THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE SCIENCES ET ENVIRONNEMENTS

Par **Eric MANDROU**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPECIALITE : ECOLOGIE EVOLUTIVE, FONCTIONNELLE ET DES COMMUNAUTES

Variabilité fonctionnelle de gènes candidats de la lignification chez l'eucalyptus.

Devant la commission d'examen formée de :

Catherine BASTIEN	Directeur de recherche, INRA, Orléans	Rapporteur
Jacques DAVID	Professeur, SupAgro, Montpellier	Rapporteur
Jean-Pierre RENAUDIN	Professeur des Universités, Bordeaux I	Examinateur
Jacqueline GRIMA-PETTENATI	Directeur de recherche, CNRS, Toulouse	Examinateur
Jean-François RAMI	Chercheur, CIRAD, Montpellier	Examinateur
Jean-Claude PROUHEZE	Société Vallourec, Aulnoye Aymeries	Examinateur
Christophe PLOMION	Directeur de recherche, INRA, Bordeaux	Directeur
Jean-Marc GION	Chercheur, CIRAD, Bordeaux	Co-directeur

Remerciements :

Tout d'abord je tiens à remercier Antoine Kremer pour m'avoir accueilli au sein de son unité. Ces années passées à l'UMR Biogeco on été très enrichissantes tant sur le plan professionnel que sur le plan personnel.

Je souhaite également remercier Mme Catherine Bastien et Monsieur Jacques David, d'avoir accepté d'être les deux rapporteurs de ce travail, Mme Jacqueline Grima-Pettenati, Mr Jean-François Ramy ainsi que Mr Jean-Pierre Renaudin d'avoir bien voulu faire partie de mon jury de thèse avec une attention toute particulière pour toi Jacqueline qui est celle qui m'a mis le pied à l'étrier ...

Je remercie la société Vallourec pour avoir financé ces travaux ainsi que ma bourse de thèse et plus particulièrement Monsieur Jean-Claude Prouhèze pour avoir encadré mon travail tout au long de ces quatre années. Entant qu'expatrié du CEV (Centre de recherche Vallourec) je voudrai remercier également Mme Geneviève Ocquident d'avoir assuré mon lien avec la société Vallourec.

Je remercie toute l'Equipe de l'UPR 39 du CIRAD, et tout particulièrement Mr Frédéric Mortier pour son aide apportée tout au long de ce travail notamment sur les aspects statistiques, Mr Philippe Vigneron pour ses nombreux conseils et sa vision experte sur les eucalyptus, Mr Gilles Chaix pour son aide et ses conseils sur la SPIR et Mme Roselyne Lannes, ma secrétaire préférée ...

Merci aussi à toute l'équipe du CRDPI Congo et l'équipe du CAPEF Brésil pour leur accueil chaleureux durant les missions de collectes du matériel végétal. Ces remerciements s'adressent tout particulièrement côté Congo à Andreas Ndeko, Andrée Mabiala, maman Juju, ainsi que Joël Polidori, Juste Akana, Dame Agnès, Maurice et Aubin Saya et côté Brésil, José Luis Lima, Leonardo Chagas, Helder Bolognani et toute sa famille, Nivaldo, Helder et tous les autres... Merci à tous pour votre aide.

Un grand merci à tous les collègues de l'UMR Biogeco. Véronique, Florence, Corinne, Chantal, Loïc, Thierry, tous ceux qui m'ont aidé dans mon travail et particulièrement Camille et

Laurent, surtout sur la fin de la thèse (toc toc toc ... je peux poser une question ??), toute l'équipe de la plateforme de séquençage pour la disponibilité, l'efficacité et l'accueil au sein de ce qui a un peu été ma deuxième maison (la salle du séquençeur), Philou, Henri, Grégoire, Jérémy, Franck merci aussi pour vos conseils dans le travail de laboratoire.

Une spéciale dédicace à tous les Djeuns de l'équipe, Emilie (Miliki), Sarah, Florian, Pierre, Jean-Paul, Christophe (Gran Gran), Emilie (Miloud), François, Erwan, ainsi que ceux qui n'ont fait que passer Frank, Tristan, Julie, Gensériс, Guillermo, Pablo, merci à tous pour vos conseils, votre aide ou simplement votre bonne humeur quotidienne (ou passagère).

Un grand merci à mes deux directeurs de thèse, Jean-Marc Gion et Christophe Plomion pour m'avoir permis de participer à cette expérience formidable qu'est la thèse. Je vous remercie également pour ces expériences inoubliables de voyages et de rencontres au travers des différentes missions et congrès auxquels j'ai pu participer et pour m'avoir guidé dans ce travail de recherche et m'avoir aidé à progresser un peu plus sur le chemin de la connaissance (même s'il reste encore du chemin à parcourir). Merci plus particulièrement à toi Jean-Marc pour m'avoir fait confiance dès le début, pour avoir cru en moi, pour m'avoir transmis ta passion communicative, merci pour ta franchise, pour avoir su me remotiver dans les moments de doutes, et bien d'autres choses encore.

Enfin, merci à ma famille, si j'en suis là c'est à vous que je le dois plus particulièrement, merci aussi à toi Julie de m'avoir supporté pendant ces 3 années, maintenant tu redeviens ma priorité number one. Merci aussi à Bibi mon frangin et tous mes potes pour m'avoir permis de prendre un peu de recul par rapport à tout ça, pour m'avoir permis de sortir un peu la tête du guidon, Merci à tous.

Sommaire

Liste des figures

Liste des tableaux

Liste des abréviations

4CL	4-coumarate:CoA ligase
ADN	Acide DésoxyriboNucléique
ADNc	Acide DésoxyriboNucléique complémentaires
AFLP	Amplified Fragment-Length Polymorphism
ARN	Acide RiboNucléique
ARNt	Acide RiboNucléique de transfert
BF	Bayes factor
BLUP	Best Linear Unbiased Prediction
C3H	p-coumarate 3-hydroxylase
C4H	cinnamate 4-hydroxylase
CAD	cinnamyl alcool dehydrogenase
CAPS	Cleaved Amplified Polymorphic Sequences
CCoAOMT	Caffeoyl coenzyme A O-methyltransferase
CCR	cinnamoyl-CoA reductase
CIRAD	Centre de Coopération International en Agronomie pour le Développement
cM	centimorgan
COMT	acide caféique O-méthytransférase
CRDPI	Centre de Recherche sur la Durabilité et la Productivité des Plantations Industrielles
DL	Déséquilibre de liaison
EST	Marqueur de Séquences Exprimées
EUCAGEN	Eucalyptus Genome Network
F5H	ferulate 5-hydroxylase
FAM	Fréquence d'Allèle Minoritaire
FAO	Food and Agriculture Organization
FDR	False Discovery Rate
FT	Facteur de Transcription
G	guaiacyl
GL	Groupe de Liaison
GWAS	Genome Wide Association Study
H	p-hydroxyphenyl

h^2	héritabilité au sens strict
HCT	p-hydroxycinnamoyl-CoA:quinate shikimate p-hydroxycinnamoyltransférase
He	Hétérozygoties attendues dans une population panmictique
HPLC	chromatographie liquide à haute pression
HRM	High Resolution Melting
ITTO	International Tropical Timber Organization
LK%	Pourcentage de lignines d éterminées avec la méthode Klason
matrice K	Kinship matrix
MFA	Angle des MicroFibrilles de cellulose
MOE	MOdule d'Elasticité
Mpb	Méga (10^6) paires de base
PAL	Phénylalanine Ammonia-Lyase
PCR	Polymerase Chain Reaction
PPA	Probabilité a Posteriori pour l'Association
QTL	Quantitative Trait Loci : régions génomiques à effet majeur
QTN	quantitative trait nucleotides
r_A	corrélations génétiques additives
RAPD	Random Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RMN	résonance magnétique nucléaire
r_P	corrélations phénotypiques
S	syringyl
SAM	Sélection Assistée par Marqueurs
SEcv	Erreur de prédiction (Standard Error)
SNP	Single Nucleotide Polymorphism
SPIR	Spectrométrie Proche Infra-Rouge
SSCP	Single Strand Conformation Polymorphism
SSR	Simple Sequence Repeat
TDT	transmission/disequilibrium test
UE	Union Européenne
UTR	régions transcrites non traduites
V&M	Vallourec et Mannesmann do Brasil
XET	Xyloglucan EndoTransglycosylases

Préambule

Dans la cadre d'un accord de coopération entre le CIRAD et la société Vallourec & Mannesman do Brasil, trois actions de recherche ont été développées afin d'améliorer la production de charbon de bois d'eucalyptus utilisé dans les hauts fourneaux pour produire de l'acier. Il s'agissait : de tester de nouveaux procédés de carbonisation du bois, de mettre en place un test d'évaluation de la qualité du bois et du charbon par spectrométrie dans le proche infrarouge, et de développer des méthodes de génétique quantitative et des outils moléculaires pour sélectionner des variétés clonales performantes sur le plan de la production de biomasse et de la qualité du bois. Je présente dans cette thèse les résultats des recherches concernant le troisième volet. Elles ont été menées au sein de l'UMR 1202 (BIOGECO) à Bordeaux et ont fait l'objet d'un contrat CIFRE entre le CIRAD et V&M France CEV débuté en Juillet 2007.

Le manuscrit s'articule autour de 5 chapitres.

Le **chapitre introductif** est très large. Il a été conçu pour donner au lecteur **les bases scientifiques** lui permettant d'aborder le sujet de la thèse. Il traite du contexte économique de l'amélioration génétique des arbres forestiers, reprend quelques fondamentaux sur la xylogenèse et l'amélioration de la qualité du bois en s'attardant sur la lignine (et sa biosynthèse), un composé majeur pour la production de pâte à papier et de charbon de bois. Il termine sur les méthodes de détection du polymorphisme de l'ADN, la caractérisation du niveau et de la structure de la diversité nucléotidique, et des considérations statistiques qui entourent les études d'association entre polymorphisme nucléotidique et variabilité phénotypique.

Le **second chapitre porte sur le matériel et les méthodes.** Il décrit l'ensemble du matériel végétal et les gènes utilisés, les mesures phénotypiques réalisées, l'approche suivie pour caractériser la variabilité moléculaire ainsi que les analyses statistiques mises en œuvre pour analyser les patrons de diversité nucléotidique, le déterminisme de la variation phénotypique et la co-variation entre ces deux niveaux de variabilité.

Dans le **troisième chapitre**, une étude du **déterminisme génétique des caractères liés aux lignines** (quantité et qualité) est menée dans trois fonds génétiques: un contexte intra-spécifique (*E.urophylla* x *E. urophylla*) et deux contextes inter-spécifiques (*E. camaldulensis*

x *E. urophylla* et *E. urophylla* x *E. grandis*). Une méthode de phénotypage à haut-débit (spectrométrie proche infrarouge) a été mise en œuvre pour caractériser la variation des propriétés du bois dans ces trois plans de croisement. A l'issu de cette étude, nous disposions des composantes de la variance et des mesures phénotypiques qui serviront de données d'entrée pour le chapitre 5.

Le **chapitre 4**, aborde le niveau de **diversité nucléotidique ainsi que l'étendue du déséquilibre de liaison** pour 10 gènes de la voie de biosynthèse des lignines avec une emphase particulière sur le gène codant la cynnamoyl CoA réductase (CCR). Cette caractérisation a été menée pour deux espèces clefs du genre *Eucalyptus* (*E. urophylla* et *E. camaldulensis*) à partir du matériel utilisé au chapitre 3. Une comparaison a également été entreprise avec les résultats obtenus dans la littérature pour d'autres espèces du même genre ainsi que d'autres espèces d'arbres forestiers. A l'issu de cette analyse, nous disposions des marqueurs moléculaires (de type SNP) qui serviront de variables d'entrée pour le chapitre 5.

Le **chapitre 5** fait la synthèse des différents jeux de données (phénotypes et marqueurs moléculaires) en traitant de **l'association entre ces deux niveaux de variabilité**. Il discute des limites de l'approche mise en œuvre dans le cadre de ce travail de thèse et ouvre des perspectives pour la sélection assistée par marqueurs qui sont discutées en **conclusion**.

Une partie des travaux réalisés dans le cadre de ce travail a fait l'objet d'un article soumis dans une revue de rang A, accessible à l'annexe 1.

Chapitre 1 : La qualité du bois : une nouvelle cible de l'amélioration génétique des arbres

1. Le bois une ressource pour l'industrie

1.1. Contexte économique

1.1.1. Le marché et les besoins en bois

Le bois est un matériau biologique complexe dont les propriétés mécaniques, thermiques et énergétiques sont exploitées par l'homme depuis des millénaires. Aujourd'hui, ce matériau à usages multiples est utilisé aussi bien de manière industrielle qu'artisanale pour répondre à différents besoins. On distingue classiquement différents types de bois en fonction de l'usage auquel ce matériau est destiné : le bois de trituration (panneaux de particules et pâtes pour le papier et le carton), le bois d'œuvre (produits de sciage), le bois de service (fabrication de poteaux et de perches), le bois d'énergie (bois de chauffe, charbon de bois)… Le marché du bois et de ses produits dérivés tient une place importante dans le commerce international. En 2006, ces produits étaient classés au 8ème rang des produits les plus échangés sur les marchés avec une valeur totale de 257 milliards de dollars soient 2,5% de la totalité des échanges internationaux. L'exploitation des ressources forestières participe à la création de nombreux emplois liés non seulement à la gestion et l'exploitation des forêts mais également à la transformation du bois en produits dérivés. En 2000, on estimait à 13 millions le nombre d'emplois dans le monde relatifs à la seule filière bois papier.

Selon la FAO, le volume de bois extrait chaque année représente environ 3,4 milliards de m^3 (FAO, 2009). Ces prélèvements sont destinés, pour moitié, à la production de bois rond pour un usage industriel. Compte tenu de l'accroissement de la population mondiale et des nouvelles voies de valorisation du bois, notamment en matière de production d'énergie verte, la production et la consommation de bois devraient connaitre une augmentation dans les années à venir. En 2030, les études prédisent une augmentation de la consommation globale de bois de 65% par rapport au niveau de 2005 (FAO, Carle et Holmgren, 2008) sans compter le secteur du bois énergie. Depuis les années 2000, poussée par les politiques mises en place pour lutter contre le réchauffement climatique et la hausse du

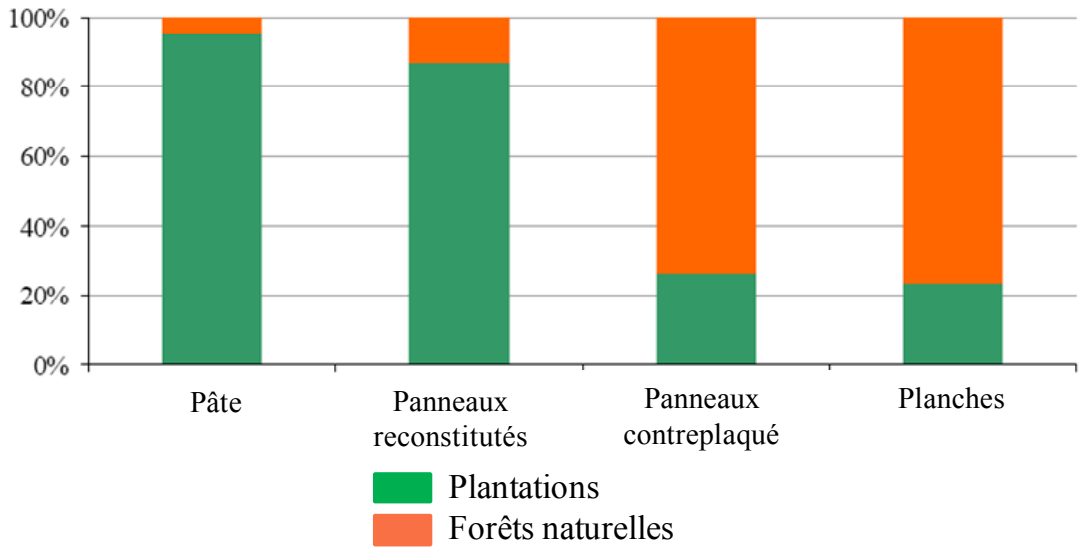


Figure 1 : importance relative des ressources de bois (plantation ou forêt naturelle) utilisées dans différents secteurs de l'industrie du bois dans les régions tropicales, (d'après ITEM 5, Global wood and wood products flow, Advisory committee on paper and wood products, FAO, 2007).

coût des énergies fossiles, l'utilisation du bois comme source d'énergie connait un essor considérable. Pour 2020, l'objectif de l'UE est de parvenir à 20% d'utilisation d'énergie « verte », basée sur l'utilisation de la biomasse de manière durable et « carboneutre ». Ces politiques ont notamment incité la mise en place de centrales à bois à différentes échelles, allant d'équipements municipaux de grande taille à la chaudière individuelle. Aujourd'hui, la production et le commerce de combustibles ligneux principalement issus du bois (billettes, granulés, buches reconstituées et autres formes) s'accroit pour satisfaire une demande en hausse. Le souci principal restant la disponibilité future de combustible liquide, les recherches s'orientent également vers la production de biocarburants à partir du bois. Le développement rapide de ces utilisations pose cependant le problème de la sécurité alimentaire du point de vue des surfaces consacrées à la culture intensive du bois. L'évolution future des marchés dans le domaine du bois énergie, bien que très prometteuse, reste par conséquent difficile à prévoir.

1.1.2. Les plantations forestières

Pour le moment, l'approvisionnement en bois reste très dépendant de l'exploitation des forêts naturelles. Dans les années 1990, on estimait que la perte nette de la surface des forêts naturelles était de 8,3 millions d'ha/an. Le niveau de ces pertes a diminué dans les années 2000 pour atteindre 5,2 millions d'ha/an dans la période 2000-2005 (FAO, 2009). Aujourd'hui, la surface totale de forêt sur la planète est estimée à un peu plus de 4 milliards d'ha. Le ralentissement des pertes de surfaces forestières mondiales est attribué en partie à l'augmentation des surfaces dédiées aux plantations forestières (de 180 millions d'ha en 1990 à 225 millions d'ha en 2005). Ce type de formation forestière répond à deux objectifs principaux : l'un de production pour 73% des surfaces et l'autre de protection (FAO, 2007).

D'après une étude de l'ITTO (International Tropical Timber Organization) réalisée en 2006, la forêt plantée productive couvre environ 187 millions d'hectares, ce qui représente 5% des surfaces forestières mondiales. De plus, on estime que la part réellement disponible pour la production industrielle de bois représente environ 50% de ces surfaces. Néanmoins, la forêt productive assure environ 50% de la production de bois rond (FAO, 2007). Certaines filières de l'industrie du bois sont aujourd'hui très dépendantes des plantations forestières pour leur approvisionnement. La Figure 1 (d'après ITTO, 2006) représente la part de bois issue de plantations et de forêts naturelles utilisée par différents secteurs de l'industrie du bois dans les régions tropicales (Asie Pacifique, Afrique, Amérique Latine, Caraïbes). La majorité de la

pâte à papier (95%) et des panneaux de particule (85%) produits dans ces régions du globe est basée sur l'utilisation de bois issu de plantations industrielles. Si la part du bois issu de ces plantations reste faible dans le cas de la production de panneaux de contreplaqué et de planches (environ 20%), elle est en constante augmentation. Le développement au niveau européen de nouvelles filières dans le domaine du bois énergie, devrait encore accroitre le développement de la forêt plantée pour répondre aux nouveaux besoins en matière d'énergie verte.

Le succès de la forêt plantée peut être principalement imputé à des raisons économiques, même si de plus en plus la dimension écologique entre en ligne de compte. La forêt plantée peut être perçue comme le résultat de la domestication des espèces forestières. Selon Libby (FAO, 2002), cette domestication repose sur un ensemble de composantes qui permettent à cette culture d'être beaucoup plus productive et donc plus rentable que la forêt naturelle. Les niveaux de productivité des forêts plantées sont de l'ordre de 15 à 40 m^3/ha/an selon le type d'espèce plantée contre 4 m^3/ha/an estimés en moyenne pour la forêt naturelle (Libby, FAO 2002).

1.2. Une domestication récentes des arbres forestiers

Les niveaux de productivité atteints par les plantations forestières résultent de la mise en place de deux composantes majeures : i/ la maîtrise partielle de l'environnement grâce à l'amélioration des pratiques culturales (itinéraires techniques de plantations) mais qui reste rapidement limitée du fait de la taille et de la longévité de ce type d'organismes, ii/ des développements techniques qui ont permis de maîtriser la pollinisation contrôlée et le bouturage herbacée, permettant la mise en place de programmes d'amélioration génétique pour la sélection des meilleurs génotypes pour la production de bois.

1.2.1. Itinéraires techniques et pratiques sylvicoles

Les composantes techniques et environnementales ont un rôle majeur dans les niveaux de productivité atteint. Elles regroupent l'ensemble des pratiques sylvicoles qui s'appliquent à la plantation depuis la préparation des plants jusqu'à leur mise en place dans les sites sélectionnés. Il s'agit par exemple du contrôle de la préparation et de la qualité des propagules en pépinières, de la sélection des meilleurs sites de plantations sur la base de différentes contraintes biotiques et abiotiques, de la préparation des sols à la mise en place des jeunes plants en termes d'humidité et de fertilisation ou encore du contrôle de la densité de plantation

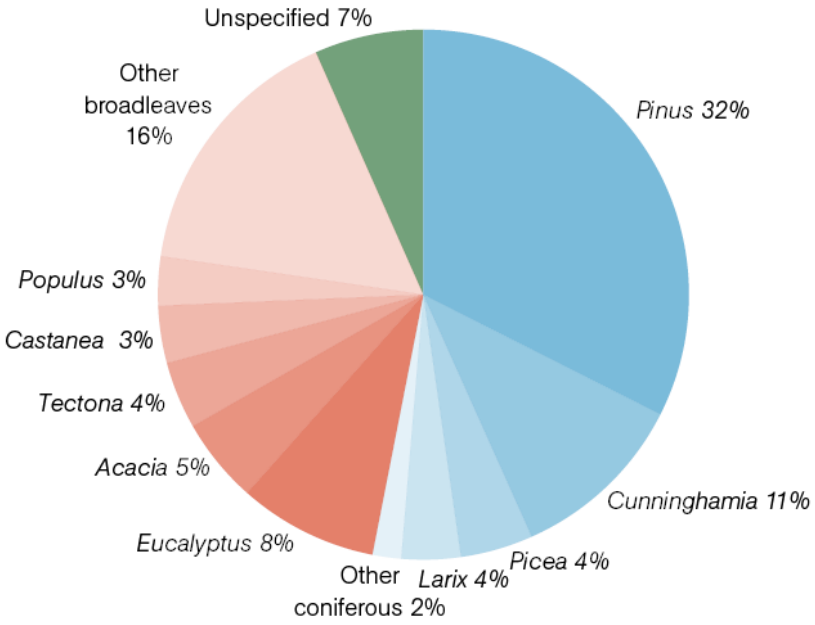


Figure 2 : importance relative des espèces forestières en plantation productive (par genres), (d'après FAO, States of the world's forests 2007).

(Libby, FAO 2002). Pallett et Sale (2004) ont évalué les gains de productivité obtenus pour différentes espèces d'*Eucalyptus* en Afrique du sud par la seule prise en compte de différents paramètres techniques et environnementaux (sites de plantation, densité de plantation et mise en place de sylvicultures intensives). Cette étude rapporte une multiplication par 4 des rendements obtenus (en m^3/ha) par la prise en compte des optimums pour ces paramètres.

Toutefois, l'amélioration de la productivité par un meilleur contrôle de l'environnement présente des limites. En effet, la taille et la durée de vie de ce type d'espèces sont deux contraintes majeures comparées à celles d'espèces annuelles de grande culture. Les cycles de révolution pour des espèces à croissance rapide utilisées en plantation restent relativement longs (30 à 40 ans chez le pin maritime, 15 à 20 ans chez le peuplier, 7 à 10 ans chez l'eucalyptus…). Dans ce contexte, l'amélioration génétique d'espèces faiblement domestiquées (proche de l'état sauvage) offre des perspectives supplémentaires en termes de gains de productivité à moyen et long terme.

1.2.2. Les programmes d'amélioration génétiques des arbres forestiers

L'amélioration génétique est basée sur l'exploitation et la mise à profit de la variabilité génétique naturelle. Des niveaux de diversité génétique importants sont observés au sein des populations d'arbres forestiers. En plus de la longévité des génotypes, trois facteurs majeurs expliquent l'importance de la variabilité génétique : i/l'hétérozygotie des génotypes, ii/ un régime de reproduction préférentiellement allogame (Hamrick *et al.* 1992) et iii/ des tailles de population généralement grandes. Cette variabilité génétique observée au sein des populations d'arbres forestiers constitue une source de gains génétiques importants.

Les espèces cibles des programmes d'amélioration appartiennent à quelques genres principaux utilisés en plantation et présentent des caractéristiques de croissance intéressantes parmi l'ensemble des espèces d'arbres forestiers. Parmi les genres les plus plantés dans le monde, les conifères représentent plus de 50% des surfaces avec le genre *Pinus* qui totalise à lui seul 32% des surfaces (*P. taeda*, *P. radiata*, *P. pinaster*, *P. caribea*…). Les feuillus représentent 37% des plantations avec quatre genres majeurs : *Eucalyptus* (8%), *Acacia* (5%), *Tectona* (4%) et *Populus* (3%) (Figure 2).

Les programmes d'amélioration génétique pour ces espèces remontent aux années 50-60 et commencent avec une sélection massale dans des peuplements naturels, dans le cas d'espèces autochtones (cas du pin maritime en France), ou suite à des campagnes de récoltes

Encadré 1 : ***Généralités sur les composantes de la variance phénotypique***

La variabilité génétique conditionne une partie de la variation des caractères phénotypiques ainsi que leur transmission de génération en génération. L'amélioration génétique vise à sélectionner une ou plusieurs combinaisons de variants génétiques qui vont maximiser la valeur des caractères phénotypiques d'intérêt. Elle nécessite de comprendre les liens qui existent entre variabilité génétique et variation phénotypique.

On distingue deux types de caractères selon leur niveau de complexité : i/ les caractères mendéliens, qui sont souvent des caractères qualitatifs et qui n'impliquent qu'un nombre limité de gènes et ii/ les caractères quantitatifs dont la variation est continue et qui impliquent généralement un grand nombre de gènes. Pour un caractère phénotypique, on parle de déterminisme génétique total si la variation phénotypique peut être reliée directement à la variabilité génétique sous jacente. Lorsque le déterminisme génétique d'un caractère est partiel, il est nécessaire de faire appel à l'utilisation de modèles génétiques pour relier les deux niveaux de variabilité phénotypique et génétique.

En population, la variation d'un caractère quantitatif est caractérisée par une valeur centrale (généralement la moyenne) et des paramètres de dispersion (écart type ou variance). La génétique quantitative vise à déterminer le mode d'expression et de transmission d'un caractère quantitatif complexe, dans une population, par l'étude du phénotype et des liens de parenté qui existent entre les individus qui la composent. Les phénotypes observés (P) pour un caractère d'intérêt peuvent être expliqués par un modèle statistique représentant la contribution d'un génotype non observé (G) et d'un ensemble inconnu de facteurs environnementaux (E) *(en l'absence d'interaction GxE)*.

$P = G + E$

La variance du phénotype observé (σ^2_P) peut être décomposée en une somme de variances non observées (σ^2_G et σ^2_E)

$\sigma^2_P = \sigma^2_G + \sigma^2_E$.

Ainsi, l'héritabilité au sens large, H^2, se définit comme un ratio de variances qui exprime la proportion de la variance phénotypique qui peut être attribuée à la variation des génotypes.

$H^2 = \sigma^2_G / \sigma^2_P$

La variance génétique peut elle-même être décomposée en la somme de la variance due aux effets génétiques additifs (σ^2_A aussi appelée « Breeding value »), de la variance due aux interactions entre allèles d'un même locus (σ^2_D pour variance de dominance) et de celle due aux interactions entre allèles de locus différents (σ^2_I pour variance d'interaction).

$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$

L'héritabilité au sens strict d'un caractère quantitatif est définie comme le ratio entre la variance génétique additive et la variance phénotypique et s'écrit

$h^2 = \sigma^2_A / \sigma^2_P$

L'héritabilité est un concept utile à la comparaison de l'importance relative des effets génétiques (totaux ou additifs) et environnementaux dans la variation de caractères quantitatifs au sein de populations et entre populations. Il permet également d'évaluer la réponse de caractères quantitatifs à la sélection en biologie évolutive et trouve également des applications dans les domaines de l'agriculture et de la médecine.

de graines pour les espèces exotiques (cas de l'eucalyptus au Brésil et au Congo ; Martin et Cossalter, 1976 a, 1976 b, 1976 c). Les premiers tests de provenances/descendances constitueront alors les populations d'amélioration de base de ces programmes. La maîtrise de la pollinisation contrôlée a ensuite permis de croiser les génotypes entre eux par hybridation intraspécifique (cas du pin maritime) ou interspécifique (cas chez le peuplier et l'eucalyptus). Lorsque la multiplication par bouturage est possible, les meilleurs génotypes obtenus par croisement sont multipliés et constituent les sorties variétales pour des plantations clonales. Dans d'autres cas, la mise en place de vergers à graines permet la production de variétés « semences » correspondant généralement à des « polycross » de géniteurs sélectionnés.

Ces programmes de sélection se font sur plusieurs générations et impliquent plusieurs cycles de croisements et de sélection des meilleurs individus. Ils sont basés sur la mise en place de dispositifs expérimentaux permettant de déterminer, pour les caractères cibles, la part des effets génétiques et environnementaux impliqués dans le contrôle de leur variation (héritabilité des caractères et composantes de la variance) (encadré 1). L'estimation de ces effets pour le(s) caractère(s) cible(s) permet de discriminer, au sein des populations, les meilleurs génotypes parentaux ainsi que les meilleures familles ou génotypes utilisables en sortie variétale. Si les gains génétique sont ainsi étalés sur plusieurs générations, on peut attendre des bénéfices importants dès les premiers cycles de sélection (encadré 2). Li *et al.* (1999) rapportent des gains de volume de 26 à 35% (volume de bois récolté) réalisés après 2 cycles d'amélioration génétique de *Pinus taeda* dans le sud des Etats Unis.

Pour le moment, la plupart des programmes d'amélioration menés chez les arbres forestiers ont pour objectif l'augmentation de la productivité. Ils sont donc basés, pour la plupart, sur l'amélioration des caractères de croissance tels que la hauteur totale ou la circonférence (volume du fût). La résistance aux bioagresseurs peut, chez certaines espèces comme le peuplier, s'avérer un critère de sélection déterminant pour la productivité. Des caractères morphologiques, tels que la rectitude du tronc ou la branchaison, sont également pris en compte dans certains cas où ils constituent des facteurs dépréciant la qualité du bois (cas du bois de sciage pour le pin maritime). Cependant, l'amélioration génétique pour la croissance présente deux contraintes majeures : d'une part une héritabilité faible à moyenne avec un fort effet de l'environnement sur la variation du phénotype notamment au jeune âge, et d'autre part de faibles corrélations juvénile-adulte rendant la sélection des meilleurs génotypes possible au tiers de la durée de rotation : 15 ans chez le pin maritime, 5 ans chez le peuplier et 3 ans chez l'eucalyptus.

Encadré 2 : ***Réponse à la sélection d'un caractère quantitatif***

La réponse à la sélection d'un caractère quantitatif correspond à la différence entre la moyenne du caractère à la génération suivante et celle à la génération initiale. On la note R et elle se calcule comme suit :

$R = h^2.S$

où h^2 est l'héritabilité au sens strict du caractère et S est la différentielle de sélection. S exprime la différence entre les individus sélectionnés pour la reproduction et la moyenne de la population initiale.

La différentielle de sélection peut se calculer comme suit :

$S = i.\sigma^2_P$

où σ^2_P est l'écart type de la distribution de la valeur des phénotypes dans la population et i est l'intensité de sélection relative au taux de sélection.

L'efficacité de la sélection est donc relative à la fois au niveau de variation phénotypique du caractère dans la population, à l'héritabilité au sens strict de ce caractère (c'est-à-dire la part des effets génétiques additifs dans la variation du phénotype) et à l'intensité de la sélection qui est appliquée pour obtenir la génération suivante.

1.3. Le cas de l'eucalyptus

En tant que première espèce feuillue plantée dans le monde (plus de 20 millions d'ha de plantations), l'*Eucalyptus* est un bon exemple d'arbres forestiers utilisés en production intensive pour lesquels différents programmes d'amélioration génétique sont menés.

1.3.1. Le genre Eucalyptus

Le genre *Eucalyptus* regroupe près de 700 espèces, en sous-séries, séries, sections et sous-genres d'inégale importance (Brooker, 2000). Le sous-genre le plus important, tant sur le plan du nombre d'espèces que sur le plan de la variété des formes et de la diversité des habitats qu'elles occupent est le sous-genre *Symphyomyrtus* avec plus de 550 espèces (revues par Vigneron et Bouvet, 2000). Les espèces de ce sous-genre sont présentes sur l'ensemble du continent Australien (incluant la Tasmanie) ainsi qu'en Nouvelle-Guinée et dans l'archipel des îles de la Sonde (Indonésie) et sont capables de pousser à des altitudes allant du niveau de la mer jusqu'à 1800 mètres. Parmi l'ensemble des espèces d'*Eucalyptus*, une dizaine, appartenant majoritairement au sous-genre *Symphyomyrtus*, est largement utilisée en plantation dans les deux hémisphères.

Le succès de ces espèces est dû principalement à leur croissance rapide, permettant d'obtenir des rendements en bois importants, et à leur faculté d'adaptation à des conditions pédoclimatiques variées. Elles sont plantées dans la plupart des régions tropicales et tempérées entre 45° de latitude sud et 40° de latitude nord et jusqu'à des altitudes de 3000 m au Pérou (Figure 3). Ces espèces présentent d'autres avantages appréciés des améliorateurs : i/ une bonne aptitude au bouturage qui permet de déployer les variétés sélectionnées en plantations clonales. ii/ certaines espèces peuvent être hybridées, permettant de créer des variétés plus vigoureuses combinant les meilleures caractéristiques de deux espèces. Cependant, il faut noter que l'hybridation entre les espèces reflète assez bien leur classification et reste limitée aux espèces d'un même sous-genre, généralement entre les séries ou au sein d'une même série (Griffin *et al.*, 1988 ; Baril *et al.*, 1997 a et b).

1.3.2. L'amélioration génétique des eucalyptus

Plusieurs programmes d'amélioration génétique des eucalyptus sont conduits au Brésil, en Afrique, en Europe, ou en Indonésie. Ces programmes d'amélioration sont menés dans la plupart des cas pour la sélection de génotypes améliorés destinés à l'industrie de la pâte à papier mais dans certains cas, ils sont orientés vers d'autres finalités comme par exemple la

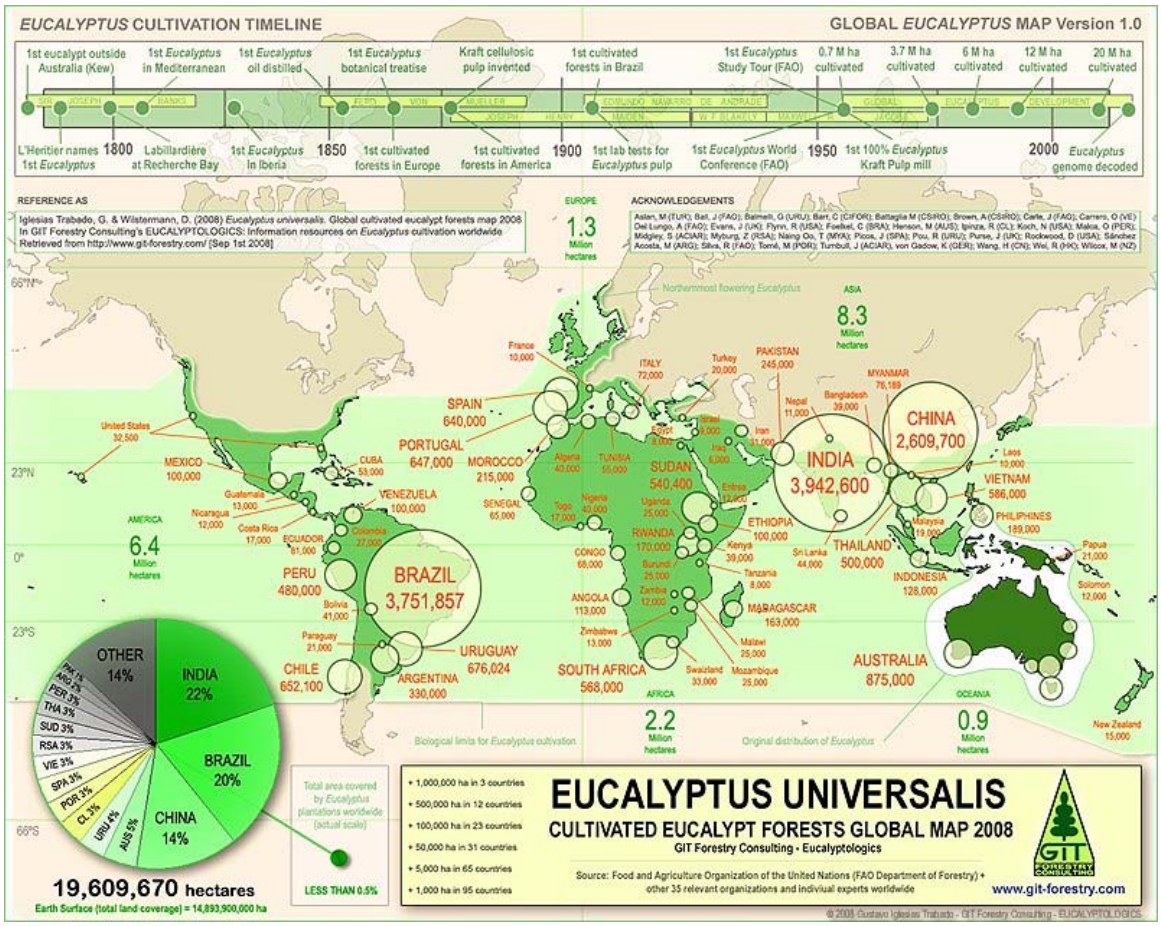


Figure 3 : répartition des surfaces plantées d'eucalyptus dans le monde (d'après Iglesias et Wiltermann, 2008).

production de variétés pour le bois d'énergie (charbon de bois destiné à des applications industrielles).

Au Congo, le programme d'amélioration des *Eucalyptus* a été mis en place dans les années 1950. Suite à la maîtrise de la technique de pollinisation contrôlée, plusieurs combinaisons hybrides interspécifiques ont été testées. L'une des combinaisons les plus performantes est le croisement entre *E. urophylla* et *E. grandis* ; *E. urophylla* étant utilisé comme parent femelle et *E. grandis* comme parent mâle (Vigneron, 1995)**.** Depuis 1989, un schéma de sélection récurrente réciproque a été développé pour ces deux espèces, avec pour objectif la production d'hybrides interspécifiques combinant la bonne adaptation d'*E. urophylla* aux conditions pédoclimatiques du Congo (climat tropical humide et sols pauvres) et les fortes potentialités de croissance d'*E. grandis*. Ce programme de sélection est mené principalement pour des caractères de croissance et permet de fournir de nouvelles variétés clonales pour 46 000 ha de plantations destinées à l'industrie papetière. Deux à trois rotations de 7 ans sont réalisées à partir d'une plantation et permettent d'obtenir des rendements moyens d'environ 20 m^3/ha/an.

Au Brésil, le programme d'amélioration mis en place par Vallourec et Mannesmann do Brasil (V&M do Brasil) date des années 1960. Il est principalement basé sur l'amélioration d'espèces hybrides obtenues du croisement entre les espèces *E. camaldulensis* et *E. urophylla*, la première étant utilisée comme parent femelle et la deuxième comme parent mâle. Ces hybrides cumulent la bonne croissance de l'espèce *E. urophylla* au Brésil avec une forte teneur en lignines pour *E. camaldulensis,* plus favorable à la production de charbon bois pour des applications industrielles comme la production d'acier. Egalement basé sur l'amélioration des caractères de croissance pour l'obtention de meilleurs rendements, ce programme d'amélioration génétique fournit les nouvelles variétés hybrides destinées à 180 000 ha de plantations répartis dans l'état du Minas Gerais. La majorité des variétés plantées sont issues de clones de première génération de sélection. Ces variétés sont également cultivées sur trois rotations de 7 ans et permettent d'atteindre des rendements moyens d'environ 30 m^3/ha/an.

1.4. De nouveaux caractères d'intérêt pour l'amélioration génétique des eucalyptus

Dans le cadre des changements globaux (accroissement de la population mondiale, évolution rapide des marchés du bois, changements climatiques…), des travaux de recherches

sont menés pour l'intégration de nouveaux caractères d'intérêt économique (qualité du bois) et/ou écologique (caractères adaptatifs) au sein des programmes d'amélioration génétique.

La « qualité du bois », qui peut se définir comme la capacité d'un bois à satisfaire des besoins en termes d'exploitation industrielle, est au cœur de ces nouveaux enjeux. Elle se définit par la mesure d'un ensemble de propriétés qui, en fonction du processus de transformation industriel ou de l'utilisation finale du bois, pourront être très différentes. C'est le cas chez l'eucalyptus, arbre à usages multiples, pour lequel le bois est utilisé aussi bien comme bois d'énergie, bois de service, bois de trituration ou encore bois d'œuvre. Les propriétés du matériau bois pourront être déclinées à des échelles macroscopiques ou microscopiques. On distingue classiquement des propriétés physiques, mécaniques, anatomiques et chimiques. Une étude récente de Schmidt (2005) évalue, par une modélisation simple, l'impact de différentes propriétés morphologiques et chimiques du bois d'*Eucalyptus* sur les coûts de l'obtention de pâte à papier. Ces modèles montrent que les caractéristiques chimiques telles que la teneur en cellulose, hémicellulose et lignines sont celles qui impactent le plus fortement les coûts de production. Dès lors, ces caractéristiques chimiques deviennent des caractères d'intérêt majeur dans le cadre des programmes d'amélioration des *Eucalyptus* menés pour la filière bois papier. La prise en compte de la « qualité du bois », ou plutôt de ses propriétés sous jacentes, comme caractère(s) cible(s) de l'amélioration génétique nécessite une bonne connaissance du matériau et des processus qui participent à sa formation et à sa mise en place dans la plante.

2. Le bois un matériau biologique complexe

2.1. Rôle biologique

La colonisation du milieu terrestre par les plantes remonte à environ 400 millions d'années et constitue l'une des étapes les plus importantes de l'histoire du monde biologique. Cette transition a notamment été rendue possible par la mise en place du système vasculaire permettant de résoudre le problème de l'approvisionnement en eau et en éléments minéraux. Le système vasculaire des arbres constitue le dernier stade de cette évolution et sa mise en place implique l'action coordonnée d'un ensemble de processus complexes.

