Nucleotide diversity in lignification genes and QTNs for lignin quality in a multi-parental population of Eucalyptus urophylla

Eric Mandrou, Marie Denis, Christophe Plomion, Franck Salin, Frédéric Mortier, Jean-Marc Gion

Nucleotide diversity in lignification genes and QTNs for lignin quality in a multi-parental population of *Eucalyptus urophylla*

Eric Mandrou · Marie Denis · Christophe Plomion · Franck Salin · Frédéric Mortier · Jean-Marc Gion

Abstract Lignin is a major chemical compound of wood and one of the most abundant organic biopolymers on earth. It accumulates in the secondary cell wall of xylem cells and is a major target for tree breeders because of its foreseen role in the emerging bioeconomy. In this study, we paved the way toward an accelerated domestication of a widely grown tree species, *Eucalyptus urophylla*, by molecular breeding. To this end, we first described the pattern of nucleotide variation occurring at seven structural and regulatory genes of the lignin biosynthesis pathway and found high levels of average nucleotide and haplotype diversity per gene (π=0.0065 and Hd=0.853). Then, taking advantage of a pre-existing factorial mating design, a candidate-gene-based quantitative trait locus (QTL) detection strategy was used to compare the variation of lignin quality (syringyl by guaiacyl ratio (S/G)) with the nucleotidic variability in these seven genes in 304 genotypes belonging to 33 connected full-sib families. Two genes, encoding cinnamoyl-CoA reductase (*CCR*) and a Rho-like GTPase (*ROP1*), were shown to be linked to the variation of S/G through different single and multi-locus single-nucleotide polymorphism (SNP)- and haplotype-based association methods. Providing that relevant candidate genes are selected and their patterns of nucleotide diversity is accurately described, we showed that quantitative trait nucleotides (QTNs) can be detected taking advantage of pre-existing field experiments and trait measurements gathered in the framework of a forest tree breeding program.

Keywords Lignin · *Eucalyptus* · QTN · Factorial design · S/G · Haplotypes

E. Mandrou
Centre de Recherche Vallourec, CEV, 59620 Aulnoye-Aymeries, France

E. Mandrou · M. Denis · J.-M. Gion
CIRAD, UMR AGAP, Equipe Génétique et Amélioration des Espèces Pérennes: Modèles Forêt et Palmier, Campus International de Baillarguet, TA A-39/C, Montpellier CEDEX 5, France

E. Mandrou · C. Plomion · F. Salin · J.-M. Gion
INRA, UMR 1202 BIOGECO, 69 route d'Arcachon, 33610 Cestas, France

E. Mandrou · C. Plomion · F. Salin · J.-M. Gion (✉)
Univ. Bordeaux, UMR 1202 BIOGECO, 33400 Talence, France
e-mail: gion@cirad.fr

F. Mortier
CIRAD, UR B&SEF, Biens et Services des Ecosystèmes Forestiers et Tropicaux, Campus International de Baillarguet, TA C-105/D, 34398 Montpellier, France

Introduction

Surveying structural variations (i.e., single-nucleotide polymorphisms (SNPs), deletion-insertion polymorphisms (DIPs), copy-number variations (CNVs), duplications, and other rearrangements) in whole genomes or specific genomic regions is currently a key step in research undertaken to decipher the genetic basis of phenotypic variation and to identify causative variants in complex evolutionary processes such as speciation, selection, and local adaptation (Barton and Keightley 2002). The availability of reference genome sequences and the democratization of next generation sequencing technologies are now changing the research paradigm and opening up many opportunities to describe DNA variation and discover its consequences at the genomic scale, especially in model organisms (Mardis 2008). On the other hand, smaller scale studies aiming to describe the landscape of nucleotide variation from

a limited number of loci or subgenic regions are still appropriate for peculiar applications, e.g., genetic identification, population structure analysis, and management of breeding populations, as well as for the analysis of genes involved in specific biosynthesis pathways or belonging to specific gene families. This is particularly true in species for which sequence information is lacking which is the case for most forest tree species.

Earlier investigations of nucleotide diversity in forest trees focused on SNP detection using sets of candidate genes known to be involved in biotic and abiotic stress responses (Ingvarsson 2005; Krutovsky and Neale 2005; Pyhäjärvi et al. 2007; Wachowiak et al. 2008), as well as developmental processes such as wood formation (Brown et al. 2004; Pot et al. 2005; González-Martinez et al. 2006) and bud phenology (Derory et al. 2009). The main objective of these studies was to understand evolutionary processes (demographic history, imprint of natural selection) underlying patterns of nucleotide variation. They revealed high levels of genetic variation and rapid decay of linkage disequilibrium, declining to negligible levels in less than 500 bp (Brown et al. 2004; Ingvarsson 2005; Heuertz et al. 2006), as expected in allogamous species with large population sizes and efficient gene flow. These studies provided the first data to conduct association mapping in trees which, given the mating characteristics and life history of those species, was believed to have great potential to accurately map mutations that contribute to trait variation (Neale and Savolainen 2004; Neale and Ingvarsson 2008). Early investigations in this domain have still been limited to nucleotide polymorphisms in candidate genes thought to impact tree phenotypes. These studies usually used prior knowledge of population structure gathered from neutral markers or results of earlier studies aimed at disentangling linkage disequilibrium (LD) due to physical linkage or other evolutionary forces. Overall, association mapping in forest trees has revealed few associated SNPs explaining only a small fraction (<5 %) of phenotypic variance, in any given trait (reviewed by Khan and Korban 2012). In addition, because of the rather small populations generally used to estimate QTL effects (in general between 100 and 300 genotypes), percentages of explained variance are certainly biased upward, and the power to detect associations is extremely low (Lepoittevin et al. 2012).

