traits. Genetics **175:** 399-409.

KUBE, P. D., C. A. RAYMOND and P. W. BANHAM, 2001 Genetic parameters for diameter, basic density, cellulose content and fibre properties for *Eucalyptus nitens*. Forest Genetics **8:** 285-294.

LANDER, E. S., and D. BOTSTEIN, 1994 Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps (Vol 121, Pg 185, 1989). Genetics **136:** 705-705.

MARTENS, H., and T. NÆS, 1989 *Multivariate calibration*. John Wiley and Sons, Chichester, UK.

MUNAFÒ, M. R., and J. FLINT, 2004 Meta-analysis of genetic association studies. Trends Genet. **20:** 439-444.

PURCELL, S., B. NEALE, K. TODD-BROWN, L. THOMAS, M. A. R. FERREIRA *et al.*, 2007 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics **81:** 559-575.

QIU, D., I. W. WILSON, S. GAN, R. WASHUSEN, G. F. MORAN *et al.*, 2008 Gene expression in *Eucalyptus* branch wood with marked variation in cellulose microfibril orientation and lacking G-layers. New Phytol. **179:** 94-103.

QUANG, T., N. KIEN, S. VON ARNOLD, G. JANSSON, H. THINH *et al.*, 2009 Relationship of wood composition to growth traits of selected open-pollinated families of Eucalyptus urophylla from a progeny trial in Vietnam. New Forests **39:** 301-312.

RAYMOND, C. A., 2002 Genetics of *Eucalyptus* wood properties. Annals of Forest Science **59:** 525-531.

SCHIMLECK, L. R., K. P.S. and C. A. RAYMOND, 2004 Genetic improvement of kraft pulp yield in *Eucalyptus nitens* using cellulose content determined by near infrared spectroscopy. Candian Journal of Forest Research **34:** 2363-2370.

SPOKEVICIUS, A. V., S. G. SOUTHERTON, C. P. MACMILLAN, D. QIU, S. GAN *et al.*, 2007 Beta-tubulin affects cellulose microfibril orientation in plant secondary fibre cell walls. Plant J. **51:** 717-726.

THUMMA, B. R., B. A. MATHESON, D. ZHANG, C. MEESKE, R. MEDER *et al.*, 2009 Identification of a *Cis*-acting Regulatory Polymorphism in a Eucalypt *Cobra*-like Gene Affecting Cellulose Content. Genetics **183:** (in press).

THUMMA, B. R., M. R. NOLAN, R. EVANS and G. F. MORAN, 2005 Polymorphisms in *cinnamoyl CoA reductase* (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics **171:** 1257-1265.

## Acknowledgements

We gratefully acknowledge the work of David Blackburn in the sampling and phenotyping of the Tarraleah population.

Appendix 1. Details of the candidate genes sequenced.