Le bois, aussi appelé « xylème secondaire », constitue donc une partie du système vasculaire des plantes. Il assure trois rôles majeurs. Tout d'abord, il permet le transport de la

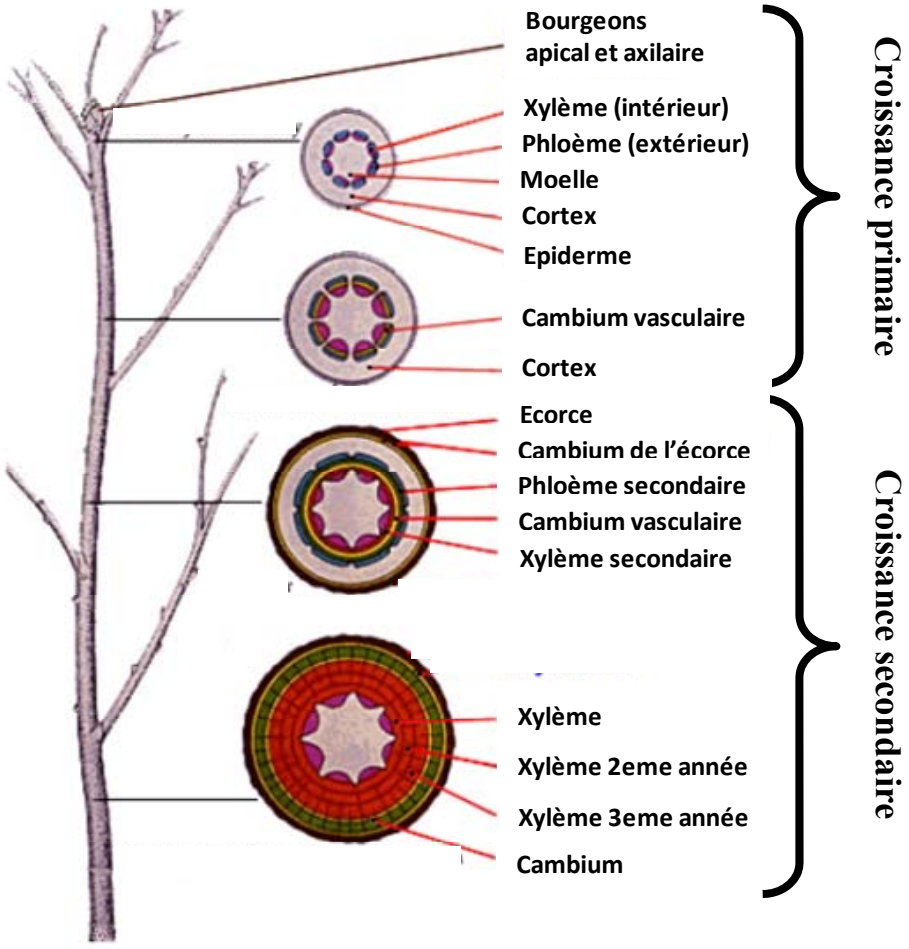


Figure 4 : schéma de coupes transversales de tiges représentant l'organisation des tissus vasculaires chez les plantes ligneuses pérennes durant les phases de croissance primaire et secondaire (Issu de http://preuniversity.grkraj.org/html/3_PLANT_ANATOMY.htm)

sève brute et des éléments minéraux qu'elle contient depuis les racines vers les apex caulinaires de la plante, ensuite, il permet à la plante de tenir un port dressé, et enfin, il lui permet d'emmagasiner des réserves nécessaires à sa croissance. Il est donc à la fois un tissu conducteur, de soutien et de réserve. Chez les angiospermes, ces différents rôles sont assurés par différents types de cellules. Les vaisseaux assurent la conduction de la sève brute, les fibres assurent le rôle de soutien, et les cellules du parenchyme assurent le transfert des nutriments depuis le phloème vers le xylème et le stockage de réserves sous forme de sucres (amidon) ou de lipides.

2.2. La formation du bois

2.2.1. Croissance secondaire chez les plantes

Après la phase de développement embryonnaire, la croissance de la plante est assurée par des tissus appelés méristèmes. Durant les premiers stades du développement de la plante, seuls les méristèmes apicaux (méristèmes primaires) vont participer à la production de nouvelles cellules. Ces cellules sont nécessaires à la formation et à la croissance des différents organes primaires de la plante (tiges, racines, feuilles). Cette phase est appelée croissance primaire et a lieu chez tous les végétaux. Chez certains d'entre eux, la phase de croissance primaire est poursuivie par une autre phase, appelée croissance secondaire, qui va leur permettre d'augmenter la circonférence des tiges et des racines. Cette croissance secondaire est alors initiée à partir de méristèmes secondaires, appelés cambium vasculaire et cambium cortical ou phellogène (Figure 4).

2.2.2. De la croissance primaire à la croissance secondaire des tissus vasculaires

La formation du xylème secondaire chez les arbres est complexe et son développement se poursuit durant toute la vie de la plante. Si l'on examine le développement des organes d'une plante depuis le haut de la tige jusqu'au bas des racines, on peut observer tous les différents stades qui conditionnent la mise en place et la croissance du xylème secondaire à partir des tissus primaires (Figure 4).

Le méristème apical ou méristème primaire, situé à l'extrémité des tiges et des racines, est composé de cellules indifférenciées appelées cellules « initiales ». En se divisant, ces cellules assurent le renouvellement du méristème avec la création d'une nouvelle cellule

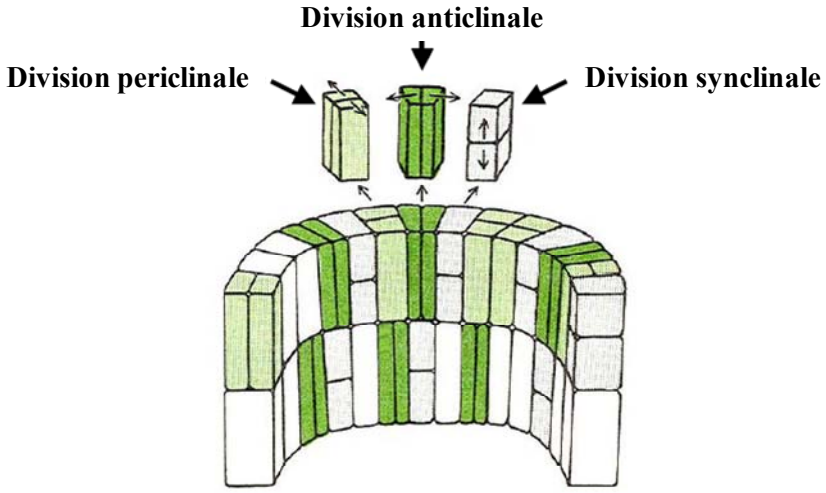


Figure 5 : polarité des divisions des initiales du procambium et du cambium vasculaire. Deux types de divisions cellulaires ont lieu au sein du procambium : periclines et synclines. Les trois types de divisions ont lieu au sein du cambium vasculaire (d'après Thibault, http://sylva.sbf.ulaval.ca/cambium/).

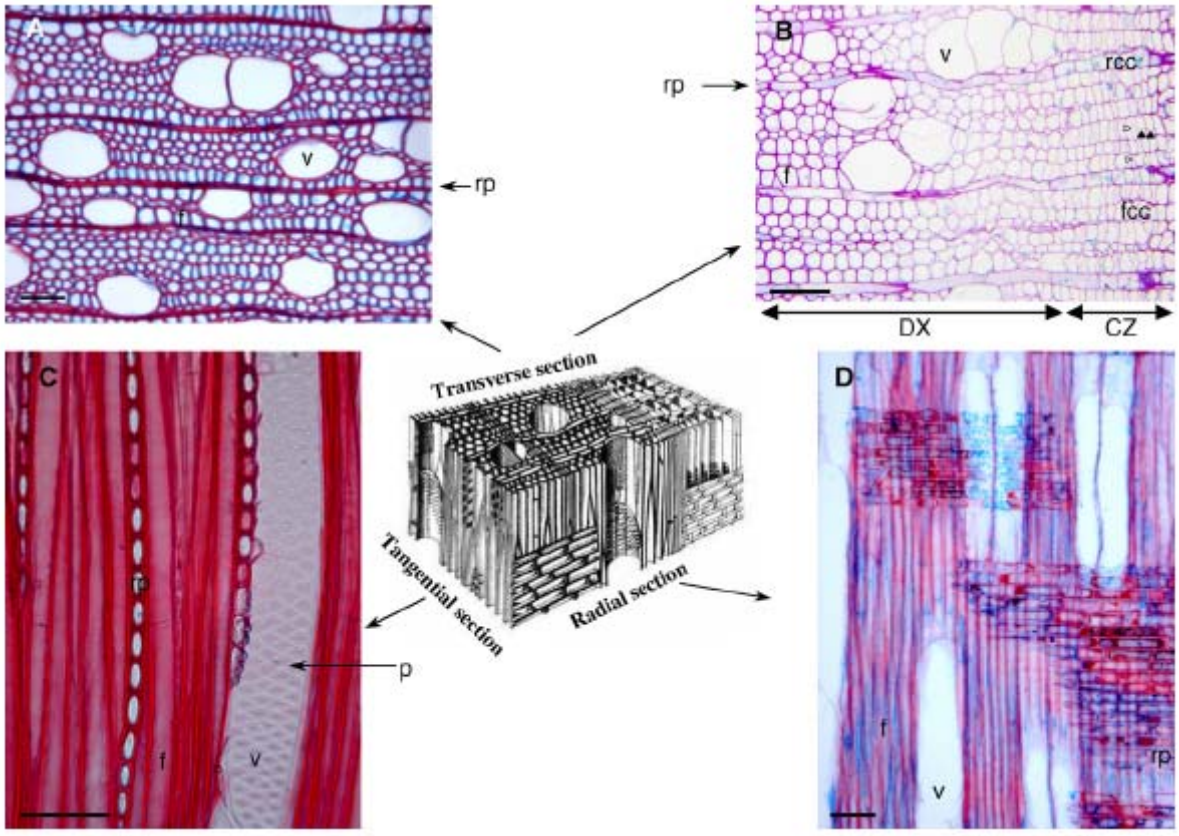


Figure 6 : représentation schématique de l'organisation tridimensionnelle du bois des angiospermes au centre et coupes de bois de peuplier en observation microscopiques en sections transversale (A et B), tangentielle (C) et radiale (D). DX : xylème en différenciation, CZ : zone cambiale, V : vaisseaux, rp : parenchyme de rayon, rcc : initiales de rayon, fcc : initiales fusiformes, f : fibres, p : pore d'un vaisseau permettant la communication avec les cellules voisines. Les barres d'échelles sur les photos de coupe en microscopie représentent 50 µm (d'après Déjardin *et al.*, 2010).

indifférenciée, et le développement d'un nouvel organe avec la création d'une cellule amenée à se différencier pour devenir une cellule spécialisée.

Chez les arbres, les tissus vasculaires primaires sont organisés en îlots (ou « bundles »), généralement disposés en cercle au sein de la tige et séparés par des cellules parenchymateuses qui forment les régions interfasciculaires. Ces îlots regroupent le procambium (méristème primaire à la base de la mise en place des tissus vasculaires primaires), le xylème primaire et le phloème primaire. L'organisation entre les différents tissus d'un îlot est variable selon les espèces de plantes mais dans la majorité des cas, les tissus sont disposés en couches parallèles; on parle d'organisation collatérale (Yé *et al.*, 2002). La Figure 4 représente la disposition et l'organisation de ces îlots chez les ligneux pérennes durant la croissance primaire de la tige. Le procambium est composé de cellules initiales vasculaires primaires qui sont capables de se diviser de manière syncline et péricline (Figure 5) donnant naissance à 3 types de cellules différentes. Les divisions périclines ont lieu vers l'intérieur et l'extérieur de la tige de part et d'autre du procambium et forment les cellules du xylème et du phloème primaire respectivement. Les divisions synclines assurent la persistance du procambium le long de l'axe de la tige et forment plus tard une partie du cambium vasculaire.

Lorsque la croissance de la plante continue, la croissance primaire connait une phase de transition qui aboutit à la formation d'un méristème secondaire : le cambium vasculaire. Il commence à se développer dans les parties les plus anciennes de la tige de la plante (bas de la tige). Lorsqu'il est formé, il est constitué d'une seule couche de cellules circulaires au sein de la tige et relie les îlots vasculaires au niveau des procambiums (Figure 4). Le cambium vasculaire est composé de deux types de cellules initiales inter-convertibles qui diffèrent par leur taille et leur forme et créent les cellules mères du xylème majoritairement par division péricline, vers l'intérieur du cylindre cambial et les cellules mères du phloème vers l'extérieur. Les cellules du xylème sont produites en quantité plus importante que les cellules du phloème. Les initiales fusiformes, allongées, vont donner côté xylème, le système cellulaire axial qui comprend les vaisseaux et les cellules qui leurs sont associées, les fibres et les cellules du parenchyme axial. Les initiales de rayon vont être elles, toujours côté xylème, à la base du système radial composé des cellules du parenchyme de rayon (Mellerowicz *et al.*, 2001) (Figure 6).

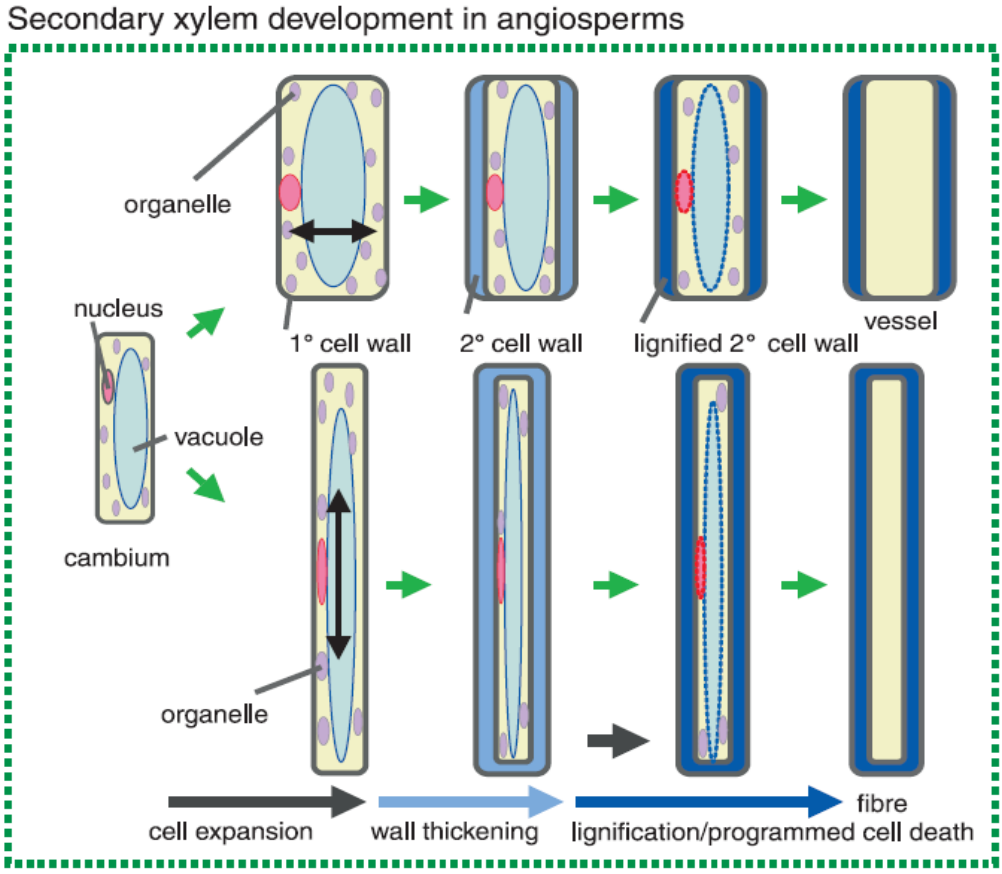


Figure 7 : représentation schématique de la xylogenèse chez les angiospermes. Les étapes d'expansion cellulaire, de mise en place, d'épaississement et de lignification de la paroi secondaire et de mort cellulaire sont présentées pour deux types de cellules : les vaisseaux et les fibres (d'après Samuels, 2006).

2.2.3. La xylogenèse

La division des initiales du cambium vasculaire donne naissance à plusieurs types de cellules. La mise en place et la différenciation de ces cellules qui constituent le xylème secondaire des ligneux pérennes passent par 3 stades : l'expansion cellulaire avec une phase d'élongation et d'élargissement, le dépôt d'une paroi secondaire et la mort cellulaire programmée (Fukuda *et al.*, 1996 ; Yé *et al.*, 2002 ; Samuels *et al.*, 2006 ; Déjardin *et al.*, 2010 ; Figure 7). Ces différents stades de la différenciation nécessitent la mise en œuvre de nombreux processus génétiques, biochimiques et morphologiques qui ne sont pas tous aujourd'hui complètement décrits. Cependant, les études menées sur des organismes modèles tels que *Arabidopsis* et *Populus* ont permis de mieux comprendre les mécanismes mis en place lors de la xylogenèse.

2.2.3.1.L'expansion cellulaire :

C'est le stade durant lequel les cellules mères du xylème vont s'allonger puis s'élargir. Dès ce premier stade, des différences importantes apparaissent entre les cellules amenées à devenir des vaisseaux conducteurs et les cellules amenées à devenir des fibres. Les futurs vaisseaux conducteurs subissent une expansion radiale alors que les futures fibres s'allongent de manière importante (Mellerowicz *et al.*, 2001). La paroi des cellules est alors composée de la lamelle moyenne qui délimite la région entre deux cellules voisines et la paroi primaire qui constitue la paroi de la cellule à proprement parlé. La paroi primaire est souple et extensible. Elle est constituée d'une structure fibrillaire de cellulose enrobée dans une matrice de polysaccharides et d'hémicellulose (Cosgrove *et al.*, 1997). Cette paroi contient également des protéines structurales en faible proportion (1% à 5%). La cellulose est un polymère de glucose constitué de chaînes linéaires de beta-1,4-D-glucanes. Ces chaines sont synthétisées et arrangées en microfibrilles par des complexes enzymatiques regroupant plusieurs unités de cellulose synthase (CesA) et appelés rosettes (Doblin *et al.*, 2002). La synthèse des microfibrilles de cellulose a lieu au niveau de la membrane plasmique de la cellule et leur dépôt se fait de manière orientée (Brett *et al.*, 2000). Plusieurs gènes *CesA* liés à la synthèse de la cellulose ont été identifiés par l'étude de mutants déficients en cellulose (Somerville, 2006). Les hémicelluloses sont des polysaccharides homo-ou hétéropolymériques constitués d'une chaîne centrale de résidus beta-D-1-,4-pyranosyl avec de courtes chaînes latérales de résidus glycosyl (xylose, galactose et fucose). Certaines hémicelluloses peuvent être plus abondantes dans la paroi primaire que dans la paroi

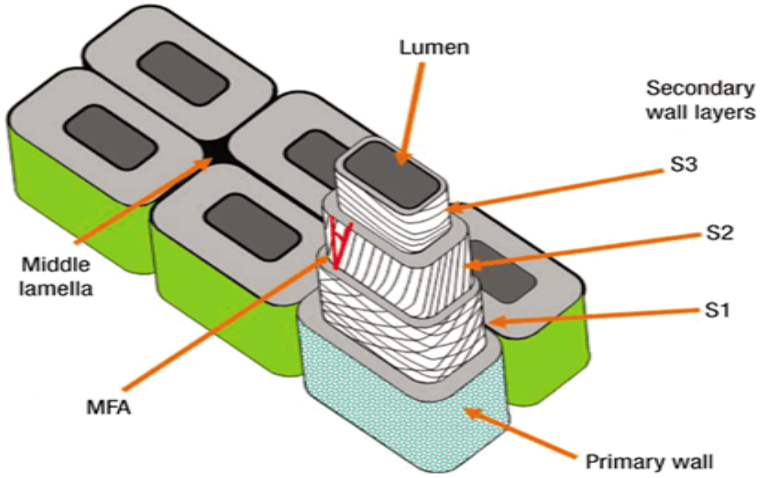


Figure 8 : représentation schématique de la paroi primaire et secondaire de la cellule végétale. La paroi primaire est à l'extérieur tandis que la paroi secondaire est composée de trois couches S1, S2 et S3 déposées de l'extérieur vers le lumen. MFA représente l'angle des microfibrilles de celluloses qui composent les couches S1, S2 et S3 de la paroi secondaire (d'après Kretschmann, 2003)

secondaire comme les xyloglucanes ou bien l'inverse dans le cas des xylanes par exemple. La biosynthèse des hémicelluloses est réalisée au niveau de l'appareil de Golgi par des enzymes de type « glycane synthase » et « glycosyltransférase ». Avant d'être intégrées à la paroi, les hémicelluloses sont transportées vers la paroi via des vésicules de sécrétion. Lors de l'expansion, le relâchement de la structure de la paroi primaire va être assisté par des expansines (Im *et al.*, 2000 ; Cosgrove *et al.*, 1999) ainsi que des enzymes de lyse pariétale telles que les xyloglucan endotransglycosylases (XET) (Bourquin *et al.*, 2002). Parmi ces enzymes, les expansines ont été montrées comme co-exprimées avec d'autres protéines du cycle cellulaire, indiquant que l'expansion cellulaire est en relation avec la fin de la mitose qui donne naissance aux cellules mères du xylème (Schrader *et al.*, 2004). Après le relâchement de la paroi primaire, c'est la pression interne de la cellule ou turgescence qui est responsable de l'expansion longitudinale et radiale. Les microtubules semblent également impliqués dans l'élongation cellulaire et pourraient diriger l'orientation du dépôt des microfibrilles de cellulose déterminant ainsi l'axe de l'élongation cellulaire (Funada *et al.*, 2000 ; Chaffrey *et al.*, 2002 ; synthétisé par Samuels *et al.*, 2006).

2.2.3.2.La mise en place d'une paroi secondaire :

Après la phase d'expansion, la différenciation des cellules du xylème se poursuit par la mise en place d'une paroi secondaire qui leur confère une grande rigidité mécanique et des propriétés hydrophobes. C'est cette paroi secondaire qui va permettre par exemple aux vaisseaux de transporter l'eau des racines vers le sommet de la plante et de résister à la pression imposée par le processus d'évapotranspiration. Cette nouvelle paroi est attenante à la paroi primaire. Elle est composée de 3 couches nommées S1, S2 et S3 qui sont déposées depuis l'extérieur de la cellule vers le lumen (Figure 8). Ces différentes couches sont composées essentiellement de microfibrilles de cellulose, d'hémicelluloses, de lignines et de quelques composés minoritaires tels que les pectines, des protéines structurales et des extractibles (Timell, 1969; Plomion *et al.*, 2001; Agarwal, 2006).

La lignine est un hétéropolymère phénolique composé majoritairement de 3 hydroxycynnamil alcools (aussi appelés monolignols) : le p-coumaryl alcool, le coniferyl alcool et le sinapyl alcool qui une fois polymérisés deviennent les unités p-hydroxyphenyl (H), guaiacyl (G) et syringyl (S) du polymère (Boerjan *et al.*, 2003). C'est lui qui confère principalement à la paroi secondaire rigidité et hydrophobie. Sa composition peut varier entre les espèces végétales ligneuses et entre les cellules du bois (fibres et vaisseaux)

(Plomion *et al.*, 2001 ; Zhong *et al.*, 2000). Au sein d'une couche, les microfibrilles de cellulose sont organisées en groupes, toutes orientées dans le même sens, selon le même angle (MFA, Figure 8). Les hémicelluloses et lignines constituent une matrice qui vient cimenter la structure. Entre les différentes couches S1, S2 et S3, la composition chimique, l'angle des microfibrilles de cellulose et donc la rigidité peuvent être variables. Plus particulièrement, l'alternance entre des angles des microfibrilles de cellulose importants et faibles renforce la rigidité de la structure (Déjardin *et al.*, 2010).

Le dépôt de la paroi secondaire des cellules du xylème est un processus complexe et organisé particulièrement au niveau des vaisseaux. Au sein de ces cellules, le dépôt de la paroi secondaire va délimiter des trous ou des plaques qui vont permettre de constituer des perforations. Ces perforations permettront aux vaisseaux matures de communiquer entre eux pour former des grandes structures tubulaires nécessaires au transport de la sève brute et d'échanger des éléments avec les autres cellules du xylème. Là encore, il semble que les microtubules soient impliqués dans l'organisation de la mise en place de la paroi secondaire (Samuels *et al.*, 2006 ; Oda et Hasezawa, 2006). Les microtubules pourraient, comme dans le cas de la paroi primaire, guider les complexes cellulose-synthases et une partie des complexes enzymatiques nécessaires à leur fonctionnement (Haigler *et al.*, 2001).

Enfin, le dépôt de lignine se fait dans les stades avancés du processus de mise en place des polysaccharides. Il existe aujourd'hui des évidences indiquant que la lignification des cellules du xylème est la résultante de l'activité des cellules du xylème en différenciation et des cellules constituant le parenchyme axial et radial (Hosokawa *et al.,* 2001 ; Li *et al.*, 2001 ; Van Raemdonck *et al.*, 2005). La plupart des enzymes qui interviennent dans la biosynthèse des lignines sont connues et leur localisation cellulaire a pu être identifiée. Si certaines d'entre elles sont présentes dans le cytoplasme des cellules du xylème en cours de lignification, d'autres semblent être associées au réticulum endoplasmique (Raes *et al.*, 2003). Même si les monolignols semblent être synthétisés dans le cytoplasme, les mécanismes qui permettent d'exporter les monolignols depuis le lieu de leur synthèse vers l'extérieur de la cellule restent inconnus.

2.2.3.3.La mort cellulaire programmée :

C'est le processus par lequel les cellules vont se vider de leur contenu pour ne conserver que leur paroi secondaire lignifiée. La mort cellulaire programmée ne touche que les fibres et les vaisseaux du bois. Les cellules du parenchyme, malgré leur paroi cellulaire lignifiée,

peuvent rester vivantes et fonctionnelles durant plusieurs années, assurant leur rôle de stockage et de transfert d'éléments depuis le phloème secondaire vers le xylème. Dans le cas des éléments conducteurs, cette étape permet de constituer les structures tubulaires perforées qui forment les vaisseaux.

Les patrons de perforation des extrémités des vaisseaux sont déterminés lors de la mise en place de la paroi secondaire. Les zones dans lesquelles la paroi primaire seule est conservée ne sont pas lignifiées et sont sensibles aux attaques des hydrolases qui dégradent la paroi cellulaire. Les patrons de perforation des extrémités des vaisseaux peuvent être plus ou moins complexes selon les espèces. La mort cellulaire s'apparente donc à l'autodigestion du protoplaste vivant par le biais d'enzymes hydrolytiques incluant des cystéine protéases, des serine protéases (Yé *et al.*, 1996 ; Beers *et al.*, 1997 ; Groover et Jones, 1999 ; Obara *et al.*, 2001), des nucléases (Aoyagi *et al.*, 1998 ; Thelen et Northcote, 1989) et probablement des cellulases même si aucune donnée n'existe pour ces dernières. La production de ces enzymes est fortement induite durant la xylogenèse et celles-ci sont accumulées dans la vacuole de la cellule (Fukuda, 2000).

La mort cellulaire est initiée par la rupture de la membrane de la vacuole et le relargage des enzymes hydrolytiques dans le cytoplasme de la cellule (Groover and Jones, 1999 ; Kuriyama, 1999 ; Obara *et al.*, 2001). De façon additionnelle, le pH du cytoplasme diminue fortement entrainant l'activation des hydrolases cytosoliques. Le noyau, les plastes, les mitochondries disparaissent, et les cellules sont vidées de leur cytoplasme. Pour le moment, peu de données sont disponibles sur les signaux qui induisent la production d'enzymes hydrolytiques et la rupture du tonoplaste. Certaines études suggèrent l'intervention de signaux calciques et de mécanismes plus complexes incluant l'oxyde nitrique (Lam, 2004).

2.3. Le bois est un matériau hétérogène

Le bois est formé d'un assemblage en 3 dimensions de différents types cellulaires dont la nature et l'organisation peuvent différer selon les espèces. L'utilisation des termes « bois tendres » et « bois durs » pour désigner les bois des gymnospermes et angiospermes dicotylédones témoignent de ces différences. Par exemple, chez les gymnospermes, le transport de la sève brute est assuré par les trachéides. Ces trachéides sont considérées comme une forme primitive des vaisseaux parfaits des angiospermes ligneux. Cependant, ils assurent un double rôle de transport de la sève brute et de soutien de la plante. Chez les angiospermes

ligneuses, les vaisseaux jouent un rôle principalement dans le transport de la sève brute. Même si les vaisseaux assurent un rôle de soutien par la présence d'une paroi secondaire lignifiée, la majorité de cette fonction est assurée par un autre type de cellules spécialisées : les fibres.

La formation du bois chez les plantes ligneuses est un processus majoritairement coordonné par un programme génétique de développement. Cependant, d'autres facteurs environnementaux de types biotiques (attaque par des pathogènes) et abiotiques (température, photopériode, vent, …) vont impacter la structure finale du matériau au niveau de l'arbre à différentes échelles. Ainsi, au sein d'un arbre, on peut distinguer différents types de bois qui ont chacun des propriétés particulières au niveau anatomique, chimique et physique (Plomion *et al.*, 2001). Par exemple en zones tempérées, le bois produit sur une année n'est pas homogène. On distingue le bois de printemps, formé dans des conditions de luminosité et températures favorables à la croissance et au développement de l'arbre et le bois d'été, formé dans des conditions d'alimentation en eau plus contraignantes. Le bois de printemps est généralement caractérisé par des vaisseaux conducteurs et des fibres de diamètre plus larges aux parois plus fines induisant une densité plus faible que le bois produit en été (Mc Millian 1968 ; Megraw, 1985, Mott *et al.*, 2002). On distingue également le bois juvénile et le bois mature. Le bois juvénile est mis en place tout le long de l'arbre, durant les premières années de l'activité cambiale. Ses propriétés sont soumises à de fortes variations environnementales et il est généralement moins dense, avec des fibres plus courtes, un angle de microfibrilles de cellulose plus important et plus riche en lignines et en hémicelluloses que le bois mature (Passialis et Kiriazalos, 2004). Enfin, on distingue le bois de réaction du bois dit « normal ». Le bois de réaction (bois de tension chez les angiospermes ou bois de compression chez les gymnospermes) est généralement formé en réponse à l'orientation non verticale de la tige qui peut elle-même être causée par différents stress environnementaux. La mise en place du bois de réaction permet à l'arbre de retrouver une position verticale plus favorable à sa croissance. Le bois de réaction est généralement plus dense et présente une composition chimique différente de celle du bois normal (Yeh *et al.*, 2005).

3. L'amélioration génétique des propriétés du bois

3.1. Des propriétés diverses

Aujourd'hui, les utilisations industrielles majeures du bois sont la production de pâte à papier, la production de panneaux de particules et de panneaux de contreplaqué, la production de planches et la production de charbon de bois pour des usages industriels. Pour chacune de ces utilisations finales, il est essentiel de déterminer quelles sont les propriétés du bois qui doivent être améliorées. Ces objectifs d'amélioration doivent être basés sur la définition des paramètres économiques clés pour les secteurs considérés. Une fois ces paramètres définis, ils doivent être mis en relation avec les différentes propriétés physiques, mécaniques, anatomiques ou chimiques du bois afin de déterminer, pour chaque utilisation, les propriétés d'intérêt majeur pour l'amélioration (Raymond, 2002).

Certaines propriétés du bois sont aujourd'hui au premier plan de l'amélioration génétique, et parmi elles, des propriétés physiques (la densité du bois), des propriétés mécaniques (le module d'élasticité, MOE), des propriétés anatomiques (l'angle des microfibrilles de cellulose MFA, ou la longueur des fibres) et enfin les propriétés chimiques comme les teneurs en lignines ou en cellulose qui suscitent un intérêt particulier pour les améliorateurs (Greaves *et al.*, 1997 ; Raymond *et al.*, 1997 ; Raymond *et al.*, 2000 ; Schmidt, 2005).

3.2. Définition des caractères et méthodes de mesure

La densité du bois correspond à la masse de bois par unité de volume. Elle peut être mesurée par différentes méthodes. La méthode classique est la mesure de l'infradensité aussi appelée gravité spécifique. Elle correspond au rapport entre la masse d'un volume de bois sec et la masse du même volume d'eau. Etant donné la variabilité de la densité du bois au sein de l'arbre, elle est souvent mesurée en différents points du tronc depuis le centre jusqu'à la périphérie et une valeur moyenne est ensuite calculée. Bien que cette méthode soit précise, elle est coûteuse en temps, destructive (elle nécessite de prélever des échantillons de bois à différents niveaux depuis le centre vers la périphérie du tronc) et pas forcément bien adaptée à l'étude de grands échantillons nécessaire dans le cadre des programmes d'amélioration génétique des arbres. En tant que caractère intégratif d'un ensemble de propriétés macroscopiques et microscopiques, la densité du bois est une propriété majeure pour la

plupart des secteurs de l'industrie (Raymond, 2002). Plusieurs méthodes de mesure indirecte de la densité ont donc été proposées pour remplacer cette méthode de mesure, parmi elles, la microdensitométrie, le pilodyn ou le résistographe sont les plus souvent utilisées (Cown, 1978 ; Taylor, 1981 ; Gough et Barnes, 1984 ; Nicholss, 1985 ; Moura *et al.*, 1987). Certaines d'entre elles ont démontré leur efficacité et remplacé la mesure classique de la gravité spécifique dans les programmes d'amélioration génétique (Isik et Li, 2003).

Le MOE et le MFA sont deux propriétés d'intérêt majeur pour le bois d'œuvre (Raymond, 2002). Le MOE traduit la rigidité d'un bois et notamment sa capacité de résistance à la flexion. Il peut être mesuré de manière directe par l'application d'une force mécanique sur une section de bois, généralement une planche, ou par méthode acoustique en étudiant la propagation d'une onde le long d'une pièce de bois. Le MFA traduit l'angle des microfibrilles de cellulose au sein de la couche S2 de la paroi secondaire des cellules du xylème. C'est l'écart de l'orientation des microfibrilles par rapport à la verticale qui est mesuré. Le MFA peut être déterminée sur un échantillon de bois par microscopie confocale ou par diffraction des rayons X.

La longueur des fibres est une propriété anatomique majeure pour l'industrie de la pâte à papier. Elle se mesure par observation microscopique d'un échantillon de fibres isolées à partir d'échantillons de bois. Au même titre que la longueur des fibres, la quantité et la composition des lignines ainsi que la quantité de cellulose sont des propriétés majeures pour l'industrie de la pâte à papier (Baucher, 2003 ; del Rio *et al.*, 2005 ; Raymond, 2002 ; Kube et Raymond, 2002). La teneur en lignine étant corrélée avec la teneur en carbone fixe du charbon de bois, elle présente aussi un intérêt dans la production de charbon pour des applications industrielles (Antal *et al.*, 2000). Teneur et composition en lignines et cellulose sont mesurées par des méthodes chimiques impliquant des méthodes d'extraction dans des solvants ou l'analyse des produits de la combustion du bois (pyrolyse). Cette dernière est plus abordable que les méthodes de mesures chimiques classiques qui ont longtemps été limitantes pour la prise en compte de ces propriétés dans les programmes d'amélioration génétique des arbres.

Comme nous l'avons vu, les méthodes de mesure classiques de ces caractères présentent deux contraintes majeures : i/ elles peuvent être coûteuses ou difficiles à mettre en œuvre même s'il existe parfois des méthodes alternatives, ii/ ce sont, pour la plupart, des méthodes destructives qui impliquent la perte des individus mesurés (ceci peut être préjudiciable dans le

Tableau 1 : valeurs de paramètres génétiques des caractères de la croissance et de la qualité du bois chez les arbres forestiers. BP : bois de printemps, BE : bois d'été. KL : lignines de Klason, ASL : lignines solubles dans l'acide, T : lignines totales.

Espèce	Variations analysées	Hauteur	Diamètre	Densité du bois	Longueur de fibres	Teneur en lignines	Teneur en celluloses	MFA	MOE	Etude
Eucalyptus globulus	5 sites		0,2							MacDonald *et al.*, 1997
Eucalyptus globulus	8 variétés (subraces)		0,2	0,44	0,16		0,84	0,27		Apiolaza *et al.*, 2005
Eucalyptus globulus	9 localités			0,24		0,13 (Klason) 0,51 (Acid-soluble) 0,29 (Total)				Poke *et al.*, 2006
Eucalyptus globulus	3 âges		0,18 (4ans) 0,22 (8ans) 0,15 (16ans)	0,52						Stackpole *et al.*, 2010
Eucalyptus nitens	3 sites / 2 âges		0,19 (6ans) 0,38 (12ans) 0,45 (entre 6 et 12)	0,7	0,58		0,79			Kube *et al.*, 2001
Eucalyptus urophylla	4 sites			0,71						Wei et Borralho, 1997
Picea glauca	3 régions	0,2		0,69 (BP) 0,13 (BE)				0,28 (BP) 0,34 (BE)	0,27 (BP) 0,41 (BE)	Lenz *et al.*, 2010
Picea abies	2régions / 2 âges / cernes	0,31-0,54	0,34-0,50	0,36-0,55 (moy cerne) 0,30-0,56 (BP) 0,33-0,51 (BE)	0,20-0,22	0,10		0,12-0,25 (BP) 0,21-0,36 (BE)		Hannrup *et al.*, 2004
Picea abies	en fonction des cernes / âge	0,18 (7ans) 0,19 (10 ans) 0,26 (26 ans)	0,26 (10ans) 0,23 (22 ans) 0,23 (25 ans)	0,48				0,41	0,5	Grans *et al.*, 2009
Pinus pinaster		0,46		0,3	0,19	0,47	0,34 (alpha)			Pot *et al.*, 2002
Pinus pinaster	par cerne			0,63 (moy cerne) 0,60 (BP) 0,26 (BE)						Gaspar *et al.*, 2008
Pinus radiata	par cerne			0,1-0,5 (BP) 0-0,35 (BE)						Zamudio *et al.*, 2005
Pinus sylvestris		0,28	0,17 (80cm)		0,3					Ericsson et Fries, 2004
Pinus sylvestris	par cerne			0,083-0,20 (BP) 0,096-0,22 (BE)						Fries et Ericsson, 2009
Pinus taeda	8 régions			0,12-1						Isik *et al.*, 2008

cas où les arbres ne peuvent pas être bouturées de manière efficace). Depuis plus récemment, la spectrométrie dans le domaine du proche infrarouge (SPIR) offre de nouvelles perspectives pour la mesure d'un grand nombre de propriétés mécaniques (Kelley *et al.*, 2004 ; Fujimoto *et al.*, 2008), anatomiques (Alves *et al.*, 2006), physiques (Tsushikawa *et al.*, 1996 ; Hein *et al.*, 2009) et chimiques (Kelley *et al.*, 2004 ; Hein *et al.*, 2010) du bois. Elle est basée sur la mise au point de modèles de prédiction permettant de mettre en relation la variation de spectres d'absorptions du bois dans le domaine du proche infrarouge avec la variation de caractères d'intérêt (Tsuchikawa, 2007) après réalisation de calibrations. La SPIR présente l'avantage de pouvoir réaliser une mesure rapide peu onéreuse et sans préparation spécifique des échantillons (poudre de bois ou échantillons massifs). Cette méthode de mesure indirecte (basée sur la prédiction des caractères) constitue une avancée importante pour l'étude du déterminisme génétique des propriétés du bois (Raymond, 2002).

3.3. Le déterminisme génétique des caractères de la qualité du bois

Du fait de l'intérêt majeur que suscitent les propriétés du bois, leur variabilité a déjà été étudiée chez différentes espèces d'arbres forestiers et dans différentes populations. De manière globale, les caractères liés aux propriétés du bois présentent des gammes de variation inférieures aux caractères de croissance. Les études indiquent des coefficients de variation phénotypiques de l'ordre de 5% à 11% pour la densité du bois, 9% pour la longueur des fibres et de l'ordre de 2% à 5% pour la composition chimique contre environ 20% pour la hauteur totale ou le diamètre de l'arbre (Hylen, 1999 ; Ivkovich *et al.*, 2002 ; Hannrup *et al.*, 2004). Cette variabilité peut être plus importante pour certaines propriétés comme pour le MFA avec des coefficients de variation phénotypiques rapportés jusqu'à 23% (Hannrup *et al.*, 2004).

Au regard des caractères de croissance, les caractères liés aux propriétés du bois semblent de manière générale plus héritables (Hamilton et Potts, 2008 ; Ukrainetz *et al.*, 2008 ; Hallingback *et al.*, 2010 ; Apiolaza *et al.*, 2004 ; Kube *et al.*, 2001 ; Stackpole *et al.*, 2010 ; Grans *et al.*, 2009). Les niveaux d'héritabilité estimés pour ces caractères peuvent aller selon les espèces et les populations étudiées jusqu'à 0,96 pour la densité du bois, 0,46 pour la longueur des fibres et le MFA et 0,83 pour les caractères chimiques comme la teneur en cellulose indiquant un fort contrôle génétique exercé sur la variation de ces caractères (Tableau 1).

3.4. Intégration des propriétés du bois dans les programmes d'amélioration

La prise en compte des propriétés du bois dans les programmes d'amélioration est prometteuse. Même si l'importance des gains génétiques espérés pour ces caractères devrait être limitée, compte tenu de leurs coefficients de variation plutôt faibles (à part quelques exceptions comme le MFA) en comparaison des caractères de croissance, les données d'héritabilité estimées chez différentes espèces et pour différentes populations indiquent que ces gains pourront être rapidement réalisés. Cependant, d'autres paramètres doivent être pris en compte pour réellement évaluer l'efficacité d'une sélection menée sur ces caractères.

Comme pour les caractères de croissance, les propriétés du bois peuvent varier avec l'âge. Il s'agit donc dans un premier temps d'étudier les corrélations juvénile-adulte pour ces caractères afin de déterminer l'âge minimum de mesure le plus représentatif de l'âge de la récolte. Bao *et al.* (2001) rapportent des différences entre bois juvénile et bois mature pour un ensemble de propriétés du bois mesurées chez 10 espèces d'arbres forestiers. Cette étude indique de faibles différences entre bois juvénile et bois mature pour les propriétés chimiques du bois mais des différences importantes pour la plupart des propriétés mécaniques et physiques. Ces différences semblaient moins prononcées pour les espèces angiospermes dicotylédones que pour les gymnospermes. Chez l'eucalyptus, Raymond *et al.* (2000) indiquent que le MFA et la teneur en lignines montrent une diminution avec l'âge jusqu'à environ 5 ans, âge à partir duquel la mesure montre une bonne corrélation avec l'âge de récolte. Pour la densité du bois, la teneur en cellulose, la longueur des fibres ou le MOE, les auteurs indiquent une augmentation avec l'âge avec une bonne corrélation des mesures effectuées entre 3 ans (densité du bois) et 5 ans (les autres propriétés) et celles à l'âge de récolte. Les faibles corrélations juvénile-adulte rapportées par ces études laissent entrevoir, comme pour les caractères de croissance, des gains génétiques par unité de temps limités pour ces caractères qui ne peuvent être évalués que tardivement.

Pour le moment, la plupart des programmes d'amélioration des arbres forestiers ont axé leurs efforts sur l'amélioration des caractères de croissance qui restent les caractères prioritaires. L'intégration des propriétés du bois dans les programmes d'amélioration génétique devra être réalisée de manière coordonnée avec celle des caractères de croissance (Raymond et Apiolaza, 2004). Les corrélations génétiques additives entre caractères de croissance et propriétés du bois constituent une donnée essentielle à prendre en compte pour

évaluer les bénéfices d'une sélection menée sur ces deux types de caractère. Pour le moment, la majorité des résultats disponibles dans la littérature porte sur les corrélations génétiques additives entre croissance et densité du bois (deux caractères majeurs dans la plupart des programmes d'amélioration génétique des arbres forestiers). Les résultats obtenus entre espèces et au sein de différentes espèces apparaissent parfois contradictoires (Bouffier *et al.*, 2009), cependant, plusieurs études indiquent des corrélations génétiques additives négatives entre ces deux caractères suggérant que la sélection des génotypes les plus performants en termes de croissance tendrait à diminuer la densité du bois dans les populations améliorées (Costa e Silva *et al.*, 2009 ; Liu et Wu, 2005 ; Kumar, 2004 ; Apiolaza *et al.*, 2005). Ces corrélations négatives pourraient indiquer, pour ces caractères, l'existence de gènes à effets pléiotropiques antagonistes. Cette corrélation négative entre croissance et densité du bois n'étant pas totale, une sélection d'individus à croissance rapide exprimant de fortes densités (comme recherché dans le cas de la production de charbon de bois pour des applications industrielles) reste possible bien que plus délicate à mettre en place que dans le cas de caractères génétiquement indépendants.

De manière générale, les stratégies proposées en matière de sélection multi-caractères s'orientent vers la détermination d'un index de sélection. Celui-ci est basé sur l'intégration des informations disponibles sur les caractères d'intérêt (héritabilité des caractères et corrélations entre caractères) et le poids économique de ces caractères pour un objectif d'amélioration (Greaves *et al.*, 1997 ; Gapare *et al.*, 2006 ; Apiolaza, 2009). Cette stratégie permet de maximiser les gains génétiques potentiels par unité de temps mais nécessite une bonne connaissance des attentes de la filière et des propriétés du bois qui sont recherchées. Elle est d'autant plus difficile à mettre en pratique que le nombre de caractères à prendre en compte dans la sélection est important (Raymond et Apiolaza, 2004).