Different approaches have been proposed to detect QTLs combining linkage information, modeled in classical linkage analysis (LA), and short-range LD accounted for in association mapping (Meuwissen et al. 2002; Farnir et al. 2002; Lund et al. 2003; Pérez-Enciso 2003; Legarra and Fernando 2009). These so-called linkage disequilibrium and linkage analysis (LDLA) approaches have been compared either on simulated and real data sets and have been proven to be efficient to detect QTLs in complex pedigrees of animals and crops (Meuwissen et al. 2002; Lee and Van der Werf 2004; Legarra and Fernando 2009; Roldan et al. 2012; Bardol et al. 2013). However, to be performed at the genome-wide level, those approaches could require an extremely large number of markers to identify identity by descent (IBD) alleles, especially in the founder population. In *Eucalyptus* as for most of forest tree species, genitors of breeding programs are mostly unrelated and directly sampled from natural populations in which LD decays rapidly. In this context, despite developments in high-throughput genotyping technology (reviewed by Grattapaglia et al. 2012), such an approach is still unrealistic at the genome-wide level in these species, and LD-based approaches (including LDLA) should be limited to the study of well-chosen candidate genes, such as those involved in the lignification pathway.

Eucalypts are among the most widely grown tree species in industrial plantations worldwide. Domestication of eucalypts by breeding is very active worldwide, and improving lignin (a biopolymer that accumulates wood) is becoming mandatory to fulfill many industrial applications such as pulp and paper, charcoal, or biofuels (Raymond 2002; Myburg et al. 2007). The identification of loci linked to the variation of lignin synthesis (quantity and/or composition) in these species could help breeders to increase genetic gains by time units. In this context, several forward genetic approaches based on genome-wide QTL analysis in bi-parental crosses of *Eucalyptus* were followed to identify such loci (Thamarus et al. 2004; Thumma et al. 2010; Gion et al. 2011; Freeman et al. 2009, 2013). They provided interesting results regarding the genetic architecture of lignin composition and identified collocations between QTLs for lignin-related traits and lignification genes (Foucart et al. 2009; Gion et al. 2011; Freeman et al. 2013). However, given the low levels of genetic variability tested in such bi-parental crosses and the quite large confidence intervals associated with the QTL regions (several cM), these results are difficult to transfer at the scale of a breeding program. Multi-parental populations are classically used to evaluate genitor performances among breeding programs. Such trials involving a larger part of genetic variability than classical bi-parental QTL populations offer great opportunities to detect quantitative trait nucleotides (QTNs) for lignin-related traits. Moreover, as those experimental designs are related to the breeding material, QTN results could be used directly in breeding through marker-assisted recurrent selection.

In this context, the objective of this study was to detect QTNs related to lignin composition through a candidate-gene-based approach in a multi-parental population of *Eucalyptus urophylla*. To achieve this objective, we used a pre-existing experimental design (8×8 factorial matting design of *E. urophylla*) to describe the landscape of nucleotide diversity of seven genes related to lignification and perform QTN detection through different single and multi-locus SNP and haplotype-based approaches.

Material and methods

Plant material

Sixteen unrelated trees belonging to the breeding population of *E. urophylla* managed by Centre de Recherches sur la Durabilité et la Productivité des Plantations Industrielles (CRDPI; Pointe-Noire, Republic of the Congo) were used to study the pattern of nucleotide variation. These genotypes originated from Flores Island in the Sunda archipelago (Indonesia, 122°–127° E, 8°–10° S) and were conserved in a seed orchard at Kissoko forestry station.

These 16 trees were used as founders of a progeny test (incomplete factorial mating design comprising eight females by eight males). A total of 304 offspring in 33 full-sib (FS) families (eight to ten in each FS family) were phenotyped for syringyl by guaiacyl ratio (S/G). Briefly, S/G was assessed by near-infrared spectroscopy (NIRS) and found to be normally distributed in the whole population. Phenotyping experiments and quantitative genetics analysis in this experimental design were fully described by Mandrou et al. (2011) where the narrow sense heritability for S/G was h^2=0.62.

Leaves were sampled from each tree (founders and offspring) and dried in silica gel. DNA was extracted according to Doyle and Doyle (1990), and samples were stored at −20 °C until use.

Candidate gene selection

Seven genes related to lignin biosynthesis were selected for this study: cinnamate 4-hydroxylase (*C4H*), ferrulate 5-hydroxylase (*F5H*), and caffeate *O*-methyltransferase 2 (*COMT2*), which encode enzymes of the common phenylpropanoid pathway, cinnamyl alcohol dehydrogenase (*CAD2*) and cinnamoyl-CoA reductase (*CCR*), which encode the two structural enzymes of the specific monolignol biosynthesis pathway, myb domain protein (*MYB2*), which is a transcription factor involved in regulation of the expression of structural genes of lignin biosynthesis (Goicoechea et al. 2005), and *ROP1*, which encodes a small GTPase Rac-like protein involved in secondary xylem differentiation in *Eucalyptus* (Foucart et al. 2009).