| Gene number | Putative gene identity | *E. nitens* clone name | Arabidopsis or eucalypt homologue | Promoter+ 5' UTR | Gene + 3' UTR |
|---|---|---|---|---|---|
| | | | | | |
| | **Cellulose/hemicellulose biosynthesis** | | | | |
| 1 | UDP-xylose synthase | **EnUXS5** | AT3G46440.2 | 800 | 963 |
| 2 | Beta-phosphoglucomutase | **EnHDH1** | AT2G38740.1 | 1128 | 2035 |
| 3 | Fructokinase | **EnFK1** | AT3G59480.1 | 400 | 1171 |
| 4 | Cellulose synthase (CesA1) | **EnCesA1** | EgrCesA1 | 771 | 3341 |
| 5 | Cellulose synthase (CesA2) | **EnCesA5** | AT5G17420.1 | 57 | 3377 |
| 6 | Cellulose synthase (CesA3) | **EnCesA3** | EgrCesA3 | 828 | 3453 |
| 7 | CesA like 2 | **EnCSLA2** | AT5G22740.1 | 298 | 2512 |
| 8 | CesA like 12 | **EnCSLA12** | AT4G07960.1 | 711 | 2816 |
| 9 | Sucrose synthase | **EnSUS1** | AT3G43190.1 | 157 | 2589 |
| 10 | Sucrose synthase (rare) | **EnSUS2** | AT3G43190.1 | 293 | 1057 |
| 11 | Glucosyltransferase-like (CesA-like A9) | **EnCSLA9** | AT5G03760.1 | 1509 | 1679 |
| 12 | beta 1,2-xylosyltransferase | **EnXLYT1** | AT5G55500.1 | 1000 | 1099 |
| 13 | 6-phosphofructokinase | **EnPPPFK1** | AT1G76550.1 | 364 | 1073 |
| 14 | glycosyl transferase | **EnGT2** | AT4G15480.1 | 1057 | 1813 |
| 15 | UGD glucose 6-dehydrogenase | **EnUGD1** | AT5G15490.1 | | 1590 |
| 16 | UDP-xylose synthase | **EnUXS5** | AT3G46440.2 | | 1023 |
| | | | | | |
| | **Lignin biosynthesis** | | | | |
| 17 | CCR | **EnCCR** | EgCCR | 700 | 3000 |
| 18 | Laccase5 | **EnLAC5** | AT2G40370.1 | 374 | 2137 |
| 19 | Peroxidase | **EnPER1** | AT1G71695.1 | 350 | 1072 |
| 20 | Alcohol dehydrogenase | **EnADH1** | AT1G77120.1 | 720 | 1400 |
| 21 | 4CL | **En4CL1** | AT1G51680.2 | 1067 | 5456 |
| 22 | CAD | **EnCAD1** | AT3G19450.1 | 1120 | 2620 |
| 23 | CCoAOMT1 | **EnCCoAMT2** | AT4G34050.1 | 987 | 1806 |
| 29 | CCoAOMT2 | **EnCCoAMT1** | AT4G34050.1 | 1703 | 981 |
| 24 | Laccase12 | **EnLAC12** | AT5G05390.1 | 374 | 2137 |
| 25 | Laccase2 | **EnLAC2** | AT2G29130.1 | 374 | 1570 |
| 26 | ascorbate peroxidase 2 | **EnAPX2A** | AT3G09640.2 | 1398 | 830 |
| 27 | S-adenosylmethionine synthase (SAM) | **EnSAM2** | AT4G01850.2 | 686 | 1180 |
| 28 | hydroxycinnamoyl transferase | **EnHCT** | AT5G48930.1 | 350 | 1191 |
| 30 | COMT | **EnOMT1** | AT5G54160.1 | | 3214 |
| 31 | p-coumaroyl shikimate 3'-hydroxylase | **EnC3H1** | AT2G40890.1 | | 717 |
| 32 | S-adenosylmethionine synthetase (SAM) | **EnSAM1** | AT1G02500.2 | | 1496 |
| 33 | Ferulate -5-hydroxylase | **EnFAH1** | AT4G36220.1 | | 1167 |
| 34 | laccase10 | **EnLAC10** | AT5G01190.1 | | 1386 |
| | | | | | |
| | **Cytoskeleton** | | | | |
| 35 | Beta-tubulin 1 | **EnTUB1** | EgTUB1 | 1400 | 3000 |