3.5. La sélection assistée par marqueurs (SAM) pour la qualité du bois

L'idée d'utiliser l'information moléculaire dans le cadre des programmes d'amélioration des arbres forestiers a été un des moteurs du développement de la génétique moléculaire chez les arbres. En effet, ces outils étaient envisagés à deux niveaux : i/ en sélection précoce pour palier au manque de corrélation juvénile – adulte, et ii/ pour mieux gérer la variabilité génétique des populations d'amélioration. Les premiers travaux sur l'architecture génétique des caractères complexes, avec des cartes génétiques construites à

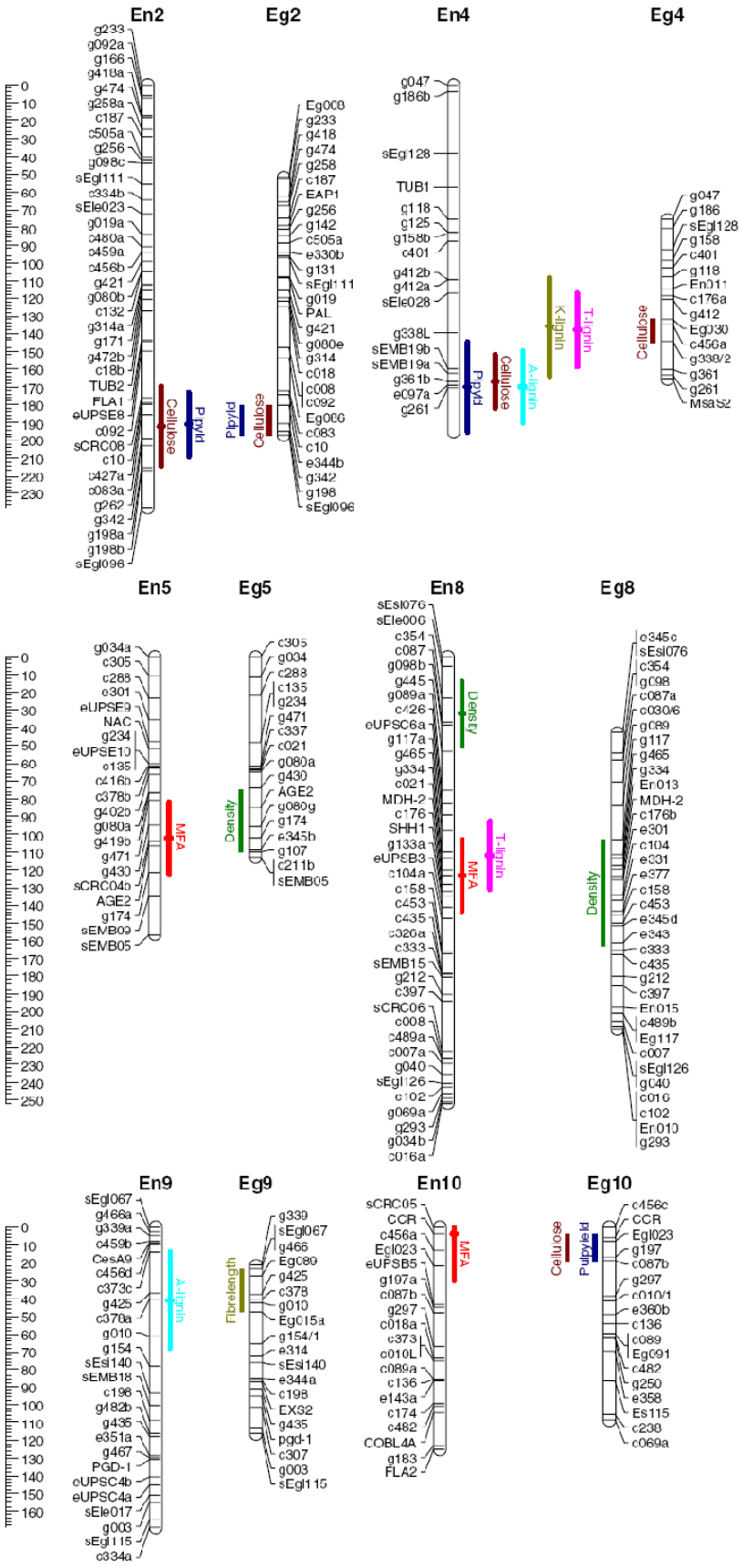


Figure 9 : comparaison des positions de QTL détectés chez *E. nitens* (En) et *E. globulus* (Eg). D'après Thumma *et al.*, 2010.

partir de familles de plein-frères et des marqueurs dominants peu transférables entre fonds génétiques différents, ont rapidement amené des doutes, et même des controverses, sur l'utilisation possible en génétique forestière d'un marqueur bordant un QTL (Strauss *et al.*, 1992.).

Dans le cas des caractères chimiques liés aux propriétés du bois, plus particulièrement dans le cas de voies de biosynthèse plutôt bien décrite comme la voie de biosynthèse des lignines, des approches combinant la détection de QTL et la cartographie de gènes candidats ont été menées sur plusieurs modèles forestiers, avec des résultats prometteurs sur l'utilisation possible de ces gènes en sélection (Gion, 2001 ; Neale *et al.*, 2002 ; Brown *et al.*, 2003 ; Thamarus *et al.*, 2004 ; Freeman *et al.*, 2009). Neale *et al.* (2002) et Thumma *et al.* (2010) ont notamment identifié des régions génomiques à effet majeur pour des propriétés du bois, communes à différentes espèces et contenant des gènes connus pour leur rôle dans des voies de biosynthèse en lien avec les propriétés ciblées (Figure 9). Ces résultats redonnent des perspectives quant à l'utilisation de cette information moléculaire pour envisager une sélection de la qualité du bois.

Cependant, le développement de marqueurs génétiques comme outils de sélection reste complexe et passe par différentes phases d'investigations qui vont du développement de ressources génomique, à la cartographie génétique et l'analyse QTL, et enfin aux études d'association entre variabilité moléculaire et variation des caractères cibles en populations complexes. Les caractères chimiques de qualité du bois, comme la teneur en lignines, sont surement ceux qui offrent le plus de perspectives quant à une sélection assistée par marqueurs à un moyen ou long terme.

4. La lignine : un caractère de choix pour la SAM chez l'eucalyptus

4.1. Une propriété chimique du bois de premier plan

4.1.1. Nature des Lignines

« Lignine » est un terme générique qui désigne un ensemble de biopolymères aromatiques obtenu par le couplage oxydatif de 4-hydroxyphenylpropanoïdes (Ralph *et al.*, 2004). Ces polymères sont déposés dans la paroi secondaire des cellules végétales et lui confèrent rigidité et imperméabilité. En plus de ce dépôt induit par des mécanismes

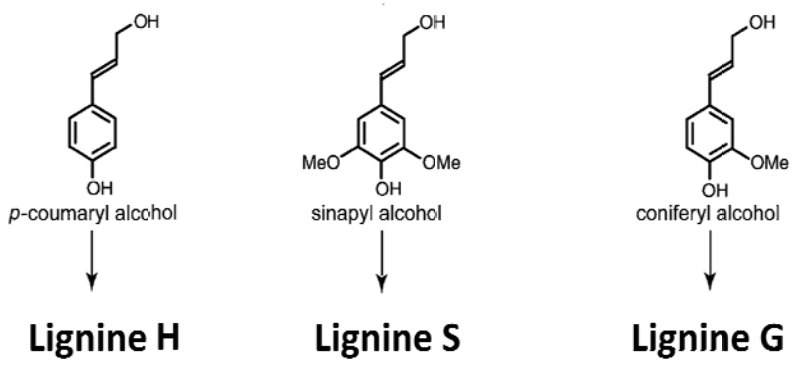


Figure 10 : représentation des 3 molécules d'hydroxy-cinnamyl alcools à la base de la structure moléculaire des polymères de lignine.

développementaux, la synthèse des lignines est également induite par des stress biotiques et abiotiques tels que des stress métaboliques, des attaques par des pathogènes, des blessures (Silva Moura *et al.*, 2010)…

La structure des lignines peut être complexe (Boerjan *et al.*, 2003), cependant, les unités de base du polymère sont les hydroxy-cinnamyl alcools aussi appelés monolignols. Il en existe 3 principaux qui sont le coniferyl alcool, le synapil alcool et le p-coumaryl alcool qui donnent respectivement les 3 phénylpropanoïdes guaiacyl (G), syringyl (S) et p-hydroxyphenyl (H) une fois polymérisés (Figure 10). La quantité relative des ces trois monomères dans les polymères de lignine peut varier selon les espèces de plantes. Ainsi, si la lignine des angiospermes est principalement composée des sous unités S et G, celle des gymnospermes est composée en majorité d'unités G avec une faible proportion de H (Campbell et Sederoff, 1996). La composition des lignines peut varier également entre différents types de cellules et entre les différentes couches de la paroi secondaire (Nakashima *et al.*, 2008 ; Ruel *et al.*, 2009 ; Shi *et al.*, 2007 ; Gou *et al.*, 2008). D'autres types de monomères d'origine phénolique peuvent également être incorporés dans le polymère, généralement de manière moins abondante que les trois principaux cités précédemment (Ralph *et al.*, 2004). Plus d'études sont cependant nécessaires pour comprendre les mécanismes à la base de leur synthèse.

Les monolignols dérivent de la phénylalanine par des processus de transformations successifs *via* la voie de biosynthèse commune des phénylpropanoïdes et la voie de biosynthèse spécifique des monolignols. La synthèse des monolignols commence par la déamination de la phénylalanine catalysée par la phénylalanine ammonia-lyase (PAL). Elle implique ensuite des hydroxylations successives du noyau aromatique catalysées par la cinnamate 4-hydroxylase (C4H), la p-coumarate 3-hydroxylase (C3H), la ferulate 5-hydroxylase (F5H), des méthylations des groupements hydroxyl par l'action de O-méthyltransférases (COMT et CCoAOMT) et la transformation du groupement carboxyl de la chaine latérale en alcool par différentes réactions chimiques impliquant la 4-coumarate:CoA ligase (4CL), la p-hydroxycinnamoyl-CoA:quinate shikimate p-hydroxycinnamoyltransférase (HCT), la cinnamoyl-CoA reductase (CCR) et la cinnamyl alcool dehydrogenase (CAD). Les routes qui sont suivies par ces différents composés et les transformations chimiques qui ont lieu *in planta* sont encore mal connues et l'enchainement des réactions chimiques qui aboutissent à la synthèse des monolignols est sujette à des remises en questions régulières (Boerjan *et al.*, 2003). La Figure 11 montre l'enchainement des réactions enzymatiques les

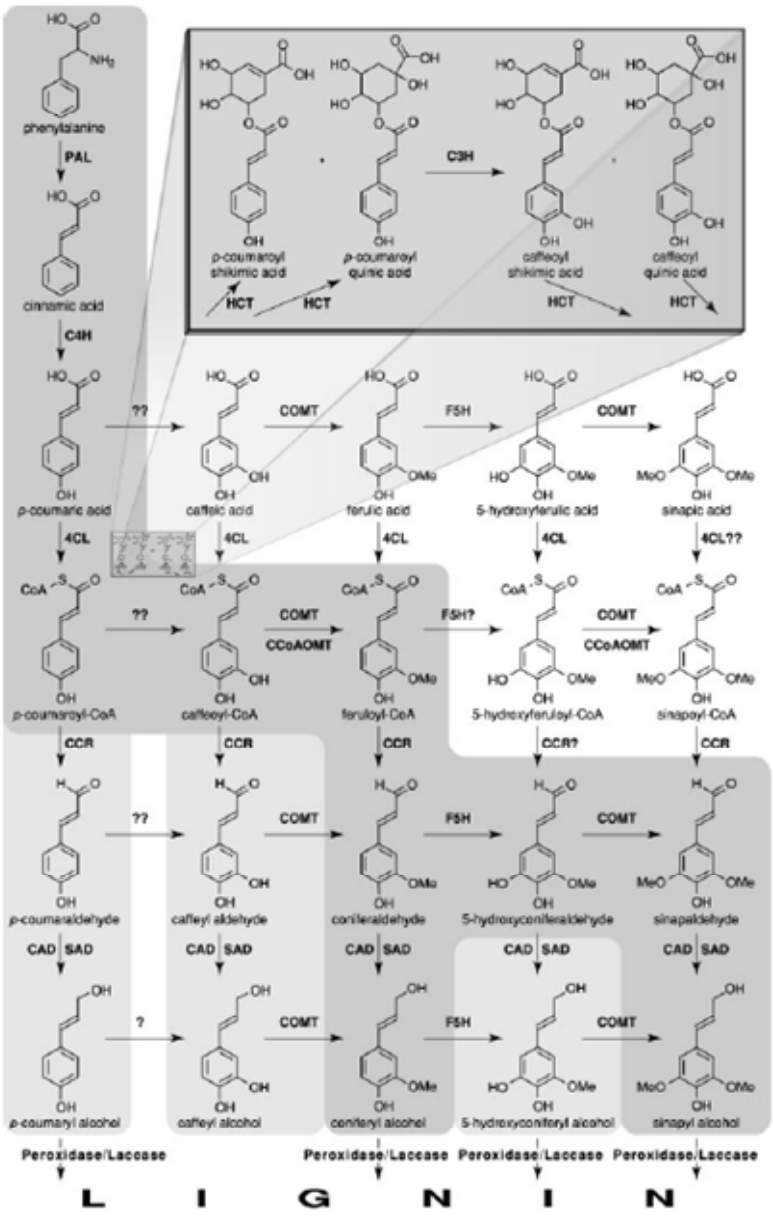


Figure 11 : Voie de biosynthèse des phénylpropanoïdes et des monolignols. Basée sur des études de composition des lignines, des essais enzymatiques et des résultats d'analyses de plantes transgéniques. Le chemin le plus foncé est celui qui parait être le plus probable et le plus favorisé chez les angiospermes. Le chemin en gris clair peut aussi être emprunté chez certaines espèces de plantes. Les parties en blanc ne semblent pas jouer un rôle dans la biosynthèse des lignines (d'après Boerjan *et al.*, 2003).

plus probables aboutissant à la synthèse des monolignols chez les angiospermes avec les composés chimiques intermédiaires et les enzymes catalysant leurs transformations depuis la phénylalanine jusqu'aux monolignols.

La lignification est le processus par lequel les monolignols sont couplées les uns aux autres. Différents types de liaisons chimiques existent entre les différents monomères au sein des polymères de lignine. La plus fréquente est la liaison β-O-4, mais il existe également des liaisons de type β-5, β-β, 5-5, 4-O-5 et β-1 (Figure 12). La principale réaction couple un nouveau monomère à un polymère déjà préformé. Ce couplage est généralement réalisé en position β et donne lieu à des liaisons de type β (le plus souvent β-O-4). Lorsque deux oligomères déjà formés se lient ensemble, les liaisons formées sont de types 5-5 et 5-O-4. La dimérisation de deux monolignols donne elle des liaisons de type β-β. La Figure 13 donne un exemple de la structure d'un polymère de lignine chez le peuplier. Ces différents types de liaisons, ainsi que les mécanismes biochimiques qui sont impliqués dans leur mise en place *in vivo,* sont largement revus par Boerjan *et al.* (2003) et Ralph *et al.* (2004) et ne seront pas exposés ici.

4.1.2. Importance économique

La lignine est le deuxième biopolymère le plus abondant sur la planète derrière la cellulose. Composant chimique majeur du bois, ce polymère est considéré avec beaucoup d'attention par les industriels dans différents domaines de la transformation du bois.

Dans le secteur de la pâte à papier, l'objectif est d'extraire les fibres de cellulose contenues dans le bois. Ces fibres sont intimement liées à une matrice de lignines qui est déposée dans la paroi secondaire des cellules du bois. Les lignines sont dans ce cas précis des composés indésirables qui doivent être éliminés. Les processus les plus utilisés pour la production de pâte à papier sont les procédés chimiques et parmi eux le procédé de « kraft pulping ». Ces procédés donnent des pâtes de qualité supérieure qui permettent de produire des papiers purs et résistants qui correspondent aux standards du marché. Cependant, ces procédés sont pénalisés par des rendements assez faibles (entre 45% et 55%) comparés aux procédés mécaniques et semi-chimiques (Baucher *et al.*, 2003). La faiblesse de ces rendements est attribuée à l'élimination des lignines et de certaines hémicelluloses qui sont récupérées dans ce que l'on appelle la liqueur noire, sous produit du procédé de « kraft pulping ». Les attentes des industriels de la filière pâte à papier en matière de qualité du bois sont donc orientées vers une diminution de la teneur en lignines du bois (Raymond *et al.*,

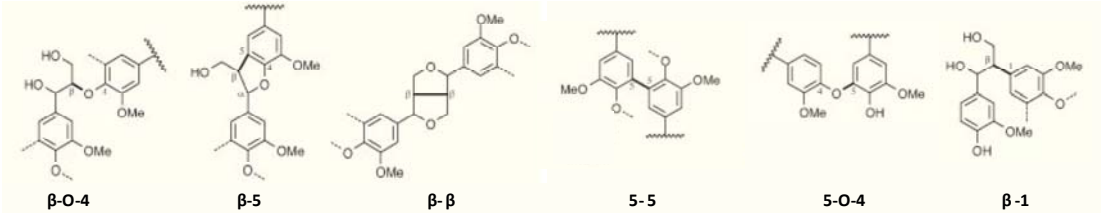


Figure 12 : différents types de liaisons entre sous unités du polymère de lignine (d'après Boerjan *et al.*, 2003).

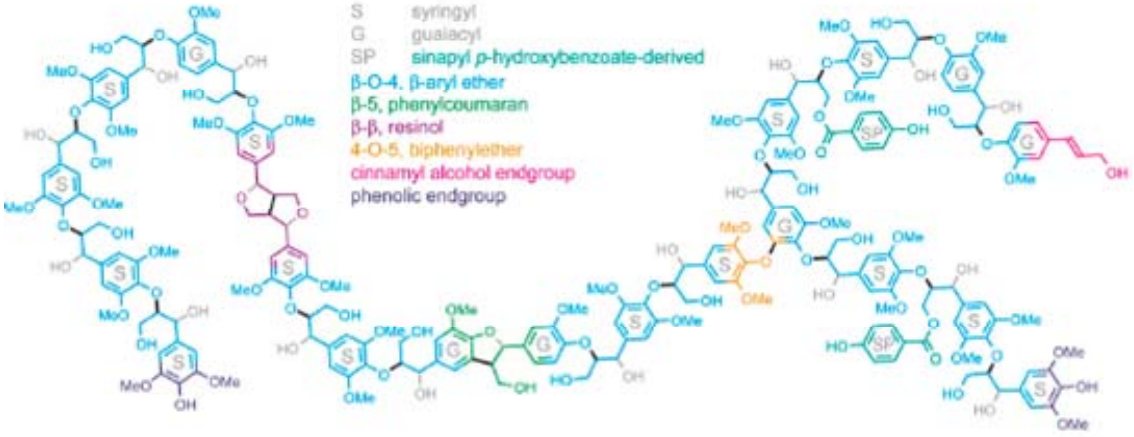


Figure 13 : représentation d'un polymère de lignine de peuplier (d'après Vanholme *et al.*, 2010).

2002 ; Raymond et Apiolaza, 2004). La qualité des lignines est également considérée avec attention. Différentes études démontrent l'importance du rapport S/G dans le processus de délignification (Chang et Sarkanen 1973 ; Chiang *et al.*, 1988). Une étude menée chez l'eucalyptus par Del Rio *et al.* (2005) montre l'importance du rapport S/G dans la détermination du rendement en pâte. Ces auteurs ont mis en évidence que les lignines riches en monomère S permettent d'obtenir de meilleurs rendements en pâte, le rapport S/G étant un paramètre plus déterminant que la teneur en lignines.

Dans le secteur de la production de charbon de bois, la lignine a aussi son importance. Le charbon de bois est obtenu par combustion incomplète du bois et possède, tout comme le bois, un ensemble de propriétés qui déterminent sa capacité à répondre à des utilisations précises (teneur en eau, friabilité, teneur en éléments minéraux, taux de cendres, teneur en carbone fixé… ; Vigneron et Gion, communication personnelle). Parmi ces propriétés, la teneur en carbone fixé est l'une des composantes qui conditionnent l'utilisation du charbon dans des secteurs majeurs comme la production d'acier. Dans ces secteurs, le charbon de bois constitue un combustible et un agent réducteur de choix compte tenu de sa pureté par rapport au coke de houille et de l'avantage écologique qu'il présente par rapport à ce dernier. Si la teneur en carbone fixé du charbon est dépendante du processus de carbonisation employé pour transformer le bois en charbon, certaines propriétés du bois comme la teneur en lignines semblent également déterminantes comme l'ont montré Antal *et al.* (2000). Ainsi, des bois exprimant des fortes teneurs en lignines sont recherchés pour la production de charbon destiné à ce type d'utilisation.

4.2. La voie de biosynthèse des lignines

4.2.1. Les gènes impliqués dans la biosynthèse des lignines

Malgré les nombreuses zones d'ombre qui persistent concernant la structure et la mise en place des polymères de lignines dans la plante, la voie de biosynthèse des lignines est l'une des voies de biosynthèse les mieux décrites au niveau moléculaire.

4.2.1.1.Les gènes de structure

La plupart des enzymes structurales de la voie commune des phénylpropanoïdes et de la voie spécifique des monolignols (citées au paragraphe 4.1.1) ont été décrites et les gènes qui les codent ont pour la plupart été identifiés par clonage ou dans des banques d'EST (Expressed Sequence Tags, pour étiquettes de séquences exprimées) chez différentes espèces

de plantes (synthèses bibliographiques par Boerjan *et al.*, 2003 ; Harakava *et al.*, 2005 ; Shi *et al.*, 2010). Les travaux de séquençage du génome menés chez des espèces modèles comme *Arabidospsis thaliana*, *Populus trichocarpa* ou *Oryza sativa* ont notamment permis de montrer que la plupart des gènes de la voie de biosynthèse des lignines appartiennent à des familles multigéniques avec pour certains gènes l'existence de nombreux paralogues plus ou moins bien conservés. Xu *et al.* (2009) se sont penchés sur les familles de gènes de structure impliqués dans la voie de biosynthèse des lignines dans une étude comparative chez 14 espèces de plantes et une espèce de champignon. Cette étude rapporte l'existence de 149 gènes chez *Populus trichocarpa*, 157 chez *Oryza sativa*, 63 chez *Arabidopsis thaliana*. Cependant, les études d'expression menées par exemple chez le peuplier, montrent qu'une faible proportion de ces gènes sont exprimés dans le xylème en différenciation (Shi *et al.*, 2010). Ceci suggère des rôles différents dans le processus de lignification pour l'ensemble des gènes d'une même famille.

4.2.1.2.Les gènes de régulation

Les facteurs de transcription de la famille des *MYBs* ont été proposés comme régulateurs de l'expression des gènes structuraux de la biosynthèse des monolignols. Certains de ces facteurs de transcription pourraient stimuler ou réprimer l'expression, dans l'espace et dans le temps, des gènes de la voie de biosynthèse des lignines en se fixant sur les éléments cis régulateur de type AC contenus dans leur promoteur. Ceci a été montré pour la première fois par Hatton *et al.* (1995) pour le promoteur du gène *PAL2* dans des plants de tabac transgéniques. La présence de ces éléments de type AC (aussi connus sous le nom de G box) a été démontrée dans les promoteurs d'autres gènes de la voie de biosynthèse des lignines comme *4CL*, *CCoAOMT*, *C4H*, *CAD* et *CCR* (revue bibliographique par Rogers et Campbell, 2004). Pour le moment, peu de gènes codant ces facteurs de transcription ont été montrés comme étant associés à la formation du bois et à la biosynthèse des lignines (revu par Vanholme *et al.*, 2010). Parmi ces gènes, *PtMYB4* chez le pin, *EgMYB1* et *EgMYB2* chez l'eucalyptus et *PtMYB21* chez le peuplier sont des activateurs ou répresseurs de l'expression de gènes de la voie de biosynthèse des lignines (revue bibliographique par Zhong et Yé, 2007 et Zhong et Yé, 2009). Un facteur de transcription de type *LIM* a également été montré par Kawaoka *et al.* (2000) comme capable de se lier de manière spécifique aux éléments de type AC des promoteurs de certains gènes de la voie de biosynthèse des lignines et d'activer leur expression. Plus récemment, des études menées chez la vigne semblent désigner un facteur de

transcription de type *WRKY* comme capable de réguler l'expression de plusieurs gènes de la voie de biosynthèse des lignines (Guillaumie *et al.*, 2010).

4.2.2. L'apport de la transgénèse dans l'étude de la relation entre gènes de la lignification et caractères quantitatifs relatifs aux lignines

Les approches de génomique inverse visent à altérer la fonction d'un ou plusieurs gènes de manière ciblée dans le but de caractériser leurs effets sur la variation des phénotypes. Beaucoup d'approches de ce type ont été menées sur les gènes impliqués dans la voie de biosynthèse des lignines et des données sont aujourd'hui disponibles chez de nombreuses espèces de plantes (synthèses bibliographiques par Boerjan *et al.*, 2003 ; Baucher *et al.*, 2003 ; Rogers et Campbell, 2004 ; Zhong et Yé, 2009).

Les résultats montrent qu'une modification d'expression impacte la quantité et la qualité des lignines pour une grande partie des gènes de structure de la voie de biosynthèse des lignines. Globalement, il semble que la diminution de l'expression des gènes *PAL*, *C4H*, *4CL*, *HCT*, *C3H*, *CCoAOMT*, *CCR* et *CAD* ait un effet sur la teneur en lignines. La diminution d'expression de certains gènes entraine également une modification de la composition des lignines. C'est le cas par exemple pour *C3H*, *F5H*, *COMT* et *CAD* (Vanholme *et al.*, 2008). Cependant, les résultats de ces études sont parfois contradictoires. Par exemple, les études menées sur le gène *COMT* chez le tabac montrent des réductions du taux de lignines dans un cas (Ni *et al.*, 1994) parmi trois (Dwivedi *et al.*, 1994 ; Atanassova *et al.*, 1995). De même, si la modification de l'expression du gène *CAD* réduit la quantité de lignines chez Arabidopsis (Sibout *et al.*, 2003), elle n'a pas d'effet sur la teneur en lignines chez le tabac (Halpin *et al.*, 1994 ; Yahiaoui *et al.*, 1998 ; Stewart *et al.*, 1997). Ceci suggère l'intervention de mécanismes de régulation complexes dans la biosynthèse des monolignols. Il existe des évidences de l'existence de tels mécanismes pouvant agir au niveau transcriptionnel ou post-traductionnel (Anterola *et al.*, 2002).

4.2.3. Les caractères quantitatifs liés aux lignines

Deux types de caractères sont classiquement étudiés pour mesurer la variation des lignines dans un échantillon. La teneur en lignines qui renseigne sur la quantité de lignines dans le bois et le rapport entre monomères S et G ou H et G qui traduit la qualité des lignines.

Il existe plusieurs méthodes pour décrire la teneur en lignines d'un échantillon de bois (synthèse bibliographique par Hatfield et Fukushima, 2005). Des méthodes sont basées sur

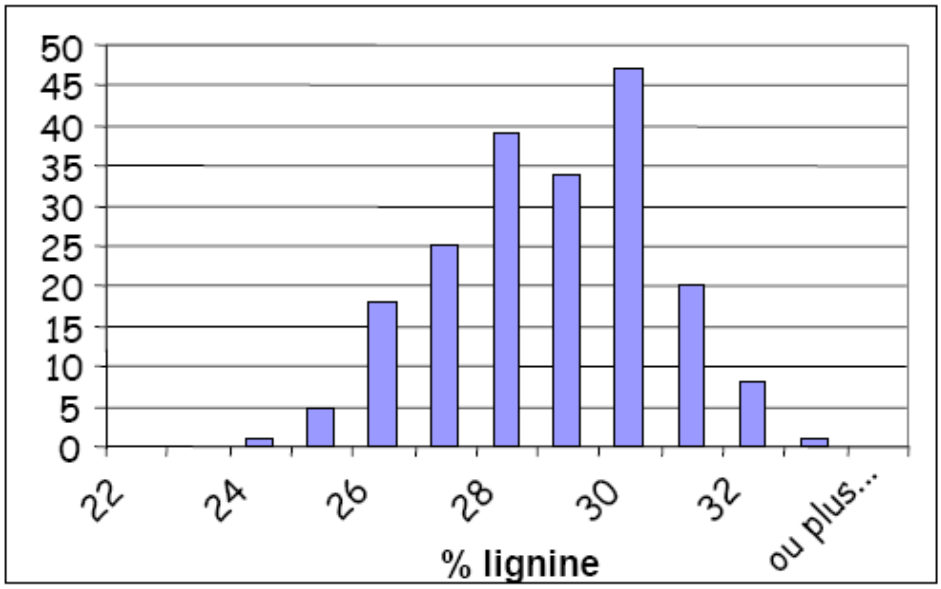


Figure 14 : distribution de la teneur en lignines (% du poids sec) dans une famille de plein-frères *E. urophylla* x *E. grandis*.

différents principes physiques, comme la gravimétrie ou la résonnance magnétique nucléaire (RMN). Mais la méthode la plus ancienne, qui reste aujourd'hui la plus communément utilisée, est une méthode chimique dite méthode de Klason décrite par Browning (1967). Cette méthode repose sur la séparation des lignines par dépolymérisation des celluloses et hémicelluloses dans de l'acide sulfurique à 72% et l'hydrolyse des polysaccharides dans de l'acide sulfurique bouillant à 3%. Pendant le déroulement de ce procédé, une partie des lignines est dissoute dans le filtrat. Cette part correspond aux lignines solubles dans l'acide. La teneur en lignines totale correspond à la somme du contenu en lignines de Klason (lignines insolubles) et en lignines solubles dans l'acide. La teneur en lignines est exprimée en pourcentage du poids sec d'un échantillon de bois. Poke *et al.* (2006) ont mis en évidence chez *E. globulus* une corrélation significative (0.93) entre la teneur en lignines totales et la teneur en lignines de Klason.

La mesure de l'abondance relative des monomères est plus récente. Elle peut être déterminée par une analyse en spectrométrie de masse couplée à de la chromatographie en phase gazeuse effectuée sur les produits de la pyrolyse du bois (Rodrigues *et al.*, 1999) ou à de la chromatographie liquide à haute pression (HPLC) sur les produits de dégradations chimiques de type thioacidolyse (Lapierre *et al.*, 1995). Le rapport entre monomères est une mesure sans dimension qui exprime des quantités relatives.

Ces mesures chimiques étant coûteuses, elles sont mal adaptées à l'étude d'un grand nombre d'échantillons. Elles peuvent aujourd'hui être prédites par l'utilisation de la méthode SPIR. De nombreuses études relatent l'efficacité de cette méthode pour la prédiction de la quantité et de la qualité des lignines chez différentes espèces d'arbres (Tsuchikawa *et al.*, 2007). Dans la majorité de ces études, de fortes corrélations ont été rapportées entre les valeurs prédites et les valeurs mesurées par les méthodes de mesure classiques. Ces nouvelles méthodes constituent une avancée technologique déterminante pour l'étude de ces propriétés du bois et leur inclusion dans les programmes d'amélioration génétique des arbres forestiers (Raymond, 2002 ; Raymond et Apiolaza 2004).

4.2.4. Les paramètres génétiques des caractères relatifs aux lignines

En population, la quantité et la qualité des lignines présentent des patrons de variation continue avec une distribution normale (Figure 14) suggérant un déterminisme polygénique pour ces deux propriétés chimiques. L'étude du déterminisme génétique de ces caractères est relativement récente chez les arbres forestiers, puisque les premiers résultats sur l'estimation

des paramètres génétiques pour la teneur en lignines ont été rapportés chez le pin maritime par Pot *et al.* (2002). Même si peu d'études sont encore disponibles, les premiers résultats chez différentes espèces indiquent des niveaux de variation de la teneur en lignines assez faibles par rapport aux caractères de croissance avec des coefficients de variations phénotypiques allant de 3% à 5,4% (Hannrup *et al.*, 2004 ; Poke *et al.*, 2006 ; Pot *et al.*, 2002). Ces études rapportent également des valeurs d'héritabilité au sens strict pouvant varier entre 0,42 et 0,56 et des valeurs d'héritabilité au sens large comprises entre 0,47 et 0,54 selon les espèces (Pot *et al.*, 2002 ; Poke *et al.*, 2006). La variation phénotypique de la teneur en lignines, bien que modérée, semble donc être dans la plupart des cas sous un contrôle génétique fort, principalement additif. De plus, une étude de Sykes *et al.* (2006) montre qu'il existe pour la teneur en lignines une relative stabilité des paramètres génétiques entre bois juvénile et bois de transition chez *Pinus taeda*.

Pour le moment, il n'existe aucune étude rapportant des estimations de paramètres génétiques pour le rapport entre monomères S et G ou H et G pour des populations de génotypes non apparentés chez les arbres forestiers. Cependant, Baillères *et al.* (2002) rapportent un coefficient de variation de 13,4% pour le rapport S/G dans une famille de pleins frères *E. urophylla* x *E. grandis* suggérant des niveaux de variabilité plus importants que ceux observés pour la teneur en lignines. Des résultats similaires ont été obtenus chez *Pinus pinaster* (Pot *et al.* 2006), pour le rapport H/G (coefficient de variation de 39% au sein d'une famille de plein-frères). Jusqu'à aujourd'hui, une seule étude rapporte une estimation de l'héritabilité pour la qualité des lignines (Novaes *et al.,* 2009). Dans cette étude, la faible héritabilité au sens large du rapport S/G chez le peuplier (reproductibilité clonale de 0,378 pour une famille de plein-frères obtenue par pseudo-backcross) suggère un contrôle génétique plus modéré pour ce caractère que pour la teneur en lignines. Ces différentes données sur la variabilité de la qualité des lignines sont limitées à l'étude de quelques familles et ne sont pas représentatives de la variabilité de ces caractères en population naturelle. Ces tendances demandent donc à être confirmées dans des contextes génétiques plus larges impliquant des échantillons plus importants et plus divers. Ceci a été l'un des objectifs de mon travail de thèse.

4.3. L'eucalyptus : un « génome forestier » modèle

En plus de ses caractéristiques sylvicoles intéressantes pour son utilisation en plantations, le génome de l'eucalyptus présente des caractéristiques qui en font un des arbres

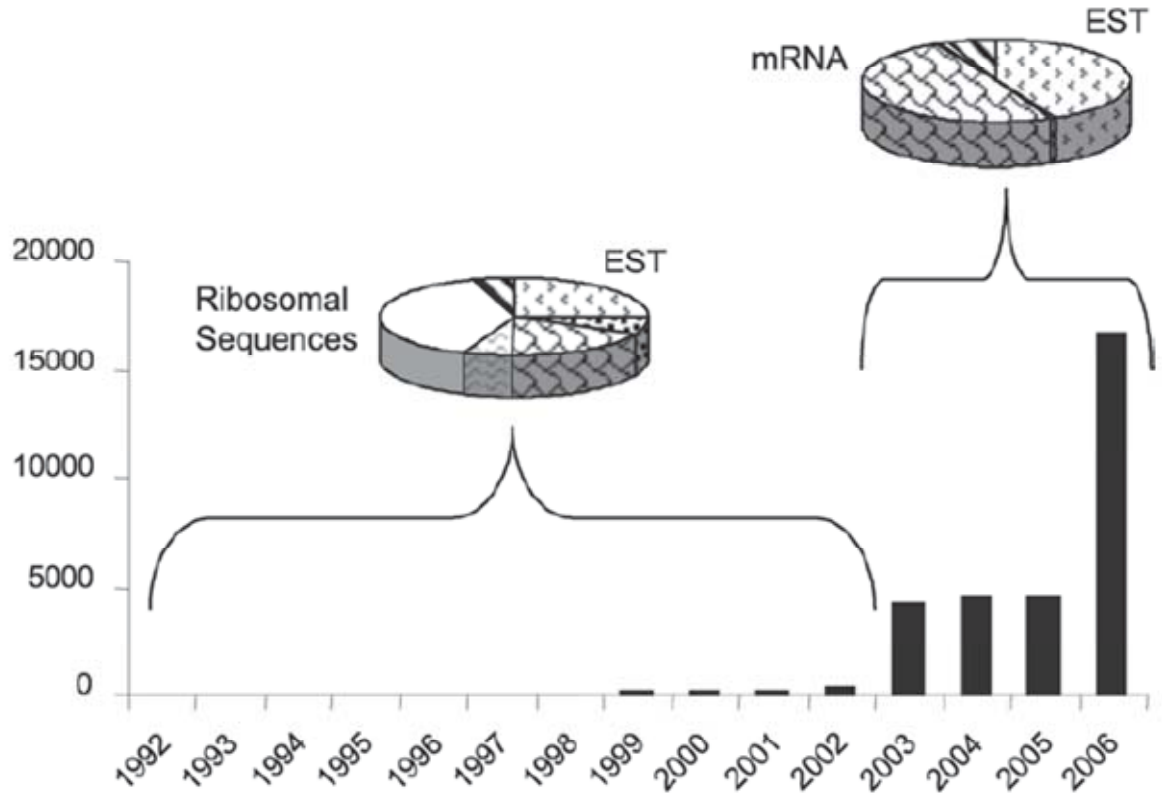


Figure 15 : quantité de séquences d'ADN d'eucalyptus déposées dans les bases de données et abondance relative des séquences issues d'EST entre 1992 et 2006. D'après Myburg *et al.*, 2007.

forestiers modèle à côté du peuplier, pour tous les travaux de génomique fonctionnelle ou structurale.

4.3.1. Caractéristiques du génome d'eucalyptus

Les espèces du genre *Eucalyptus* sont diploïdes avec un génome nucléaire constitué de 11 paires de chromosomes homologues (Eldridge *et al.*, 1993 ; Potts et Wiltshire, 1997). La taille de ce génome (en génome haploïde) a été estimée par cytométrie en flux et comparaison avec des érythrocytes de poulet pour plusieurs espèces pures et hybrides (Grattapaglia et Bradshaw, 1994). Selon les espèces les tailles estimées variaient entre 370 et 700 millions de paires de base (Mpb) avec, pour les espèces du sous genre *Symphyomyrtus*, une moyenne de 650 Mpb. Une estimation plus récente de la taille du génome haploïde d'*E. globulus*, réalisée avec un standard interne végétal, indiquait une taille moyenne de 644 Mpb chez cette espèce. Au regard de ces données, le génome nucléaire des eucalyptus apparait plus important que celui d'*Arabidopsis thaliana* (125 Mpb ; The Arabidopsis genome initiative, 2000), et de l'ordre de celui d'*Oryza sativa* (420 - 466 Mbp ; Yu *et al.*, 2002 ; Goff *et al.*, 2002) ou de *Populus trichocarpa* (410 - 485 Mpb ; Tuskan *et al.*, 2006), trois espèces végétales dont le génome a été entièrement séquencé. Il reste largement inférieur aux tailles de génome des gymnospermes. En effet, Bogunic *et al.* (2003) ont estimé la taille des génomes de cinq espèces de pin entre 20830 et 26920 Mpb, dues principalement à de grandes régions d'ADN répétées chez ces espèces (Ahuja, 2001). La taille du génome de l'eucalyptus constitue un atout majeur pour des travaux de biologie moléculaire et le place, comme le peuplier, comme « genre modèle » chez les ligneux.

4.3.2. Les ressources génomiques disponibles

Chez l'eucalyptus comme chez les arbres forestiers en général, les données de séquences publiques ont été développées moins rapidement que chez d'autres espèces modèles comme *Arabidopsis* ou les espèces de grandes cultures comme le riz et le maïs. Depuis le dépôt de la première séquence de gène d'eucalyptus (CAD ; Feuillet *et al.*, 1993), le nombre de séquence de gènes d'eucalyptus n'a cessé d'augmenter dans les bases publiques (Figure 15). A partir de 2003, l'augmentation du nombre de séquences est principalement due à la publication de marqueurs de séquence exprimée ou Expressed Sequence Tag (EST) qui sont des séquences issues d'ARN messager. En 2006, environs 15000 EST d'eucalyptus étaient disponibles, issues de séquençage Sanger. Ce chiffre est passé en 2010 à 55619 séquences (36981 référencées comme EST et 18638 comme séquences issues d'ADN

génomique). La majorité de ces séquences proviennent de quelques espèces du sous genre *Symphyomyrtus* : *E. camaldulensis*, *E. grandis*, *E. gunnii*, *E. globulus* et *E. tereticornis*.

Avec l'avènement des nouvelles générations de séquenceurs, le nombre de séquences de génomes forestiers est amené à considérablement augmenter (ces différents outils seront détaillés dans la partie 5.1). L'eucalyptus suit cette tendance avec actuellement 1,04 millions de séquences disponibles (correspondant à 166,4 Gb). Les séquences générées sont en général plus courtes que celles issues de séquençage Sanger, mais elles sont générées en beaucoup plus grand nombre. Novaes *et al.* (2008) ont récemment réalisé un pyroséquençage (454, Roche) chez *E. grandis* qui a permis de générer plus d'un millions de séquences (entre 100 et 200 pb) qu'ils ont pu regrouper en 71 000 contigs environ (119 000 séquences étant identifiées comme singletons). En général, pour 50% des séquences générées une identification fonctionnelle est possible. A titre d'exemple Paux *et al.* (2004), sur 224 unigenes générés à partir d'ADNc issus d'une banque soustractive de xylème, ont mis en évidence que 39% des séquences correspondaient à différentes catégories fonctionnelles traduisant la complexité du xylème secondaire dans le bois d'angiospermes.

En 2010, une première version du génome total d'*E. grandis* a été établie (http://eucalyptusdb.bi.up.ac.za ; les données devraient être complètement accessibles au début de l'année 2011). Ce génome de référence ainsi que les nouveaux outils technologiques haut débit de séquençage et de génotypage SNP offrent de nouvelles perspectives pour le développement et l'utilisation des ressources génomiques chez l'eucalyptus. Une initiative internationale (*Eucalyptus* Genome Network ; EUCAGEN) est d'ailleurs en cours pour coordonner le développement de ces ressources.

4.3.3. Architecture génétique des caractères d'intérêt

Aujourd'hui, l'étude de l'architecture génétique des caractères complexes chez l'eucalyptus est devenue un enjeu majeur à l'aube du séquençage complet de son génome (Grattapaglia, 2008). Une des étapes clef permettant de disséquer les bases génétiques des caractères complexes est la cartographie de régions génomiques à effet majeur (QTL) sur la variation des caractères d'intérêt (Doerge, 2002).

4.3.3.1. Cartographie génétique

Les premières cartes génétiques d'*Eucalyptus* ont été construites majoritairement avec des marqueurs dominants (Random Amplified Polymorphic DNA, RAPD ; ou Amplified

Fragment-Length Polymorphism, AFLP) et plus rarement codominants (izozymes et/ou Restriction Fragment Length Polymorphism, RFLP), grâce à une stratégie de double pseudo test cross, à partir de croisements inter-spécifiques ou intra-spécifiques (Grattapaglia et Sederoff, 1994 ; Byrne *et al.*, 1996 ; Verhaegen et Plomion, 1996 ; Marques *et al.*, 1998). Ces cartes ont ensuite été densifiées avec des marqueurs codominants de type microsatellites (« Simple Sequence Repeat », SSR, développés par Brondani *et al.*, 1998 ; Marques *et al.*, 2002 ; Brondani *et al.*, 2006), ou de type Single Strand Conformation Polymorphism, SSCP (Gion *et al.*, 2000 ; Thamarus *et al.* 2002). Parmi ces marqueurs, les SSR se sont révélés particulièrement bien transférables entre différentes espèces du sous-genre *Symphyomyrtus* (Van Der Nest *et al.*, 2000 ; Brondani, 2002 ; Marques *et al.*, 2002 ; Brondani *et al.*, 2006) suggérant une bonne conservation des génomes au sein de ce sous-genre. Aujourd'hui, des cartes génétiques saturées avec 11 groupes de liaison sont disponibles pour six espèces d'Eucalyptus utilisées en plantations comme *E. camaldulensis* (Agrama *et al.*, 2002), *E. globulus* (Bundock *et al.* 2000; Thamarus *et al.* 2002; Freeman *et al.*, 2006), *E. grandis* (Grattapaglia et Sederoff, 1994 ; Verhaegen et Plomion, 1996), *E. tereticornis* (Marques *et al.*, 1998), *E. urophylla* (Grattapaglia et Sederoff, 1994; Verhaegen et Plomion, 1996), *E. nitens* (Byrne *et al.*, 1995). Dans certains cas, des gènes candidats ont aussi été cartographiés (Gion *et al.*, 2000, Thamarus *et al.*, 2002). Cependant, en 2006 cette comparaison entre génomes restait encore limitée à quatre espèces maximum avec très peu de marqueurs communs entre cartes (Myburg *et al.*, 2007), ce qui ne permettait pas réellement une analyse détaillée de la synténie ou même de la colinéarité entre génomes.

Les développements récents en matière de génotypage de marqueurs SNP devraient permettre de comparer les cartes génétiques de différentes espèces d'*Eucalyptus* de façon beaucoup plus poussée que les travaux rapportés jusqu'à aujourd'hui, et ceci permettra aussi de comparer les QTL localisés séparément chez ces différentes espèces et donc de les valider.

4.3.3.2.Détection de QTL

Plusieurs études QTL ont été rapportées pour différents types de caractères chez les eucalyptus comme la qualité du bois, la floraison, la croissance, la tolérance au froid et la résistance aux pathogènes (revue par Myburg *et al.*, 2007 ; Poke *et al.*, 2005 ; Freeman *et al.*, 2009 ; Thumma *et al.*, 2010).