PCR amplification and DNA sequencing

A total of six gene fragments (one fragment per gene except for *CCR* for which the full CDS was obtained by Mandrou et al. 2011) were amplified for the 16 highly heterozygous genitors of the experimental design (all sequences of primer pairs are given in Online Resource 1). Diploid DNA was amplified using a Tetrad 2 PTC-0240 Thermo Cycler (MJ Research, Whaltam, MA, USA). A 20-μl final reaction volume comprising 2 μl of 10× PCR reaction buffer (Invitrogen, Carlsbad, CA, USA), 0.8 μl dNTPs (5-mM stock solution), 0.8 μl $MgCl_2$ (50 mM), 0.4 μl of each primer solution (10 μM), 20 ng of genomic DNA, and 0.8 U of native Taq polymerase (Invitrogen) and dH_2O was used to amplify each gene fragment from each DNA sample. A three-step PCR cycle was used with an initial denaturation step of 4 min at 94 °C, 35 cycles of 30 s at 94 °C, 1 min and 20 s at primers Tm, 1 min at 72 °C, followed by a final 10 min extension step at 72 °C. For the six genes (*C4H*, *F5H*, *COMT2*, *CAD2*, *MYB2*, and *ROP1*), PCR products were cloned independently using the TOPO-TA cloning kit for sequencing (Invitrogen). A total of 16 transformed clones were collected from each cloning product, and plasmid inserts were sequenced in both forward and reverse directions using T3 and T7 universal primers. All sequences were performed using Big Dye Terminator V 1.1 cycle sequencing kit (Applied Biosystems, Foster City, CA, USA), and electrophoreses were run on an ABI 3730 DNA Analyzer (Applied Biosystems). For each gene fragment, sequences were aligned and checked for base calling errors using CodonCode Aligner software version 1.6.1 (CodonCode, Deadham, MA, USA). For each set of clone sequences obtained from one cloning product, Taq polymerase amplification errors and chimeric allele products (creating artificial technical variants) were removed by comparing nucleotides at each sequence position. Thanks to this sequencing method, phased haplotypes were obtained for the six studied genes. Previously obtained haplotype sequence data for the full-length *CCR* gene in the same 16 genotypes of *E. urophylla* (Mandrou et al. 2011) were added to this study. More details on sequencing data are provided in Online Resource 2.

Analysis of genetic diversity

Nucleotide and haplotype diversity were estimated with DnaSP v. 5.0 (Rozas et al. 2003). Nucleotide diversity was estimated by π, the average number of differences between two sequences in the population sample. Haplotype diversity was measured as $Hd = \frac{n\left(1-\sum p_i^2\right)}{n-1}$ (Nei 1987).

Genotyping assay

For *C4H*, *F5H*, *COMT2*, *CCR*, *CAD2*, *MYB2*, and *ROP1*, the 304 offspring of the factorial matting design were genotyped. As previously described by Mandrou et al. (2011), genotyping of the *CCR* gene was achieved taking advantage of an SSR locus located in intron #4 of the gene. For the six other genes, a total of 32 SNPs were available after genotyping the experimental population through the Sequenom MassARRAY SNP Genotyping technology (Sequenom, Hamburg, Germany). Those 32 SNPs represented the fraction of SNPs with good quality results and conformity to the expectations according to

parental allelic states and Mendelian segregation. More details on the SNP genotyping array are given in Online Resource 3. Finally, for the six genes, haplotypes of the parent trees were imputed in the progeny based on both SNP genotypes and pedigree information.

Statistical analysis for gene-trait associations

SNP-based approaches

In order to detect associations between lignification genes and S/G, two SNP-based approaches were applied. On the one hand, a SNP by SNP approach based on a mixed model correcting for relatedness between offspring (Yu et al. 2006) was used (MLM (K)). This approach was already applied on *CCR* data and fully described in the "Methods" section of Mandrou et al. (2011). Briefly, the approach was based on the following model:

$$y = \mu + X\beta + Zu + \varepsilon$$

where y is an $n\times1$ observation vector of phenotypes, μ is the mean of the population, X is a $n\times1$ observation vector of genotypes at a given SNP, β is the fixed SNP effect, Z is a $n\times1$ design matrix linking observations to random effect u, u is a $n\times1$ vector of random breeding values, and ε is the vector of random errors such that $Var(e)=\sigma_e^2 Id$ For the random effect u, the associated variance is equal to $Var(u)=\sigma_a^2 A$ with A the kinship matrix computed from a pedigree file that takes into account all the familial relationships between the 304 individuals of the association population.

On the other hand, the multi-locus mixed model approach (MLMM) proposed by Segura et al. (2012) was used. This approach includes SNPs as cofactors in the model to finally select the best model by a stepwise selection approach based on a Bayesian information criterion (BIC) (Schwarz 1978) or an extended BIC (ext BIC) criterion (Chen and Chen 2008).

Both approaches were applied to a total set of 26 biallelic SNPs in *CCR* (12 SNPs), *CAD2* (2 SNPs), *C4H* (3 SNPs), *COMT2* (1 SNP), *MYB2* (2 SNPs), *ROP1* (4 SNPs), and *F5H* (2 SNPs) selected according to minimal MAF of 5 % and LD threshold of $r^2<0.5$ between SNPs to avoid too much redundancy in the SNP data set.

Haplotype-based approaches

First, a single-locus haplotype-based approach was applied to test each gene independently against S/G. This approach was based on the same model as the SNP by SNP approach except that the design matrix related to SNP alleles was replaced by a $n\times n_k$ design matrix related to gene haplotypes. This matrix was filled with 0, 1, and 2 as in the multi-locus haplotype-based approach described below. For this approach, all the haplotypes with frequencies <5 % were grouped in a single class. The most frequent haplotype was assigned as the reference level. The class grouping all rare haplotypes was never used as the reference class in the model unless it was detected as nonsignificant.