| | | | | | |
|---|---|---|---|---|---|
| 36 | Beta-tubulin 4 | **EnTUB4** | EgTUB4 | 991 | 1569 |
| 37 | Alpha-tubulin 1 | **EnTUA1** | EgTUA1 | 340 | 1581 |
| 38 | Alpha-tubulin 3 | **EnTUA3** | EgTUA3 | 502 | 1633 |
| 39 | Myosin heavy chain protein | **EnMHC1** | AT5G52280.1 | 1890 | 1013 |
| 40 | Actin | **EnACT1** | AT5G09810.1 | 1700 | 1452 |
| 41 | Actin depolymerizing factor | **EnADF2** | AT2G31200.1 | Thumma | Thumma |
| 42 | Actin-depolymerizing factor | **EnADF1** | AT5G59890.1 | 857 | 828 |
| 43 | kinesin motor protein | **EnK1B** | AT3G44730.1 | 900 | 1181 |
| 44 | Myosin heavy chain, putative | **EnUP15** | AT2G40480.1 | | 1655 |
| | | | | | |
| | **Cell expansion/wall loosening** | | | | |
| 45 | endo xyloglucan transferase (XET) | **EnEXGT1** | AT5G13870.1 | 400 | 1225 |
| 46 | Korrigan (cellulase) | **EnKOR** | AT5G49720.1 | 177 | 1980 |
| 47 | Expansin | **EnEXPA4** | AT2G39700.1 | 723 | 1787 |
| 48 | Pectinesterase | **EnPE1** | AT5G66920.1 | 1300 | 1783 |
| 49 | Pectate lyase | **EnPel1** | AT4G13710.1 | 1500 | 1041 |
| 50 | Pectate lyase | **EnPel2** | AT3G53190.1 | 500 | 662 |
| 51 | Alpha-expansin | **EnEXP1** | AT1G69530.2 | 900 | 1279 |
| 52 | Extensin | **EnEXT1** | no sequence | 2200 | 1312 |
| 53 | Glucoside hydrolase | **EnGH28A** | AT3G61490.1 | | 1429 |
| 54 | Glycosyl hydrolase | **EnGH5** | AT2G20680.1 | | 1226 |
| 55 | glycosyl hydrolase 17 | **EnGH17** | AT1G11820.1 | | 963 |
| 56 | Cellulase | **EnCel2** | AT4G02290.1 | | 1275 |
| | | | | | |
| | **Cell wall/membrane proteins** | | | | |
| 57 | Fasciclin-like arabinogalactan (FLA1) | **EnFLA1** | EgrFLA1 | 1678 | 1179 |
| 58 | Fasciclin-like arabinogalactan (FLA2) | **EnFLA2** | EgrFLA2 | 1949 | 1124 |
| 59 | Fasciclin-like arabinogalactan (FLA3) | **EnFLA3** | EgrFLA3 | 1650 | 1033 |
| 60 | Fasciclin-like AGP (17) | **EnFLA17A** | AT5G06390.1 | 100 | 819 |
| 61 | AGP | **EnAGP1** | AT2G33850.1 | 841 | 770 |
| 62 | GPI-anchored protein | **EnGPI1** | AT3G06035.1 | 815 | 859 |
| 63 | Cobra-like (COBL4) | **EnCOBL4A** | AT5G15630.1 | 83 | 1518 |
| 64 | Cobra-like (COBL4) | **EnCOBL4B** | AT5G15630.1 | 207 | 2092 |
| 65 | PAAPA | **EnPAAPA** | No hits | 600 | 1202 |
| 66 | Integral membrane protein LIH | **EnIMP1** | AT2G35760.1 | 1320 | 817 |
| 67 | LRR transmembrane kinase | **EnLRRK2** | AT2G24230.1 | 284 | 470 |
| 68 | LRR transmembrane kinase | **EnLRRK1** | AT2G36570.1 | 1050 | 1591 |
| 69 | vesicle-associated membrane protein | **EnVAMP7** | AT4G32150.1 | 862 | 1111 |
| 70 | major latex protein (ficus) allergen | **EnMLP28** | AT1G70830.1 | 1000 | 959 |
| 71 | Aquaporin pip5 | **EnPip2** | | 1185 | 804 |
| 72 | Aquaporin (Pip2;5) | **EnPIP1** | AT3G54820.1 | 49 | 1817 |
| 73 | Nac domain protein 008 | **EnNAC8** | AT1G25580.1 | | 901 |
| 74 | LRR protein | **EnLRRK3** | AT3G20820.1 | | 857 |
| | | | | | |
| | | | | | |
| | **Transcription, gene activation** | | | | |
| 75 | Zinc finger protein | **EnZnf1** | | 351 | 1441 |
| 76 | CCH zinc finger | **EnCCHZ1** | AT1G66810.1 | 100 | 868 |