Concernant les propriétés du bois, les études ont été principalement réalisées sur la densité du bois (Grattapaglia *et al.*, 1996 ; Verhaegen *et al.*, 1997 ; Bundock *et al.*, 2008).

Thamarus *et al.* (2004) ont publié les premiers résultats significatifs sur l'architecture des propriétés du bois en utilisant des méthodes de phénotypage haut débit qui ont permis de détecter des QTL pour la longueur de fibre, la teneur en cellulose, le rendement papetier et le MFA chez *E. globulus*. Trois QTL (fibre, densité du bois et rendement papetier) ont été détectés à partir de deux descendances demi-frères indiquant un effet stable possible du contrôle génétique de cette région à travers plusieurs fonds génétiques. Rocha *et al.* (2007) ont trouvé des colocalisations entre QTL de rendement papetier et de teneur en lignines à partir d'un croisement interspécifique *E. grandis* x *E. urophylla*. Des résultats similaires avaient été mis en évidence pour le rendement papetier et la teneur en cellulose chez *E. globulus* (Thamarus *et al.*, 2004). Plus récemment, Freeman *et al.* (2009) ont publié plusieurs régions du génome d'*E. globulus* qui affectent des caractéristiques physiques (densité) et chimique (teneurs en cellulose et en lignines) et ont mis en évidence des colocalisations entre QTL qui peuvent expliquer les corrélations génétiques entre ces propriétés du bois.

La disponibilité d'EST et/ou de séquences génomiques d'eucalyptus dans les banques de données publiques (Paux *et al.*, 2004) a permis de développer des approches gènes candidats avec une connaissance *a priori* des gènes de structure et de régulation impliqués dans les voies métaboliques liées aux propriétés du bois. Des colocalisations entre QTL de qualité du bois et des gènes candidats ont aussi été rapportées chez l'*Eucalyptus* : Goicoechea *et al.* (2005) ont ainsi montré une coïncidence entre un gène de régulation, *Myb2,* cartographié chez *E. grandis* et un QTL de teneur en lignines. Le même type de résultat a été obtenu entre un gène Rac-like GTPase vasculaire d'*E. urophylla*, *ROP1*, et un QTL de qualité des lignines (S/G) et de morphologie des fibres (Foucart *et al.*, 2009). Gion (2000) et Thamarus *et al.* (2004) ont observé une colocalisation entre un gène de structure de la voie de biosynthèse codant la Cinnamoyl CoA Reductase (CCR) et un QTL de teneur en lignines et en cellulose, respectivement chez *E. urophylla* et *E. globulus*.

Ces différents résultats sur l'architecture des propriétés du bois chez l'*Eucalyptus* doivent être encore développés pour offrir de réelles perspectives en termes d'amélioration génétique. En effet, compte tenu de leur domestication très récente, les populations d'amélioration d'eucalyptus présentent une grande diversité génétique et des niveaux de déséquilibre de liaison très faibles (les allèles existants à différents locus polymorphes sont associés aléatoirement) (Grattapaglia et Kirst, 2008). Dès lors, les résultats d'une analyse QTL, obtenus dans une population synthétique à base génétique étroite (deux individus

fondateurs de la population de cartographie dans le cadre d'une stratégie de type double pseudo-test-cross), sont difficilement généralisables à l'échelle de populations. Même si certaines études QTL réalisées avec différents fonds génétiques (Thamarus *et al.* 2004), ou dans différents environnements (Gion, communication personnelle) sont très encourageantes. De plus, des études de cartographie génétique comparée ont déjà permis, à petite échelle, d'identifier des QTL conservés entre différentes espèces d'eucalyptus (Marques *et al.*, 2002 ; Thumma *et al.*, 2010). Le potentiel d'utilisation des résultats des analyses QTL en sélection est également limité par la faible résolution avec laquelle les régions d'intérêt peuvent être identifiées. Ceci est dû à un faible niveau d'indépendance entre marqueurs génétiques proches. Il s'explique par le fait qu'un faible nombre d'événements de recombinaisons est pris en compte dans une descendance de petite taille obtenu à partir du croisement de deux individus. Ainsi, les intervalles de confiance qui sont associés aux QTL sont généralement de l'ordre de 10 à 20 cM (centimorgans), ce qui chez l'*Eucalyptus* représente une distance physique de l'ordre de 4 à 9 Mpb (Gion *et al.*, 2000).

Les colocalisations entre gènes connus et QTL offrent néanmoins une information précieuse pour sélectionner des gènes candidats dont l'effet pourra être validé dans des populations plus complexes (populations d'amélioration ou populations naturelles) et révéler ainsi si le gène est porteur ou non de la « variabilité fonctionnelle » contrôlant la variation du caractère d'intérêt.

4.3.4. De la diversité neutre à la diversité fonctionnelle

Malgré le très large spectre d'espèces que représente le genre *Eucalyptus*, seules quelques une ont été étudiées en termes de diversité génétique. La majorité d'entre elles pour lesquelles des études de diversité ont été réalisées sont des espèces d'intérêt commercial. En effet, la connaissance de la diversité génétique et de son organisation à l'échelle d'une espèce présente un intérêt majeur pour la mise en place et la gestion des populations d'amélioration.

La plupart des données proviennent pour le moment de l'utilisation des marqueurs moléculaires neutres (*e.g.* supposés non soumis aux effets de la sélection naturelle) et notamment des microsatellites. Les études rapportent des niveaux de diversité génétique globalement importants avec des valeurs d'hétérozygoties attendues dans une population panmictique (H_e) comprises entre 0.5 et 0.9 chez différentes espèces aux aires de répartitions continues (Mc Gowen *et al.*, 2001 ; Jones *et al.*, 2002 ; Holman *et al.*, 2003 ; Tripiana *et al.*, 2007 ; Payn *et al.*, 2008, Butcher *et al.*, 2009 ; Shepherd *et al.*, 2010). Les données obtenues

reflètent également un système de reproduction majoritairement allogame (Byrne *et al.*, 1998 ; Butcher *et al.*, 2002 ; Tripiana *et al.*, 2007 ; Butcher *et al.*, 2009). Chez certaines espèces, des flux de pollens et des tailles de populations importants permettent le maintien d'une faible différentiation entre populations. Ces faibles niveaux de différentiations ont été rapportés pour différentes espèces d'eucalyptus aux aires de répartition naturelles parfois très différentes : *E. urophylla* (Tripiana *et al.*, 2007 ; Payn *et al.*, 2008), *E. camaldulensis* (Butcher *et al.*, 2002), *E. gobulus* (Jones *et al.*, 2002) et *E. pilularis* (Shepherd *et al.*, 2010). Des niveaux de différenciation plus forts entre populations ont cependant été observés chez des espèces ayant une distribution régionale avec des populations isolées comme *E. curtisii* (Smith *et al.*, 2003) ou *E. nitens* (Byrne *et al.*, 1998).

Des études rapportent l'existence d'hybridation en milieu naturel entre espèces parapatriques (espèces aux territoires distincts mais adjacents). Il existe par exemple des évidences d'une zone d'introgression entre *E. camaldulensis* et *E. tereticornis* dans le nord ouest de l'Australie (Butcher *et al.*, 2002). Ces données vont en faveur d'une bonne conservation des génomes entre les espèces d'*Eucalyptus*. Cette hypothèse est corroborée par les études de phylogénie moléculaire et de variabilité génétique menées chez plusieurs espèces d'*Eucalyptus* qui relatent des niveaux de divergence relativement faibles entre espèces phylogénétiquement proches (Steane *et al.*, 1998 ; Balasaravanan *et al.*, 2006 ; Kulheim *et al.*, 2009). Ces données sont cohérentes avec les résultats des études de Ladiges *et al.* (2003) et Crisp *et al.* (2004) estimant sur la base de données climatiques, tectoniques et moléculaires, une séparation récente des espèces d'*Eucalyptus* (entre 5 et 70 millions d'années selon les taxons).

Aujourd'hui, peu de données ont été publiées e, ce qui concerne l'effet de la variabilité génétique sur la variation des caractères quantitatifs et plus précisément ceux liés à la qualité du bois d'eucalyptus. Or pour faciliter la mise à profit de cette diversité génétique au sein des programmes d'amélioration grâce aux marqueurs moléculaires, il est nécessaire de tester l'association entre variabilité génétique et la variation des caractères d'intérêt. Mon travail de thèse s'est donc également attaché à tester l'association entre des variants alléliques de gènes candidats de la voie de biosynthèse des lignines et la qualité et quantité des lignines chez l'eucalyptus.

5. La génétique d'association pour la recherche des polymorphismes contrôlant la variation des caractères liés aux lignines chez l'eucalyptus

La « génétique d'association » est une approche de génétique directe qui s'attache à mettre en évidence la variabilité moléculaire qui explique la variation de caractères quantitatifs. Les études de génétique d'association ont débuté chez l'homme dans les années 1980 avec pour objectif de mettre en évidence les locus impliqués dans des pathologies (synthèse bibliographique par Hirschhorn *et al.*, 2002). Au départ ciblées sur l'étude de quelques locus, ces approches ont rapidement évolué dans les dix dernières années et notamment avec le séquençage du génome humain (Lander *et al.*, 2001 ; Venter *et al.*, 2001). Aujourd'hui l'évolution des technologies de séquençage et de génotypage, la compréhension de l'organisation des polymorphismes au sein du génome (synthèse bibliographique par Ardlie *et al.*, 2002), le développement rapide des ressources génomiques (Gibbs *et al.*, 2003) et des tests statistiques (synthèses bibliographiques par Balding, 2006 ; Stephens et Balding, 2009), permettent de réaliser ce type d'étude à l'échelle du génome entier. Chez les plantes, ces approches ont été initiées plus récemment (Thornsberry *et al.*, 2001).

Les études d'association utilisant des populations plus ou moins complexes et structurées constituent aujourd'hui une approche complémentaire des études traditionnelles de détection de QTL utilisées chez les plantes jusqu'alors. Basées sur l'exploitation des événements de recombinaison historiques qui ont lieu au sein des populations (Nordborg et Tavare, 2002), ces approches présentent deux avantages majeurs par rapport aux études de liaison génétique classiques : i/ une meilleure résolution par l'augmentation des événements de recombinaison pris en compte dans l'analyse, et ii/ la prise en compte d'une variabilité génétique plus large que celle analysée dans des descendances utilisées pour la détection de QTL (en général, un seul croisement issu de deux parents) (Yu et Buckler, 2006). Chez les plantes modèles comme *Arabidopsis* ou certaines espèces de grandes cultures, ces approches ont connu une avancée rapide ces dernières années notamment avec l'arrivée de nouvelles méthodes de génotypage et de séquençage à haut débit (Zhu *et al.*, 2008).

5.1. Diversité nucléotidique

5.1.1. Définition des marqueurs SNP

Le terme « SNP » (pour Single Nucleotide Polymorphism) est employé pour désigner un polymorphisme moléculaire de l'ADN porté par une seule base. Les SNP représentent donc la forme la plus fine de variabilité qui peut être détectée au niveau de l'ADN. Les SNP ont pour origine la substitution d'un nucléotide par un autre et sont classés en deux catégories, selon les changements que cette substitution implique. Les transitions d'une part, correspondent au changement d'une base purique en une autre base purique (A/G) ou d'une base pyrimidique par une autre base pyrimidique (C/T), les transversions, d'autre part, impliquent les autres changements de type purine-pyrimidine (A/C, A/T, T/G, G/C). La fréquence de ces deux événements de mutation n'est pas équivalente et les transitions constituent la forme de SNP la plus communément rencontrée (Garg *et al.*, 1999 ; Deutsch *et al.*, 2001 ; Batley *et al.*, 2003). Les transitions constituent 67% des SNP détectés chez l'humain et ceci est également rapporté chez les plantes avec jusqu'à 75% des SNP détectés entre deux cultivars de riz correspondant à des transitions (Hayashi *et al.*, 2004). Certains auteurs considèrent également les indels (insertions ou délétions) d'une paire de base comme des SNP bien qu'ils dérivent probablement d'autres mécanismes mutationnels.

Les SNP peuvent en principe présenter quatre allèles différents pour un site particulier. Cependant, la faible fréquence globale des événements mutationnels qui les produisent (1.10^{-9} mutations par pb et par an pour les sites non sélectionnés chez les mammifères (Martinez-Arias *et al.*, 2001), ainsi que la plus faible fréquence d'apparition des transversions en comparaison des transitions, font que la probabilité d'occurrence de deux événements mutationnels indépendants sur un même nucléotide est très faible (Vignal *et al.*, 2002). Les SNP sont donc en grande majorité bi-alléliques. Cette propriété leur confère un désavantage en comparaison d'autres marqueurs moléculaires multialléliques comme les SSR, qui sont plus résolutifs que les SNP pour l'étude de la variabilité génétique. Ce désavantage est cependant compensé par le fait que, les SNP sont présents en très grand nombre dans le génome des organismes (1 SNP pour 100 à 300 pb en moyenne chez les plantes, environ 1 SNP tous les 1200 pb chez l'humain ; Edwards *et al.*, 2007). L'information de plusieurs SNP peut ainsi être combinée à l'échelle d'un chromosome pour constituer des haplotypes (combinaisons d'allèles pour plusieurs SNP situés sur le même chromosome). L'abondance des marqueurs SNP ainsi que leur bonne conservation au cours des générations (Lopez *et al.*,

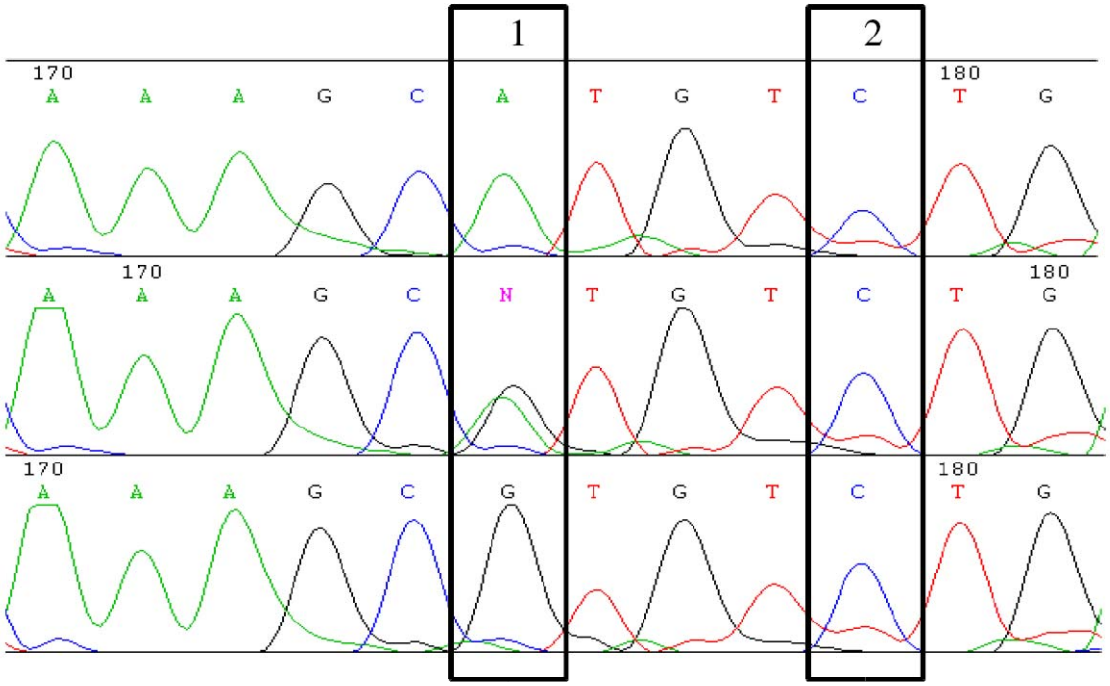


Figure 16 : identification de SNP par séquençage direct d'amplicons. Le cadre 1 indique une variation SNP observée par un individu homozygote AA, un individu hétérozygote AG (pics superposés) et un individu homozygote GG. Le cadre 2 montre une superposition de pics due à un artéfact de la réaction dans le cadre de 3 individus homozygotes CC.

2005) en ont fait des outils de choix dans le cadre d'études nécessitant la mise en évidence de marqueurs moléculaires à haute densité. Ils sont aujourd'hui largement utilisés pour l'étude du déterminisme génétique des caractères (cartographie génétique fine de QTL, clonage positionnel, études d'association) mais également pour comprendre les mécanismes d'évolution des génomes à l'échelle du gène (Syvanen, 2001).

5.1.2. Détection des SNP

Comme pour la plupart des marqueurs moléculaires, une des limites à l'utilisation des marqueurs SNP est due au coût associé à leur développement. L'identification et la caractérisation de nouveaux SNP est basée sur une étude comparative de séquences obtenues au niveau d'un gène, d'un transcriptome voire d'un génome. Il existe aujourd'hui plusieurs méthodes pour mettre en évidence ces marqueurs moléculaires.

5.1.2.1.Méthodes de séquençage d'amplicons

Les méthodes de détection par séquençage d'amplicons sont basées sur l'analyse comparative de séquences obtenues par la méthode de séquençage Sanger, pour une ou plusieurs portions d'un gène. Cette méthode de détection de SNP nécessite la construction d'amorces spécifiques de la région génomique ciblée pour l'amplification par PCR de courts fragments (de l'ordre de 500 à 1000 pb) adaptés à l'utilisation de la méthode de séquençage Sanger. La génération de ces amplicons peut ensuite être suivie d'une étape de séquençage direct, en utilisant une des amorces conçues pour l'amplification, ou par le clonage des amplicons dans un vecteur de type plasmide et le séquençage spécifique de clones obtenus en utilisant une amorce spécifique du vecteur. Le choix de l'une ou l'autre de ces deux approches est généralement déterminé par des contraintes d'ordre technique ou logistique et par des contraintes liées à la biologie de l'organisme étudié (système de reproduction, niveau de ploïdie).

Le séquençage direct de produits PCR est sans doute la stratégie la moins coûteuse à mettre en œuvre. Cependant, son succès dépend de sa capacité à détecter les SNP dans un mélange d'amplicons issus de deux haplotypes (dans le cas d'espèces diploïdes) ou plus (dans le cas d'espèces polyploïdes) (Figure 16). Selon la région du génome ciblée, la présence d'indels cause des décalages d'électrophorégrammes qui peuvent considérablement compliquer la détection des polymorphismes. De plus, il s'avère difficile dans le cadre de ce type d'approche de différencier les SNP de variations apportées par l'amplification parasite de

locus paralogues lors de la PCR. Enfin, la donnée de séquençage obtenue ne permet pas de déterminer directement la phase de liaison gamétique entre les allèles des SNP. L'obtention des haplotypes, au locus considéré, doit alors être reconstruite par l'utilisation de méthodes statistiques comme celles proposées par Stephens *et al.* (2001).

Le séquençage spécifique de clones demande plus de temps car il nécessite l'ajout de l'étape de clonage, souvent longue et fastidieuse dans le cas de l'étude d'un grand nombre d'individus et de locus. De plus, cette stratégie pose la question du nombre de clones qui doivent être étudiés, pour chaque produit de réaction PCR, afin de palier aux problèmes du tirage aléatoire des allèles au locus ciblé parmi les clones sélectionnés, à l'amplification préférentielle d'allèles, aux erreurs d'amplification de la Taq polymérase (1 erreur toute les 1 000 paires de bases répliquées selon Palumbi et Baker, 1994), à l'amplification parasite de multiples paralogues et à la recombinaison technique (produits PCR chimères, Cronn *et al.*, 2002), tous dus à la réaction de PCR. La prise en compte d'un nombre insuffisant de clones (<8) peut aboutir à la mise en évidence de « faux » SNP.

Ces méthodes ont été largement utilisées pour mettre en évidence des SNP, dans le but d'étudier la diversité nucléotidique de gènes ou l'association entre polymorphismes SNP et la variation de caractères en populations. Dans la plupart des cas, ces études étaient limitées à quelques gènes ou portions de gènes pour quelques individus (Edwards *et al.*, 2007).

5.1.2.2.Méthodes de détection in silico

Les approches de détection de SNP *in silico*, restent réservées aux quelques espèces pour lesquelles des ressources génomiques de type banque d'EST ou génome entier sont disponibles. Par exemple, dans le cas d'Arabidopsis, les données obtenues des banques d'EST représentent plus d'un million de séquences Sanger de gènes exprimés dans différentes conditions (organes, tissus, stress…) souvent obtenues sur la base de quelques individus. Ces ressources peuvent être exploitées pour identifier de nouveaux SNP (Schmid *et al.*, 2003 ; Pavy *et al.*, 2006). Cependant, bien que peu coûteuse, cette approche présente certains désavantages. La mauvaise qualité des séquences obtenues dans le cadre de séquençage d'EST et la difficulté de discriminer les séquences relatives à des gènes paralogues, sont autant de facteurs qui diminuent l'efficacité de ces approches (Ganal *et al.*, 2009). Les taux de validation des SNP obtenus par ce type d'approches varient ente 50% et 85% selon les études (Kota *et al.*, 2003 ; Le Dantec *et al.*, 2004 ; Lepoittevin *et al.*, 2010).

5.1.2.3.Méthodes de séquençage nouvelle-génération, une révolution ?

Aujourd'hui, les attentes en matière d'identification de SNP à haut débit sont focalisées sur les technologies de séquençage nouvelle-génération. Capables de générer l'information de plusieurs centaines de millions de paires de bases en une seule expérimentation, ces méthodes devraient permettre d'identifier de grandes quantités de SNP chez de nombreuses espèces et à moindre coût (Mardis, 2007). Cette méthode a été utilisée pour le reséquençage complet du premier génome d'Arabidopsis (Ossowsky *et al.*, 2008) mais son utilisation chez des espèces non modèles, dont le génome n'a pas été séquencé, n'est encore qu'à ses débuts. Les premières études indiquent qu'au-delà du nombre de paires de base séquencées dans une expérimentation, c'est le nombre de copies associé à une redondance technique (Ueno *et al.*, en préparation) qui sont limitantes pour la détection de « vrais » SNP. Barbazuk *et al.* (2007) ont utilisé cette méthode de séquençage pour la mise en évidence de SNP chez deux lignées de maïs. Les auteurs se sont focalisés sur l'analyse du transcriptome de méristèmes apicaux (plus de 250000 EST) et ont pu identifier 7000 SNP au sein de 2400 gènes avec un taux de faux positifs estimé à 15%. Chez l'eucalyptus, Novaes *et al.* (2008) ont également utilisé cette méthode pour séquencer des ADNc obtenus à partir de multiples tissus sur 21 génotypes d'*E. grandis*. Les auteurs ont pu identifier, en trois expérimentations, près de 24000 SNP avec un taux de faux positifs estimé à 17%. Plus récemment, Kulheim *et al.* (2009) ont utilisé la même méthode pour séquencer 23 gènes impliqués dans la synthèse de métabolites secondaires chez 1764 individus représentant 4 espèces d'eucalyptus. Les auteurs ont ainsi identifié plus de 8600 SNP qu'ils ont utilisé pour étudier la diversité nucléotidique de ces espèces. Ces premiers résultats sont prometteurs et ces nouvelles méthodes devraient bientôt devenir des standards en matière de découverte de SNP. Cependant, ce séquençage « nouvelle génération » pose le problème de la gestion des grandes quantités de données générées par ce type d'approche. Les nouveaux défis de la bioinformatique s'orientent vers la mise au point d'outils d'analyse adaptés à ces données (Parkhill *et al.*, 2010).

5.1.3. Echantillonnage pour la détection de SNP

Le développement des marqueurs SNP reste pour le moment encore coûteux, il est souvent limité à un sous échantillon d'une population, représentatif de la variabilité génétique de la population étudiée. Cet échantillon est appelé panel de détection de SNP. La taille et la représentativité de ce panel déterminent la possibilité de mettre en évidence des SNP fréquents ou des SNP rares dans les populations de l'espèce étudiée. On parle de fréquence de

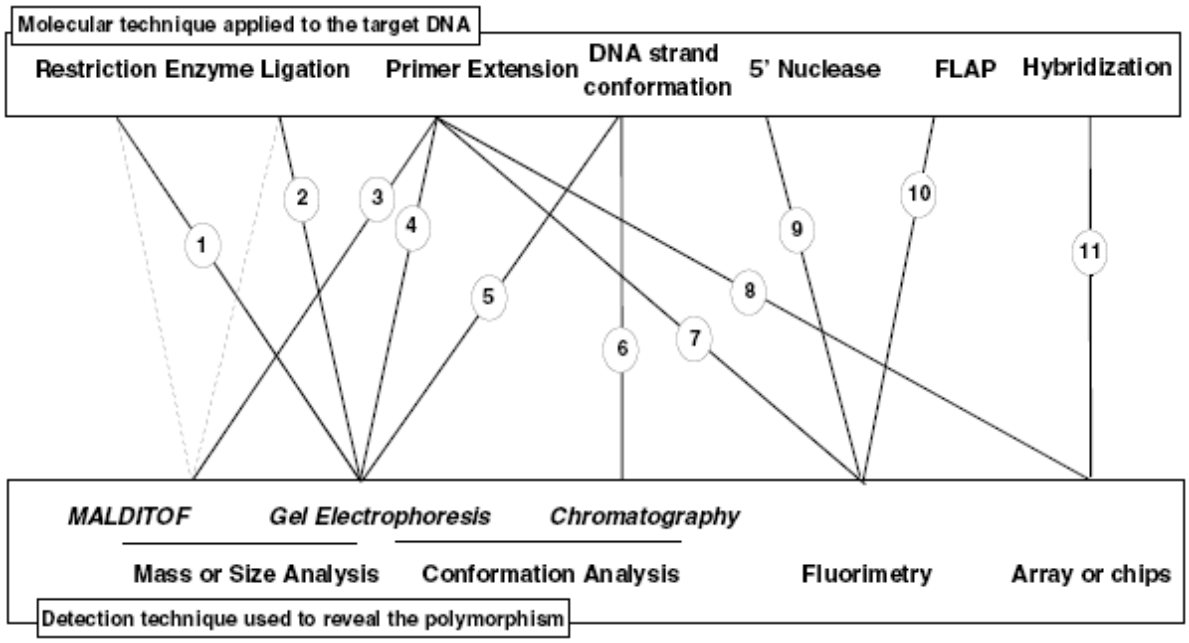


Figure 17 : diversité des méthodes permettant le génotypage de SNP en fonction des méthodes moléculaires appliquées à l'ADN et des techniques de détection utilisées pour révéler le polymorphisme. D'après Vignal *et al.* (2002).

l'allèle minoritaire (MAF pour Minor Allele Frequency) pour désigner, la fréquence de l'allèle d'un SNP qui apparait le moins souvent dans l'échantillon. Selon la taille de l'échantillon, on parle de SNP rare si la MAF est inférieure à 1%, 5 % ou 10 %. Sur la base de ces données, tout ou partie des SNP identifiés peuvent être sélectionnés pour être mis en évidence au sein de populations plus larges par l'utilisation de méthodes de génotypage. On peut envisager raisonnablement que dans les années à venir, les contraintes liées à l'échantillon utilisé pour la mise en évidence des SNP seront contournées par un séquençage direct de tous les individus de la population étudiée.

5.1.4. Génotypage de SNP

Les méthodes de génotypage permettent d'obtenir le génotype d'un individu pour un SNP ou un ensemble de SNP connus. Il existe aujourd'hui un grand nombre de méthodes (Figure 17) qui se distinguent par leur « débit » d'analyse (nombre d'échantillons et d'individus dont le génotype peut être déterminé en même temps), leur coût d'analyse et leur facilité d'accès en termes d'équipements.

5.1.4.1.Les méthodes de génotypage bas et moyens débits

Les méthodes de génotypage bas et moyens débits font partie des premières méthodes utilisées pour le génotypage. Ces méthodes sont basées essentiellement sur la mise en évidence de différents types de polymorphismes : polymorphisme de taille après utilisation d'enzymes de restriction (RFLP, Botstein *et al.*, 1980 et CAPS, Konieczny et Ausubel, 1993), polymorphisme de conformation de l'ADN (simple ou double brin) dans un gradient de température ou dans un gradient chimique (D/TGGE, Myers *et al.*, 1988 ; SSCP, Orita *et al.*, 1989 ; dHPLC, Kota *et al.*, 2001 ; HRM, Wittwer *et al.*, 2003). Ces méthodes présentent l'avantage d'être facilement accessibles et des coûts de mise en œuvre généralement bas. Au départ limitées à un faible débit d'analyses, certaines de ces méthodes se sont développées notamment par l'utilisation de l'électrophorèse capillaire pour la séparation des molécules d'ADN (Hebenbrock *et al.*, 1995) et des marqueurs fluorescents pour leur visualisation et permettent aujourd'hui de génotyper quelques dizaines de SNP sur des échantillons de quelques centaines d'individus (Hsia *et al.*, 2005 ; Kuhn *et al.*, 2005 ; Krypuy *et al.*, 2006).

5.1.4.2.Les méthodes de génotypage haut débit

Avec l'accumulation des données de séquence, qui devrait s'accroître par l'utilisation des méthodes de séquençage nouvelle-génération, le nombre de marqueurs SNP disponibles

chez certaines espèces de plantes permet aujourd'hui de réaliser des études à grande échelle notamment dans le domaine de la génétique d'association. Ce type d'étude nécessite l'utilisation d'outils de génotypage à haut débit, adaptés à la mise en évidence d'un grand nombre de SNP sur un grand nombre d'individus. Il existe aujourd'hui plusieurs méthodes de génotypage à haut débit. Ces méthodes sont basées sur différents principes dont l'extension d'amorces et l'hybridation de sondes allèle spécifiques.

Les méthodes d'extension d'amorces (Syvänen *et al.*, 1990 ; Syvänen, 1999) sont basées sur la conception d'une amorce de détection capables de s'hybrider sur une séquence cible directement en amont du SNP à génotyper. L'extension d'amorce est ensuite réalisée en 3' par une ADN polymérase utilisant des nucléotides marqués (quatre marquages différents pour les quatre nucléotides). Elle permet le génotypage simultané de plusieurs SNP en mélange (multiplex) et a été adaptée à différents types de plateformes de détection (microarray, électrophorèse capillaire, spectrométrie de masse, lecteur de fluorescence). Elle est aujourd'hui l'une des méthodes de génotypage à haut débit les plus utilisées et regroupe notamment les technologies Illumina GoldenGate, Sequenom MassARRAY (adaptées au génotypage d'échantillons de l'ordre de quelques centaines de SNP sur quelques centaines d'individus) et la technologie Illumina GoldenGate et infinium (permettant de génotyper de 1536 SNP à des centaines de millier) (Edenberg et Liu, 2009). Ces méthodes bien que très puissantes présentent quelques contraintes : le mélange d'amorces dans la réaction nécessite de minimiser les interactions entre amorces et ne permet pas toujours de génotyper les SNP souhaités. De plus, ces méthodes ne permettent pas de génotyper les SNP pour lesquels la zone de fixation des amorces présente une variabilité. La réussite de l'expérience de génotypage est donc largement conditionnée par une phase de conception reposant sur le choix des SNP qui seront génotypés.

Il existe aujourd'hui des méthodes de génotypage plus puissantes basées sur le principe d'hybridation de sondes spécifiques. Ces technologies utilisent l'interaction entre des oligonucléotides fixés sur support solide (lame) et la matrice d'ADN à génotyper. La différence de stabilité thermique entre la sonde et la matrice d'ADN testée est conditionnée par l'existence de mésappariements et permet de différencier les allèles d'un SNP considéré. La détection des SNP se fait par lecture de fluorescence. Cette méthode est cependant réservée aux espèces pour lesquelles des ressources importantes de marqueurs SNP ont été développées. Pour le moment, elle permet chez l'humain, de génotyper plusieurs centaines de milliers de SNP sur des milliers d'individus (Zeggini *et al.*, 2009). Nul doute que cette

méthode sera bientôt accessible aux espèces d'arbres forestiers, et notamment les espèces à fort intérêt commercial comme l'eucalyptus.

5.1.5. Caractérisation de la diversité génétique

La connaissance de la diversité génétique d'une espèce est primordiale pour conserver, gérer et valoriser les ressources génétiques. Différents types de marqueurs moléculaires, dominants ou codominants, ont été utilisés pour caractériser cette variabilité génétique, au cours des deux dernières décennies. Aujourd'hui, les marqueurs SNP sont aussi largement utilisés pour caractériser la diversité génétique des arbres forestiers (Neale et Savolainen, 2004 ; Savolainen et Pyjähärvi, 2007).

L'étude de la diversité génétique en utilisant des SNP passe par la détermination i/ de leur densité pour la région du génome considérée, et ii/ de la diversité nucléotidique (θ) qui permet d'appréhender l'histoire de groupes d'individus (espèces, populations) ainsi que les mécanismes évolutifs qui déterminent leur variabilité génétique.

5.1.5.1.Densité des SNP

La densité de SNP se mesure par le nombre de paires de bases moyen qui doit être séquencé pour détecter un SNP. Elle est la mesure la plus simple pour quantifier le polymorphisme au sein d'une séquence ou d'une région du génome. Même si les SNP sont reconnus pour être abondants au sein des génomes, leur densité varie significativement selon les espèces, les populations et les régions du génome étudiées. Chez l'Homme par exemple, la densité moyenne de SNP le long du génome est de l'ordre de 1/1200 pb avec des variations entre zones géniques et intergéniques (Zhao *et al.*, 2003). Chez plusieurs espèces de plantes, Edwards *et al.* (2007) rapportent une densité moyenne de SNP de l'ordre de 1 SNP tous les 100 à 300 pb. La majorité des donnés proviennent d'études menées sur des plantes modèles ou de grande culture pour lesquelles des quantités importantes de données de séquences sont disponibles permettant d'estimer la densité globale moyenne de SNP avec plus de précision. Par exemple, chez le soja, Zhu *et al.*, (2003) rapportent une densité de SNP de 1/273 pb, basée sur l'étude de 76 kpb pour 25 génotypes. Chez le riz, les densité moyennes rapportées varient entre 1/170 pb et 1/248 pb selon les études (Yu *et al.*, 2002 ; Hayashi *et al.*, 2004). L'étude de Hayashi *et al.* (2004) montre également une variation de cette densité selon les cultivars comparés. Chez le maïs, Ching *et al.* (2002) rapportent une densité de SNP de 1/31 pb dans

les régions non codantes et 1/124 pb dans les régions codantes pour 18 gènes étudiés au sein de 36 lignées élites.

La taille des populations, le taux de mutation, les flux de gènes entre populations, la sélection naturelle sont autant de paramètres qui influent sur le niveau de polymorphisme présent au sein des espèces ou des populations et entre les différentes régions du génome (Buckler et Thornsberry, 2002 ; Rafalski et Morgante, 2004 ; Ingvarsson *et al.*, 2008).

Dans le cas des arbres, la majorité des données disponibles ont été obtenues sur la base de l'étude de quelques régions géniques. Chez le chêne, des niveaux de variabilité importants ont été rapportés pour 11 gènes, avec une densité moyenne de 1/25 pb (Quang *et al.*, 2008). Chez le peuplier les niveaux de variabilité détectés sont également importants avec une densité de SNP moyenne pour 9 gènes de 1/26 pb Chu *et al.*, 2009), 1/60 pb pour 5 gènes selon Ingvarsson *et al.* (2005) et 1/130 pb au sein de 9 gènes pour Gilchrist *et al.* (2006). Une étude menée récemment chez l'eucalyptus, compare les niveaux de polymorphismes détectés par pyroséquençage sur 23 gènes impliqués dans la synthèse des métabolites secondaires chez 1764 individus représentant 4 espèces. Cette étude rapporte des densités de SNP variant de 1/33 pb à 1/16 pb selon l'espèce avec des variations entre zones introniques et exoniques (Külheim *et al.*, 2009). Novaes *et al.* (2008) rapportent une densité moyenne du génome exprimé d'un pool de 21 génotypes d'*E. grandis* à 1/192. De façon générale, ces résultats sont cohérents avec les caractéristiques énoncés précédemment des génomes forestiers : ils présentent un fort niveau d'hétérozygotie et une diversité génétique importante.

5.1.5.2.Estimation de la diversité nucléotidique θ

Même si de nombreux facteurs influencent les niveaux de polymorphismes (Buckler et Thornsberry, 2002), la théorie neutre de l'évolution suggère que la diversité nucléotidique θ est égale au produit de la taille efficace de la population Ne et du taux de mutation μ ($\theta = 4Ne\mu$; Kimura, 1969). Le modèle neutre standard considère des populations non subdivisées, de tailles finies, dans lesquelles les croisements se font au hasard et les individus ont la même chance de survivre et de se reproduire. Les sites de mutation synonymes sont supposés suivre ce modèle d'évolution. Ainsi pour ces sites, la dérive génétique (relative à la taille des populations) et le taux de mutation sont les processus qui déterminent la quantité de polymorphisme observée dans un échantillon. Toutes conditions étant égales par ailleurs, on espère des quantités de polymorphisme plus importantes dans les populations de grande taille

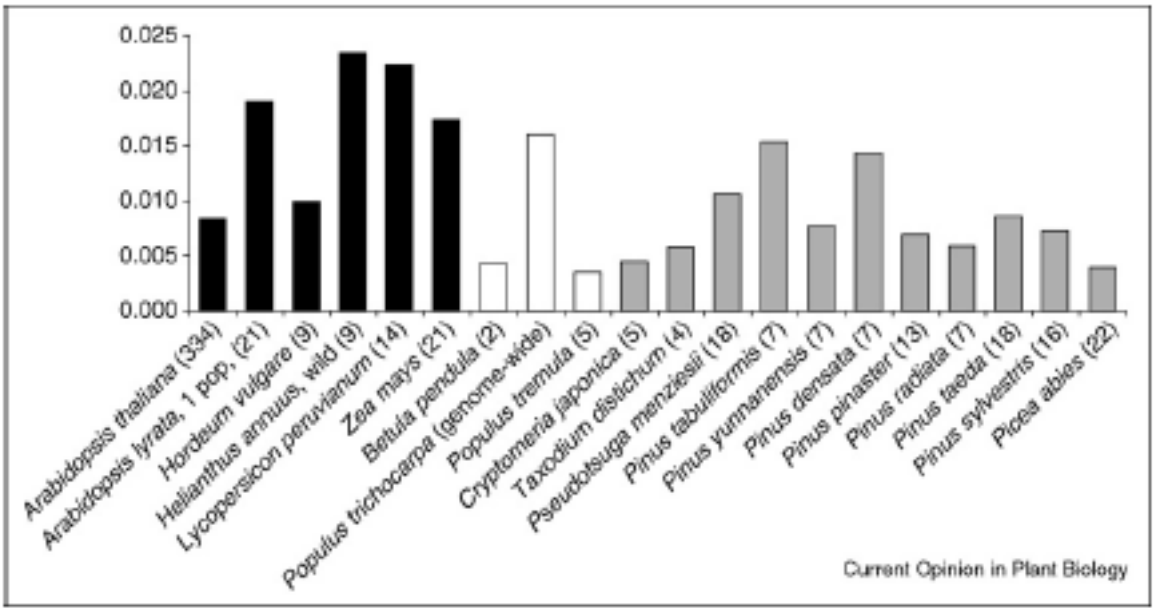


Figure 18 : Estimation de la diversité nucléotidique pour les sites silencieux (pars site) chez différentes espèces de plantes : en blanc chez les arbres angiospermes, en gris chez les conifères et en noir chez les autres espèces de plantes (d'après P. Garnier-Géré, publié dans Savolainen et Pyhajarvi, 2007).

au sein desquelles l'effet de la dérive génétique est moins important (Savolainen et Pyhäjärvi, 2007).

Les taux de mutations et la taille efficace des populations n'étant généralement pas connus, la diversité génétique θ peut être estimée de différentes manières sur la base de données de séquences. Un des estimateurs « θ_π » (Nei, 1987) est obtenu par la différence moyenne entre séquences prises deux à deux. Il dépend donc à la fois de la quantité de polymorphisme et de la fréquence des allèles (MAF) de ces polymorphismes. Un autre estimateur « θ_w » (Watterson, 1975) considère le nombre de sites polymorphes dans l'échantillon au lieu des fréquences. Ces deux estimateurs sont utilisés pour comparer les niveaux de diversité nucléotidique entre espèces, populations ou régions du génome. Ils se distinguent par l'importance accordée aux variants rares et ceux en fréquence intermédiaire dans l'estimation de θ. Différents tests permettent de comparer ces deux estimateurs dont le D de Tajima (Tajima, 1989) et le Fs de Fu (Fu, 1997). La différence entre θ_π et θ_w mise en évidence par ces tests permet par exemple de détecter un excès de variants rares (ou de variants en fréquence intermédiaire) au sein d'un échantillon traduisant un écart au modèle neutre d'évolution. Ils permettent d'identifier des régions du génome soumises aux effets de la sélection, mais sont sensibles aux effets démographiques (changement de taille des populations ou structuration de la population en sous populations).

Chez les plantes, la majorité des données proviennent de l'étude du maïs et d'*Arabidopsis*. Les autres plantes pour lesquelles des données importantes sont disponibles incluent l'orge, la tomate, le sorgho, le riz, le coton et pour les espèces forestières, certains conifères et le peuplier (synthèses bibliographiques réalisées par Wright et Gaut, 2005 et Savolainen et Pyhäjärvi, 2007). Les objectifs de ces études étaient divers : étudier l'impact de la sélection pour quelques gènes cibles, inférer l'histoire évolutive des populations ou simplement identifier les niveaux de polymorphismes au sein des espèces à partir des données de diversité nucléotidique. Pour toutes ces espèces de plantes, les niveaux de diversité génétique observés sont supérieurs à ceux estimés chez l'Homme (de l'ordre de 0,001 selon Zwick *et al.* (2000) et Frisse *et al.* (2001)). Chez les arbres, les niveaux de diversité nucléotidique sont globalement inférieurs à ceux détectés chez les plantes modèles (Savolainen et Pyhäjärvi, 2007, Figure 18). En plus de mettre en évidence la variabilité des niveaux de diversité nucléotidique entre différentes espèces et différentes populations, ces études ont également permis de mettre en évidence des différences entre les locus étudiés. Une étude récente chez le peuplier (Olson *et al.*, 2010), menée sur 881 fragments de gènes

répartis le long du génome, rapporte une valeur moyenne de θ de 0,0049 pour les sites synonymes, avec d'importantes variations entre les fragments. Les auteurs indiquent que parmi les fragments étudiés, 191 ne présentaient pas de polymorphisme.

La caractérisation de la diversité génétique des arbres forestiers grâce aux SNP n'est encore qu'à ses débuts. Même si peu de données de variabilité SNP sont aujourd'hui disponibles chez les arbres, les nouvelles méthodes de séquençage et de génotypage à haut débit offrent de réelles perspectives pour développer ce type d'études chez ces organismes. Cette caractérisation de la diversité génétique, passe aussi par une description de son organisation au sein des gènes et des génomes.

5.2. Le déséquilibre de liaison

5.2.1. Définition

Equilibre et déséquilibre de liaison sont des notions de génétique des populations utilisées pour décrire la probabilité de cooccurrence des allèles de locus différents dans une population. Le déséquilibre de liaison (DL) traduit la dépendance d'allèles à différents locus autrement dit l'association préférentielle d'allèles à des locus différents. Il s'agit donc d'une corrélation entre polymorphismes de locus différents causée par une histoire partagée en termes de mutation et de recombinaison. Dans une population idéale de Hardy-Weinberg (population de taille infinie dans laquelle les croisements se font au hasard, en l'absence de sélection, migration et mutation), tous les locus polymorphes ségrègent de manière indépendante seraient en équilibre de liaison. La liaison physique, la mutation, le régime de reproduction, la sélection notamment épistatique, les variations de taille des populations et les flux de gènes sont autant de paramètres qui influent sur la structure du DL (Flint Garcia *et al.*, 2003).