Finally, a multi-locus haplotype-based QTL detection method was developed to detect the statistical relationship between the nucleotide variability of the seven genes together and the variation of S/G. This approach followed the two steps described below.

Step 1, model selection To determine which of the seven genes to include in the statistical model, a forward model selection approach was used based on the following linear mixed model:

$$y = \mu + X^1\beta^1 + \ldots + X^k\beta^k \ldots + X^N\beta^N + Zu + \varepsilon$$

where y is an $n\times1$ observation vector, u is the mean of the population, X^k for k=1 to N is $n\times n_k$ design matrix allocating records to the haplotype effects of gene k, $\beta^k = \left(\beta_1^k, \ldots, \beta_{n_k}^k\right)$ is a $(n_k)\times1$ vector of fixed haplotype effects associated with gene k, Z is an $n\times q$ design matrix linking observations to random effect u, u is a $q\times1$ vector of random breeding values, and ε is the vector of random errors such that $Var(e)=\sigma_e^2 Id$ For the random effect u, the associated variance is equal to $Var(u)=\sigma_a^2 A$ with A the kinship matrix computed from a pedigree file that takes into account all the relationships between the individuals.

The design matrix X^k associated to each gene was made of 0, 1, and 2. Indeed, the haplotype effect was assumed the same regardless the paternal or maternal origin. Then, for an individual carrying twice the same haplotype at a given gene, only one effect was estimated, and the effect associated with this genotype was twice the estimated effect.

The components of the X^k matrix were determined by

$$\begin{cases} X_{ij}^k = 0 \text{ if the individual } i \text{ has the genotype } (H_l, H_{l'}) \text{with } l\neq j \text{ and } l'\neq j \\ X_{ij}^k = 1 \text{ if the individual } i \text{ has the genotype } (H_l, H_j) \text{ with } l\neq j \\ X_{ij}^k = 2 \text{ if the individual } i \text{ has the genotype } (H_j, H_j) \end{cases}$$

An example is given below to explain the structure of the design matrix X^k for gene k.

Let H_1^k, H_2^k, and H_3^k, three haplotypes of gene k with (H_1^k, H_1^k), (H_2^k,H_1^k), and (H_3^k,H_3^k), the haplotypic composition of three offspring; X^k is then defined by

$$X^k = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The variance component was estimated by maximum likelihood (ML), and the selection criterion used for the fixed

effects was the corrected Akaike information criterion (AICc) (Hurvich and Tsai 1989).

The principle of the forward model selection approach was to start with no variable in the model. This model is referred to as the null model (*MNull*). Then, each gene with all corresponding haplotypes was added to the model one by one. The model with the smallest AICc criterion was selected. At each step, each gene that was not considered in the previous model was tested for its inclusion depending on the AICc. The most significant of these genes according to the AICc was added to the model. This process continued as long as the AICc was smaller than the AICc at the previous step.

For example, let X^1, X^2, and X^3 be the design matrix associated to three genes. We first computed the AICc of the null model *MNull*. Then, the three genes X^1, X^2, and X^3 were added separately, and the corresponding AICc was computed. The gene with the smallest AICc was selected. Assuming gene X^1 is the gene selected for model M_1, in the second step, we separately added X^2 and X^3 to M_1. The model with the smallest AICc was selected and added to model M_1 to obtain model M_2. This approach was repeated as long as the addition of a new gene led to a smaller AICc. Thus, a model with a subset of genes having the smallest AICc was retained as the full model (*MFull*).

Step 2, haplotype-based QTL detection For each gene included in the model, a backward selection method was applied to select haplotypes statistically linked to trait variation. For each haplotype of each gene, the full model (*MFull*) was tested according to a p value from a t distribution. The haplotypes associated with a p value above the threshold value of $\alpha=0.05$ were assigned to the reference level. The same procedure was applied until all p values were significant.

For example, let X^1 and X^2 be the two selected genes and $H_1^1, H_2^1, H_3^1, H_4^1$, and H_1^2, H_2^2, H_3^2 their associated haplotypes. Let H_1^1 be the reference haplotype (the most frequent in the factorial design assigned to the reference level) for gene 1 and H_3^2 the reference haplotype for gene 2 and $(\beta_1^1, \beta_2^1, \beta_3^1, \beta_4^1)$ and $(\beta_1^2, \beta_2^2, \beta_3^2)$ the haplotype effects associated to genes 1 and 2, respectively.

First, the full model *MFull* was fitted, giving a p value at each haplotype. If β_3^1 and β_2^2 coefficients had p values above the threshold, the haplotype H_3^1 of gene 1 was moved to the reference haplotype of gene 1, and the haplotype H_2^2 of gene 2 was moved to the reference haplotype of gene 2. Model M_2 then comprised both genes X^1 and X^2 with haplotypes H_1^1, H_2^1, H_4^1 and H_1^2, H_3^2, respectively. The model was fitted according to p values associated with haplotypes until all p values were significant.

At each step of selection, the likelihood-ratio-based coefficient of determination (R^2) indicated the proportion of phenotypic variance explained by a given gene and its haplotypes. As for the single-locus haplotype-based approach, all haplotypes with frequencies <5 % were grouped in a single class. This class, grouping all rare haplotypes, was never used as the reference class in the model unless it was detected as nonsignificant after the first round of haplotype selection.