| | | | | | |
|---|---|---|---|---|---|
| 77 | C3HC4 zinc finger | **EnC3HC4Z1** | AT5G55970.2 | 700 | 1016 |
| 78 | Protein kinase | **EnPK1** | AT3G15220.1 | 600 | 815 |
| 79 | LIM domain protein | **EnLIM1** | AT1G10200.1 | 1172 | 1878 |
| 80 | LIM domain protein | **EnLIM2** | AT1G10200.1 | 786 | 1018 |
| 81 | MYB83 (MYB1) | **EnMYB83** | AT3G08500.1 | 514 | 1086 |
| 82 | MYB4 (MYB2) | **EnMYB4** | AT4G38620.1 | 453 | 1790 |
| 83 | MYB82 | **EnMYB82** | AT5G52600.1 | 323 | 1780 |
| 84 | MYB1 | **EnMYB1** | | 729 | 1655 |
| 85 | EnNAM1(NAC domain protein) | **EnNAM1** | | 801 | 1881 |
| 86 | NAC domain protein | **EnNAC1** | AT2G46770.1 | 1600 | 1417 |
| 87 | SND1 (NAC protein) | **EnSND1** | AT2G46770.1 | 172 | 1417 |
| 88 | BTF3 (NAC domain) | **EnBTF1** | AT1G73230.1 | 911 | 759 |
| 89 | Steroid recepter/transcription factor | **EnSRT1** | AT2G40320.1 | 500 | 3310 |
| 90 | Mitochondrial transcription term. factor | **EnTERF1** | AT4G14605.1 | 372 | 2285 |
| 91 | Rab GTPase | **EnRABA1** | AT5G45750.1 | 857 | 1007 |
| 92 | **HB15 transcription factor** | **EnHB1** | AT1G52150.1 | 566 | 7171 |
| 93 | Rac GTPase activating protein | **EnRAC1** | AT5G22400.1 | 721 | 1805 |
| 94 | Rac-like GTP binding protein | **EnRAC7** | AT4G28950.1 | 1000 | 1053 |
| 95 | bHLH transcription factor, putative | **EnBHLH2** | AT5G48560.1 | 1000 | 818 |
| 96 | Ubiquitin inteaction motif-/LIM domain | **EnLIM3** | AT1G19270.1 | 717 | 720 |
| 97 | bzip transcription factor | **EnBzip1** | AT1G75390.2 | 1226 | 453 |
| 98 | Squamosa promoter binding protein | **EnSPL2** | AT5G43270.3 | 688 | 1926 |
| 99 | Zinc binding family protein | **EnZB1** | AT1G32700.1 | 2390 | 1473 |
| 100 | HD-zip transcription factor | **EnHB8** | AT4G32880.1 | 80 | 1250 |
| 101 | COP 1 interacting protein (CIP7) | **EnCIP7A** | AT4G27430.1 | | 1710 |
| 102 | COP 1 interacting protein (CIP7) | **EnCIP7B** | AT5G43310.1 | | 1655 |
| 103 | bHLH transcription factor (ethylene resp.) | **EnBHLH1** | AT4G29100.1 | | 1878 |
| 104 | Rab GTPase | **EnRAB6** | AT2G44610.1 | | 1236 |
| | | | | | |
| | **Unknown or other function** | | | | |
| 105 | Unknown protein | **EnUP1** | AT4G39870.1 | 1176 | 1839 |
| 106 | Histidinol-phosphate aminotransferase | **EnHPA** | AT5G10330.3 | 1550 | 1276 |
| 107 | Metallothionein | **EnMT1** | AT5G02380.1 | 930 | 721 |
| 108 | Unknown protein | **EnUP2** | AT4G19080.1 | 900 | 1722 |
| 109 | Unknown protein | **EnUP3** | AT5G42710.1 | 800 | 2500 |
| 110 | Unknown protein | **EnUP4** | AT2G46890.1 | 802 | 2868 |
| 111 | Unknown protein (Eskimo protein) | **EnUP5** | AT5G01360.1 | 595 | 1684 |
| 112 | Unknown protein | **EnUP13** | AT1G78430.1 | 759 | 1488 |
| 113 | Unknown protein | **EnUP7** | AT5G65470.1 | 500 | 816 |
| 114 | UDP-galactose transporter putative | **EnUGT1** | AT5G59740.1 | 950 | 1658 |
| 115 | Disease resistance-respomsive protein | **EnUP8** | AT5G42510.1 | 862 | 1170 |
| 116 | Nodulin | **EnNOD1** | AT1G75500.1 | 309 | 1879 |
| 117 | Phagocytosis and cell mobility protein | **EnELMO1** | AT1G03620.1 | 877 | 1388 |
| 118 | Spermidine synthase (ACAULIS 5) | **EnSPDSY2** | AT5G19530.2 | 686 | 1248 |
| 119 | Unknown | **EnUP10** | AT3G61750.1 | 1400 | 1653 |
| 120 | Unknown protein (Gibberellin | **EnUP16** | AT5G14920.1 | 600 | 947 |