5.2.2. Estimations du déséquilibre de liaison

Il existe différents estimateurs du DL. Toutes les méthodes d'estimations sont basées sur la mesure du degré d'indépendance entre allèles de locus polymorphes pris deux à deux. Les méthodes les plus utilisées sont basées sur l'étude des fréquences des haplotypes aux deux locus considérés. Un haplotype représente une association d'allèles à plusieurs locus polymorphes. Ce qui est mesuré c'est l'écart entre les fréquences haplotypiques observées en population et l'attendu sous l'hypothèse d'indépendance en considérant les fréquences marginales des différents allèles à chaque locus observées dans la même population. Ainsi le

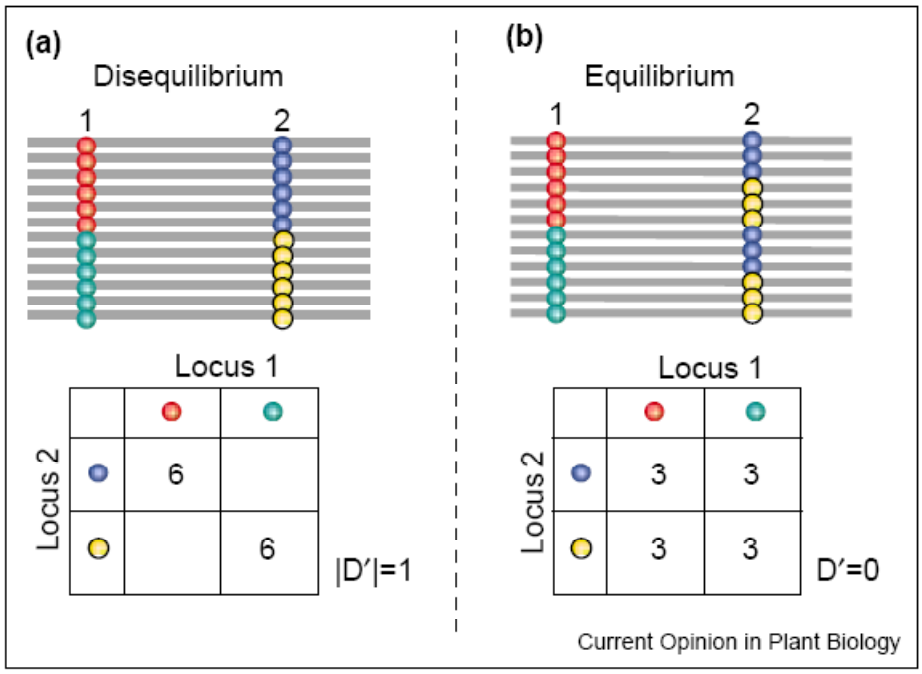


Figure 19 : représentation schématique de l'association entre allèles de 2 SNP au sein d'un échantillon de séquences dans le cas d'un déséquilibre de liaison total (D'=1), ou d'un équilibre de liaison total (D'=0). D'après Rafalski, 2002.

déséquilibre de liaison D peut être mesuré par la formule suivante proposée par Lewontin et Kojima (1960) :

$$D = \Pr(A_1, B_1) - \Pr(A_1)\Pr(B_1)$$

Où $\Pr(A_1,B_1)$ représente la fréquence de l'association entre les allèles 1 aux locus *A* et *B* observée dans la population (haplotype A_1B_1) et $\Pr(A_1)$ et $\Pr(B_1)$ les fréquences des allèles 1 observées au locus *A* et *B* respectivement. Sous l'hypothèse d'indépendance des allèles aux deux locus, D doit prendre la valeur 0. Il y a équilibre de liaison, ou en d'autres termes, la fréquence de l'association entre les allèles 1 des locus *A* et *B* est égale au produit des fréquences de ces allèles. En général, D est reporté en valeur absolue ($|D|$) et prend la même valeur que l'on considère l'association entre les allèles 1, 2 ou 1 et 2 des deux locus *A* et *B*. D'une génération à la suivante, le DL décroit en fonction du taux de recombinaison « c » entre les deux locus selon la relation :

$$D_t = (1-c)^t D_0$$

où D_t est le DL à la génération t et D_0 le DL initial.

Lorsqu'il y a déséquilibre de liaison (D≠0), la mesure de l'amplitude du DL est dépendante des fréquences allèliques aux locus considérés. D'autres mesures standardisées ont donc été proposées pour permettre la comparaison du DL entre paires de locus différents. Les deux mesures standardisées les plus utilisées aujourd'hui sont le D' (Lewontin, 1964) ou le r (ou sa valeur au carré r^2) (Hill et Robertson, 1968). Si on considère deux locus *A* et *B*, chacun ayant deux allèles 1 et 2, avec p_1 et p_2 les fréquences des allèles 1 et 2 au locus *A* et q_1 et q_2 les fréquences des allèles 1 et 2 au locus *B*, D' et r s'écrivent alors comme suit :

$$D' = \frac{D}{\min(p_1q_2, p_2q_1)} \text{ si } D>0 \qquad \text{ou} \qquad D' = \frac{D}{\min(p_1q_1, p_2q_2)} \text{ si } D<0$$

le dénominateur traduit la valeur du produit des fréquences des allèles minoritaires aux deux locus *A* et *B*,

et $$r^2 = \frac{D^2}{(p_1 p_2 q_1 q_2)}$$.

Ces deux mesures du DL sont bornées entre 0 et 1. Cependant, $|D'|$ prendra la valeur 1 si seuls 2 ou 3 des 4 haplotypes possibles formés par l'association des allèles 1 et 2 aux locus

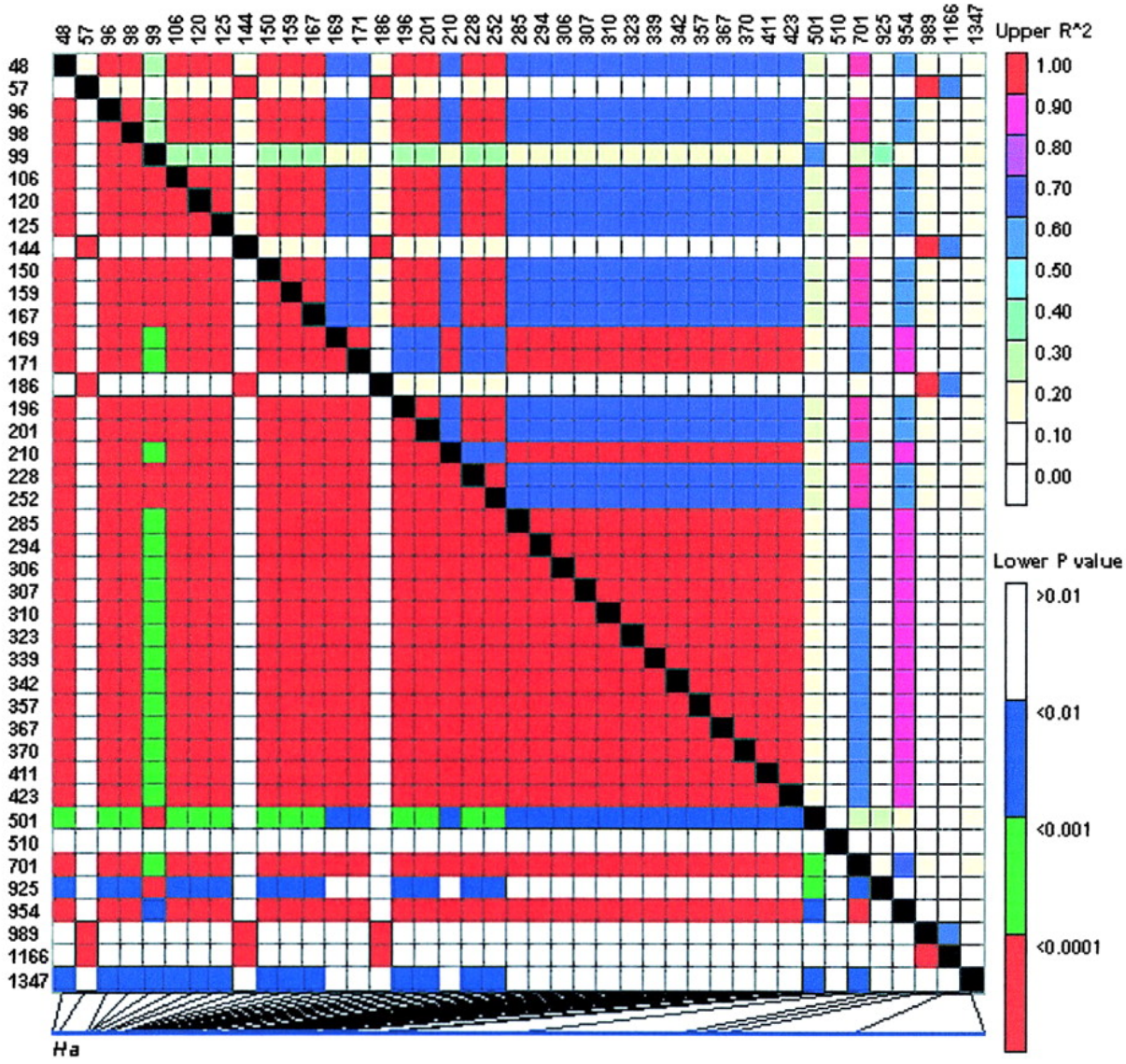


Figure 20 : exemple d'une représentation matricielle du DL obtenue par l'utilisation du logiciel TASSEL.

A et *B* (A_1B_1, A_1B_2, A_2B_1, A_2B_2) sont détectés dans la population (Figure 19). La valeur de 1 ne pourra-t-être obtenue pour r^2 que si les fréquences des allèles mineurs sont les mêmes aux deux locus et que donc, seuls 2 des 4 haplotypes possibles sont présents dans la population. On parle alors de déséquilibre complet. Ces deux métriques standardisées du DL ne sont pas équivalentes. r^2 résume l'histoire de l'échantillon en terme de mutation et de recombinaisons alors que D' mesure seulement l'histoire des recombinaisons. Ces mesures sont toutes les deux peu performantes pour estimer le DL dans de petits échantillons cependant D' est plus affecté que r^2 par l'utilisation d'un échantillon de petite taille et l'étude de polymorphismes en faible fréquence. En effet, dans ces deux cas, la probabilité d'observer les 4 combinaisons haplotypiques est faible et D' peut prendre une valeur de 1 même en l'absence de liaison physique entre les deux polymorphismes. Ainsi, D' sera préféré pour étudier les niveaux de recombinaison et r^2 pour étudier la corrélation entre tous les sites polymorphes et estimer la persistance du DL en fonction de la distance physique entre sites (décroissance du DL).

Il y a deux manières classiques de représenter le DL : i/ les valeurs de DL en fonction du nombre de pb entre sites polymorphes qui traduit alors la décroissance du DL en fonction d'une distance physique ou génétique, et ii/ la matrice des DL entre polymorphismes qui renseigne sur sa structure (arrangement linéaire) au sein de la région étudiée (Figure 20).

5.2.3. Structure du DL en populations

5.2.3.1. Facteurs influençant le DL

Comme nous l'avons vu plus tôt, le DL dépend des fréquences des allèles aux locus considérés et de la recombinaison entre ces locus. Même si la mutation et la recombinaison sont les deux facteurs qui influent le plus sur le DL, la plupart des processus observés en génétique des populations (mutation, recombinaison, dérive, migration, sélection) vont avoir un impact sur sa structure (Rafalski et Morgante, 2004). Les patrons de DL observés à l'échelle du génome d'un organisme en population naturelle sont la résultante complexe de tous ces processus qui peuvent agir de manière simultanée avec plus ou moins d'intensité selon le type de population étudiée.

Liaison et DL sont des notions qui sont souvent confondues en génétique, pourtant elles traduisent des phénomènes distincts. En génétique, la liaison fait référence à la coségrégation d'allèles à deux locus polymorphes. Cette coségrégation est due à la proximité physique des deux locus sur le même chromosome. Cette proximité physique rend un événement de

recombinaison entre les deux sites d'autant plus improbable qu'ils sont proches. Ainsi, un DL fort est communément rencontré entre allèles de locus physiquement liés pour lesquels la recombinaison n'a pas eu le temps de casser les associations créées par mutation. Cependant, un fort DL entre sites polymorphes ne traduit pas forcément une liaison physique étroite. D'autres phénomènes relatifs à l'histoire et aux traits de vie d'une population vont influer sur la mise en place et la cassure du DL. Ainsi, si deux mutations éloignées (même sur des chromosomes différents) sont soumises aux mêmes effets de la sélection (par exemple dans le cas de gènes en interaction épistatique), ils peuvent être en fort DL.

5.2.3.2.Etendue et structure du DL : l'apport des espèces modèles

L'étendue du DL est une mesure qui permet d'évaluer la persistance du DL en fonction de la distance physique ou génétique entre paires de sites polymorphes d'une ou plusieurs régions du génome (Remington *et al.*, 2001). Cette mesure permet d'obtenir une vision globale du niveau de DL observé chez une espèce ou dans une population. Elle est un indicateur de la diversité haplotypique moyenne observée le long du génome et peut traduire des différences relatives à l'histoire de vie des espèces et des populations au sein de ces espèces.

Sur la base des données obtenues chez plusieurs espèces, des différences très nettes apparaissent dans les estimations de l'étendue du DL. Il peut être maintenu sur des distances de plusieurs centaines de kb ou au contraire n'être significatif qu'à de très faibles distances (inférieures à 1 kb). Au sein des espèces, d'importantes variations peuvent être également observées selon le type de population étudiée.

Chez l'Homme, les données collectées indiquent que l'étendue du DL se situe entre quelques dizaines à quelques centaines de kb (Abecasis *et al.*, 2001 ; Reich *et al.*, 2001 ; Stephens *et al.*, 2001) même si il existe une variabilité importante des niveaux de DL observés entre différentes régions du génome. Reich *et al.* (2001) ont étudié le maintien du DL au sein de 19 régions du génome pour 2 populations américaines de descendants d'Europe du Nord et une population du Nigeria. Ils ont mis en évidence une étendue du DL moins importante (d'environ un facteur 10) au sein de la population africaine (de 60 kb dans les populations de descendants d'Europe du Nord à 5 kb dans la population des Yorubas du Nigeria pour une valeur de D' diminuée de moitié). Ils ont également mis en évidence que la plupart des haplotypes identifiés dans la population africaine étaient contenus au sein des haplotypes plus longs détectés dans la population de descendants d'Europe du Nord. Les

auteurs ont alors suggéré que ces patrons de décroissance du DL étaient compatibles avec des effets démographiques de type goulots d'étranglements (« bottlenecks ») chez les descendants d'Europe du Nord et expansions de population dans la population nigériane. Ils suggèrent également une origine africaine commune pour toutes ces populations.

Chez le maïs, la première étude publiée a été basée sur les données de variabilité de 21 régions le long du chromosome 1 (Tenaillon *et al.*, 2001). Dans cette étude, les auteurs ont étudié la décroissance du DL en fonction de la distance sur 25 individus représentant un échantillon de l'ensemble de l'espèce et un sous échantillon de lignées élites américaines. Cette étude a montré que le DL entre sites polymorphes était généralement maintenu sur de très courtes distances avec des valeurs de r^2 égales à 0,20 après 400 pb. Les auteurs ont montré que le DL était maintenu sur de plus longues distances (de l'ordre de 1 kb) au sein du sous échantillon de lignées élites américaines. Ces données sont cohérentes avec le fait que l'obtention de ces lignées repose sur le maintien d'une sélection artificielle appliquée sur un échantillon de population naturelle qui constitue la base des programmes d'amélioration. Ces lignées ont donc une base génétique étroite et représentent une petite proportion de la diversité génétique et haplotypique de l'espèce (moins de 1% selon Hoisington *et al.*, 1999).

Dans le cas d'*Arabidopsis*, les données recueillies montrent une décroissance plus lente du DL que chez le maïs. Nordborg *et al.* (2002) ont étudié le DL au sein d'une région de 250 kb localisée autour du locus *FR1* impliqué dans la phénologie de la floraison. Ils ont séquencé 13 régions de 0,5 à 1 kb dans un échantillon de 20 individus représentatifs d'une partie de la diversité de l'espèce. Sur la base de ces données, les auteurs ont mis en évidence le maintien du DL entre site polymorphes sur une distances d'environ 1 cM ce qui représente environ 250 kb. Ils ont ensuite étudié la décroissance du DL dans une autre région du génome autour du locus de résistance *RPM1* au sein de différentes populations locales du Michigan. Ils ont mis en évidence au sein de cette région et dans ces populations des niveaux de DL plus importants avec un maintien du DL sur des distances génétiques de 50 à 100 cM. Les auteurs indiquent que le locus *RPM1* pourrait être impliqué dans l'adaptation de l'espèce à son environnement ce qui pourrait expliquer un plus fort niveau de DL. Ces données sont compatibles avec le système de reproduction de l'espèce à 99 % autogame caractérisée par une distribution en patch de populations fortement apparentées (Abbott and Gomes, 1989 ; .Bergelson *et al.*, 1998 ; Todokoro *et al.*, 1995). Plus récemment, Kim *et al.* (2007) ont étudié le DL à l'échelle du génome entier en comparant deux types de jeux de données. Ces deux jeux de données consistaient en plus de 1000 séquences représentant environ 1% du génome obtenues sur 95

individus représentatifs de l'espèce et 341 602 SNP génotypés au sein d'un échantillon de 19 individus sélectionnés parmi les 95. Les auteurs ont montré que le DL était maintenu sur une distance de 10 kb environ. L'étendue du DL rapportée ici est proche de celle mise en évidence chez l'Homme et peut paraitre surprenante. Les auteurs expliquent que la taille efficace très importante de l'espèce peut expliquer ce niveau relativement faible de DL observé chez *Arabidopsis thaliana* compte tenu de son régime de reproduction largement autogame.

Ces données indiquent un fort impact des paramètres populationnels sur la structure du DL. Les données obtenues chez les espèces modèles les plus étudiées indiquent que la structure du DL n'est pas linéaire mais plutôt irrégulière avec alternance des zones de fort et faible DL. Ces blocs de DL sont séparés par des zones de faible DL qualifiés de points chauds de recombinaison (« recombination hot spots »). L'existence d'une structure en blocs de DL a d'abord été décrite chez l'Homme (Huttley *et al.*, 1999 ; Daly *et al.*, 2001 ; Jeffreys *et al.*, 2002 ; Gabriel *et al.*, 2002 ; Kauppi *et al.*, 2003, 2004 ; McVean *et al.*, 2004). Entre blocs de DL et « hot spots » de recombinaison, le taux de recombinaison peut varier d'un facteur 1000 (Mc Vean *et al.*, 2004). Certaines études montrent l'importance de la biologie de l'espèce mais également des traits de vie des populations dans la mise en place de ce type de structure (Wang *et al.*, 2002 ; Zhang *et al.*, 2003). Chez les plantes, l'organisation du génome en bloc n'est pas bien documentée. Cependant, certaines études révèlent des variations importantes du taux de recombinaison méiotique au sein du génome de plusieurs espèces comme le blé, le maïs ou Arabidopsis (Gill *et al.*, 1996 ; Xu *et al.*, 1995 ; Dooner *et al.*, 1997 ; Okagaki *et al.*, 1997 ; Eggleston *et al.*, 1995). On peut donc faire l'hypothèse qu'une structure en mosaïque existe également chez les plantes.

5.2.3.3.Etendue du DL chez les arbres forestier

Peu d'études de DL ont été réalisées chez les arbres forestiers. Les données recueillies proviennent, pour la plupart, d'études menées sur quelques gènes candidats pour quelques génotypes représentatifs de populations ou d'espèces majoritairement commerciales. Ces données rapportent globalement des niveaux de DL assez faibles, de l'ordre de ceux mis en évidence chez le maïs. Ces niveaux de DL bas sont cohérents avec les régimes de reproduction et l'histoire des populations de ces espèces peu domestiquées principalement allogames avec des tailles de populations importantes. Chez *Pinus taeda*, le DL (r^2) devient inférieur à 0,2 au-delà de 1500 pb à 2000 pb (synthèses bibliographiques par Neale et Savolainen, 2004 ; Brown *et al.*, 2004), chez *Pinus sylvestris* le DL mesuré au sein du gène

pal1 est très faible au-delà de quelques centaines de paires de bases (Dvornyk *et al.*, 2002), chez *Picea abies*, le DL n'est pas maintenu au-delà de 100 à 200 pb (Rafalski et Morgante, 2004, Heuertz *et al.*, 2006), chez *Populus tremula* il décroit également rapidement en quelques centaines de paires de bases (Ingvarsson *et al.*, 2005).

5.2.4. DL et identification des polymorphismes causaux

Dans une population d'association, la puissance de détection des polymorphismes causaux (responsables de la variation du caractère) dépend fortement de l'étendue du DL. Lorsque le DL est de l'ordre de plusieurs kb, il est alors possible de mettre en évidence des liens entre la variabilité des gènes et la variation des caractères en considérant un nombre de SNP limités. L'analyse QTL exploite ces forts DL présents dans des populations en ségrégation. Un des inconvénients du DL fort est son manque de résolution pour identifier le polymorphisme causal. Il est impossible de discriminer l'effet d'une mutation causale de celui des marqueurs qui lui sont liés (information qui n'est pas forcement nécessaire pour le sélectionneur). Une faible étendue du DL, de l'ordre de quelques centaines de bases, donne accès à une meilleure résolution, mais elle requiert un nombre plus important de marqueurs pour détecter ceux qui sont responsables de la variation des caractères phénotypiques. Une faible étendue de DL implique donc des efforts plus importants en termes de génotypage pour la réalisation d'études d'associations et la recherche des polymorphismes causaux.

5.3. Mise en évidence de la variabilité « fonctionnelle »

5.3.1. Association directe ou indirecte

Dans une situation idéale, il faudrait pouvoir tester l'ensemble de la variabilité d'une espèce (tous les sites polymorphes : SNP, short indel et indel) pour être certain de compter parmi eux le ou les variants causaux recherchés. En pratique, il n'est pas encore possible de réaliser cette analyse exhaustive pour des raisons de coûts et de quantité de données nécessaires. Aujourd'hui, il est nécessaire de choisir un ensemble limité de sites polymorphes à tester qui permette de trouver le meilleur compromis entre coût et efficacité. Une bonne connaissance du niveau de DL chez l'organisme étudié permet d'optimiser ce choix.

L'objectif d'une étude d'association est d'identifier les variants causaux de manière directe (le polymorphisme observé correspond à la variabilité causale ou fonctionnelle, responsable d'une partie de la variation du phénotypique étudié) ou indirecte (il est physiquement lié et/ou en fort DL avec la variabilité fonctionnelle). Dans le cadre d'une

association indirecte, si un variant causal contribue à une fraction de la variation totale du caractère étudié h^2_q, un variant en déséquilibre de liaison r^2 avec lui ne pourra être détecté qu'avec un effet $r^2 \times h^2_q$ (Zhu *et al.*, 2008).

5.3.2. Deux grands types d'approches

5.3.2.1.L'approche pan génomique

L'approche pan génomique, aussi connue sous le nom de « genome-wide associaton study » (GWAS), peut être qualifiée d'approche sans *a priori*. Comme les approches de cartographie de QTL, elles ne ciblent aucune région particulière et considèrent un ensemble de marqueurs répartis sur l'ensemble du génome. L'étendue du DL étant plus faible et les niveaux de diversité génétique plus fort dans les populations d'associations par rapport aux pedigrees de détection de QTL, des densités de marqueurs plus importantes sont nécessaires dans les approches d'études d'associations pour caractériser la diversité haplotypique présente au sein de la population d'étude. En pratique, ces approches nécessitent le développement de ressources importantes en termes de marqueurs SNP (généralement facilitée par la disposition d'un génome de référence) et ne sont réalisables que pour des espèces présentant des niveaux de DL relativement élevés.

Chez l'Homme, des études prédisent l'existence de 5 à 15 millions de SNP dispersés le long du génome avec des fréquences d'allèles minoritaires (MAF) supérieures à 1% (Carlson *et al.*, 2003 ; Wang *et al.*, 2005 ; Kruglyak *et al.*, 2008). Cependant, étant donnée l'étendue moyenne du DL le long du génome dans les différentes populations humaines, 500000 à 1000000 de SNP sont nécessaires pour assurer une bonne couverture du génome pour la réalisation d'une approche pan-génomique. Un tel nombre de marqueurs devrait permettre de détecter les locus impliqués dans de maladies communes avec des effets modérés (Kruglyak *et al.*, 2008). Les résultats du projet HapMap permettent d'accéder aujourd'hui à plus 3 millions de SNP avec des MAF supérieures à 5% (The International HapMap Consortium, Frazer *et al.*, 2007). Etant donné que l'étendue du DL est hétérogène le long du génome humain, des méthodes ont été proposées pour sélectionner un ensemble de marqueurs SNP qui décrivent au mieux la diversité haplotypique présente au sein du génome. On parle de sélection de « tag SNP » ou « haplotype tagging ». La sélection appliquée par ces méthodes est basée sur la MAF des SNP (sélection des SNP les plus fréquents) et sur le niveau de DL entre les SNP. Généralement, un seuil de $r^2=0.8$ est défini pour estimer des blocs de DL au sein desquels seuls quelques SNP sont nécessaires pour représenter la diversité haplotypique

présente. Ces approches nécessitent une bonne connaissance de la structure du DL à l'échelle des populations (Hinds *et al.*, 2005 ; The International HapMap Consortium, Altshuler *et al.*, 2005). Plusieurs études incluant plusieurs dizaines de milliers de marqueurs SNP ont été réalisées chez l'Homme et ont permis d'identifier des variants génétiques impliqués dans la susceptibilité à des maladies (synthèse bibliographique par Kingsmore *et al.*, 2008).

De telles études sont difficilement réalisables chez les plantes allogames présentant des niveaux de DL très faibles et pour lesquelles les ressources en termes de nombre de SNP identifiés et connaissance des niveaux de diversité nucléotidique en population restent encore embryonnaires. L'approche GWAS a pour le moment été menée chez Arabidopsis et le blé. Ces deux espèces présentent des ressources génomiques importantes en termes de marqueurs SNP et des niveaux de DL en population suffisamment importants (Zhu *et al.*, 2008). Chez Arabidopsis, un catalogue de la diversité génétique contenant plus d'un million de SNP (1/166 pb en moyenne) est disponible et les premières approches d'études d'association pan-génomiques commencent à voir le jour (Aranzana *et al.*, 2005 ; Rostoks *et al.*, 2006 ; Atwell *et al.*, 2010). Des projets ambitieux d'étude d'association pan-génomique chez les plantes sont en cours de montage et devraient aboutir dans les cinq prochaines années.

5.3.2.2.L'approche « gènes candidats »

L'approche « gènes candidats » est une approche avec « *a priori* » basée sur la sélection de gènes cibles. Ces gènes sont sélectionnés en fonction de leur potentielle implication dans la variation des caractères étudiés sur la base de données obtenues dans le cadre d'approches i/ de génétique directe (colocalisation gènes-QTL, expression dans des conditions particulières), ii/ biochimiques (implication dans une voie métabolique), et iii/ physiologiques (phénotypes particuliers observés chez des mutants) obtenues chez des plantes modèles ou non (Hattersley et McCarthy, 2005). Dans ce cas, la variabilité nucléotidique des gènes candidats est décrite sur tout ou portion du gène. Cela nécessite donc la mise en évidence des polymorphismes par séquençage au sein d'un panel de détection et le génotypage des SNP au sein de la population d'association. Des panels de détection de petite taille (<20 individus) sont généralement suffisants pour identifier les SNP les plus fréquents alors que des panels plus importants (>50 individus) sont nécessaires pour identifier les variants les plus rares.

L'objectif de ces approches est généralement d'identifier les variants causaux de la variation du caractère étudié. Dès lors, les variants susceptibles de causer des changements dans la structure de la protéine (non synonymes) ou de l'expression des gènes (sites de

régulation) deviennent des cibles privilégiées pour le génotypage en populations d'association (Tabor *et al.*, 2002). Ces sites n'étant pas toujours connus, des approches de sélection de « tag SNP » sont également employées, comme pour les approches de GWAS et la détection de « quantitative trait nucleotides » (QTN) devient alors indirecte. L'approche « gènes candidats » est privilégiée dans le cadre de voies de biosynthèses bien caractérisées, chez des espèces dont les ressources génomiques sont encore peu développées et dont l'étendue du DL est trop faible pour permettre des approches pan génomiques. C'est le cas de la majorité des espèces de plantes allogames et notamment des arbres forestiers. Cette approche a montré son efficacité chez différentes espèces de plantes et les principaux résultats obtenus par la mise en place de ce type d'approches ont été synthétisés par Zhu *et al.* (2008). C'est cette approche que j'ai mis en œuvre dans le cadre de ma thèse.

5.3.3. Puissance d'une étude d'association

La puissance d'une étude d'association (probabilité de détecter une association vraie) dépend de plusieurs facteurs tels que : la dimension de l'effet (fort ou faible) porté par le marqueur, la fréquence des allèles au marqueur dans la population, la taille de la population utilisée (population d'association), le type de population utilisée (structure, apparentement, …), l'héritabilité du caractère, la méthode statistique employée, le seuil de significativité fixé pour le test statistique. Généralement, la taille de la population, le type de population, le test statistique employé et le seuil de significativité du test sont fixés par l'expérimentateur (Hattersley et McCarthy, 2005). En fonction de ces paramètres fixés, il sera possible de détecter des QTN plus ou moins fréquents et à effets plus ou moins forts dans la population d'association. Plusieurs études comparent ainsi, par simulation, la puissance de différentes méthodes statistiques pour détecter des polymorphismes de différentes fréquences avec des effets de différentes dimensions dans des types de populations d'associations spécifiques (Long *et al.*, 1999 ; Amin *et al.*, 2007 ; Schaid, 2005 ; Thomas *et al.*, 2004).

Si le nombre et la sélection des marqueurs génétiques et le modèle statistique utilisé prennent une place importante dans le succès d'une étude d'association, la taille de l'échantillon utilisé est un facteur déterminant. La puissance d'un test d'association est proportionnelle à la taille de l'échantillon utilisé et un petit échantillon sera très mal adapté pour la détection de QTN peu fréquents et à effets faibles (Hattersley et McCarthy, 2005 ; Zhu *et al.*, 2008). En pratique, il est parfois difficile de disposer de populations idéales avec de grands effectifs. La plupart des études réalisées chez les plantes impliquent des population

diverses (lignées, clones, populations naturelles, cultivars, …) avec des effectifs faibles à modérés (de l'ordre de 30 à 600 individus selon Zhu *et al.*, 2008). Pour le moment, les études d'association n'ont le plus souvent permis de mettre en évidence que des QTN à effet faibles (comptant pour la majorité pour moins de 5% de la variation des caractères) malgré les fortes héritabilités estimées pour certains caractères. Ceci suggère pour la plupart des caractères quantitatifs complexes l'intervention de multiples locus à effet faible ou de locus à effets forts mais en trop faible fréquences pour être identifiés de manière efficace dans des populations de petite taille (Manolio *et al.*, 2009).

5.3.4. Populations et principales méthodes statistiques utilisées pour l'analyse des caractères en génétique d'association

5.3.4.1.Populations idéales :

Dans une situation idéale, les populations d'associations sont constituées d'un ensemble d'individus non apparentés, en nombre suffisant pour pouvoir détecter des QTN d'intérêt (au moins ceux à forte fréquence et effets forts, forte fréquence et effets faibles, faible fréquence et effets forts). Dans ce cas, une analyse de variance (ANOVA) est un test statistique adapté à l'étude de l'effet des marqueurs sur la variation des caractères quantitatifs. Ce type de population a été utilisé chez l'eucalyptus par Thumma *et al.* (2005) et a permis de mettre en évidence, par une approche gène candidat, 2 SNP associés avec la variation de l'angle des microfibrilles de celluloses. Ces variants expliquaient 4,6 % de la variance phénotypique du caractère.

5.3.4.2.Structure et apparentement : comment les prendre en compte

Dans la pratique, ces populations d'association idéales ne sont pas toujours disponibles et il est plus fréquent d'avoir affaire à des populations structurées en sous populations, des populations contenant des individus apparentés ou encore des populations structurées contenant des individus apparentés (Yu *et al.*, 2006) comme dans un programme d'amélioration géré sur plusieurs générations.

La présence d'une structure génétique dans la population d'association peut causer de fausses associations (c'est-à-dire la mise en évidence d'une liaison entre un marqueur génétique et le caractère d'intérêt sans qu'il n'y ait un lien de causalité) autrement dit augmenter le taux de faux positifs. Ce type de problème est rencontré lorsque le caractère d'intérêt est différencié entre les sous-populations. Les marqueurs qui sont en fréquence

importante au sein d'une sous-population, alors surreprésentés, pourront être associés avec le caractère d'intérêt (Ewens et Spielman, 1995 ; Pritchard et Rosenberg, 1999). Chez l'Homme, Knowler *et al.* (1988) ont montré que seule la structure de la population était en cause dans l'association entre un variant génétique et la sensibilité au diabète de type 2. La première approche proposée pour s'affranchir des effets de la structure génétique en population d'association était le « transmission/disequilibrium test » (TDT) (Spielman *et al.*, 1993). Cette méthode fait partie de la famille des tests d'associations basés sur l'étude familiale. Elle était initialement basée sur l'étude de trios parents descendants (constituant la population d'association) qui permet, par l'utilisation des génotypes de parents des individus affectés, de s'assurer que l'association entre marqueur et phénotype n'est détectée que si le marqueur est lié à un variant causal. Un grand nombre de tests statistiques basés sur ces structures familiales se sont ensuite développées, et permettent aujourd'hui de traiter des caractères quantitatifs (synthèse bibliographique par Laird et Lange, 2006). La mise en place de ce type de tests nécessite la mise à disposition de nombreuses familles indépendantes (non apparentées) et de petite taille (quelques descendants). Ce type de population n'est pas toujours disponible et implique un coût plus élevé pour la collecte des données (les données des génotypes parentaux étant nécessaires).

Pritchard *et al.* (2000 a, b) ont proposé de déterminer cette structure (Q) dans les populations classique sur la base d'informations de marqueurs génétiques neutres et de la prendre en compte dans les tests statistiques. Grâce aux marqueurs génétiques, le nombre de sous-populations qui composent la population d'association est d'abord estimé, puis la probabilité d'appartenance des individus à ces sous-populations est déterminée pour être incluse dans le modèle statistique. Chez les plantes, Thornsberry *et al.* (2001) ont été les premiers à utiliser cette méthode pour la mise en évidence de l'effet de polymorphismes du gène *dwarf8* dans la précocité de floraison chez le maïs. Plusieurs méthodes permettent aujourd'hui de prendre en compte la structure d'une population dans les études d'association. On parle de « structured association » (SA) pour désigner ces méthodes (Pritchard *et al.*, 2000 ; Camus-Kulandaivelu *et al.*, 2006 ; Price *et al.*, 2006 ; Yu *et al.*, 2006 ; Zhao *et al.*, 2007 ; Stich *et al.*, 2008 ; Weber *et al.*, 2008). Même si ces études permettent de réduire le nombre de faux positifs dans les études d'association, elles ne permettent pas de les éliminer complètement comme l'ont montré Aranzana *et al.* (2005) et Zhao *et al.* (2007) chez *Arabidopsis*.

La présence d'apparentement entre les individus de la population d'association peut également augmenter le taux de faux positifs et entraîner une perte de puissance statistique (Yu *et al.*, 2006). Ce type de population peut être rencontré dans le cadre de populations d'amélioration chez les animaux et les plantes. Yu *et al.* (2006) ont proposé de prendre en compte cet apparentement par l'inclusion d'une matrice d'apparentement (matrice K, pour kinship matrix) dans un modèle mixte. Ce modèle explique la variation d'un caractère par sa moyenne, l'effet des QTL non testés, l'effet du marqueur SNP testé et un effet résiduel. La matrice K permet de définir le degré de covariance génétique entre les individus qui composent la population d'association. Elle peut être déterminée par des marqueurs génétiques ou sur la base des informations connues d'apparentement, lorsque celles-ci sont disponibles. Ces méthodes permettent de prendre en compte des niveaux d'apparentement différents entre les individus des populations d'association. Il existe également des méthodes basées sur l'utilisation du modèle linéaire prenant en compte les effets parentaux. Ces approches sont exposées et comparées par Sahana *et al.* (2010).

Enfin, dans le cas de populations structurées contenant des individus apparentés, il est possible de combiner les approches en incluant dans le modèle à la fois la probabilité d'appartenance des individus à une sous-population de la population d'association (Q) et l'apparentement entre les individus qui composent la population (K) (Yu *et al.*, 2006).

5.3.4.3. Tests d'association et inférences statistiques

L'objectif des études d'associations est de déterminer, parmi un ensemble de marqueurs génétiques testés, lesquels sont liés à la variation d'un caractère. Le problème est de distinguer parmi ces marqueurs quels sont ceux qui présentent une grande probabilité d'avoir un effet sur la variation du caractère phénotypique étudié de manière à prendre la meilleure décision quant à leur implication dans cette variation.

Les approches fréquentistes, incluant les modèles exposés ci-dessus, sont basées sur un test d'hypothèse. Ils permettent de rejeter une hypothèse H_0 de « non association entre le marqueur étudié et le phénotype ». Cette hypothèse est rejetée avec un risque α (aussi appelé erreur de type I) qui est la probabilité de faire une erreur en rejetant H_0. Généralement, on accepte un risque α de 5% de se tromper en rejetant H_0 alors qu'elle est vraie (faux positifs). Si la p-value du test (probabilité que la valeur du test statistique soit supérieure ou égale à la valeur du test sous l'hypothèse H_0 avec un risque α) est inférieure à ce seuil, alors l'hypothèse H_0 est rejetée avec un risque de 5%. Pour être valide et traduire la meilleure évidence d'une

non-indépendance entre marqueur et phénotype, le seuil α doit être déterminé expérimentalement. Si le seuil de 5% parait être un bon seuil pour le risque d'erreur commis sur l'étude d'un seul marqueur, lorsque 100 marqueurs sont testés de manière indépendante, 5 d'entre eux peuvent présenter des p-values inférieures au seuil de 5% simplement par l'effet du hasard. On distingue donc deux types de risques α : α[PT] (par test) et α[PF] (par famille de tests) aussi appelés « testwise » et « experimentwise » alpha respectivement.

Il existe des méthodes qui mettent en relation α[PT] et α[PF] permettant ainsi de déterminer le seuil α[PF] qui doit être considéré pour maintenir un risque d'erreur α[PT] constant dans le cas de tests multiples (lorsque plusieurs marqueurs sont analysés indépendamment dans la même population et pour le même caractère par exemple). La plus simple de ces corrections est la correction de Bonferroni (Sidàk, 1967 ; voir aussi Abdi, 2007). Bien qu'elle soit couramment utilisée, elle est souvent jugée trop conservative surtout quand le nombre de tests est important et quand ces tests ne sont pas indépendants (comme dans le cas de marqueurs génétiques en DL). Cette correction résulte dans la mise en évidence de faux négatifs (des associations probables jugées improbables). D'autres méthodes comme le FDR (False Discovery Rate) (Benjamini et Hochberg, 1995) ou des tests basés sur les permutations ou le ré-échantillonnage peuvent être utilisées. Cependant, le FDR est mal adapté dans le cas de tests non indépendants et bien que les tests de permutations donnent de bons résultats, ils nécessitent des temps de calcul supérieurs aux autres méthodes et peuvent être impraticables dans le cas des approches pan-génomiques par exemple.

Parallèlement aux approches fréquentistes, le développement des outils de calcul rendent aujourd'hui l'utilisation des méthodes Bayesiennes possibles pour les études d'association. Les approches mises en place proposent de déterminer, *a priori*, une probabilité de l'hypothèse H_1 (π) pour chaque marqueur testé (le marqueur génétique et le phénotype sont associés) sur la base de données telles que la fréquence du SNP dans la population, son impact sur la structure primaire de la protéine (SNP synonyme ou non synonyme), sa conservation entre les espèces, son appartenance ou sa relative proximité par rapport à un gène de fonction connue. Pour chaque SNP, la probabilité des données sous l'hypothèse H_1 est ensuite comparée avec la probabilité des données sous l'hypothèse H_0 ($1-\pi$) par l'utilisation du « Bayes factor » (BF) qui permet donc de comparer la probabilité de deux modèles : le marqueur est associé (H_1) ou non (H_0). Plus le BF est important, plus les données supportent l'hypothèse H_1 par rapport à H_0. Le BF est similaire à un rapport de vraisemblance mais compare des modèles à la place de paramètres dans un modèle. Le BF et la probabilité *a*

priori de H_1 (π) permettent ensuite de déterminer, dans un cadre Bayesien, une probabilité *a posteriori* pour l'association (PPA). Ces méthodes sont aujourd'hui accessibles pour des approches gènes candidats ou des approches pan génomiques et sont discutées dans une synthèse bibliographique récente (Stephens et Balding, 2009). Même si elles nécessitent des temps de calcul plus importants que les méthodes fréquentistes, les méthodes Bayesiennes sont reconnues comme plus puissantes et tendent à se développer de plus en plus.

5.3.4.4.D'autres méthodes d'analyse

Les méthodes citées ci-dessus ont souvent été utilisées dans le cadre d'approches dites marqueur par marqueur. Ces approches testent indépendamment l'effet de chaque marqueur sur la variation du caractère dans la population d'association. L'accès à la phase de liaison entre les allèles de plusieurs marqueurs SNP permet de constituer les haplotypes (marqueurs multialléliques obtenus par la combinaison des allèles de plusieurs SNP) et de tester l'effet de ces combinaisons de marqueurs SNP sur le caractère phénotypique étudié. Les haplotypes étant plus informatifs que les marqueurs SNP quant à l'histoire des individus constituant les populations d'association, ces approches sont considérées comme plus puissantes que les approches marqueur par marqueur, bien que la littérature soit divisée par rapport à cette question (Nielsen et Zaykin, 2001). Cependant, tester l'effet de tous les allèles pour une combinaison de marqueurs nécessite des effectifs d'autant plus grands que le nombre de marqueurs en combinaison est grand et que le DL entre ces marqueurs est faible. C'est la principale limite de ces approches qui ont été développées pour différents types de populations (individus non apparentés, populations structurées et/ou contenant des individus apparentés). Les combinaisons de marqueurs utilisées peuvent être sélectionnées en fonction de leur position sur la séquence (fenêtres déroulantes), de manière empirique, ou de manière plus complexe en fonction de l'histoire évolutive révélée par la diversité nucléotidique (mutations et recombinaisons) des régions étudiées (Meuwissen et Goddard, 2000 et 2001 ; Liu *et al.*, 2001 ; Molitor *et al.*, 2003 ; Zöllner et Pritchard, 2005).

Plus récemment, des approches basées sur la régression multiple, ont été développées pour tester simultanément l'effet de plusieurs marqueurs sur la variation de caractères quantitatifs en population d'association. Ces méthodes sélectionnent, parmi un ensemble de marqueurs, ceux qui, en combinaison, expliquent au mieux la variation du caractère phénotypique étudié. Elles sont basées sur la comparaison de modèles incluant ou non les

différents marqueurs testés et permettent aujourd'hui d'inclure des données d'apparentement ou de structure (De los Campos *et al.*, 2009).

5.3.5. Les résultats chez les arbres forestiers

Chez les arbres forestiers, les études d'association ne sont qu'à leurs débuts. Cependant, elles ont déjà permis d'identifier, chez différentes espèces, des liens entre polymorphismes nucléotidiques de gènes et variation de caractères quantitatifs relatifs à la qualité du bois, la résistance au froid, le débourrement ou la discrimination isotopique du carbone (Eckert *et al.*, 2009 ; Ingvarsson *et al.*, 2008 ; Gonzalez Martinez *et al.*, 2007 ; Gonzalez Martinez *et al.*, 2008 ; Thumma *et al.*, 2005 ; Thumma *et al.*, 2009). Chez l'eucalyptus, deux études ont permis de mettre en évidence des liens entre des SNP du gène *CCR* (un gène de structure de la voie de biosynthèse des lignines) et un gène *COBRA-like* (impliqué dans le dépôt de la cellulose) et des propriétés du bois : l'angle des microfibrilles de cellulose dans la paroi S2 et la teneur en cellulose respectivement (Thumma *et al.*, 2005 ; Thumma *et al.*, 2009). Même si les effets associés à ces marqueurs son plutôt faibles (<5% de la variation du caractère), ces marqueurs constituent *a priori* de bons critères diagnostics pour la mise en place de la SAM chez l'*Eucalyptus*. Ces études, toutes basées sur l'approche « gène candidat », ont été réalisées dans différentes populations (structurées, apparentées, sans apparentement ni structure) et différentes méthodes statistiques ont été utilisées. Même si la plupart des locus identifiés sont porteurs d'effets faibles (inférieur à 5 % de la variance phénotypique expliquée par les marqueurs) les résultats obtenus indiquent que l'approche gènes candidats permet d'identifier des polymorphismes qui sous-tendent la variation de caractères quantitatifs d'intérêt.

6. Les objectifs de ce travail de thèse

Au vue des données qui ont été présentées plus haut, l'objectif de ce travail de thèse était donc dans un premier temps d'évaluer la variabilité de caractères de la chimie du bois relatifs aux lignines chez l'eucalyptus dans différents contextes de croisement intra et interspécifiques. Il a été décidé de se focaliser sur la quantité et la qualité des lignines, ces deux caractères présentant un intérêt démontré pour les industries de la pâte à papier et du charbon de bois. Il s'agissait d'estimer, pour la teneur en lignines et le rapport des monomères S et G, les composantes de la variance phénotypique et notamment la variance génétique additive dans différents dispositifs expérimenatux pour évaluer l'intérêt de la prise en compte

de ces caractères dans des programmes d'amélioration génétique. Pour cela nous disposions de trois plans de croisement impliquant les espèces *E. urophylla*, *E. camaldulensis* et *E. grandis*, trois espèces d'intérêt économique majeur pour les programmes d'amélioration génétique menés au Brésil par Vallourec et Mannesman et en République du Congo par le CRDPI. La variation des caractères relatifs aux lignines devait ensuite être mise en relation avec la variation des caractères plus classiquement étudiés et pris en compte dans les programmes de sélection chez ces espèces (croissance et densité du bois) notamment par l'étude des corrélations phénotypiques et génétiques additives.