Results

Nucleotide and haplotype diversity

A total of 330 SNPs were detected including results from the *CCR* gene already described by Mandrou et al. (2011). Among these, 301 corresponded to silent mutations (not modifying the primary structure of proteins) and 29 were nonsynonymous mutations. Additional features (presence in coding vs noncoding regions, silent vs nonsynonymous mutations) are provided in Online Resource 4.

Levels of nucleotide diversity were estimated as π, for total, silent, and nonsynonymous positions (Table 1). Overall, the average nucleotide diversity of these seven genes was 0.0065, with values ranging from 0.0131 in *CCR* to 0.0027 in *COMT2*. This average was lower in nonsynonymous positions (0.0016) with estimated values ranging from 0.0024 in *COMT2* to 0.0008 in *CAD2* and higher in silent positions (0.0094) with values ranging from 0.0173 in *CCR* to 0.0029 in *COMT2*. At the haplotype level, an average Hd value of 0.853 was observed in the sample, ranging from 0.738 in *COMT2* to 0.958 in *CCR*.

Haplotype inference from parents to offspring

Complete resolution of segregating haplotypes was obtained for the *CCR* gene by genotyping a hypervariable SSR marker located in intron #4 of the gene. The segregation and genotyping results have been described by Mandrou et al. (2011) and are not reported here. For the other six genes, 32 genotyped SNPs enabled a partial resolution of haplotype variability in the progeny of the factorial design. The level of resolution of parental haplotypes depended on the ability of genotyped SNPs to allow differentiating the parental haplotypes obtained by sequencing in each FS family. For the six genes, the 32 genotyped SNPs allowed to achieve a complete resolution of high-frequency haplotype classes in the factorial design. Those haplotypes could then be directly inferred from the founders to their progeny. However, for some parents carrying two rare haplotypes (only one copy identified in the 16 founders), the SNP data set did not allow to discriminate them precluding the inference of those haplotypes in the offspring. For the haplotype-based association approaches, those unresolved rare haplotypes were grouped together with others rare haplotypes in a single class. More information on the experimental design, the parental haplotypes, and the segregation of genotyped haplotypes for testing associations is given in Online Resource 5.

Table 1 Nucleotide diversity (π) and haplotype diversity (Hd) of seven lignification gene fragments

	Nucleotide diversity									Haplotype diversity
	Total			Silent			Nonsynonymous			Hd
Locus ID	Length	*S*	π	Length	*S*	π	Length	*S*	π	
C4H	1,119	48	0.0075	747	43	0.0103	370	5	0.0020	0.940
F5H	835	14	0.0034	200	8	0.0079	631	6	0.0020	0.881
COMT2	941	19	0.0027	481	13	0.0029	458	6	0.0024	0.738
CCR	2,942	156	0.0131	2,187	152	0.0173	755	4	0.0009	0.958
CAD2	1,180	30	0.0064	694	27	0.0103	486	3	0.0008	0.764
MYB2	873	14	0.0050	442	10	0.0080	431	4	0.0020	0.752
ROP1	1,359	49	0.0071	1,047	48	0.0090	312	1	0.0009	0.938
Mean	1,321	47	0.0065	828	43	0.0094	492	4	0.0016	0.853

Sequence lengths are in base pairs and *S* is the number of mutation sites accounted for in the estimates of π for total, silent, and nonsynonymous mutations and Hd

QTN detection

SNP-based approaches

Over the 26 SNP tested, six were significantly associated with S/G at a *p* value <0.05: three in *CCR*, already reported by Mandrou et al. 2011, two in *C4H*, and one in *ROP1* (Table 2). Only two SNPs remained significant after correcting for multiple testing (Bonferroni's correction at the experiment wise error rate of 0.05), one in *CCR*, and one in *ROP1*. SNP#11 of *CCR* and SNP #4 of *ROP1* were found at frequencies of 38 and 17 %, respectively, in the multi-parental population. The MLMM approach provided quite similar results selecting one to three SNPs depending on the model selection criterion BIC or ext BIC. The model including only SNP #11 of *CCR* resulted in the lowest value of ext BIC (113.1505); however, the model including SNP #11 of *CCR* together with SNP #4 of *ROP1* and SNP #2 of *C4H* resulted in the lowest BIC value (100.7417). In this second model, a maximal *p* value of 0.009 was obtained for SNP #2 of *C4H*. After multiple testing correction at the Bonferroni's threshold of 0.05, only SNP #11 of *CCR* and SNP #4 of *ROP1* remained significant.

Haplotype-based approach

Results of the single locus haplotype-based approach are presented in Table 3. Model *M1* including *ROP1* as factor

Table 2 Results of QTN detection by single-locus (MLM (K)) and multi-locus (MLMM) SNP-based approaches in the 8×8 factorial matting design of *E. urophylla*

Method	SNP_ID	*p* value	Bonferroni	MAF
MLM (K)	*CCR_SNP4*	0.0067	NS	0.13
	CCR_SNP6	0.0069	NS	0.06
	CCR_SNP11	<1e–04	S	0.38
	ROP1_SNP4	0.0013	S	0.17
	C4H_SNP1	0.0104	NS	0.19
	C4H_SNP2	0.0030	NS	0.21
Method	Model	Max *p* value	BIC	ext BIC
MLMM	*MNull*	–	115.9816	115.9816
	MFull_1=MNull + *CCR*_SNP11	<1e–04	106.3481	113.1505
	*M2=MFull_1+ROP1*_SNP4	0.0016	101.9124	114.063
	*MFull_2=M2+C4H*_SNP2	0.009	100.7417	117.3596