| | | | | | |
|---|---|---|---|---|---|
| | regulated) | | | | |
| 121 | Fibrillarin2 | **EnFIB2** | AT4G25630.1 | 1315 | 720 |
| 127 | ERECTA | **EnERE** | AT2G26330.1 | 1000 | 6000 |
| 122 | Proteasome regulatory subunit (RPN) | **EnRPN1** | AT2G20580.1 | | 1270 |
| 123 | ATP binding/protien ser/thr kinase | **EnBRL1** | AT2G01950.1 | | 3117 |
| 124 | EF hand calcium binding protein | **EnEFH1** | AT1G53210.1 | | 4511 |
| 125 | Disease resistance protein | **EnUP14** | AT3G07040.1 | | 1075 |
| 126 | subtilisin-like serine protease | **EnSLP1** | AT4G34980.1 | | 1589 |
| | | | | | |
| | **Novel proteins** | | | | |
| 128 | No hits | **EnNOV1** | | 790 | 568 |
| 129 | No hits | **EnNOV2** | | 500 | 833 |
| 130 | No hits | **EnNOV3** | | 400 | 1340 |
| 131 | No hits | **EnNOV4** | | 800 | 1268 |
| 132 | No hits | **EnNOV10** | | 2100 | 832 |
| 133 | No hits | **EnNOV11** | | 750 | 823 |
| 134 | No hits | **EnNOV6** | | | 2950 |
| 135 | No hits | **EnNOV8** | | | 2411 |
| 136 | No hits | **EnNOV5** | | | 1581 |

Appendix 2. SNP associations with wood quality traits at nominal significance level in Meunna association population.

| Gene | SNP | Trait | *P* |
|---|---|---|---|
| VAMP | SNP-7 | KPY | 0.03 |
| EXP1 | SNP-4 | KPY | 0.05 |
| TUB1 | SNP-12 | KPY | 0.04 |
| UP10 | SNP-1 | KPY | 0.03 |
| CIP7A | SNP-02 | cellulose | 0.03 |
| CNX1 | SNP-2 | Total-lignin | 0.027 |
| TUB1 | SNP-17 | extractives | 0.02 |
| Expa1 | SNP-10 | Extractives | 0.02 |
| UP3 | SNP-19 | cellulose-KPY | 0.02 |
| EXGT1 | SNP-1 | celllulose | 0.015 |
| CNX1 | SNP-11 | Total-lignin | 0.015 |
| EXGT1 | SNP-14 | cellulose | 0.016 |
| NOV10 | SNP-1 | cellulose | 0.015 |
| CIP7A | SNP-04 | cellulose | 0.01 |
| ACT1 | SNP-15 | cellulose | 0.012 |
| RABA1 | SNP-14 | KPY | 0.01 |
| ACT1 | SNP-1 | cellulose | 0.01 |
| ACT1 | SNP-5 | cellulose - MFA | 0.0085 |
| PIP2 | SNP-12 | cellulose | 0.005 |
| PIP2 | SNP-15 | cellulose | 0.007 |
| EXGT1 | SNP-26 | cellulose | 0.008 |
| EXGT1 | SNP-11 | cellulose | 0.007 |
| LIM1 | SNP-2 | Klason-lignin | 0.009 |
| 4CL | SNP-14 | MFA | 0.014 |
| EXGT1 | SNP-25 | Density | 0.013 |
| UXS5 | SNP-3 | mfa | 0.005 |
| LIM1 | SNP-27 | Density | 0.01 |
| *CAD | SNP-07 | KPY | 0.01 |
| *CAD | SNP-02 | cellulose | 0.02 |
| *CCR | SNP-21 | MFA | 0.0002 |
| *COBL4 | SNP-07 | cellulose | 0.007 |
| *COBL4 | SNP-02 | KPY | 0.02 |
| *FLA3Pr | SNP-03 | cellulose - KPY | 0.04 |
| *EXP | SNP-10 | KPY | 0.02 |

*SNPs identified in earlier studies