Dans un deuxième temps, il s'agissait de révéler et décrire la variabilité de gènes candidats relatifs à la biosynthèse des lignines chez l'eucalyptus. L'objectif de ces travaux était d'une part d'obtenir des données sur la variabilité des gènes chez plusieurs espèces du genre, notamment en termes de diversité nucléotidique et haplotypique et de l'étendue du déséquilibre de liaison. Au démarrage de cette thèse très peu de données de variabilité moléculaire de gènes étaient disponibles chez l'eucalyptus et ces données devaient faire référence chez les espèces du genre. Elles seraient comparées avec celles mise en évidence pour d'autres espèces d'arbres forestiers et notamment celles du genre *Pinus* et *Populus*. D'autre part, ce travail devait permettre de disposer d'un catalogue de polymorphismes pour tester l'association entre variabilité moléculaire de gènes candidats de la lignification et variation de caractères quantitatifs en population, et plus particulièrement pour les caractères relatifs aux lignines.

Enfin, dans un troisième temps, il s'agissait de mettre en évidence des polymorphismes expliquant tout ou partie de la variation de caractères quantitatifs d'intérêt agronomique chez l'eucalyptus. La réalisation de tests d'association devait permettre de révéler les liens statistiques entre la variation des caractères de la croissance, de la densité du bois, de la quantité et de la qualité des lignines mise en évidence dans la première partie de ce travail et la variabilité moléculaire des gènes candidats de la biosynthèse des lignines décrite dans la deuxième partie. L'objectif était également d'évaluer la possibilité de détecter ces liens dans des plans de croisement déjà mis en place dans le cadre des programmes d'amélioration génétique des eucalyptus au Brésil et en République du Congo. Ces plans de croisement se distinguent des populations d'association classiquement utilisées pour ce type d'analyse par le fait qu'ils sont composés d'individus apparentés, pour lesquels les relations d'apparentement sont connues. Ils se rapprochent des pédigrées utilisés pour la cartographie génétique et la détection de QTL par le fait qu'ils sont composés de plusieurs familles

obtenues du croisement de quelques individus parentaux. Ces dispositifs peuvent être utilisés pour la détection de QTL dans un contexte multi-parental mais, dans notre cas, les données disponibles en termes de marqueurs génétiques ne nous permettaient pas de réaliser ce type d'analyse. Une approche de génétique d'association basée sur la variabilité nucléotidique de gènes candidats pour lesquels des colocalisations avec des QTLs de caractères relatifs à la croissance ou à des propriétés anatomiques et chimiques du bois nous est apparue comme la plus pertinente compte tenue des ressources disponibles chez ces espèces au démarrage de la thèse.

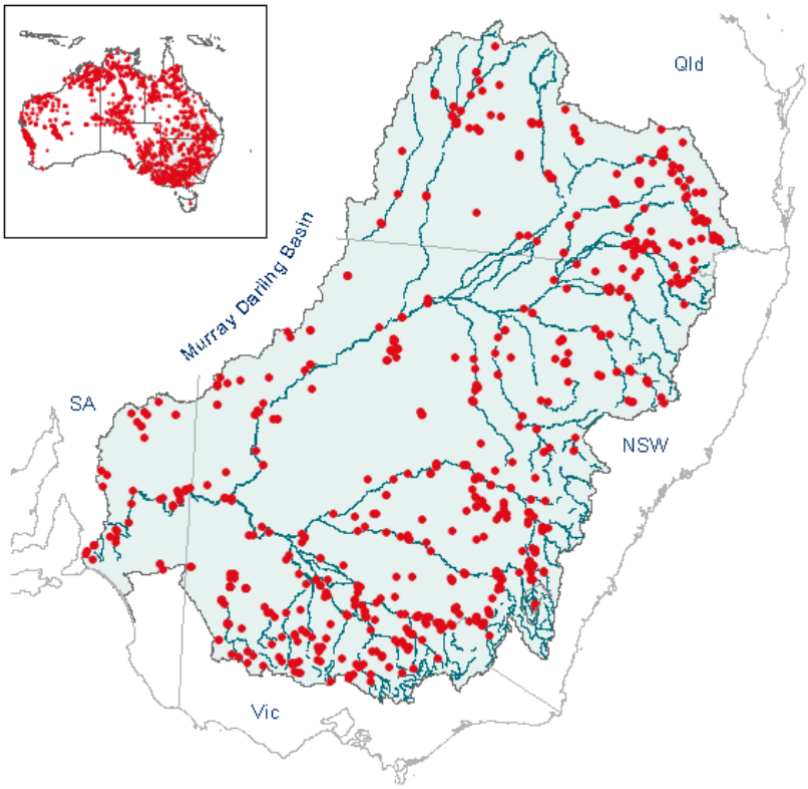


Figure 21 : aire de répartition naturelle d'*E. camaldulensis*. Les populations d'*E. camaldulensis* sont réparties sur la majeure partie de l'Australie excepté le sud ouest. Il se rencontre fréquemment à proximité des cours d'eau.

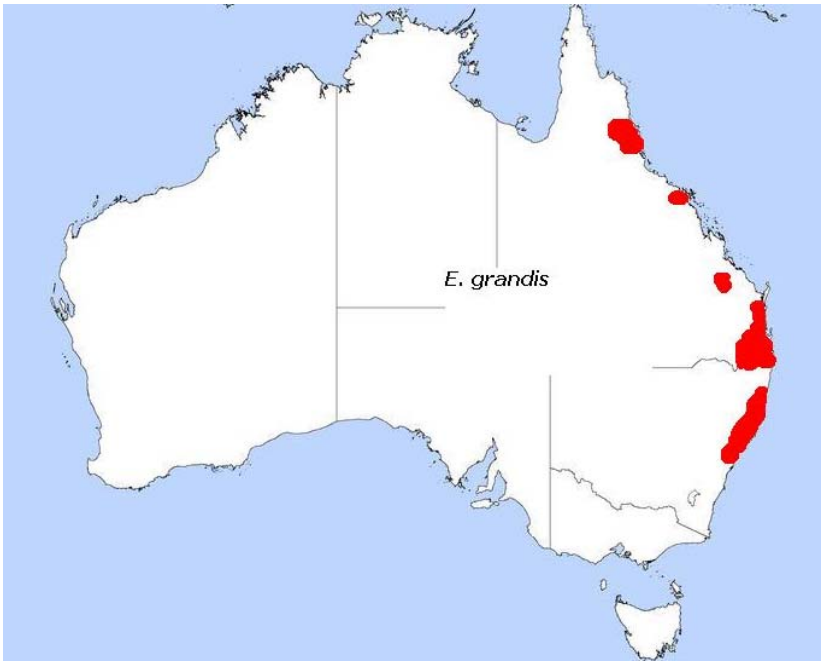


Figure 22 : aire de répartition naturelle d'*E. grandis*. Les populations d'*E. grandis* sont réparties à l'est de l'Australie.

Chapitre 2 : Matériel et méthodes

1. Les espèces étudiées

Dans le cadre des programmes d'amélioration génétique des *Eucalyptus* menés en République du Congo par le CRDPI (structure de recherche sous la triple tutelle du ministère congolais de la recherche, d'une société de reboisement EFC - pour Eucalyptus Fibres du Congo - et du CIRAD) et au Brésil par la société Vallourec et Mannesmann do Brasil (V&M do Brasil), 3 espèces sont majoritairement utilisées : *E. urophylla*, *E. camaldulensis* et *E. grandis*. Ces trois espèces d'*Eucalyptus* appartiennent au sous genre *Symphyomyrtus* mais à des sections différentes. *E. urophylla* et *E. grandis* appartiennent à la section des *Latoangulatae* et *E. camaldulensis* appartient à la section *Exsertaria* et constitue une super-espèce dans la sous série des *Camaldulensosae*.

Ces trois espèces ont des aires de répartition naturelles différentes. *E. camaldulensis* possède la plus large des aires de répartition parmi les espèces d'*Eucalyptus*. Elle est dispersée sur plus de 5 millions de km² dans la quasi-totalité de l'Australie (Figure 21) mais occupe une niche écologique assez étroite puisqu'elle se rencontre principalement à proximité des cours d'eau (Butcher *et al.*, 2002). Elle est capable de s'adapter à des conditions climatiques assez différentes supportant des niveaux moyens de précipitions annuelles de 110 mm à 1400 mm selon les régions dans lesquelles elle se développe et des températures annuelles moyennes de 10 à 25°C (Jovanovic et Booth, 2002). Grâce à des racines à haut niveau de conductivité, elle peut se développer dans les milieux arides et semi-arides. Les populations naturelles d'*E. grandis* sont réparties à l'est de l'Australie depuis le New South Wales jusqu'au nord du Queensland (Figure 22). *E. grandis* se rencontre généralement en bordure des forêts subtropicales humides et se développe dans des zones aux conditions climatiques assez variées avec des niveaux de précipitations annuelles moyennes de 750 mm à 3750 mm et des températures annuelles moyennes de 12 à 24°C (Jovanovic et Booth, 2002). L'espèce *E. urophylla* occupe une aire de répartition particulière. Elle se développe exclusivement sur l'archipel des îles de la sonde au sud de l'Indonésie. Elle est distribuée en petites populations disjointes réparties sur sept îles de l'archipel. Les populations les plus importantes sont rencontrées sur les îles de Timor et Wetar, les autres îles étant Adonara, Alor, Flores, Lomblen et Pantar (Figure 23). Elle est l'espèce la plus largement répartie en

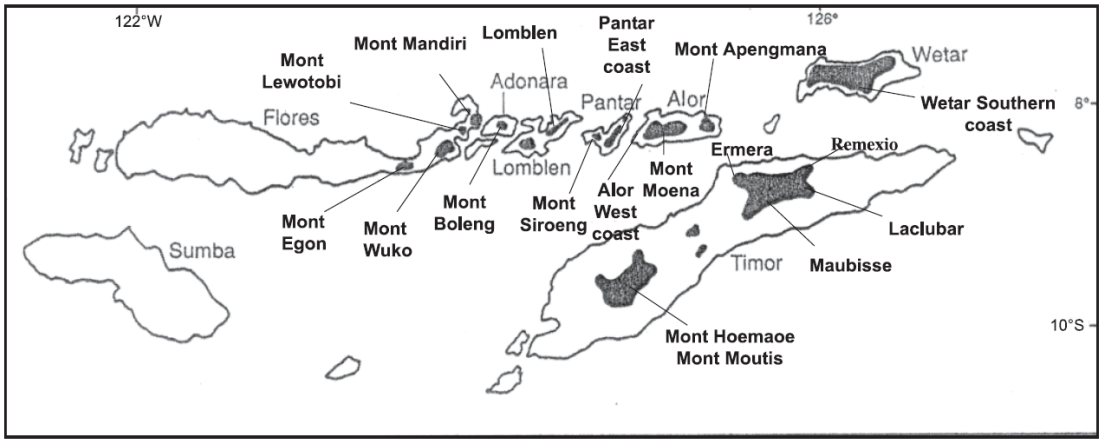


Figure 23: aire de répartition d'*E. urophylla* au sein de l'archipel des îles de la Sonde. Cette carte représente également les provenances qui ont été sélectionnées pour une étude de diversité génétiques réalisée par Tripiana *et al.* (2007).

termes d'altitude. Elle se développe entre 0 m et 3000 m au dessus de niveau de la mer avec un optimum de croissance atteint entre 500 m et 2200 m. Elle supporte également des conditions climatiques variables avec des durées de saison sèche pouvant s'étaler sur 2 à 8 mois selon les régions (Tripiana *et al.*, 2007 ; Payn *et al.*, 2008).

E. urophylla est une espèce majeure pour les programmes d'amélioration génétique des eucalyptus menés par le CRDPI et V&M do Brasil. Elle est particulièrement appréciée pour la création de clones hybrides utilisés en plantation car elle est parfaitement adaptée au développement en zone tropicale et son bois est de bonne qualité pour la production de bois de trituration (Vigneron et Bouvet, 1997). Dans le cadre du programme d'amélioration génétique mené en République du Congo, la majorité des clones commerciaux sont issus de croisements entre *E. urophylla* et *E. grandis* alors qu'au Brésil, il s'agit de croisements entre *E. urophylla* et *E. camaldulensis*.

2. Dispositifs expérimentaux

2.1. Plan de croisement factoriel *E. urophylla* x *E. urophylla*

Le dispositif expérimental principal utilisé pour ce travail de thèse est un test de descendances intra-spécifique *E. urophylla* x *E. urophylla* mis en place en 1992 dans le cadre du programme d'amélioration génétique mené par le CRDPI en République du Congo. Ce dispositif expérimental est un plan de croisement factoriel incomplet quasi-équilibré. Il est issu du croisement de 16 géniteurs *E. urophylla* non apparentés collectés au sein l'ile de Flores (Figure 23). La majorité des génotypes *E. urophylla* des populations d'amélioration du CRDPI et de V&M do Brasil est issue de cette île de l'archipel des îles de la Sonde.

Ce plan de croisement 8x8 fait intervenir huit géniteurs mères et huit autres utilisés comme pères. Le dispositif expérimental comprend 33 familles de plein-frères pour un total de 328 individus avec 31 familles de 10 individus et 2 familles de 9 individus (Figure 24). Pour chacun des géniteurs, 3 à 5 familles de plein-frères ont été produites par croisements contrôlés. Au sein des descendants, trois contextes d'apparentements sont présents : les individus peuvent être plein-frères (issus du même croisement), demi-frères (un seul parent en commun) ou encore non apparentés.

Le dispositif expérimental a été installé à la station forestière de Kissoko (Pointe Noire, République du Congo, 4°45'S, 12°00'E, 50 m d'altitude). La pluviométrie annuelle est de

x		Pères *E. urophylla*							
		1	2	3	4	5	6	7	8
Mères *E. urophylla*	9		FS	FS	FS		FS		
	10	FS	FS				FS	FS	
	11	FS		FS	FS	FS			FS
	12	FS	FS		FS			FS	FS
	13						FS	FS	FS
	14		FS	FS		FS	FS	FS	
	15				FS	FS			FS
	16	FS				FS		FS	FS

Figure 24: schéma du dispositif expérimental majeur de la thèse. Ce schéma représente les croisements qui ont été effectués entre les 16 géniteurs non apparentés *E. urophylla* récoltés au sein de l'île de Flores. Les pères sont numérotés de 1 à 8 et les mères de 9 à 16. La mention FS indique les 33 familles de pleins frères obtenus du croisement des différents géniteurs utilisés comme pères et mères.

X		Pères *E. urophylla*					
		1	2	3	4	5	6
Mères *E. camaldulensis*	1				FS	FS	
	2			FS	FS	FS	
	3	FS	FS	FS			
	4	FS		FS			FS
	5		FS			FS	FS
	6					FS	FS

Figure 25: Schéma du dispositif *E. camaldulensis* x *E. urophylla*. Le dispositif est un plan de croisements factoriel incomplet. Le schéma représente les croisements qui ont été effectués entre les 6 géniteurs *E. camaldulensis* et les 6 géniteurs *E. urophylla* pour l'obtention des 16 familles de descendants plein-frères. La mention FS indique les 16 familles de pleins frères obtenus du croisement des différents géniteurs utilisés comme pères et mères.

1200 mm avec 4 mois de saison sèche. La température annuelle moyenne est de 24°C et les sols sableux sont chimiquement pauvres. Une fertilisation de 150 g de NPK (13-13-21) a été apportée à chaque arbre 15 jours après plantation. La parcelle unitaire est de 16 arbres, correspondant à une famille de plein-frères, et plantés à 667 tiges/ha. Le dispositif expérimental est initialement composé de 4 blocs de répétition au sein desquels les familles sont disposées de manière randomisée. Cependant, pour des raisons pratiques, un seul de ces 4 blocs à pu être utilisé dans le cadre de ce travail. Les familles de plein-frères obtenues pour chacun des 16 parents sont donc bien randomisées au sein de ce bloc mais ces familles ne sont pas répétées dans le dispositif. Les arbres ont été abattus à l'âge de 14 ans (169 mois) pour réaliser les mesures de croissance et récolter les échantillons de feuilles et de bois nécessaires aux analyses génétiques et à l'étude des propriétés chimiques du bois.

2.2. Autres dispositifs de terrain

Deux dispositifs complémentaires ont également été utilisés dans le cadre de ce travail de thèse. Ces deux dispositifs, du CRDPI et de V&M do Brasil, sont également des plans de croisements factoriels impliquant des croisements interspécifiques.

Le premier dispositif de V&M do Brasil est un plan de croisement factoriel incomplet *E. camaldulensis* x *E. urophylla*. Il a été planté en 2001 sur le site d'itapoa (Curvelo, Minas Gerais, Brésil, 19º17'S et 44º29'W à une altitude de 700 m). La pluviométrie annuelle est de 1350 mm pour une température variant de 16°C à 26°C. Il comporte 182 descendants, obtenus à partir du croisement de 6 géniteurs *E. urophylla* non apparentés de l'île de Flores (utilisés comme pères) et 6 géniteurs *E. camaldulensis* non apparentés provenant de la région de Petford dans le Queensland en Australie (utilisés comme mères). Les 182 descendants sont organisés en 16 familles de plein-frères. Chaque parent est impliqué dans 2 à 4 croisements (Figure 25). Comme dans le dispositif précédent les descendants sont plein-frères, demi-frères ou encore non apparentés. Chacune des familles est répétée dans trois blocs au sein desquels elles ont été plantées en groupes de 5 arbres disposés en lignes. Les groupes ont été répartis de manière aléatoire au sein de chaque bloc de répétition. La densité de plantation est de 1111 tiges/ha. Ce dispositif a souffert d'un taux de mortalité important. Initialement composées de 15 individus, les familles de plein-frères étudiées comportent entre 4 et 15 individus. Les prélèvements de bois et de feuilles nécessaires à l'étude des caractères chimiques et aux analyses génétiques ont été effectués 80 mois après la plantation.

		Pères *E. grandis*				
		1	2	3	4	5
Mères *E. urophylla*	1	FS	FS	FS	FS	FS
	2	FS	FS	FS	FS	FS
	3	FS	FS	FS	FS	FS
	4	FS	FS	FS	FS	FS

Figure 26: Schéma du dispositif *E. urophylla* x *E.grandis*. Le dispositif est un plan de croisements factoriel complet. Le schéma représente les croisements qui ont été effectués entre les 4 géniteurs *E. urophylla* et les 5 géniteurs *E. grandis* pour l'obtention des 20 familles de descendants plein-frères (FS).

Le deuxième dispositif est un plan de croisement factoriel complet *E. urophylla* x *E. grandis*. Ce dispositif expérimental à été mis en place par le CRDPI. Il a été planté en 1998 à la station forestière de Kissoko (Pointe Noire, République du Congo, 4°45'S, 12°00'E, 50 m d'altitude) dans les mêmes conditions de sol et de climat que le plan de croisement factoriel *E. urophylla* x *E. urophylla*. Une fertilisation de 150 g de NPK 13-13-21 a également été apportée à chaque arbre, 15 jours après plantation Le dispositif expérimental comporte 197 descendants obtenus par le croisement de 4 géniteurs *E. urophylla* utilisés comme mères et 5 géniteurs *E. grandis* utilisés comme pères. Tous ces géniteurs sont non apparentés et ont été récoltés sur les aires de répartition naturelle des deux espèces. Les *E. urophylla* sont originaires de l'île de Flores et de l'île de Lomblen dans l'archipel des îles de la sonde (Indonésie). Les *E. grandis* sont originaires de la région de Atherton Taberland dans le Queensland (Australie). Les descendants sont organisés en 20 familles de plein-frères de 9 ou 10 individus chacune (Figure 26). Comme pour le dispositif *E. urophylla* x *E. urophylla*, la parcelle unitaire est de 16 arbres représentant une famille de plein-frères. Ce dispositif ne comporte qu'un seul bloc de répétition au sein duquel la disposition des familles à été randomisée. Les prélèvements de bois nécessaires à l'étude des caractères chimiques ont été effectués 63 mois après la plantation.

3. Mesure des caractères phénotypiques

3.1. Croissance

Pour mesurer la croissance, deux types de données ont été collectées : la hauteur totale de l'arbre et la circonférence du tronc à 1,3 m de hauteur. La hauteur, a été mesurée après l'abattage de l'arbre. Les mesures ont été réalisées au moment de l'abattage à 169 mois pour le dispositif *E. urophylla* x *E. urophylla.*

3.2. Densité du bois

La densité du bois a été mesurée par la méthode d'infrandensité (poids anhydre de l'échantillon rapporté au volume d'eau déplacé). Pour chaque arbre, un disque de 10 cm d'épaisseur environ a été prélevé à 1,30 m de hauteur. Trois éprouvettes ont été taillées dans ce disque de la moelle vers la périphérie. Pour chaque éprouvette *i*, la distance depuis le centre jusqu'à la moelle de l'arbre à été mesurée ($dist_i$). Ces éprouvettes étaient de dimension approximative 20 mm x 40 mm x 40 mm dans le sens radial, longitudinal et tangentiel

Tableau 2 : échantillons de calibration et qualité des modèles de prédiction utilisés pour les caractères chimiques au sein des dispositifs expérimentaux *E. urophylla* x *E. urophylla*, *E. camaldulensis* x *E. urophylla* et *E. urophylla* x *E. grandis*. R^2_{cv} est le coefficient de détermination du modèle de prédiction, SE_{cv} est l'erreur standard de prédiction associée à la mesure chimique dans l'échantillon de calibration, CVp% est le coefficient de variation phénotypique en %.

		Dispositif expérimental				
		E. urophylla x *E. urophylla*		*E. urophylla* x *E. grandis*		*E. camaldulensis* x *E. urophylla*
	Espèce	*E. u*		*E. u* et hybrides *E. u* x *E. g*		*Hybrides E. c* x *E. u*
	n° d'échantillon	1	2	3	4	5
	Caractère	LK %	S/G	LK %	S/G	LK %
	Moyenne	28.5	2.4	30.2	3.6	28.4
Caractéristiques de l'échantillon	Ecart type	1.37	0.21	3.42	0.86	2.1
	Max	31.9	3.0	39.4	5.8	32
	Min	25.4	1.7	24.5	1.7	22.4
	CV_P %	4.8	14.5	11	24	7.4
	Taille de l'échantillon	60	60	75	231	123
	Préparation	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm
Modèles de prédiction	R^2_{cv}	0.86	0.86	0.89	0.86	0.89
	SE_{cv}	0.48	0.12	1.12	0.12	0.91

respectivement. Chaque éprouvette a été bouillie pendant 1 heure afin d'obtenir un état saturé, puis stabilisée par séchage à l'étuve à 65°C, et enfin déshydratée après séchage à l'étuve à 103°C. La dimension radiale à l'état saturé à été mesurée pour chaque éprouvette *i* (r_sat_i). L'infradensité à été calculée pour chaque éprouvette par la formule :

$$den_i = \frac{m\_anh}{v\_sat},$$

où *m_anh* est la masse de l'éprouvette anhydre et *v_sat* est la volume de l'éprouvette à l'état saturé (mesuré par la poussée d'Archimède). La densité moyenne par arbre à été calculée en appliquant la formule suivante :

$$den\_r = \frac{1}{3}\sum_{i=1}^{3} den_i \times coef_i,$$

où *den_r* est la densité moyenne par rondelle (arbre), et $coef_i$ est un coefficient appliqué à chaque éprouvette en fonction de sa distance à la moelle de l'arbre ($dist_i$).

La mesure de densité a été réalisée au sein des descendances du plan de croisement factoriel *E. urophylla* x *E. urophylla* par le CRDPI (protocole de mesure interne).

3.3. Teneur en lignines et rapport S/G

Les mesures de teneur en lignines de Klason et rapport S/G ont été réalisées par méthode indirecte. Elles ont été prédites par la méthode SPIR (Spectrométrie Proche InfraRouge) sur les arbres qui composent les descendances des plans de croisement factoriel *E. urophylla* x *E. urophylla*, *E. urophylla* x *E. grandis* et *E. camaldulensis* x *E. urophylla.* Pour chacun des arbres, un disque de bois a été prélevé à 1,30 m de hauteur, broyé à 4,0 mm, puis un aliquot représentatif du broyat a été prélevé. Cet aliquot a été utilisé pour effectuer les mesures d'absorbance dans le domaine du proche infrarouge. Les spectres SPIR (spectres d'absorbance des échantillons entre les longueurs d'ondes 800 et 2850 nm) ont été obtenus à partir d'un spectrophotomètre de marque BRUKER (modèle Vector 22/N, Bruker Optik GmbH, Ettlingen, Germany). Pour chaque échantillon, les mesures SPIR ont été répétées 3 fois et moyennées pour obtenir un seul spectre moyen par arbre. A partir de ces données, les valeurs des caractères relatifs aux lignines ont été obtenues par l'utilisation d'un modèle de prédiction.

Dans notre cas, 5 modèles de prédiction différents ont été utilisés en fonction des caractères (teneur en lignines et rapport S/G) et des plans de croisement. Le Tableau 2 présente les échantillons de calibration qui ont été utilisés pour développer les modèles de prédiction ainsi que la qualité des différents modèles déterminée par validation croisée. Brièvement, ces modèles ont été développés sur la base de 4 échantillons de calibration de taille différente et établis sur la base d'espèces différentes (espèces pures, hybrides ou mélange des deux). Les méthodes chimiques de référence utilisées pour la réalisation des mesures au sein de ces échantillons sont la méthode de Klason pour la mesure de la teneur en lignines (Dence, 1992) et la méthode de thioacydolyse pour le rapport S/G (Lapierre, 1995). Les spectres SPIR ont été obtenus de la même manière que pour les descendances des plans de croisement (3 spectre moyennés par arbre sur des poudres de 4,0 mm). Les données de mesure chimique et de variation des spectres SPIR ont été mises en relation par méthode de régressions PLS. La qualité des modèles ainsi mis au point a été déterminée par validation croisée et est exprimée par les valeurs de R^2_{cv} (coefficient de détermination du modèle de prédiction sur la base des données prédites et des données mesurées) et SE_{cv} (erreur standard de prédiction du modèle). Ces modèles de prédiction ont été mis au point dans les laboratoires de l'UPR40 (CIRAD) par Gilles Chaix et Paulo Ricardo Gherardi Hein (modèles de prédiction n°1, 2 et 3) et du CAPEF (V&M do brasil) par Leonardo Chagas (modèle de prédiction n°4).

Plus de détails sur les modèles mis en œuvre dans le cadre de la prédiction de la teneur en lignines et du rapport S/G au sein des descendances du plan de croisement factoriel *E. urophylla* x *E. urophylla* sont exposés dans l'article de Hein *et al.* (2010).

4. Méthodes statistiques pour l'estimation des paramètres génétiques des caractères

4.1. Estimation des paramètres génétiques

Les analyses de génétique quantitative ont été réalisées à l'aide du logiciel ASReml version 3.0 (Gilmour *et al.*, 2002). Les caractères de croissance, la densité du bois et les caractères relatifs aux lignines suivaient une distribution normale dans les descendances des trois dispositifs expérimentaux étudiés.

Les caractères ont d'abord été analysés de manière indépendante (analyse univariée) pour estimer les composantes de la variance. Le modèle linéaire mixte suivant a été considéré :

$$y = X.b + Z_1.a + Z_2.f + e$$

où, y est le vecteur des observations (phénotype), b est le vecteur des effets fixes, a est le vecteur des effets génétiques additifs aléatoires, f est le vecteur des effets aléatoires liés aux familles de plein-frères, e est le vecteur des résidus et X, Z_1 et Z_2 sont les matrices d'incidence liant les observations aux effets. Les effets aléatoires du modèle suivent une distribution normale de paramètres :

$$E\begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ et } \qquad Var\begin{bmatrix} a \\ f \\ e \end{bmatrix} = \begin{bmatrix} G & 0 & 0 \\ 0 & F & 0 \\ 0 & 0 & R \end{bmatrix}$$

Les matrices de variance-covariance sont définies comme suit :

$$G = A.\sigma^2_A, \quad F = I.\sigma^2_f \quad \text{et} \quad R = I.\sigma^2_e$$

où 0 est une matrice nulle, A est la matrice de parenté obtenue à partir du pedigree et qui prend en compte l'ensemble des apparentements entre individus du dispositif, I est la matrice d'identité, σ^2_A est la variance génétique additive, σ^2_f est la variance non additive entre les familles de plein-frères et σ^2_e la variance résiduelle.

Les variances associées aux effets aléatoires ont été estimées en résolvant les équations du modèle mixte d'Henderson (Henderson, 1975) par la méthode du maximum de vraisemblance restreinte (méthode REML) en utilisant ASReml (Gilmour *et al.*, 2002).

Comme les variances sont supposées indépendantes, la variance phénotypique totale σ^2_P a été calculée comme suit :

$$\sigma^2_P = \sigma^2_A + \sigma^2_f + \sigma^2_e .$$

L'effet famille n'étant jamais significatif dans le cadre de notre étude, il a été exclu du modèle pour les analyses.

Tableau 3: données bibliographiques associées à la sélection des gènes candidats pour la recherche de la variabilité fonctionnelle impliquée dans la variation des caractères quantitatifs. [1] d'après la synthèse bibliographique de Baucher *et al.*, 2003 ; [2] d'après Gion *et al.*, 2010 (soumis), dans une famille de plein-frères *E. urophylla* x *E. grandis* (P=propriétés physiques, C=propriétés chimiques, A=propriétés anatomiques, T=propriétés technologiques et G=croissance).

Gène	Locus	Catégorie fonctionnelle	Famille multigénique	Expression (tissus)	Modification d'expression [1]		Colocalisation QTLs [2]
					Effet sur teneur en lignines	Effet sur rapport S/G	
cinnamate 4-hydroxylase	*C4H*	Enzyme biosynthèse des phénylpropanoïdes	Famille multigénique (Cytochrome P450-dependent monooxygenase), 2 clusters chez l'*Eucalyptus* (Harakava *et al.*, 2005)	xyleme (Harakava *et al.* , 2005)	oui	oui	?
p-coumarate 3-hydroxylase	*C3H*	Enzyme biosynthèse des phénylpropanoïdes	Famille multigénique (Cytochrome P450-dependent monooxygenase), 2 clusters chez l'*Eucalyptus* (Harakava *et al.*, 2005)		oui	oui	P,T,M,G,C,A
ferulate 5-hydroxylase	*F5H*	Enzyme biosynthèse des phénylpropanoïdes	Famille multigénique (Cytochrome P450-dependent monooxygenase), 3 clusters chez l'*Eucalyptus* (Harakava *et al.*, 2005)		oui	oui	?
caffeate O-méthyltransférase	*COMT2*	Enzyme biosynthèse des phénylpropanoïdes	Famille multigénique, au moins 14 clusters chez l'*Eucalyptus* (Harakava *et al.* , 2005)		oui	oui	G,M,T,P
4-coumarate:CoA ligase	*4CL*	Enzyme biosynthèse des phénylpropanoïdes	Famille multigénique, classe I et classe II, au moins 10 clusters chez l'*Eucalyptus* (Harakava *et al.*, 2005)		oui	oui	G, M, C, P
cinnamoyl-CoA reductase	*CCR*	Enzyme biosynthèse des monolignols	Famille multigénique, superfamille des Dyhydroflavonol 4-Reductase (DFR), un seul cluster chez l'*Eucalyptus* (Harakava *et al.* , 2005)	xyleme (Grima-Pettenati *et al.* , 1993; Lacombe *et al.* , 1997; Harakava *et al.* , 2005)	oui	oui	M,C,A,P,G
cinnamyl alcool dehydrogenase	*CAD2*	Enzyme biosynthèse des monolignols	Famille multigénique, superfamille des Dyhydroflavonol 4-Reductase (DFR), de nombreux cluster chez l'Eucalyptus (Harakava *et al.* , 2005), répartis en 3 classes		non	oui	M,T,P,C
facteur de transcription MYB	*MYB1*	Facteur de transcription	Famille multigénique, MYB transcription factor	xyleme (Goicoechea *et al.* , 2005; Legay *et al.* , 2007)	?	?	?
facteur de transcription MYB	*MYB2*	Facteur de transcription	Famille multigénique, MYB transcription factor		?	?	non
Rac-like GTPase	*ROP1*	GTPase	Famille multigénique, superfamille des Rho et famille des Rac small GTPase	xyleme (Foucart *et al.* , 2009)	?	?	M,C,A,P,G

La variation de chaque caractère a été exprimée par le coefficient de variation phénotypique, le coefficient de variation génétique additif et l'héritabilité au sens strict calculés comme suit :

$$CV_P = \frac{\sigma_P}{\overline{X}}, \qquad CV_A = \frac{\sigma_A}{\overline{X}} \qquad \text{et} \qquad h^2 = \frac{\sigma^2{}_A}{\sigma^2{}_P}.$$

Pour estimer les corrélations phénotypiques et génétiques additives (r_P et r_A respectivement), nous avons réalisé une analyse bi-variée. Ainsi, r_P et r_A ont été estimés comme suit :

$$r_P = \frac{Cov_p(x,y)}{\sqrt{\sigma^2{}_{Px}.\sigma^2{}_{Py}}} \qquad \text{et} \qquad r_A = \frac{Cov_A(x,y)}{\sqrt{\sigma^2{}_{Ax}.\sigma^2{}_{Ay}}}.$$

Les erreurs associées à h^2, σ^2_A, σ^2_P, r_P et r_A ont été estimées avec ASReml en utilisant l'approximation par les séries standard de Taylor (Gilmour *et al.*, 2002).

La résolution du modèle mixte permet d'obtenir les BLUP (Best Linear Unbiased Predictor) c'est-à-dire les valeurs génétiques additives de chaque génotype.

4.2. Calcul de gains génétiques

Les gains génétiques attendus pour les caractères hauteur et densité par sélection directe ont été obtenus en utilisant l'équation :

$$GG_d = i.h^2{}_x.\sigma_{Px},$$

où x fait référence au caractère sélectionné, h^2x est l'héritabilité au sens strict de ce caractère, i est l'intensité de sélection, σ_{Px} est l'écart type de la mesure du caractère dans la population.

Dans le cas d'une sélection indirecte d'un caractère y basée sur le caractère x, cette équation devient :

$$GG_i = i.h_x.h_y.r_A.\sigma_{Py},$$

où h_x et h_y sont les racines carrées des héritabilités au sens strict des caractères x et y, r_A est le coefficient de corrélation génétique additif entre ces caractères et σ_{Py} est l'écart type de la mesure du caractère indirectement sélectionné dans la population. Cette formule a été

Tableau 4: effets des facteurs de transcription *MYB1* et *MYB2* sur la variation d'expression de gènes de structure de la voie de biosynthèse des lignines. D'après Vanholme *et al.*, 2010.

	Effets sur l'expression des gènes de la voie de biosynthèse des lignines										
	PAL	*C4H*	*4CL*	*HCT*	*C3H*	*CCoAOMT*	*CCR*	*F5H*	*COMT*	*CAD*	référence
MYB1							↓			↓	Legay *et al.* , 2007
MYB2			↑	↑	↑	↑	↑	↑	↑	↑	Goicoechea *et al.* , 2005

appliquée pour estimer les conséquences d'une sélection appliquée sur la hauteur et la densité pour la teneur en lignines et le rapport S/G.

L'estimation des gains génétique ne concerne dans le cadre de ce travail de thèse que le plan de croisements *E. urophylla* x *E. urophylla*.

5. Sélection des gènes candidats

Un total de 10 gènes candidats a été sélectionné pour ce travail de thèse (Tableau 3). Ils ont été choisis sur la base des données bibliographiques indiquant leur potentielle implication dans la variation des caractères relatifs à la quantité et la qualité des lignines. Ces données sont apportées par les études d'expression, de génomique fonctionnelle (avec l'étude du phénotype de plantes mutantes) et la colocalisation des gènes avec des QTLs de teneur en lignines, rapport S/G ou d'autres caractères de la qualité du bois. La disponibilité de données de séquence complète chez l'eucalyptus a également orienté le choix de ces gènes. Celle-ci facilite d'une part le travail de laboratoire pour la mise en évidence de la variabilité des gènes. D'autre part elle permet de dissocier les zones introniques et exoniques des gènes et ainsi d'étudier *a priori* les polymorphismes les plus pertinents (mutations non synonymes) pour mettre en évidence la variabilité fonctionnelle des gènes candidats. Les gènes *C4H*, *C3H*, *F5H*, *COMT2*, *4CL*, *CCR*, *CAD2*, *MYB1*, *MYB2* et *ROP1*, ont été sélectionnés sur la base de ces données. Parmi ces gènes, cinq codent des enzymes de structure de la voie de biosynthèse des phénylpropanoïdes (*C4H*, *C3H*, *F5H*, *COMT2*, *4CL*), deux codent les enzymes de la voie de biosynthèse des monolignols (*CCR*, *CAD2*), deux codent des facteurs de transcription impliqués dans la régulation de l'expression de gènes de structure de la voie de biosynthèse des lignines (*MYB1*, *MYB2* ; Tableau 4) et un dernier gène code une GTPase impliquée dans la différenciation du xylème secondaire chez l'eucalyptus (*ROP1*).

6. Mise en évidence de la variabilité nucléotidique des gènes candidats

Un total de 18 mois de travaux de laboratoire et d'analyse des données de séquençage ont été nécessaires à l'obtention des données de variabilité de ces gènes au sein des échantillons de détection de la variabilité.

Tableau 5: Séquences des amorces utilisées pour l'amplification des fragments des gènes candidats de la lignification chez *E. urophylla* et *E. camaldulensis*. Les amorces SSR ont été utilisées pour le génotypage du gène *CCR* dans la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla*.

Gène	Fragment PCR		Amorce (5' → 3')	Taille attendue
CCR	F1	sens	CACCTCCTGAACCCCTCT	397 bp
		anti-sens	CGCACCCCTTGATGGCTTCT	
	F2	sens	GCGAGGAACCGTCAGGAAC	619 bp
		anti-sens	TTTCCTCCCCAATCGTCTG	
	F3	sens	AAGAATGTGCGATGGCGAACC	474 bp
		anti-sens	GTCCCGATCACCGCTGGCT	
	F4	sens	ACGTAAGAAAGAGGGACCG	672 bp
		anti-sens	ACTTGAGGATGTGGATGATG	
	F5	sens	GCTACGGCAAGGCAGTGG	631 bp
		anti-sens	AACCGACAACCCACACCTG	
	F6	sens	CTTAGATAGATAGTCCCGC	649 bp
		anti-sens	CAAAGGGATTCAAGACAGG	
	F7	sens	CGTCATCATCGTTCTCTCT	695 bp
		anti-sens	TGACAACTTCCATTCCAA	
	SSR	sens	AGGTGTGGGTTGTCG	112 bp
		anti-sens	ATTTCCTCCCTTTTGCCC	
4CL	F1	sens	AATACCATAGACAAAGAAGG	893 bp
		anti-sens	GTTAGAGACAGATTGAGTTA	
C3H	F1	sens	GGACAAATACGACCTCAGCG	776 bp
		anti-sens	TCCACGACGATGACAAACTTC	
C4H	F1	sens	CGTCGTCTTCGACATCTTCA	529 bp
		anti-sens	ACTTCAGCCCCGTCGTTGTC	
F5H	F1	sens	CCCTCCTCCTCTCCGTCGT	977 bp
		anti-sens	TCGTTCACCTTCGCTTCGTCGC	
COMT2	F1	sens	TTACCTGAAAGATGCGGTCC	592 bp
		anti-sens	CCAATGCCTCGAACTCCTTC	
CAD2	F1	sens	CTAGTTGGCGAACTTGAAGG	1171 bp
		anti-sens	ACAATGATACAACAGAGGGC	
MYB1	F1	sens	CAAGGAGGAGGACGACAAGC	965 bp
		anti-sens	CTGCCTCATCTCTTCGCTGC	
MYB2	F1	sens	TCCAATCCACAAGACATAGC	1409 bp
		anti-sens	TCACATTTCTCAACAACGC	
ROP1	F1	sens	TCTGGGATTGTGGGATACTG	776 bp
		anti-sens	GGGTCTTGTTAGTTGCGATG	

6.1. Echantillons de mise en évidence de la variabilité nucléotidique

Deux échantillons ont été utilisés dans le cadre de ce travail de thèse pour la mise en évidence de la variabilité des gènes candidats de la lignification. Le premier comprend 20 individus non apparentés de l'espèce *E. urophylla* échantillonnés au sein de l'île de Flores dans l'Archipel des îles de la Sonde en Indonésie (Figure 23). Cet échantillon a été utilisé pour mettre en évidence, chez *E. urophylla*, la variabilité des 10 gènes candidats sélectionnés. Parmi ces 20 individus, 16 correspondent aux géniteurs des familles du plan de croisement factoriel *E. urophylla* x *E. urophylla* (Figure 24). Le deuxième échantillon est composé de 8 individus non apparentés de l'espèce *E. camaldulensis* échantillonnés dans la région de Petford en Australie (Queensland). Parmi ces 8 individus, 6 correspondent aux géniteurs des familles du plan de croisement factoriel *E. camaldulensis* x *E. urophylla* (Figure 25). Cet échantillon a été utilisé pour mettre en évidence la variabilité du gène *CCR* chez *E. camaldulensis*.

6.2. Détection de la variabilité nucléotidique de gènes candidats par séquençage d'amplicons

6.2.1. Amplification des gènes ciblés

La stratégie de séquençage d'amplicons de PCR clonés a été choisie pour décrire la variabilité des 10 gènes sélectionnés. Les amorces d'amplification par PCR ont été conçues sur la base des données de séquences contenues dans les bases de données publiques Genbank et EUCAWOOD (http://www.polebio.scsv.ups-tlse.fr/eucalyptus/eucawood). Les amorces conçues pour l'amplification des fragments de gènes et les températures d'hybridation de ces amorces sont décrites au Tableau 5. En tant que gène candidat principal pour ce travail de thèse, le gène *CCR* a été entièrement séquencé. Un total de sept couples d'amorces a été conçu pour l'amplification de sept fragments chevauchants représentant 94% de la séquence du clone *CCR* pleine longueur d'*E. gunnii* disponible dans les bases de données publiques (Genbank accession : X97433). Ces sept couples d'amorces ont été utilisés pour l'amplification du gène au sein des deux échantillons de mise en évidence de la variabilité nucléotidique chez *E. urophylla* (16 géniteurs) et *E. camaldulensis* (8 géniteurs). Les gènes *C4H*, *C3H*, *F5H*, *COMT2*, *4CL*, *CAD2*, *MYB1*, *MYB2* et *ROP1* n'ont été séquencés que partiellement et un seul couple d'amorce a été conçu pour l'amplification de chacun de ces gènes. Ces neuf couples d'amorces ont été utilisés pour l'amplification partielle de ces gènes

chez 16 génotypes *E. urophylla*, correspondant en partie seulement aux géniteurs du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Les conditions de la réaction de PCR (composition des mélanges réactionnels et programmes d'amplification) sont les mêmes pour tous les fragments des gènes ciblés, sauf dans le cas de la température d'hybridation des amorces. Ces conditions sont décrites dans l'Annexe 1.

6.2.2. Clonage des produits d'amplification et stratégie de séquençage

En théorie, si pour un individu diploïde donné, la PCR présente la même efficacité d'amplification pour les deux allèles du fragment d'intérêt, ces deux allèles se retrouvent en même quantité dans le produit de la réaction PCR. Si le clonage se fait pour ces deux allèles avec la même efficacité, à la fin de l'étape de clonage les clones correspondant à chacun des allèles sont attendus dans les mêmes proportions. Sélectionner n clones parmi l'ensemble des clones transformés devrait conduire à $(1/2)^n$ chances de n'obtenir que l'un des deux allèles pour le fragment d'intérêt lors du séquençage. Un total de 8 clones (n=8) a été séquencé pour chacun des produits de PCR cloné. Dans le cas où le nombre de clones séquencés n'était pas suffisant pour détecter les 2 allèles, une autre amplification et un autre clonage ont été réalisés.