For results obtained from the MLM + K approach, the Bonferroni column indicates which of the significant SNPs remain significant (*S*) or not (*NS*) after multiple testing correction at the experiment-wise error rate of 5 %. The MAF column indicates the minor allele frequency of the SNP in the multi-parental population. For the MLMM approach, *MNull* stands for the null model with no declared fixed effect and *MFull_1* or *MFull_2* stands for the best selected model according to BIC or ext BIC value, respectively. Max *p* value indicates the highest *p* value obtained for the fixed effects included in the model

Table 3 Results of haplotype-based associations of lignification genes with S/G in an 8×8 factorial mating design of *E. urophylla*

Model	AICc	PEV	Haplotypes	Freq	*P* value	Effect
MNull	67.05	–	–	–	–	–
M1 = MNull + ROP1	52.06	0.06	*ROP1_H2*	0.07	0.0002	0.1961
			ROP1_H6	0.07	0.014	0.1397
M2 = MNull + F5H	64.08	0.02	*F5H_H2*	0.12	0.021	−0.0951
MFull = MNull + ROP1 + CCR	46.02	0.13	*ROP1_H2*	0.07	<1e–04	0.2230
			CCR_H4	0.07	0.0110	0.1403
			CCR_H5	0.11	0.0045	−0.1541

S/G corresponds to the relative abundance of syringyl and guaiacyl monomers. For each model, the value of the AICc and the percentage of phenotype variance explained by the fixed effects (PEV) are given as well as the significantly associated haplotypes (Haplotypes), the *p* values of the corresponding *t* tests, the haplotype frequencies in the factorial design (Freq), and the haplotype effect

presented a better AICc than model *MNull* only including the random polygenic effect. The same occurred with model *M2* including *F5H* as factor. For *M1*, haplotypes *H2* and *H6* were significant at the error rate of 5 %. Both *ROP1_H2* and *ROP1_H6* haplotypes were detected at the frequency of 7 % in the experimental population. For *M2*, only haplotype *H2* of *F5H* was significant at the error risk of 5 %. This haplotype was detected in the experimental population at the frequency of 12 %. Out of these associations, *H2* of *ROP1* was the most significant. *ROP1* and *F5H* genes explained 6 and 2 % of the variation of S/G, respectively. The model including *ROP1* and *CCR* as fixed effects (*MFull*) was the best selected model based on AICc (Table 3). The haplotype selection procedure resulted in three significantly associated haplotypes with one in *ROP1* (*ROP1_H2*) and two in *CCR* (*CCR_H4* and *CCR_H5*). The cumulated effect of these haplotypes explained 13 % of the phenotypic variance of S/G. Those three haplotypes segregated at frequencies of 7 % (*ROP1_H2* and *CCR_H4*) and 11 % (*CCR_H5*) in the population.

Discussion

Nucleotide and haplotype diversity at lignification genes in *E. urophylla*

Overall, high levels of nucleotide diversity were detected. Our SNP density estimates are comparable to previous reports from *E. urophylla* (Denis et al. 2013) and other *Eucalyptus* species (Poke et al. 2003; Novaes et al. 2008; Külheim et al. 2009; Thavamanikumar et al. 2011), despite differences in sample size and/or the number of genes in each study (Online Resource 4). The haplotype diversity indexes obtained in the present study were moderate (*COMT2*, *CAD2*, *MYB2*) to high (*C4H*, *CCR*, *ROP1*) compared to values observed in other forest tree species. In different gymnosperm and angiosperm species such as *Pinus taeda* (González-Martinez et al. 2006), *Pinus radiata* and *Pinus pinaster* (Pot et al. 2005), *Picea abies* (Heuertz et al. 2006), *Pseudotsuga menzesii* (Krutovsky and Neale 2005), and *Populus nigra* (Chu et al. 2009), estimated values of Hd ranged from 0.376 to 0.931 for different sets of candidate genes. To the best of our knowledge, there are no published estimates of haplotype diversity at candidate genes in *Eucalyptus* species. However, our results are not surprising given the high levels of nucleotide diversity and rapid decay of LD with distance between SNPs described in forest tree species (Brown et al. 2004; Rafalski and Morgante 2004; Ingvarsson 2005; Heuertz et al. 2006) and in the current study (Online Resource 6). Such a high genetic diversity at both nucleotide and haplotype levels is promising for the improvement of lignin through breeding, provided that QTLs controlling lignin variability could be identified.

A validation of a major region for S/G in *Eucalyptus*

Out of the seven lignification genes studied, *C4H, CCR, ROP1*, and *F5H* showed significant effects on S/G in the *E. urophylla* factorial mating design. All these loci are present in an 8-Mb window flanked by *C4H* and *CCR* on scaffold #10 (Online Resource 7) of the reference genome (http://www.phytozome.net/). Within this window, *F5H*, *ROP1*, and *CCR* are linked consistent with their physical location. The closest linkage was found between *F5H* and *ROP1* (2 cM), as demonstrated by our mapping of *F5H* (using the mapping population of Gion et al. 2011, data not shown). *ROP1* and *CCR* genes are more distantly linked (11 cM, in Gion et al. 2011). Unfortunately, *C4H* could not be mapped in this population because of a lack of polymorphic markers. But, the physical location suggests that *C4H* would be linked to this gene cluster "*F5H-ROP1-CCR*". The segregation of these four genes in the present progenies also suggests that they are linked, consistent with genetic mapping and their physical location.