L'ensemble des clones a été séquencé par la méthode de séquençage Sanger. Les clones correspondant aux fragments 1, 3, 4, 5 et 7 du gène *CCR* n'ont été séquencés que dans un seul sens. Les clones correspondant aux produits d'amplification des fragments 2 et 6 du gène *CCR* et des fragments uniques des autres gènes ont été séquencés dans les deux sens. Les protocoles de clonage et de séquençage utilisés sont les mêmes pour l'ensemble des fragments de gènes étudiés et sont décrits dans l'Annexe 1.

6.2.3. Expertise des séquences et mise en évidence de la variabilité nucléotidique

Pour les produits de PCR clonés, les séquences obtenues représentant un fragment de gène chez un individu ont été alignées grâce au logiciel CodonCode Aligner v.2.0.4 (Codon Code Corporation, Dedham, MA, USA). Chacun de ces alignements était composé d'un mélange de clones correspondant aux deux allèles d'un individu pour chaque fragment de gène considéré. Au sein d'un alignement, chacun des allèles était répété plusieurs fois (entre 4 et 8 fois). Pour chaque alignement, les correspondances entre les assignations des nucléotides et les pics des électrophérogrames ont été manuellement vérifiés. La répétition des clones a également permis de différencier les erreurs d'amplification de la Taq polymérase (événement

aléatoire n'affectant pas toutes les répétitions pour un allèle d'un fragment) des vrais SNP (répétés entre les clones d'un même allèle pour un même fragment). Les séquences correspondant à des « allèles chimères » (produits de la recombinaison entre allèles d'un même individu induite par la réaction de PCR, Cronn *et al.*, 2001) ont été identifiées en comparant les phases de liaison gamétique entre allèles aux SNP identifiés pour les clones de chaque alignement (la recombinaison induite par la réaction de PCR étant également un évènement aléatoire). Pour le gène *CCR*, l'ensemble des 7 fragments chevauchants a été aligné pour chaque individu séparément en utilisant CodonCode Aligner. Cette étape a permis de reconstituer les allèles du gène pleine longueur pour chaque génotype. L'information de la variabilité des zones chevauchantes a été utilisée pour reconstituer la phase de liaison gamétique entre les allèles aux différents fragments du gène pour chacun des individus séquencés. Pour chaque gène, les séquences correspondantes aux différents allèles mis en évidence pour chaque individu ont été alignées par l'utilisation du logiciel BioEdit sequence editor (Hall, 1999) pour détecter l'ensemble de la variabilité nucléotidique.

6.2.4. Cas des gènes C3H et MYB1

Dans le cas des gènes *C3H* et *MYB1*, la stratégie de séquençage d'amplicons clonés n'ayant pas été efficace (des difficultés ont été rencontrées lors de l'étape de clonage), les données de variabilité ont été obtenue par le séquençage direct des amplicons de la réaction de PCR. Le séquençage a été réalisé dans les deux sens, par la méthode de séquençage Sanger, en utilisant les amorces d'amplification des fragments utilisées pour la PCR. L'ensemble des séquences obtenues a été aligné dans CodonCode Aligner et l'attribution des bases en fonction des pics des électrophérogrames a été vérifiée.

7. Méthodes statistiques pour l'étude de la diversité nucléotidique et du déséquilibre de liaison

7.1. Diversité nucléotidique et écart à la neutralité

Deux estimateurs de la diversité nucléotidique ont été calculés : le θ_w de Watterson, basé sur le nombre de sites polymorphes en ségrégation dans l'échantillon (Watterson, 1975) et le θ_π, basé sur le nombre de différences moyen entre les séquences de l'échantillon prises deux à deux (Nei, 1987). Les valeurs de θ_w et θ_π ont été estimées par site en divisant les valeurs estimées par le nombre de sites inclus dans l'analyse de la diversité nucléotidique.

Sous le modèle neutre standard de Wright-Fisher (Wright, 1931), θ_w et θ_π sont des estimateurs du paramètre mutationnel de la population θ ($4N_e\mu$, où N_e est la taille efficace de la population et μ est le taux de mutation par gène et par génération). Les estimateurs de la diversité nucléotidique et la diversité haplotypique H_d (Nei, 1987), ont été calculés avec le logiciel DNAsp version 5 (Rozas *et al.*, 2003). Le nombre minimum d'événements de recombinaisons (R_M) a également été mesuré au sein de chaque gène par le test des 4 gamètes (Hudson et Kaplan, 1985) à l'aide de DNAsp version 5. Les INDEL n'étant pas soumis aux mêmes mécanismes mutationnels que les SNP, ils ont été exclus des analyses de diversité. Pour tester l'écart de l'évolution des gènes à la neutralité sélective, nous avons utilisé deux tests. Le D de Tajima (Tajima, 1989) et le Fs de Fu (Fu, 1997). Le premier est un test basé sur l'observation du spectre des fréquences alléliques aux sites observés et représente une différence standardisée entre θ_π et θ_w. A l'équilibre entre mutation et dérive génétique, la valeur du D de Tajima est porche de 0. Le deuxième est basé sur les haplotypes et reflète le nombre d'haplotypes attendus étant donnée la diversité nucléotidique observée. La valeur de ces deux statistiques est sensible aux effets démographiques tels que les changements de taille de populations (Fu, 1997 ; Sano et Tachida, 2005).

7.2. Déséquilibre de liaison

Le déséquilibre de liaison entre paires de SNP bialléliques a été mesuré au sein des gènes candidats pour lesquels l'information des haplotypes était disponible. La statistique r^2 (Hill et Robertson, 1968) a été calculée et son niveau de significativité a été estimé par un test exact de Fisher unilatéral à l'aide du logiciel TASSEL v. 2.1 (Bradbury *et al* ., 2007). Le DL a été mesuré entre les allèles de paires de SNP informatifs dont les fréquences d'allèles minoritaires (FAM) étaient supérieures à 15%. La décroissance du DL avec la distance physique entre paires de sites polymorphes a été mesurée de manière indépendante au sein de chaque gène puis de manière globale pour obtenir un estimateur de l'étendue du DL sur une courte distance pour l'ensemble du génome. Pour cette mesure nous avons utilisé une régression non linéaire de r^2 (Remington *et al*., 2001) en ajustant nos données pour chaque gène, ou pour l'ensemble des données, à une estimation expérimentale du taux de recombinaison C ($C=4N_ec$ où N_e est la taille efficace de la population et c est la fraction de recombinaison observée entre sites polymorphes) à l'aide de la fonction régression non linéaire du logiciel R v. 2.6.2 (R Development Core Team, 2010). L'espérance de r^2 selon Hill

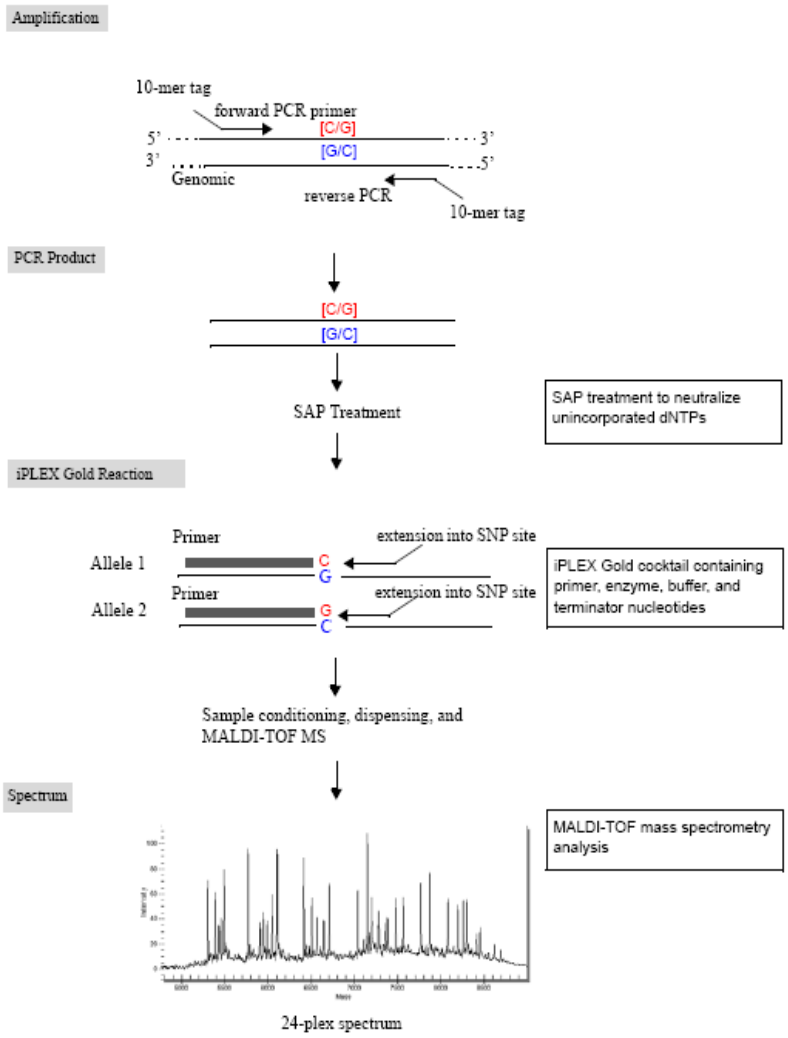


Figure 27: principe de la méthode de génotypage Sequenom iPLEX Gold par extension d'amorces et spectrométrie de masse

et Weir (1988) pour un faible taux de mutation et ajusté à une taille d'échantillon *n* a été utilisée :

$$E(r^2) = \left[\frac{10 + C}{(2 + C)(11 + C)}\right]\left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)}\right].$$

8. Génotypage des gènes chez les descendants

8.1. Génotypage microsatellite

La génotypage du gène *CCR* dans la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla* a été réalisée par la mise en évidence d'un polymorphisme microsatellite. Ce type de polymorphisme se caractérise par un motif répété de 2, 3 ou plus paires de bases. Le polymorphisme se traduit par une différence du nombre de répétition de ces unités et peut être mis en évidence par l'étude de la taille d'un fragment PCR contenant le microsatellite. Dans notre cas, le motif à été amplifié par PCR au sein des 16 génotypes parentaux et des descendants du plan de croisement factoriel. Les amorces utilisées pour l'amplification sont données au Tableau 5. L'analyse de la ségrégation des allèles du gène a été réalisée par famille en comparant les données des parents avec celles des descendants. Le mélange réactionnel ainsi que les méthodes utilisées en laboratoire pour l'étude du polymorphisme microsatellite dans le plan de croisement factoriel *E. urophylla* x *E. urophylla* sont indiqués dans l'Annexe 1.

8.2. Génotypage Sequenom iPLEX Gold

Pour génotyper la variabilité des gènes *4CL*, *C4H*, *C3H*, *COMT2*, *F5H*, *CAD2*, *MYB1*, *MYB2* et *ROP1* dans la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla* et la variabilité du gène *CCR* dans la descendance du plan de croisement factoriel *E. camaldulensis* x *E. urophylla*, la méthode de génotypage iPLEX Gold de SEQUENOM a été appliquée. Cette méthode permet de détecter des INDEL, des SNP ainsi que d'autres polymorphismes dans de l'ADN amplifié par PCR (Figure 27). C'est une méthode, de moyen à haut débit, basée sur le principe d'extension d'amorces positionnées en 5' du polymorphisme à génotyper. Brièvement, sur la base de données de séquences, et pour chaque SNP à génotyper, un couple d'amorces est conçu permettant l'amplification d'une région de 80 à 120 pb contenant le polymorphisme d'intérêt. Une autre amorce est conçue

(amorce d'extension) capable de se positionner un nucléotide en amont du polymorphisme d'intérêt. Cette amorce permet de réaliser une extension d'une base, différente en fonction de l'allèle au polymorphisme étudié. Les fragments d'ADN obtenus (amorce d'extension plus une base fonction de l'allèle au site polymorphe) sont séparés en fonction de leur masse (la masse des quatre nucléotides étant différente) dans un spectromètre de masse MALDI-TOF. Les profils de spectre obtenus permettent de différencier les génotypes des individus pour les polymorphismes étudiés. Les fragments contenant les polymorphismes à génotyper peuvent être amplifiés et analysés en mélange (multiplexage de 36 polymorphismes). Les polymorphismes qui peuvent être pris en compte par la méthode sont sélectionnés sur la base de critères de variabilité des régions bordantes. Celles-ci ne doivent pas être variables pour que le site polymorphe puisse être pris en compte. De plus, pour le multiplexage, les polymorphismes sont sélectionnés de manière à limiter les interactions entre amorces. Le génotypage à été réalisé *via* la plate forme génome transcriptome du centre de génomique fonctionnelle de Bordeaux.

9. Méthodes statistiques pour l'étude d'association

9.1. Analyse de variance à deux facteurs

Au sein du plan de croisement factoriel *E. urophylla* x *E. urophylla*, l'effet des haplotypes du gène *CCR* a été testé pour chacun des 16 parents de manière indépendante sur les caractères teneur en lignines et rapport S/G. Pour cela, une analyse de variance à deux facteurs a été réalisée selon le modèle suivant :

$$y_{fki} = \mu + F_f + H_k + \varepsilon_{fki}$$

où y est la variable à expliquer (teneur en lignines et rapport S/G), μ est la moyenne générale du caractère, F est la variable explicative relative à l'effet « famille » (effet des pères ou des mères en croisement avec le parent considéré), H est la variable explicative relative à l'effet haplotype (2 modalités par parent testé), et ε est la résiduelle. Seuls les niveaux de significativité associés aux effets haplotypes et famille ont été reportés pour l'ANOVA de type 3.

9.2. Analyses marqueur par marqueur

L'approche marqueur par marqueur a été conduite en utilisant un modèle linéaire mixte (MLM) ajusté de manière indépendante pour chaque marqueur et chaque caractère et implémentée dans le logiciel TASSEL. Cette approche permet de prendre en compte la structure de l'échantillon en sous populations et l'apparentement entre les individus de la population d'association. Elle est adaptée à l'étude du dispositif expérimental *E. urophylla* x *E. urophylla* (Yu *et al.*, 2006). En ce qui concerne la structure de la population d'association, les études de diversité génétique menées chez *E. urophylla* indiquent une très faible différenciation entre populations de toute l'aire de répartition naturelle (F_{ST}=0.03-0.04 dans Tripiana *et al.*, 2007 et Payn *et al.*, 2008). Cette différenciation n'est pas significative entre les individus des populations de l'île de Flores. Cette structure n'a donc pas été étudiée, ni prise en compte dans le modèle et tous les individus inclus dans le plan de croisement ont été considérés comme issus de la même population. L'information du pédigrée a été utilisée pour construire la matrice d'apparentement (*K*) implémentée dans le modèle pour contrôler la covariation génétique entre individus apparentés (QTLs du fond génétique). Le modèle statistique est le suivant :

$$y = X\beta + Zu + e ,$$

où *y* est le vecteur des observations phénotypiques, β est le vecteur des effets fixes pour les différents génotypes au SNP considéré, *u* est le vecteur des effets polygéniques aléatoires individuels, *e* est le vecteur des effets résiduels et *X* et *Z* sont les matrices d'incidence liant les observations aux effets. Les effets aléatoires suivent une distribution normale de paramètres

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{et} \quad Var\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} K\sigma^2_a & 0 \\ 0 & I\sigma^2_e \end{bmatrix},$$

où *K* est la matrice d'apparentement entre les individus de la population d'association (Identité par Descendance obtenue sur la base des informations de pédigrée prenant les valeurs de 0,5 pour les plein-frères, 0,25 pour les demi-frères et 0 pour les individus non apparentés), *I* est la matrice d'identité, σ^2_a est la variance génétique additive (variance des effets polygéniques) et σ^2_e est la variance résiduelle. Les variances associées aux effets aléatoires ont été estimées par la méthode du maximum de vraisemblance restreinte (REML) dans TASSEL.

La correction pour les tests multiples a été réalisée en appliquant la méthode FDR (false discovery rate ; Storey, 2002 ; Storey et Tibshirani, 2003). Les associations ont été considérées significatives au seuil de Q-value<0,05.

Chapitre 3 : Etude des paramètres génétiques des caractères liés aux lignines chez l'eucalyptus

L'étude des paramètres génétiques des propriétés chimiques du bois est relativement récente. Durant la dernière décennie, le développement des méthodes de spectrométrie dans le domaine du proche infrarouge (SPIR) pour la prédiction de ces caractères a constitué une avancée majeure pour leur étude en population. Si ces caractères suscitent aujourd'hui un grand intérêt au niveau industriel, trop peu d'études sont encore disponibles pour réellement évaluer leur niveau de variabilité et la contribution de la composante génétique dans la variation observée en population, deux paramètres essentiels pour évaluer l'intérêt de leur prise en compte dans les programmes d'amélioration génétique des arbres forestiers.

L'objectif de ces travaux était donc de mieux comprendre le déterminisme génétique des caractères relatifs aux lignines chez *E. urophylla*, et ceci dans différents contextes génétiques (intraspécifique et interspécifique). Dans un premier temps, les paramètres génétiques de qualité et de quantité des lignines ont été estimés sur les bases de données prédites par spectrométrie proche infrarouge (SPIR). Ces prédictions ont été réalisées dans les descendances de trois plans de croisements factoriels : un premier intraspécifique *E. urophylla* x *E. urophylla*, et deux autres interspécifiques *E. urophylla x E. grandis* et *E. camaldulensis x E. urophylla*. Les corrélations génétiques entre ces caractères chimiques et les caractères classiques de croissance et de densité du bois ont été analysées dans le contexte intraspécifique *E. urophylla* x *E. urophylla*. Cette étude de génétique quantitative menée sur des données de chimie du bois fait partie des toutes premières chez l'eucalyptus. En ce qui concerne l'estimation des paramètres génétiques du rapport S/G, aucune étude n'a pour le moment été publiée à notre connaissance.

Les mesures de la croissance, de la densité du bois, de la teneur en lignines et du rapport S/G obtenues dans le cadre de cette étude ont également été utilisées pour l'étude d'association entre la variabilité nucléotidique de gènes candidats de la lignification et la variation de ces caractères. Ces derniers résultats sont exposés dans le chapitre 3 de ce manuscrit.

Tableau 6: information sur les échantillons de calibrations et modèles de prédiction utilisés pour prédire les caractères liés aux lignines au sein des 3 dispositifs expérimentaux. Les caractéristiques de l'échantillon traduisent les valeurs obtenues de la mesure chimique de référence. R^2_{cv} et SE_{cv}% sont le coefficient de détermination du modèle de prédiction et l'erreur de prédiction estimés lors de l'étape de validation croisée.

	Dispositif expérimental				
	E. urophylla x *E. urophylla*		*E. urophylla* x *E. grandis*		*E. camaldulensis* x *E. urophylla*
Caractères	LK %	S/G	LK %	S/G	LK %
Préparation	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm	Poudre 4.0 mm
R^2_{cv}	0.86	0.86	0.89	0.86	0.89
SE_{cv} %	0.48	0.12	1.12	0.12	0.91

Tableau 7: dispositifs expérimentaux et caractères étudiés au sein de ces dispositifs (N est le nombre d'individus mesuré pour chacun des 3 dispositifs expérimentaux). Pour chacun des caractères sont indiqués l'unité de mesure, la moyenne, le minimum et le maximum et le coefficient de variation phénotypique.

Caractère	Unité	Dispositif	N	Moyenne	Min	Max	CV_P %
LK %	%	*E. urophylla x E. urophylla*	328	28,15	24,82	31,76	4,33
LK %		*E. urophylla x E. grandis*	197	27,85	23,50	31,10	5,02
LK %		*E. camaldulensis x E. urophylla*	182	31,31	26,10	33,70	4,10
S/G		*E. urophylla x E. urophylla*	328	2,41	1,67	3,13	12,28
S/G		*E. urophylla x E. grandis*	197	2,86	1,40	4,00	18,28
Hauteur	m	*E. urophylla x E. urophylla*	328	21,15	7,80	29,60	17,31
C 1.30	cm			52,77	24,00	84,00	21,21
Densité	$g.cm^{-3}$			0,52	0,36	0,64	7,96

1. Résultats

1.1. Qualité des prédictions par SPIR

Les caractères de chimie du bios (teneur en lignines et rapport S/G) ont été estimés par spectrométrie proche infrarouge (SPIR). Pour ce faire, les calibrations et les prédictions ont toutes été réalisées à partir de poudres de bois issues du broyage (particules de 4 mm) d'une rondelle prélevée à 1,3 m de hauteur. Les modèles de prédiction utilisés pour estimer les valeurs des caractères ont été établis sur la base d'échantillons de calibration différents en fonction des dispositifs considérés (cf. chapitre 2 : Matériel et méthodes). Les caractéristiques de ces modèles de prédiction (déterminées par validation croisée) sont présentées au Tableau 6. Pour les différents modèles, les coefficients de détermination (R^2cv) des régressions PLS obtenues entre les mesures prédites et les mesures chimiques sont du même ordre de grandeur (de 0,86 à 0,89) pour les deux caractères de composition chimique du bois au sein des 5 échantillons de calibration. En revanche, l'erreur de prédiction (SEcv) n'est pas du même ordre. Les prédictions de la quantité de lignine (LK%) à partir des dispositifs *E. urophylla* x *E. grandis* (SEcv=1,12%) et *E. camaldulensis* x *E. urophylla* (SEcv=0,91%) sont moins bonnes que celle réalisée à partir du dispositif *E. urophylla* x *E. urophylla* (SEcv=0,48%). En ce qui concerne le rapport S/G (S/G) la qualité des prédictions est satisfaisante pour les deux dispositifs *E. urophylla* x *E. urophylla* et *E camaldulensis* x *E. urophylla* (SEcv=0,12%).

1.2. Variabilité phénotypique des caractères

Ces modèles de prédiction ont été utilisés pour obtenir les valeurs des caractères LK% et S/G pour les 328 descendants du dispositif *E. urophylla* x *E. urophylla*, les 197 descendants hybrides du dispositif *E. urophylla* x *E. grandis* et enfin les 182 descendants hybrides du dispositif *E. camaldulensis* x *E. urophylla*. Les statistiques descriptives de ces caractères sont données au Tableau 7. Il est important de noter ici que l'amplitude de la variation incluse dans les échantillons de calibration utilisés pour prédire les caractères chimiques était parfois inférieure à celle des échantillons qui ont été prédits dans le cadre des plans de croisement étudiés. Ceci peut induire une erreur de prédiction plus importante que celle qui a été estimée par validation croisée dans l'échantillon de calibration. Cependant, seuls quelques individus (<10) avait des valeurs prédites extérieures à la gamme de variation de l'échantillon de calibration, ce qui à l'échelle des plans de croisement peut être considéré comme négligeable.

Tableau 8: estimation des composantes de la variance et des paramètres génétiques pour les caractères de croissance, la densité du bois et les caractères relatifs aux lignines au sein des 3 dispositifs expérimentaux selon le modèle individuel. Les estimations des différents paramètres sont notées en gras et les écarts types associés à ces estimations sont notés à leur droite. σ^2_A est la variance additive, σ^2_P est la variance phénotypique, h^2 est l'héritabilité au sens stricte et CV_A% est le coefficient de variation génétique additif. C 1.30 est la circonférence du tronc mesurées à 1,30 m, LK % est la teneur en lignines, S/G est le rapport entre les monomères S et G des lignines.

Dispositif	Caractère	σ^2_A		σ^2_P		h^2		CV_A %
E. urophylla x E. urophylla	Hauteur	**4.40**	2.16	**13.65**	1.40	**0.32**	0.14	**9.91**
	C 1.30	**18.47**	11.69	**125.95**	10.87	**0.15**	0.09	**8.14**
	Densité	**9.82E-04**	4.35E-04	**1.82E-03**	2.42E-04	**0.54**	0.18	**6.03**
	LK %	**1.39**	5.72E-01	**1.63**	2.95E-01	**0.85**	0.21	**4.19**
	S/G	**5.90E-02**	2.56E-02	**9.50E-02**	1.38E-02	**0.62**	0.19	**10.05**
E. urophylla x E. grandis	LK %	**8.54E-01**	5.38E-01	**2.05**	3.12E-01	**0.42**	0.21	**3.32**
	S/G	**2.23E-01**	1.28E-01	**2.99E-01**	6.67E-02	**0.74**	0.27	**16.47**
E. camaldulensis x E. urophylla	LK %	**6.72E-01**	3.91E-01	**1.64**	2.37E-01	**0.41**	0.19	**2.62**

Pour la quantité de lignines, les moyennes obtenues varient de 27,85% pour le dispositif *E. urophylla* x *E. grandis* à 31,31% pour le dispositif *E. camaldulensis* x *E. urophylla*. La quantité de lignines globalement plus forte au sein du dispositif *E. camaldulensis* x *E. urophylla* est sans doute liée au croisement avec *E. camaldulensis* qui est justement plus riche en lignines et donc bien adaptées à la production de charbon. Au sein des 3 dispositifs expérimentaux et pour des estimations réalisées à trois âges différents (169 mois pour le dispositif *E. urophylla* x *E. urophylla*, 63 mois pour le dispositif *E. urophylla* x *E. grandis* et 80 mois pour le dispositif *E. camaldulensis* x *E. urophylla*), la teneur en lignines montre une variabilité globalement faible avec des coefficients de variation phénotypiques de 3,8% à 5% environ. La qualité des lignines (mesurée par le rapport S/G) montre une variation plus importante que la teneur en lignines au sein des deux dispositifs *E. urophylla x E.urophylla* et *E. urophylla x E. grandis* avec des coefficients de variation phénotypique de 12,3% et 18,3% respectivement.

Les coefficients de variations phénotypiques obtenus, dans le plan de croisement factoriel *E. urophylla* x *E. urophylla*, pour les caractères de croissance, sont plus importants que pour les caractères relatifs aux lignines (17.31% et 21.21% pour la hauteur et la circonférence respectivement). Dans le même plan de croisement, le coefficient de variation phénotypique mesuré pour la densité du bois indique un niveau de variation intermédiaire (entre les caractères de la croissance et de la composition chimique du bois) pour ce caractère (environ 8%).

1.3. Estimation des paramètres génétiques

1.3.1. Croissance et densité du bois

Les résultats concernant l'estimation des composantes de la variance et des paramètres génétiques pour les caractères de croissance (hauteur et circonférence à 1,30 m) et la densité du bois au sein du dispositif *E. urophylla* x *E. urophylla* sont présentés au Tableau 8.

La hauteur (h^2=0,32) et la circonférence (h^2=0,17) sont sous contrôle génétique modéré dans ce dispositif. Les effets génétiques qui contrôlent la variation de ces deux caractères sont principalement additifs. En effet, les effets famille inclus dans le modèle génétique étaient globalement faibles et non significatifs suggérant des effets de dominance faibles au sein de ce dispositif et pour cet âge de mesure (169 mois). Pour ce qui est de la densité du bois (mesurée ici par la méthode de l'infradensité), un contrôle génétique plus fort (h^2=0,54) a été

Tableau 9: corrélations phénotypiques (au dessus de la diagonale) et génétiques additives (au dessous de la diagonale) entre caractères de croissance, densité, quantité et qualité des lignines et rapport S/G au sein du plan de croisement factoriel *E. urophylla* x *E. urophylla*. L'erreur associée à chacune des valeurs estimées de corrélation est indiquée entre parenthèses. Les valeurs significatives sont indiquées en gras (* seuil de 5 %, ** seuil de 1 %, * seuil de 0.1 %).**

	Corrélations phénotypiques									
	Hauteur		C 1.30		Densité		LK %		S/G	
Hauteur			**0.77 *****	*(0.03)*	-0.36	*(0.29)*	-0.1	*(0.09)*	0.02	*(0.09)*
C 1.30	**0.68 ****	*(0.21)*			-0.09	*(0.07)*	0.09	*(0.08)*	-0.05	*(0.07)*
Densité	-0.07	*(0.08)*	-0.47	*(0.32)*			.	.	**-0.25 ****	*(0.09)*
LK %	**-0.53 ***	*(0.24)*	-0.28	*(0.35)*	.	.			**-0.23 ***	*(0.11)*
S/G	0.31	*(0.31)*	0.22	*(0.37)*	**-0.55 ***	*(0.22)*	-0.25	*(0.28)*		
	Corrélations génétiques additives									

mis en évidence. Comme dans le cas des caractères de croissance, l'effet familial inclus dans le modèle reste faible et non significatif indiquant un contrôle génétique majoritairement additif dans le contexte étudié.

1.3.2. Quantité et Qualité des lignines

Le Tableau 8 fournit également les valeurs des estimations des composantes de la variance ainsi que des paramètres génétiques des caractères relatifs aux lignines, pour les 3 dispositifs expérimentaux. De manière générale, les résultats montrent des valeurs d'héritabilité plus fortes que la croissance et comprises entre 0,41 pour LK% dans le plan de croisement *E. camaldulensis* x *E. urophylla* et 0,85 pour LK% dans le plan de croisement *E. urophylla* x *E. urophylla*. Pour les trois dispositifs expérimentaux, et pour les deux caractères de composition chimique du bois, les effets familles du modèle génétique utilisé sont globalement faibles et non significatifs suggérant un contrôle génétique de leur variation majoritairement additif.

La valeur d'héritabilité la plus forte (h^2=0,85) est obtenue pour la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Elle indique une part très importante des effets génétiques dans la variation de la teneur en lignines dans ce contexte intraspécifique. Dans le cas des deux croisements interspécifiques, les valeurs d'héritabilité sont plus modérées (0,40 environ). De manière globale, ce caractère est celui qui présente les coefficients de variation génétique additif les plus faibles parmi l'ensemble des caractères étudiés (entre 2,62% et 4,19%).

Pour S/G, les valeurs d'héritabilité individuelles les plus fortes sont obtenues dans le contexte interspécifique *E. urophylla* x *E. grandis* (0,74) même si le contrôle génétique de ce caractère reste fort dans le cadre du plan de croisement *E. urophylla* x *E. urophylla* (0,62). Parmi l'ensemble des caractères étudiés, S/G est celui qui présente les coefficients de variation génétiques additifs (CV_A%) les plus importants et ce, que ce soit dans le contexte génétique intraspécifique (*E. urophylla* x *E. urophylla*) ou interspecifique (*E. urophylla* x *E. grandis*).

1.3.3. Corrélations entre les caractères relatifs aux lignines, caractères de croissance et densité du bois

L'ensemble des estimations pour les corrélations phénotypiques et génétiques additives entre caractères de croissance, densité et caractères chimiques relatifs aux lignines obtenues

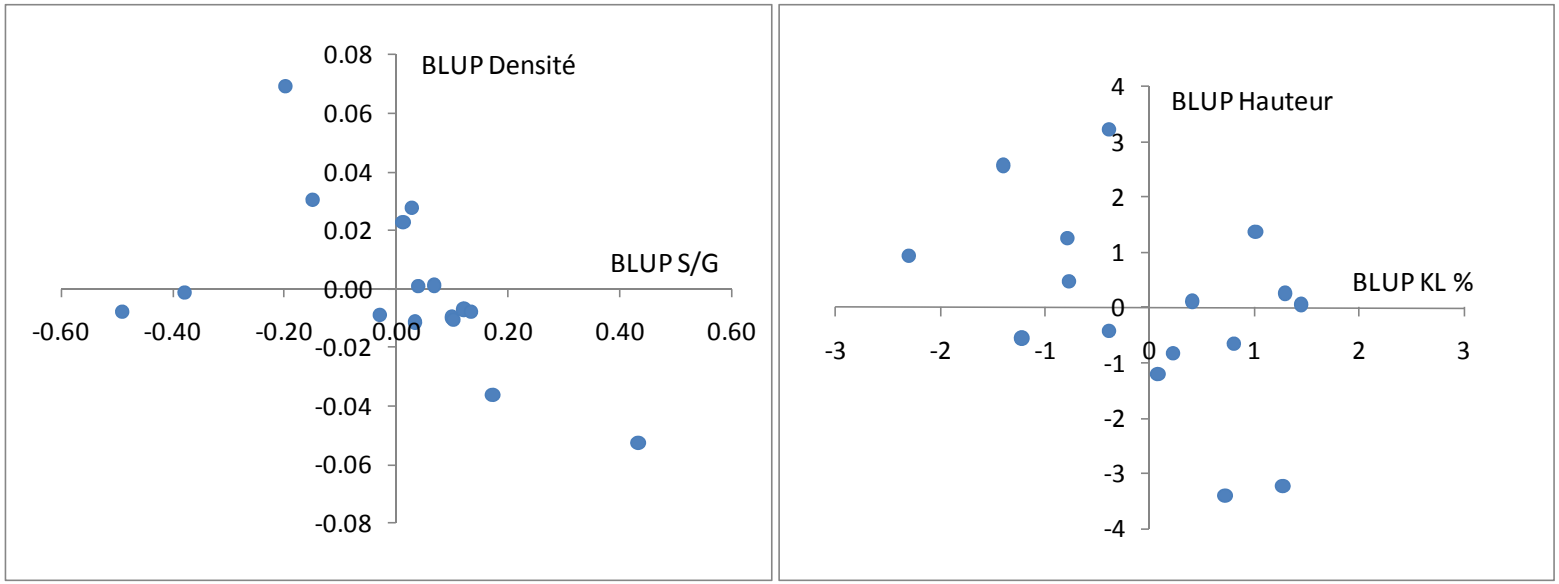


Figure 28: BLUP pour la densité en fonction du rapport S/G ainsi que pour la hauteur en fonction de la teneur en lignines chez les 16 parents *E. urophylla* du plan de croisement factoriel *E. urophylla* x *E. urophylla*.

pour le plan de croisement factoriel *E. urophylla* x *E. urophylla* sont présentés au Tableau 9. Compte tenu de la taille du dispositif expérimental utilisé pour l'analyse, les valeurs des corrélations génétiques additives sont généralement assorties d'erreurs importantes. Ces données montrent cependant des corrélations phénotypiques significatives entre certains des caractères étudiés. La corrélation phénotypique la plus forte est observée entre les caractères de croissance en hauteur et la circonférence (r_P=0,77). Des corrélations phénotypiques significatives sont également observées entre la densité du bois et le rapport S/G et la teneur en lignines et le rapport S/G. Ces deux corrélations sont négatives et du même ordre (-0,25 environ). Au niveau génétique additif, une corrélation positive entre les caractères de croissance, et négative entre le rapport S/G et la densité du bois est retrouvée. Si cette corrélation est inférieure à la corrélation phénotypique dans le cas du couple de caractères hauteur et circonférence (r_A=0,68), elle est plus forte dans le cas du couple de caractères S/G et densité du bois (r_A=-0,55). Les données indiquent également une corrélation génétique additive négative importante entre la hauteur et la teneur en lignines (r_A=-0,53). Ces corrélations génétiques additives fortes suggèrent un contrôle génétique en partie commun pour ces couples de caractères. La Figure 28 montre la répartition des BLUP des 16 parents du plan de croisement pour les couples de caractères densité du bois et rapport S/G d'une part et hauteur et teneur en lignines d'autre part. Cette figure montre que malgré une corrélation génétique additive négative entre la hauteur et la teneur en lignines, certains individus présentent de bonnes aptitudes générales à la combinaison pour ces deux caractères. Ces individus sont particulièrement intéressants dans le cadre d'une sélection pour la production de charbon de bois à usage industriel.

1.3.4. Impact d'une sélection dirigée sur la hauteur et la densité sur les caractères relatifs aux lignines

Les efforts menés en matière de sélection sont, à l'heure actuelle, principalement orientés sur l'augmentation des caractères de croissance. Si les propriétés du bois sont encore peu considérées, elles présentent un intérêt croissant pour les améliorateurs. Etant données les corrélations génétique additives importantes mise en évidence au sein du dispositif expérimental *E. urophylla* x *E. urophylla*, entre hauteur et teneur en lignines de Klason et entre densité du bois et rapport S/G, nous avons cherché à évaluer la réponse corrélée des caractères relatifs aux lignines à une sélection ciblée sur la hauteur ou la densité. Pour ce faire, un taux de sélection de 5 % a été considéré (l'indice de sélection i=2.063). Les gains génétiques obtenus par sélection directe (sur la hauteur et la densité) et indirecte (pour la

Tableau 10: gains génétiques espérés sur LK% par la mise en place d'une sélection basée sur la hauteur et sur S/G par la mise en place d'une sélection basée sur la densité pour un taux de sélection de 5 %. Cd : caractère directement sélectionné. Ci : caractère indirectement sélectionné. r_A : coefficient de corrélation génétique additif. GG Sd : gains génétiques obtenus sur Cd par sélection directe. GG Si : gains génétiques obtenus sur Ci par sélection indirecte. r_A est la corrélation génétique additive estimée entre les caractères Cd et Ci.

Taux de sélection	5%	5%
Cd Ci	Hauteur LK%	Densité S/G
r_A	-0,53	-0,55
GG Sd %	11,61	9,13
GG Si %	-2,60	-8,83

teneur en lignines et le rapport S/G) ont été exprimés en pourcentage de la moyenne du caractère dans la population parentale (Tableau 10). Les données indiquent qu'une sélection directe de 5% appliquée sur la hauteur ou la densité du bois donne des gains génétiques de 11,6% et 9.1% pour ces caractères respectivement. Une telle sélection appliquée sur la hauteur induit une diminution sensible de la teneur en lignines dans la population améliorée (-2,6%). En revanche, une telle sélection appliquée sur la densité a pour conséquence indirecte une diminution importante du rapport S/G (-8.4%).

2. Discussion

2.1. L'utilisation de la SPIR pour la prédiction des caractères relatifs aux lignines

Les données présentées dans cette étude ont été obtenues par l'utilisation de la méthode SPIR. Etant donné que l'application de ces méthodes à la prédiction des propriétés chimiques du bois est récente, peu d'études relatant la qualité des modèles de prédiction mis au point sont disponibles, notamment pour la teneur en lingines et le rapport S/G. Cependant, les premiers résultats obtenus sont encourageants. Chez plusieurs espèces de pin, les valeurs rapportées pour les coefficients de détermination des modèles développés pour la prédiction de la teneur en lignines varient entre 0.57 et 0.91 avec des erreurs sur les valeurs prédites comprises entre 4,9% et 0,38% (Meder *et al.*, 1999 ; Hodge et Woodbridge, 2004 ; Yeh *et al.*, 2004). Chez l'eucalyptus, les valeurs des coefficients de détermination des modèles de prédiction décrits par Baillères *et al.* (2002) et Poke *et al.* (2006 a) pour le même caractère sont de 0,87 et 0,78 respectivement (les erreurs de prédiction associées à ces modèles étant de 1,02% et 0.37% respectivement). Concernant le rapport S/G, les deux seules études disponibles à notre connaissance indiquent des valeurs des coefficient de détermination des modèles de prédiction de 0,89 (Alves *et al.*, 2006) et 0,90 (Baillères *et al.*, 2002). Ces valeurs sont assorties de faibles erreurs de prédiction (0,22% et 0,0054% respectivement pour ces deux études). Une étude plus récente menée chez le peuplier (Robinson et Mansfield, 2009) indique que les quantités de monomères syringyl et guaiacyl peuvent également être estimées séparément avec une bonne précision (R^2=0,96 et REcv=0,202% pour les deux monomères G et S). Sur la base de ces données, les modèles de prédiction utilisés dans le cadre de nos travaux sont comparables à ceux développés par d'autres groupes pour la prédiction des caractères relatifs aux lignines. Différentes études rapportent l'utilisation de tels modèles de

prédiction, avec des caractéristiques similaires en termes de coefficients de détermination et d'erreurs de prédiction, pour l'estimation des paramètres génétique de différentes propriétés chimiques du bois (Raymond *et al.*, 2002 ; Poke et Raymond, 2006 b ; Gaspar *et al.*, 2009) ou la détection de QTLs (Markussen *et al.*, 2003 ; Freeman *et al.*, 2009). Ces études tendent à démontrer l'efficacité de la méthode SPIR pour la mesure rapide des propriétés chimiques du bois avec une précision suffisante pour l'estimation des paramètres génétiques ou l'identification de régions génomiques impliquées dans leur variation.

Dans certains cas, il semble également possible d'établir des modèles de prédiction de bonne qualité pour les propriétés chimiques du bois sur la base d'échantillons solides de bois (voir Poke *et al.*, 2006 a, pour la teneur en lignines de Klason, la teneur en lignines totales, la teneur en lignines soluble dans l'acide, la teneur en cellulose et le taux d'extrait). De tels modèles présentent un avantage dans le cadre du phénotypage de grands échantillons car ils permettent de supprimer l'étape de broyage pour la prédiction des caractères. Dans notre cas, ce sont des poudres de bois qui ont été utilisées. La qualité des modèles de prédiction associés à ce type d'échantillon était meilleure que dans le cas d'échantillons solide. Hein *et al.* (2010) montrent par une étude comparative de différentes méthodes de préparation des échantillons, que celle-ci influence la qualité des modèles de prédiction de manière significative. Les auteurs indiquent que la qualité des modèles développés pour la prédiction de la teneur en lignines et du rapport S/G sur des échantillons de poudres de bois est toujours meilleure en comparaison des échantillons solides. Ils indiquent cependant une bonne corrélation entre les deux types de mesures (corrélation croisée supérieures à 0,80 pour les meilleurs modèles de prédiction obtenus à partir de poudres ou d'échantillons solides).

Dans notre cas, pour améliorer la précision des valeurs prédites par SPIR pour la teneur en lignines de Klason et le rapport S/G, il apparait important de développer les échantillons de calibration. L'objectif à terme serait de pouvoir disposer d'échantillons représentatifs de la gamme de variation de ces caractères chez les espèces pures et hybrides étudiées. De tels échantillons permettraient de disposer de modèles de prédiction mieux adaptés à l'étude de populations comme celles qui sont gérées dans le cadre des programmes d'amélioration génétique des arbres forestiers.

2.2. Variation phénotypique et génétique des caractères relatifs aux lignines chez l'eucalyptus

Sur la base des données obtenues dans le cadre de ce travail, les caractères relatifs aux lignines présentent des gammes de variation phénotypique inférieures (surtout pour la teneur en lignines) aux caractères de croissance mais présentent également des niveaux d'héritabilité plus importants que ces derniers. Ces résultats sont en accord avec les résultats d'autres études indiquant que la variation des propriétés du bois est globalement sous contrôle génétique plus fort que celle des caractères de croissance (Raymond *et al.*, 2002). Les niveaux d'héritabilité au sens strict estimés pour la teneur en lignines dans les populations d'hybrides interspécifiques sont de l'ordre de ceux rapportés chez *Pinus pinaster* par Pot *et al.* (2002). En revanche, ces niveaux contrastent avec les résultats obtenus chez *Picea Abies* par Hannrup *et al.* 2004 (h^2=0,10) et chez *E. globulus* par Poke et Dungey, 2006 (h^2=0,13). Dans ce contexte, la valeur d'héritabilité au sens strict estimée au sein du plan de croisement factoriel *E. urophylla* x *E. urophylla* apparait comme exceptionnellement forte (h^2=0,85). Cette forte valeur d'héritabilité obtenue pour les descendances intraspécifiques du plan de croisement factoriel *E. urophylla* suggère également que le déterminisme génétique de la teneur en lignines pourrait être différent en fonction des espèces d'eucalyptus. Une étude récente de Volker *et al.* (2008) indique qu'il existe des différences dans le déterminisme génétique de la croissance et de la densité chez *E. nitens* et *E. globulus*. Les auteurs montrent notamment que l'héritabilité au sens strict estimée pour le diamètre des arbres, diminue avec l'âge chez *E. globulus* alors qu'elle augmente chez *E. nitens* (à 10 ans, l'héritabilité au sens strict estimée chez *E. globulus* est cinq fois plus faible que chez *E. nitens*). Même si la transmission des caractères dans le cas d'espèces hybrides est moins renseignée que dans le cas d'espèces pures, les études indiquent, chez l'hybride, un mode de transmission des caractères généralement intermédiaire à celui des parents (Potts *et al.*, 2004). Les valeurs d'héritabilité obtenues dans le cadre de notre étude, dans un contexte intraspécifique et deux contextes interspécifiques, pourraient indiquer un contrôle génétique additif plus important chez *E. urophylla* que chez les deux autres espèces *E. camaldulensis* et *E. grandis*. Cependant, ces résultats doivent être temporisés par la taille des dispositifs étudiés qui n'impliquent que peu de parents et ne permettent de tester la variabilité des caractères que dans un seul environnement. Dans ce cadre, les valeurs d'héritabilité qui ont été estimées ici ne peuvent pas être considérées comme relatives au contrôle génétique de la variation des caractères pour ces espèces dans un contexte multi-environnemental. D'autres études impliquant des

dispositifs plus larges plantés dans des environnements plus variés seraient nécessaires pour mieux comprendre le déterminisme génétique des caractères relatifs aux lignines et cette étude doit être considérée comme une étude préliminaire.