Our results validate the major role of this genomic region in the genetic control of S/G in a broader genetic background than the bi-parental crosses used for QTL detection. Although many genomic regions (QTLs) have been found to influence lignin-related traits in different *Eucalyptus* species, this region of chromosome #10 has consistently been identified as a major region for wood chemical traits (Thamarus et al. 2004; Thumma et al. 2010; Gion et al. 2011; Freeman et al. 2013), including S/G (Gion et al. 2011; Freeman et al. 2013). Moreover, Denis et al. (2013) reported a significant association between *C4H* and S/G in several provenances of *E. urophylla*. QTL may reflect the effect of a single gene or a cluster of genes. The identification of multiple genes from the lignin biosynthesis pathway (three structural genes *CCR*, *F5H*, and *C4H* and one regulatory gene *ROP1*) suggests that a cluster of genes could underlie the major S/G QTLs in this region. An interaction between the functional variability of *ROP1* and *CCR* could also be possible, as these two genes are known to interact at an expressional level (Foucart et al. 2009).

Two putative QTNs explained effects of *CCR* and *ROP1*

In this study, the effect of gene variability on S/G was tested at the SNP or haplotype level using both single- and multi-locus models. This approach using a multi-parental trial provided robust associations by taking into account familial relationship.

The single-locus SNP approach allowed us to detect six QTNs (three QTNs for *CCR*, two for *C4H*, and one for *ROP1*). However, the multi-locus analysis suggested a probable redundancy in the effects of some the QTNs on S/G. The QTNs (*CCR_SNP11* and *ROP1_SNP4*) with the highest significance are the best candidates to explain independent additive effects on S/G in *E. urophylla*. Association studies previously reported significant effects of *CCR* on microfibril angle *in Eucalyptus nitens* (Thumma et al. 2005) and S/G in *E. urophylla* (Mandrou et al. 2011). To date, this is the first report of QTNs from two different linked genes explaining different part of the variation of S/G.

Using the multi-locus haplotype approach, the high level of phenotypic variance (13 %) explained by three haplotypes (*CCR_H4, CCR_H5*, and *ROP1_H2*) confirmed two distinct sources of variation in S/G. Moreover, this analysis allowed us to assess the effect of each specific haplotype, revealing *CCR_H4* and *ROP1_H2* as favorable alleles for S/G and *CCR_H5* as unfavorable. Although we did not detect a significant interaction between the haplotypes (*CCR_H4*, *CCR_H5*, and *ROP1_H2*), the likelihood of the model was higher when taking into account an interaction effect, suggesting that their effects are not totally additive, i.e., possible epistasis.

Comparison of the SNPs discriminating the two haplotypes of *CCR* (*CCR_H4* and *CCR_H5*) revealed only synonymous mutations. Therefore, the apparent effect of *CCR* on S/G may reflect causal mutations within *CCR* or adjacent gene/locus in LD with the identified synonymous SNPs. Alternatively, we cannot exclude the possibility that the identified QTNs are causative, as it was shown for the *Tb1* gene of maize that noncoding DNA could be an important source of variation for quantitative trait (Clark et al. 2006).

Following validation of the two QTNs (*CCR_SNP11* and *ROP1_SNP4*) and the possible interaction of the corresponding haplotypes (*CCR_H4*, *CCR_H5*, and *ROP1_H2*) in independent samples of *E. urophylla*, they could be used for breeding of lignin composition.

Acknowledgments This article is a part of Eric Mandrou's PhD thesis supervised by Jean-Marc Gion and Christophe Plomion. EM was supported by a CIFRE contract between Vallourec CEV and CIRAD. This research was also supported by grants from Vallourec (Services agreement 2006 between CIRAD and VMB), from Bureau des Ressources Génétiques (2005_2006 No.25), from Agence Nationale de la Recherche, Plates-Formes Technologiques du Vivant (BOOST-SNP project, 07PFTV002), the Aquitaine Region ("ABIOGEN" FEDER project, No. presage: 33973), and from CIRAD. The field experiments were carried out at the CRDPI station (Pointe-Noire, Republic of the Congo). We thank also Jules Freeman for reviewing the final manuscript and two anonymous reviewers for their helpful comments. The founders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data archiving statement Reference sequences for all the genes as well as SNP data described in this study will be deposited to the NCBI repositories (GenBank and NCBI SNP database) except for the *CCR* gene. Sequences and SNP described for the *CCR* gene were already made available and can be retrieved from GenBank under accession number JN639535 and the NCBI SNP database [http://www.ncbi.nlm.nih.gov/SNP] under accession numbers 469270958 to 469271109, respectively.

References

Bardol N, Ventelon M, Mangin B, Jasson S, Loywick V, Couton F, Derue C, Blanchard P, Charcosset A, Moreau L (2013) Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (*Zea mays* L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. Theor Appl Genet (in press)

Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. Nat Rev Genet 3:11–21

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A 101:15255–15260

Chen JH, Chen ZH (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95: 759–771

Chu Y, Su X, Huang Q, Zhang X (2009) Patterns of DNA sequence variation at candidate gene loci in black poplar (*Populus nigra* L.) as revealed by single nucleotide polymorphisms. Genetica 137: 141–150

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38:594–597

Denis M, Favreau B, Ueno S, Camus-Kulandaivelu L, Chaix G, Gion J-M, Nourrisier-Mountou S, Polidori J, Bouvet J-M (2013)

Genetic variation of wood chemical traits and association with underlying genes in *Eucalyptus urophylla*. Tree Genet Genomes 9:927–942

Derory J, Scotti-Saintagne C, Bertocchi E, Dantec LL, Graignic N, Jauffres A, Casasoli M, Chancerel E, Bodénès C, Alberto F, Kremer A (2009) Contrasting relationships between the diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. Heredity 104:438–448

Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L, Mni M, Moisio S, Simon P, Wagenaar D, Vilkki J, Georges M (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161(1): 275–287

Foucart C, Jauneau A, Gion J-M, Amelot N, Martinez Y, Panegos P, Grima-Pettenati J, Sivadon P (2009) Overexpression of EgROP1, a *Eucalyptus* vascular-expressed Rac-like small GTPase, affects secondary xylem formation in *Arabidopsis thaliana*. New Phytol 183: 1014–1029

Freeman JS, Whittock PS, Potts BM, Vaillancourt RE (2009) QTL influencing growth and wood properties in *Eucalyptus globulus*. Tree Genet Genomes 5:713–722

Freeman JS, Pott BM, Downes GM, Pibeam D, Thavamanikumar S, Vaillancourt RE (2013) Stability of quantitative trait loci for growth and wood properties across multiple pedigrees and environments in *Eucalyptus globulus*. New Phytol 198(4):1121–1134

Gion J-M, Carouché A, Deweer S, Bedon F, Pichavant F, Charpentier J-P, Baillères H, Rozenberg P, Carocha V, Ognouabi N, Verhaegen D, Grima-Pettenati J, Vigneron P, Plomion C (2011) Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *Eucalyptus*. BMC Genomics 12:301

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J (2005) EgMYB2, a new transcriptional activator from *Eucalyptus xylem*, regulates secondary cell wall formation and lignin biosynthesis. Plant J 43:553–567

González-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915–1926

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, Potts BM, Myburg A (2012) Progress in Myrtaceae genetics and genomics: eucalyptus as the pivotal genus. Tree Genet Genomes 8:463–508

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics 174: 2095–2105

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus Tremula* L., Salicaceae). Genetics 169:945–953

Khan MA, Korban SS (2012) Association mapping in forest trees and fruit crops. J Exp Bot 63:4045–4060

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171:2029–2041

Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. BMC Genomics 10:452

Lee SH, van der Werf JH (2004) The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. Genet Sel Evol 36(2):145–161

Legarra A, Fernando RL (2009) Linear models for joint association and linkage QTL mapping. Genet Sel Evol 29:41–43

Lepoittevin C, Harvengt L, Plomion C, Garnier-Géré P (2012) Association mapping for growth, straightness and wood chemistry traits in the *Pinus pinaster* Aquitaine breeding population. Tree Genet Genomes 8:113–126

Lund MS, Sorensen P, Guldbrandtsen B, Sorensen DA (2003) Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. Genetics 163(1):405–410

Mandrou E, Hein PRG, Villar E, Vigneron P, Plomion C, Gion J-M (2011) A candidate gene for lignin composition in *Eucalyptus*: cinnamoyl-CoA reductase (CCR). Tree Genet Genomes 8:353–364

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

Meuwissen T, Karlsen A, Lien S, Olsaker I, Goddard ME (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379

Myburg AA, Potts BM, Marques CM, Kirst M, Gion J-M, Grattapaglia D, Grima-Pettenatti J (2007) Eucalypts. In: Kole C (ed) Forest trees. Springer Berlin Heidelberg, Berlin, pp 115–160

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325–330

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11: 149–155

Nei I (1987) Molecular evolutionary genetics. Columbia University Press, New York

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312

Pérez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163(4):1497–1510

Poke FS, Vaillancourt RE, Elliott RC, Reid JB (2003) Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). Mol Breed 12:107–118

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Géré P, Cato S, Plomion C (2005) Nucleotide variation in genes involved in wood formation in two pine species. New Phytol 167:101–112

Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177:1713–1724

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111

Raymond CA (2002) Genetics of *Eucalyptus* wood properties. Ann For Sci 59:8

Roldan DL, Gilbert H, Henshall JM, Legarra A, Elsen JM (2012) Fine-mapping quantitative trait loci with a medium density marker panel: efficiency of population structures and comparison of linkage disequilibrium linkage analysis models. Genet Res 94(4):223–234

Rozas J, Sánchez-Del Barrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830

Thamarus KA, Groom K, Bradley A, Raymond CA, Schimleck LR, Williams ER, Moran GF (2004) Identification of quantitative trait loci for wood and fibre properties in two full-sib pedigrees of *Eucalyptus globulus*. Theor Appl Genet 109:856–864

Thavamanikumar S, McManus LJ, Tibbits JFG, Bossinger G (2011) The Significance of single nucleotide polymorphisms (SNPS) in "*Eucalyptus globulus*" breeding programs. Aust For 74:23

Thumma BR, Nolan MF, Evans R, Moran GF (2005) Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257–1265

Thumma BR, Baltunis BS, Bell JC, Emebiri LC, Moran GF, Southerton SG (2010) Quantitative trait locus (QTL) analysis of growth and vegetative propagation traits in *Eucalyptus nitens* full-sib families. Tree Genet Genomes 6:877–889

Wachowiak W, Balk PA, Savolainen O (2008) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). Tree Genet Genomes 5:117–132

Yu J, Pressoir G, Briggs W, Vroh BI, Yamasaki M et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208