Quoi qu'il en soit, dans le cas du rapport S/G, les résultats obtenus ici sont tout à fait intéressants puisqu'ils mettent en évidence, pour ce caractère et dans le contexte génétique et environnemental étudié, un niveau de variation important associé à une forte héritabilité au sens strict. Ces deux caractéristiques (une forte variabilité sous un contrôle génétique fort et principalement additif) sont des composantes très favorables à l'amélioration génétique et indiquent que des gains génétiques importants peuvent être attendus de la prise en compte de ce caractère de chimie du bois dans les programmes d'amélioration génétique.

2.3. Corrélations génétiques et gains génétiques espérés

Dans le contexte intraspécifique, la corrélation génétique additive négative entre croissance en hauteur et LK% est également intéressante. Elle indique qu'une sélection pour la croissance favoriserait de manière globale la diminution du taux de lignines. Si cette corrélation présente un intérêt pour la production de pâte à papier (Raymond *et al.*, 2002), elle est dommageable pour la production de charbon de bois (Vigneron et Gion, communication personnelle). Cette corrélation reste faible et l'étude des gains génétiques espérés de l'amélioration de la croissance indiquent un impact modéré sur la teneur en lignines (-2,5% pour un taux de sélection de 5% sur la croissance). Cependant, étant donnés les volumes de bois traités chaque année par ces deux industries, des gains génétiques même faibles peuvent présenter un intérêt économique pour les industriels. Ces données indiquent que, pour la production de charbon de bois, la teneur en lignines doit être contrôlée dans le cadre d'une amélioration sur la croissance. Une corrélation génétique additive négative est également observée entre la densité du bois et le rapport S/G. Dans le cadre de la production de pâte, l'augmentation de ces deux caractères est associée à de meilleurs rendements en pâte (Miranda *et al.*, 2001 ; Del Rio *et al.*, 2005). Les gains génétiques négatifs qui sont obtenus sur le rapport S/G en réponse à une sélection de 5% appliquée à la densité sont importants (-8,5%). Ces résultats indiquent que dans le cadre de la production de pâte, l'amélioration de la densité du bois nécessite de contrôler l'évolution du rapport S/G dans la population.

2.4. Paramètres génétiques et études d'association

Les résultats obtenus dans le cadre de ce travail font partie des tous premiers sur l'étude des paramètres génétiques des caractères relatifs aux lignines. Ces résultats sont en faveur d'un contrôle génétique additif fort pour la teneur en lignines et le rapport S/G et indiquent des corrélations génétiques additives avec la hauteur (pour la teneur en lignines) et la densité du bois (pour le rapport S/G). Certains de ces résultats contrastent avec ceux obtenus pour d'autres études menées sur ces caractères. Cependant, les dispositifs considérés dans le cadre de cette étude n'incluent que peu de parents et peu d'individus et ne sont en aucun cas représentatifs des populations qui sont utilisées dans les programmes d'amélioration des eucalyptus menés par le CRDPI au Congo ou V&M do Brazil. Les quelques familles étudiées au sein des 3 dispositifs expérimentaux (entre 16 et 33) n'ont permis d'utiliser qu'un modèle génétique très simple pour l'étude des caractères et n'ont pas permis d'estimer la part des effets environnementaux dans la variatino de ces carcatères (les dispositifs n'ayant pas été répétés). L'ensemble de ces résultats doit donc être perçu comme une première analyse des paramètres génétiques qui leurs sont associés. Dans le cadre d'études plus poussées des caractères relatifs aux lignines, la méthode SPIR présente une bonne alternative aux méthodes de mesure classiques et pourrait permettre de prendre en compte des dispositifs plus importants, plantés dans différents sites, incluant des âges de mesures différents pour ces caractères et éventuellement des dispositifs intra et interspécifiques connectés pour une meilleure connaissance des effets génétiques et environnementaux qui contrôlent leur variation. Dans le cadre de notre étude, les analyses indiquent un contrôle génétique fort de la teneur en lignines et du rapport S/G dans les différents dispositifs étudiés avec des niveaux de variabilité importants pour le rapport S/G et particulièrement dans le cas du plan de croisement *E. urophylla* x *E. urophylla*. Ces résultats sont favorables à l'utilisation de ces dispositifs expérimentaux dans le cadre d'une étude d'association entre la variabilité des gènes et la variation de ces propriétés chimiques du bois.

Chapitre 4 : Diversité nucléotidique, étendue du déséquilibre de liaison chez *E. urophylla* et comparaison avec d'autres espèces d'*Eucalyptus*

L'étude de la diversité nucléotidique s'attache à mettre en évidence et décrire le polymorphisme de la séquence d'ADN au sein de groupes d'individus (populations). Si elle permet dans certains cas d'émettre des inférences quant à l'histoire des populations ou l'impact des forces évolutives sur différentes régions du génome, l'étude de la diversité nucléotidique est également à la base des études d'association. En effet, la compréhension de l'organisation des polymorphismes, incluant la description des haplotypes et l'évaluation de l'étendue du DL en population, constitue un pré-requis essentiel pour la mise en évidence des polymorphismes (QTN, quantitative trait nucleotide) impliqués dans la variation phénotypique. Il s'agira d'une part, d'évaluer le niveau de résolution qui peut être atteint pour la détection de ces QTN, et d'autre part de choisir les polymorphismes les plus pertinents pour les détecter.

Dans cette partie, nous nous sommes intéressés au niveau et à la structure de la variabilité nucléotidique de 10 gènes candidats impliqués dans le processus de lignification chez *E. urophylla*. Nous avons ensuite comparé les données obtenues pour un gène, codant la *CCR*, avec deux espèces du même sous-genre utilisées en plantation : *E. camaldulensis*, par la mise en évidence de sa variabilité dans un échantillon de 8 individus, et *E. globulus*, par la comparaison des résultats obtenus dans l'étude de Poke *et al*. (2003). Notre étude fait partie des toutes premières réalisée chez l'*Eucalyptus* visant à acquérir des données en termes de niveau de diversité nucléotidique et d'étendue du DL à faible (intragène) et longue (intergène) distance. Ces résultats ont également permis de disposer d'un ensemble de marqueurs génétiques pour l'étude de la variabilité fonctionnelle des gènes de la lignification chez *E. urophylla* et *E. camaldulensis* (chapitre 5), deux espèces majeures des programmes d'amélioration des eucalyptus menés au Congo par le CRDPI et au Brésil par V&M do Brasil.

1. Résultats

1.1. Diversité nucléotidique de gènes de la lignification chez *E. urophylla*

1.1.1. Séquençage et mise en évidence de la variabilité de gènes de la lignification

1.1.1.1.Obtention des séquences et contraintes techniques

Pour décrire la variabilité de 10 gènes candidats de la lignification chez *E. urophylla*, 2444 séquences ont été générées. La méthode de séquençage de produits PCR clonés a été utilisée pour 8 de ces gènes (*4CL*, *C4H*, *F5H*, *COMT2*, *CAD2*, *CCR*, *MYB2* et *ROP1*) donnant accès à l'information des haplotypes présents dans l'échantillon pour une partie ou la totalité (cas du gène *CCR*) du gène. Dans le cadre de cette approche, 8 clones indépendants ont été séquencés en moyenne pour 249 produits PCR clonés. Les données de variabilité des deux autres gènes (*C3H* et *MYB1*) ont été obtenues par séquençage direct de produits PCR et correspondent à des données génotypiques au sein des individus du panel de recherche de polymorphisme nucléotidique. L'analyse des séquences a permis de mettre en évidence un certain nombre de contraintes liées à la technique employée. Les biais les plus importants sont causés par les erreurs de réplication de la Taq polymérase et la production d'allèles « chimères » lors de la réaction PCR.

Les erreurs de réplication de la Taq polymérase sont des événements aléatoires qui se distinguent facilement des polymorphismes ponctuels de type SNP, par le fait qu'elles ne sont généralement pas répétées. Ainsi, au sein d'un ensemble de séquences alignées, elles apparaissent comme des événements de mutation isolés. Au sein de 1291 séquences générées pour l'étude de la variabilité du gène *CCR*, 1147 erreurs de réplication de la Taq polymérase ont été détectées. Sur la base de la somme totale de la taille des séquences obtenues pour le gène, un taux d'erreurs par paire de base a pu être estimé à $1{,}48.10^{-3}$/pb (une erreur pour 657 pb séquencées en moyenne).

Les séquences « chimères » correspondent à des amplicons combinant une partie de la séquence de deux matrices différentes. Elles sont produites au cours de la réaction PCR lorsque plusieurs matrices homologues sont en mélange dans la réaction (comme dans le cas d'individus diploïdes). Lors d'un cycle de la réaction PCR, lorsque la réplication (qui a lieu

Tableau 11: données de séquençage recueillies pour les 10 gènes impliqués dans la lignification chez *E. urophylla*. Pour la colonne Indels le nombre d'Indels est indiqué à gauche et entre parenthèses est indiqué le nombre de paires de bases correspondant à la somme des Indels.

			Paires de bases étudiées						
Gène	Locus	Groupe de liaison	Total	5'UTR	Exon	Intron	3'UTR	Indels	Type de données
cinnamate 4-hydroxylase	*C4H*	11	1199		518	681		2 (4)	haplotypes
p-coumarate 3-hydroxylase	*C3H*	9	619		526	93			génotypes
ferulate 5-hydroxylase	*F5H*	6	835		835				haplotypes
Caffeate O-méthyltransférase	*COMT2*	7	941		593	348			haplotypes
4-coumarate:CoA ligase	*4CL*	11	929		291	638		1 (2)	haplotypes
cinnamoyl-CoA reductase	*CCR*	6	3222	23	1011	2076	112	17 (67)	haplotypes
cinnamyl alcool dehydrogenase	*CAD2*	10	1183		636	459	88	2 (3)	haplotypes
MYB transcription factor	*MYB1*	2	796		564	232		1 (9)	génotypes
MYB transcription factor	*MYB2*	2	937		549		388	2 (25)	haplotypes
Rac GTPase like	*EgROP1*	6	1391		405	854	132	2 (21)	haplotypes
			12052	23	5928	5381	720	27 (110)	

Tableau 12: nombre et densité de SNP identifiés au sein d'un échantillon d'individus non apparentés de l'espèce *E. urophylla* pour 10 gènes impliqués dans la lignification. En bas du tableau sont indiqués les totaux des nombres de SNP détectés par région (codantes et non codantes) et l'impact sur la structure primaire de la protéine (mutations silencieuses ou non synonymes). Les densités moyennes de SNP totales, et par région (codantes et non codantes) sont également indiquées en nombre de paires de bases par SNP. Les zones transcrites non traduites (UTR) ont été considérées comme non codantes pour l'attribution des SNP et les calculs de densité. [a] : nombre de SNP total et entre parenthèses nombre de singletons. [b] : exons sans les UTR. [c] : introns avec UTR.

		SNPs					Densité de SNPs 1/x pb		
Locus	N	Total [a]	régions codantes [b]	régions non codantes [c]	Silencieux	Non synonymes	Total	régions non codantes [c]	régions codantes [b]
C4H	32	48 (18)	8	40	43	5	25.0	17.0	64.8
C3H	28	22 (2)	15	7	19	3	28.1	13.3	35.1
F5H	32	14 (3)	14	0	8	6	59.6		59.6
COMT2	32	19 (14)	9	10	13	6	49.5	34.8	65.9
4CL	14	21 (3)	3	18	21		44.2	35.4	97.0
CCR	32	156 (49)	34	122	152	4	20.7	18.1	29.7
CAD2	32	30 (6)	8	22	27	3	39.4	24.9	79.5
MYB1	32	14 (5)	9	5	9	5	56.9	46.4	62.7
MYB2	32	14 (2)	9	5	10	4	66.9	77.6	61.0
EgROP1	30	49 (14)	9	40	48	1	28.4	24.7	45.0
		387 (116)	118	269	350	37	41.9	32.5	60.0

sur une des deux matrices d'ADN) n'est pas complète, le produit de réplication incomplet peut, au cycle suivant, se fixer sur l'autre matrice pour continuer son élongation. Si les deux matrices présentent des différences (comme dans le cas de deux allèles pour un fragment de gène chez un individu diploïde), le produit d'amplification combine une partie de la variabilité des deux allèles et est appelé « chimère ». Les séquences « chimères » peuvent être détectées par l'observation d'un changement de la phase de liaison gamétique entre les deux allèles attendus pour un individu diploïde. Le caractère aléatoire de cet événement permet également de l'identifier assez facilement dans un alignement de séquences de clones obtenues à partir d'un produit de réaction PCR chez un individu diploïde. Dans le cas du gène *CCR*, 94 séquences parmi les 1291 générées (7,3%) correspondaient à une « chimère ».

1.1.1.2.Détection d'un deuxième locus CCR

Pour deux des génotypes sélectionnés pour la mise en évidence de la variabilité nucléotidique du gène *CCR*, l'amplification des fragments 3, 5 et 6 du gène a permis, dans certains cas, le clonage et le séquençage de 3 produits d'amplification différents. Parmi ces 3 produits, 2 correspondent aux 2 allèles recherchés au locus *CCR* et le troisième correspond à un locus homologue à celui recherché. Ce locus homologue (*CCR2*) a été détecté chez deux individus uniquement pour le fragment 5. Par ailleurs, ce locus a été détecté pour l'un ou l'autre des deux génotypes pour les fragments 3 et 6 du gène. Deux clones sont disponibles pour le locus *CCR2* au sein du fragment 3, neuf clones au sein du fragment 2 et trois clones au sein du fragment 6. Aucune différence n'a été détectée entre les clones du locus *CCR2* issus de l'amplification du fragment 5 pour ces deux génotypes. Ceci suggère qu'il pourrait s'agir, pour toutes les séquences, d'un seul et même haplotype. Sur la base des données de séquence obtenues pour le fragment 5, *CCR2* présente 88% d'identité avec le locus *CCR* étudié. Un total de 34 polymorphismes exclusifs a été détecté pour ce deuxième locus (30 SNPs et 4 INDELS) parmi lesquels une délétion de 28 pb correspondant à une zone de l'exon 4 du locus *CCR*. Cette délétion provoque un décalage du cadre de lecture et des changements importants dans la protéine prédite pour *CCR2* (sur la base des données d'épissage connues pour le premier locus *CCR*). Des études d'expression ont été réalisées pour tenter de mettre en évidence l'expression du locus *CCR2* chez *E. urophylla* dans le cadre du travail de thèse d'E. Villar (communication personnelle). Des extractions d'ARN totaux ont été réalisées sur différents tissus (feuilles, tiges, apex). Des PCR quantitatives ont été réalisées mais celles-ci n'ont pas permis de détecter le locus *CCR2*, ce qui suggère que ce gène ne serait pas transcrit dans ces tissus ou qu'il pourrait s'agir d'un pseudo-gène.

Tableau 13: diversité nucléotidique, diversité haplotypique et tests d'écart à la neutralité pour les 10 gènes impliqués dans la lignification chez *E. urophylla*. La longueur des séquences est exprimée en pb. [1] : seulement des sites synonymes codants. [a] : R_M pour nombre minimum d'événements de recombinaison. [b] : nombre d'haplotypes K. [c] : valeur exprimée par site. * : significatif au seuil de 5%.

	Diversité nucléotidique												Diversité haplotypique			Tests d'écart à la neutralité	
	Total				Silencieux				Non synonymes								
Locus	Longueur de séquence	S	θw [c]	π [c]	Longueur de séquence	S	θw [c]	π [c]	Longueur de séquence	S	θw [c]	π [c]	K [b]	Hd (SD)	R_M [a]	Tajima's D	Fu's Fs
C4H	1119	48	0.01065	0.00755	747	43	0.01429	0.01033	370	5	0.00336	0.00198	17	0.94 (< 0.001)	6	-1.07	-1.900
C3H	619	22	0.00913	0.01483	220	19	0.02224	0.03686	398	3	0.00193	0.00273				2.22 *	
F5H	835	14	0.00416	0.00341	200	8	0.01118 [1]	0.00787 [1]	631	6	0.00236	0.00202	12	0.881 (0.001)	1	-0.78	-3.238
COMT2	941	19	0.00501	0.00265	481	13	0.00827	0.00293	458	6	0.00163	0.00238	9	0.738 (0.004)	0	-1.61 *	-1.194
CCR	2942	156	0.01317	0.01309	2187	152	0.01772	0.01729	755	4	0.00131	0.00093	20	0.958 (< 0.001)	20	-0.12	2.962
CAD2	1180	30	0.00631	0.0064	694	27	0.00966	0.01031	486	3	0.00153	0.00082	10	0.764 (0.005)	1	0.05	2.393
MYB1	763	14	0.00456	0.0037	322	9	0.00772	0.0063	441	5	0.00281	0.0018				-0.81	
MYB2	873	14	0.00398	0.00503	442	10	0.00617	0.00795	431	4	0.00231	0.00203	10	0.752 (0.005)	0	0.59	0.123
EgROP1	1359	49	0.0091	0.00712	1047	48	0.01206	0.00897	312	1	0.00081	0.00092	16	0.938 (< 0.001)	2	-0.93	-0.958
Moyenne	10631		0.00734	0.00709	6338		0.012266	0.012618	4284		0.00201	0.00173	13	0.853		-0.61	1.258

1.1.1.3.Mise en évidence de la variabilité nucléotidique

L'ensemble des 10 gènes étudiés représente 12 kbp de séquence. Ces données représentent la variabilité de 5928 pb de régions exoniques, 5381 pb de régions introniques et 743 pb de régions transcrites non traduites (UTRs) (Tableau 11). Parmi l'ensemble des gènes étudiés, le gène *CCR* est le seul qui a été séquencé entièrement. Pour ce gène, 94% de la séquence totale (3222 pb) englobant 100% du CDS (région du gène codant pour la structure primaire de la protéine) a été obtenue. Pour les autres gènes, les données de séquence représentent un fragment du gène variant de 796 pb à 1391 pb. Pour la plupart des gènes, les données de variabilité ont été obtenues sur un échantillon de 16 individus non apparentés correspondant à 32 copies indépendantes du gène ou de la portion génique étudiée (Tableau 12). Les données de séquences des gènes *ROP1*, *C3H*, et *4CL* n'ont été obtenues que pour des échantillons de 30, 28, et 14 copies indépendantes respectivement. Dans le cas de *4CL*, étant donnée la faible taille de l'échantillon disponible, les données n'ont été utilisées que pour la mise en évidence de SNP et ont été exclues des analyses de la diversité nucléotidique et du DL.

1.1.2. Variabilité des gènes et densité de SNP

Au total, 387 SNPs ont été identifiés sur la base de ces données de séquençage (Tableau 12). Parmi ces SNPs, 116 correspondent à des singletons (l'allèle minoritaire au SNP n'est porté que par une seule copie du gène ou de la portion du gène étudié dans l'échantillon soit une fréquence d'apparition inférieure à 5%). Un total de 156 SNP a été détecté au sein du seul gène *CCR*. Sur la base de ces données, la densité moyenne de SNP, toutes régions géniques confondues, est estimée à 1 SNP/42 pb. Ces données ont également révélé une densité de SNP moyenne presque 2 fois plus importante au sein des régions non codantes (1SNP/33 pb) qu'au sein des régions codantes (1 SNP/60 pb). Les différences de densité de SNP entre zones non codantes (incluant les UTRs) et zones codantes sont variables pour les gènes étudiés. Ainsi, pour le gène *C4H*, les régions non codantes sont presque 4 fois plus variables que les régions codantes (1 SNP/17 pb et 1 SNP/65 pb respectivement). Inversement, pour le gène *MYB2*, les régions non codantes (n'incluant que de l'UTR) sont les moins variables (1 SNP/78 pb contre 1 SNP/61 pb pour les régions codantes). L'étude de la densité des SNP a également permis de mettre en évidence des différences de niveau de variabilité au sein des gènes étudiés. Parmi ces 10 gènes, le gène *CCR* est le plus variable au sein de l'échantillon avec une densité de SNPs, toutes régions géniques confondues, de 1 SNP/21 pb. A l'inverse, le gène *MYB2* est le

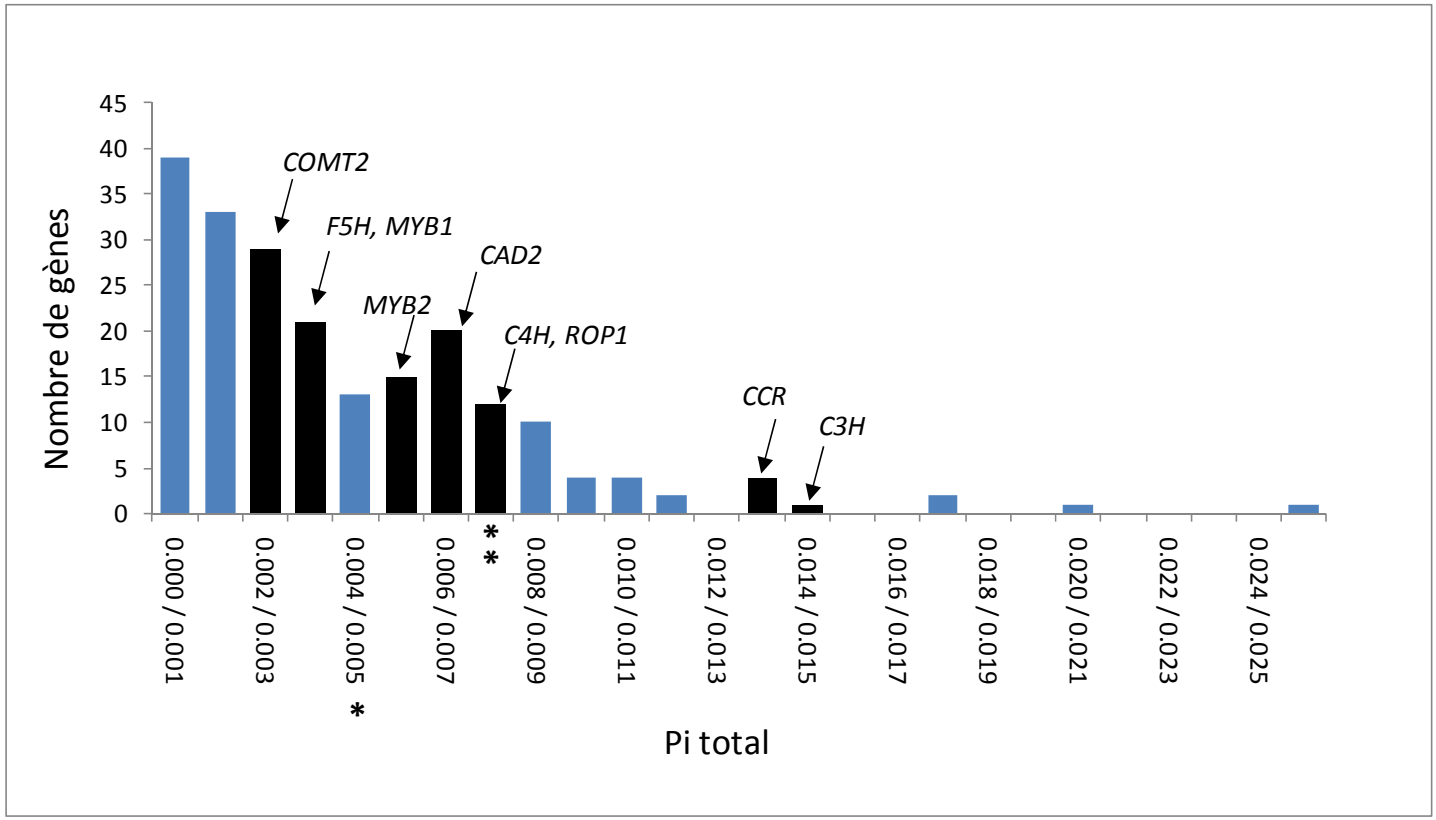


Figure 29: comparaison des valeurs estimées de la diversité nucléotidique totale (π_{total}) mesurée par site au sein de différents gènes ou portions de gènes chez les espèces d'arbres forestiers. Basé sur les études de Kado *et al.* (2003), Fujimoto *et al.* (2008), Pot *et al.* (2005), Brown *et al.* (2004), Gonzales Martinez *et al.* (2006), Ma *et al.* (2006), Palmé *et al.* (2008), Wachowiak *et al.* (2009), Krutovsky and Neale (2005), Heuertz *et al.* (2006), Ingvarsson *et al.* (2005), Breen *et al.* (2009), Quang *et al.* (2008) et les résultats de cette étude. Au total, 211 valeurs de π_{total} ont été retenues. Le nombre de gènes est indiqué en fonction des classes de valeur de π_{total}. Les gènes étudiés dans ce travail de thèse ont été rajoutés et les classes qui les incluent ont été représentées en noir. * : la classe 0.004-0.005 contient la valeur moyenne de $\theta_{\pi\ total}$ estimée sur l'ensemble des gènes considérés dans cette analyse. * * : la classe 0.007-0.008 contient la valeur moyenne de $\theta_{\pi\ total}$ estimée sur l'ensemble des locus étudiés dans ce travail (sauf *4CL*).

moins variable avec 1 SNP/67 pb. Enfin, 37 SNPs non synonymes ont été détectés au sein de l'échantillon, tous gènes confondus.

1.1.3. Cartographie génétique de gènes candidats

Parmi les 10 locus étudiés, 7 étaient déjà cartographiés chez *E. urophylla* (Gion *et al.*, 2000) par l'étude de leur variabilité dans une famille de plein-frères *E. urophylla* x *E. grandis* (croisement élite utilisé pour la cartographie génétique au CIRAD). La variabilité détectée au sein des trois autres gènes (*C4H*, *F5H* et *MYB1*) a été mise en évidence au sein de 94 descendants de cette famille. Deux SNP par gène ont été sélectionnés sur la base de leur variabilité au sein du parent *E. urophylla* et ont été utilisés (méthode iPLEX Gold de Sequenom) pour positionner ces 3 gènes sur la carte génétique d'*E. urophylla*. Parmi les 11 groupes de liaisons que compte l'eucalyptus, les 3 locus étudiés ont été positionnés sur les groupes de liaison 11 (*C4H*), 6 (*F5H*) et 2 (*MYB1*) (Tableau 11).

1.1.4. Diversité nucléotidique, diversité haplotypique et tests d'écart à la neutralité

La diversité nucléotidique moyenne (π par site) est égale à $7{,}09.10^{-3}$ pour l'ensemble des gènes, toutes régions géniques confondues. La diversité génétique est globalement plus importante pour les sites silencieux (dont la mutation potentielle n'induit pas de changement dans la structure primaire de la protéine) que pour les sites non synonymes (dont la mutation potentielle peut induire un changement sur la structure primaire de la protéine) avec des valeurs de π par site de $12{,}62.10^{-3}$ et $1{,}7.10^{-3}$ respectivement. Les niveaux de diversité nucléotidique mesurés montrent des variations d'un ordre 1000 entre les gènes et les régions géniques étudiées avec des valeurs de π par site maximales obtenues pour les sites silencieux du gène *C3H* ($36{,}86.10^{-3}$) et minimales pour les sites non synonymes du gène *CAD2* ($0{,}82.10^{-3}$). En terme de diversité nucléotidique totale, le gène *CCR* présente le niveau le plus fort ($\pi = 13{,}17.10^{-3}$) et le gène *MYB2* le plus faible ($\pi = 3{,}98.10^{-3}$) (Tableau 13). La Figure 29 montre la répartition des effectifs par classe de valeurs de π_{total} rapportées pour un ensemble de gènes étudiés chez les arbres forestiers. Sur la base de ces résultats, les gènes *CCR* et *C3H* présentent un niveau de diversité nucléotidique totale (π_{total} par site) important et en moyenne, la diversité nucléotidique totale des gènes de la lignification est élevée chez *E. urophylla*.

Le nombre d'haplotypes moyen détectés, basé sur l'étude de 7 de ces 10 gènes (ceux dont la variabilité a été révélée par la méthode de séquençage de produits PCR clonés), est de 13 avec un maximum de 20 haplotypes pour *CCR* et un minimum de 9 haplotypes pour

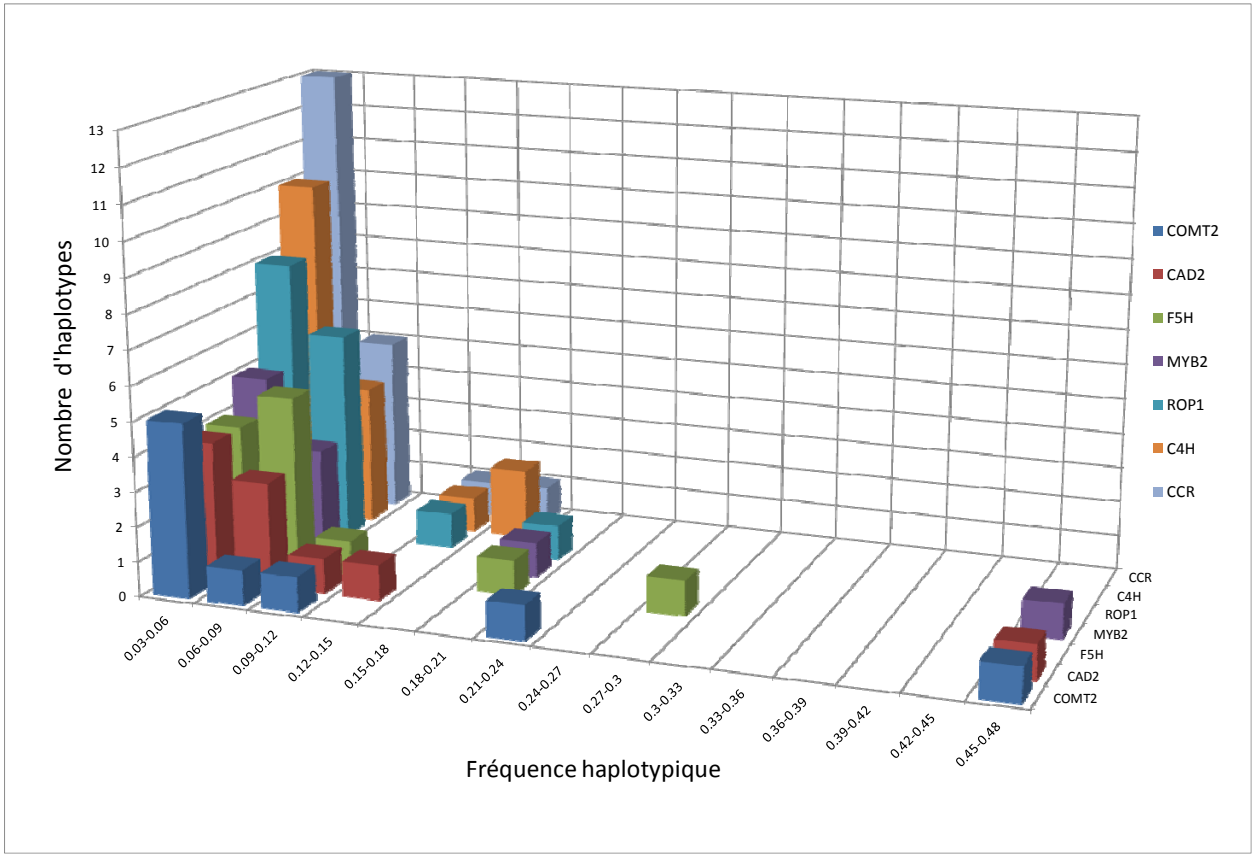


Figure 30: nombre d'haplotypes en fonction de leur fréquence d'apparition dans l'échantillon de séquençage chez *E. urophylla* pour les gènes *COMT2*, *CAD2*, *F5H*, *MYB2*, *ROP1*, *C4H* et *CCR*.

COMT2. Les nombres d'événements minimums de recombinaison associés à l'histoire de chacun des gènes dans l'échantillon étudié varie entre 0, pour les gènes *MYB2* et *COMT2*, et 20 au sein du gène *CCR*. La diversité haplotypique moyenne estimée sur la base de ces données est de 0,853 avec des valeurs supérieures obtenues pour *CCR* (maximum), *C4H*, *ROP1* et *F5H* et des valeurs inférieures obtenues pour *COMT2* (minimum), *MYB2* et *CAD2* (Tableau 13). Pour certains de ces gènes (*MYB2*, *CAD2* et *COMT2*), un haplotype se retrouve largement majoritaire (fréquence proche de 0.5) au sein de l'échantillon étudié. Pour les gènes *C4H*, *COMT2*, *CCR*, *MYB2* et *ROP1*, les haplotypes en faibles fréquences (<6%) représentent plus de 50% des haplotypes détectés dans l'échantillon (Figure 30).

Sur la base de ces données, la valeur moyenne de D de Tajima estimée pour l'ensemble des gènes étudiés était faiblement négative (-0,11). Parmi les gènes candidats, les gènes *C3H* et *COMT2* présentaient des valeurs de D de Tajima significativement différentes de 0 indiquant un écart entre les patrons de diversité nucléotidique observés et ceux attendus sous le modèle neutre standard d'évolution (MNSE). Aucun des gènes étudiés ne présentait des valeurs de Fs du Fu significativement différentes de 0 (Tableau 13).

1.2. Déséquilibre de liaison au sein des gènes de la lignification chez *E. urophylla*

1.2.1. Etendue du DL

L'étendue du DL au sein de gènes de la lignification chez *E. urophylla* a été estimée sur la base des SNP détectés au sein des gènes *C4H*, *F5H*, *COMT2*, *CAD2*, *CCR*, *MYB2* et *ROP1* dont la fréquence de l'allèle minoritaire est supérieure à 15% au sein de l'échantillon. Ce critère nous a permis de conserver 139 SNP pour un total de 3376 mesures de DL entre paires de sites d'un même gène sur l'ensemble des gènes. Parmi ces mesures, 986 montraient un écart significatif à l'hypothèse d'indépendance, donc l'existence d'un déséquilibre de liaison (Figure 31). Etant donné le faible nombre de sites conservés pour l'étude du DL au sein de certains gènes, l'ensemble des valeurs de DL estimées entre paires de SNP au sein de chacun des gènes a été regroupé pour évaluer la décroissance du DL globale en fonction de la distance entre SNP pour l'ensemble des gènes étudiés (Figure 32). Les résultats de cette analyse montraient une décroissance rapide du DL avec des valeurs moyennes de r^2 inférieures à 0,2 au-delà de 1000 pb.

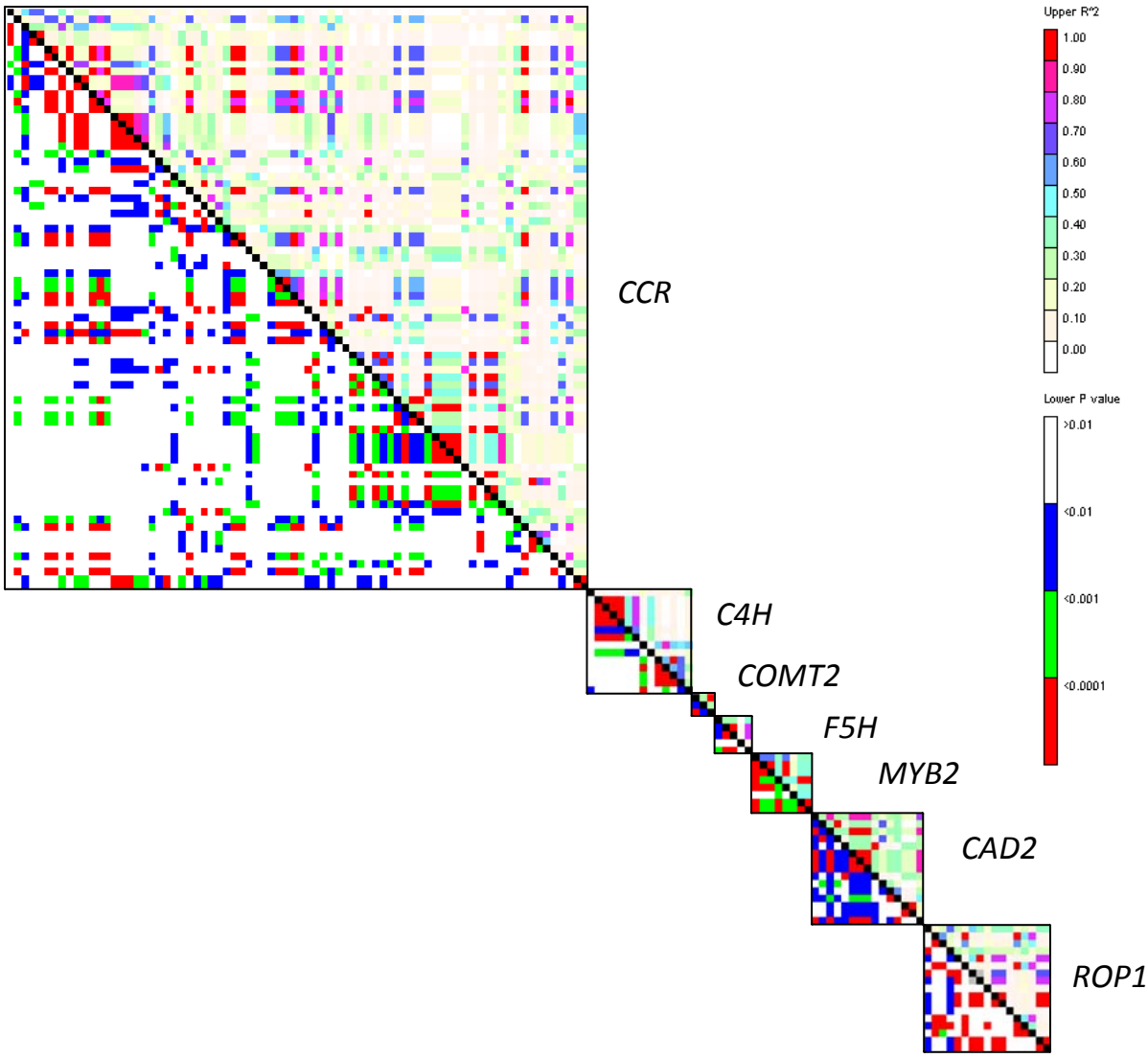


Figure 31: matrices de déséquilibre de liaison intra-gène pour les gènes *CCR*, *C4H*, *COMT2*, *F5H*, *MYB2*, *CAD2* et *ROP1* au sein de l'échantillon d'individus non apparentés d'*E. urophylla*. Ces matrices représentent, pour chacun des gènes, le déséquilibre de liaison mesuré par la métrique r^2 entre allèles par paires de SNP détectés. Seuls les SNP pour lesquels la fréquence de l'allèle minoritaire était supérieure à 15% ont été représentés. La diagonale représente l'enchainement des SNP sélectionnés (MAF > 15%) dans l'ordre de leur position sur la séquence du gène de 5' en 3'. Au dessus de la diagonale, les valeurs de DL entre paires de sites SNP sont représentées par un carré dont la couleur dépend de la valeur de r^2 obtenue entre les sites étudiés. En dessous de la diagonale, et de la même manière, ce sont les P-values relatives au niveau de significativité du DL mesuré entre paires de SNP qui sont représentées.

1.2.2. Evaluation du DL entre gènes

Malgré une décroissance rapide du DL mise en évidence de manière globale au sein des 7 gènes étudiés, l'examen des résultats indique néanmoins la persistance d'un DL fort ($r^2 \approx 0,8$) entre certains sites éloignés de plus de 2 kb. Afin d'évaluer la persistance d'un DL sur de longue distances (métrique génétique), nous avons étudié le DL entre polymorphismes SNP des gènes *ROP1* et *CCR*, deux des gènes situés sur le même groupe de liaison (GL6), à une distance génétique de 11 cM. Les données haplotypiques phasées ont pu être reconstituées pour 8 individus *E. urophylla* non apparentés constituant l'échantillon de détection de SNP. Les données de phase de liaison gamétique entre les haplotypes des deux gènes portés par ces 8 individus ont été obtenues par l'étude de la ségrégation des haplotypes de ces gènes dans les descendances du plan de croisements factoriel *E. urophylla* x *E. urophylla*. L'échantillon ainsi obtenu pour l'étude du DL entre sites polymorphes pour ces deux gènes représente 16 copies indépendantes d'une portion chromosomique de 11 cM correspondant à une partie du groupe de liaison 6. L'étude du DL a été réalisée en considérant les sites SNP dont la MAF (fréquence de l'allèle minoritaire) est supérieure à 15%, soit pour 17 SNP du gène *ROP1* et 75 SNP du gène *CCR* (Figure 33). Conformément aux résultats obtenus pour l'étude de l'étendue du DL sur de faibles distances, les résultats obtenus n'ont pas permis de mettre en évidence un DL significatif entre paires de sites polymorphes appartenant à ces deux gènes.

1.3. Séquençage, diversité nucléotidique et haplotypique et DL au sein du gène *CCR* chez *E. camaldulensis*

Les données générées par séquençage (726 séquences) ont permis de mettre en évidence la variabilité du gène *CCR* au sein d'un échantillon de 8 génotypes d'*E. camaldulensis,* représentant 16 copies indépendantes du gène. Les données de variabilité ont été obtenues, comme pour *E. urophylla*, sur 94% de la séquence complète incluant 100% du CDS de *CCR*. Au sein de l'échantillon étudié, ces données représentent une longueur totale de 3130 pb avec 19 INDEL représentant une longueur totale de 90 pb (Tableau 14). Au total, 145 SNP ont été identifiés le long du gène dont 56 singletons. La densité de SNP détectés est de 1 SNP/20 pb au sein des régions introniques et de 1 SNP/27 pb au sein des régions exoniques du gène avec 102 et 43 sites SNP détectés au sein de ces régions respectivement. Parmi les 43 SNP exoniques, 4 sont non synonymes.

Les données de séquence obtenues n'ont pas permis de reconstituer la phase entre les polymorphismes pour le gène *CCR* pleine longueur. Deux portions du gène ont donc été

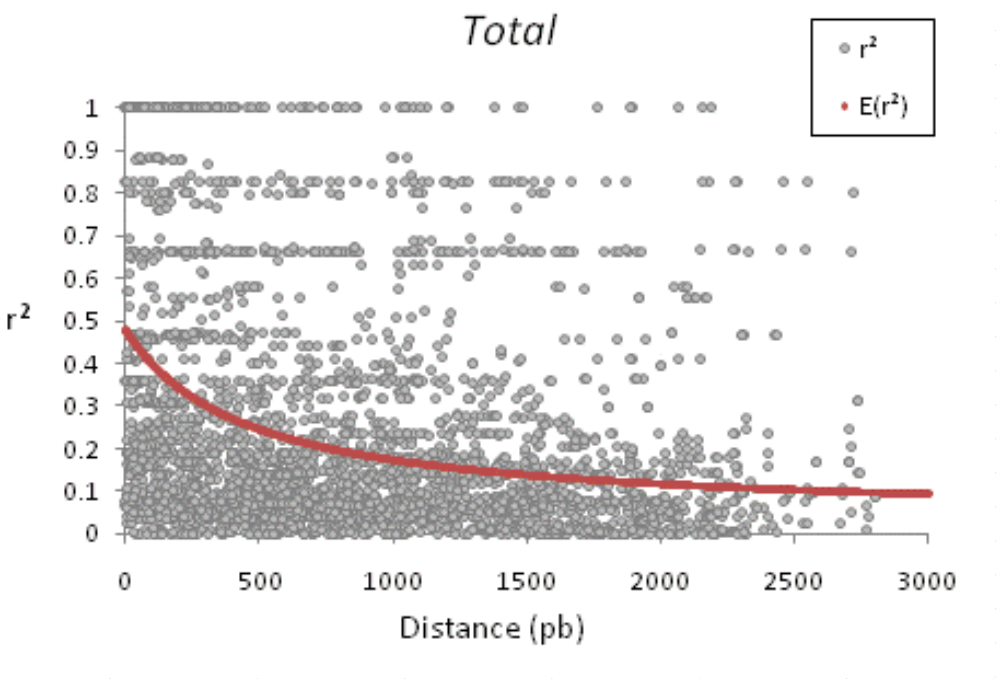


Figure 32: étendue du déséquilibre de liaison (E(r^2) en fonction de la distance entre paires de SNP) au sein de 7 gènes de la lignification. La courbe de décroissance du DL (en rouge) à été obtenue sur la base de 3376 comparaisons impliquant 139 SNP détectés au sein des gènes *CCR*, *C4H*, *COMT2*, *F5H*, *MYB2*, *CAD2* et *ROP1*.

étudiées indépendamment : la région 1, englobant les fragments d'amplification PCR 1 et 2 et la région 2 englobant l'ensemble des autres fragments d'amplification. La valeur de la diversité nucléotidique (π par site) estimée pour le gène complet chez *E. camaldulensis* est de $14{,}3.10^{-3}$ (Tableau 15). Un total de 11 haplotypes a été détecté au sein du fragment 1 pour une diversité haplotypique de 0,925. Pour le fragment 2, 13 haplotypes ont été mis en évidence pour une diversité haplotypique de 0,975. Enfin, nous avons détecté 3 et 14 événements de recombinaisons minimum au sein des fragments 1 et 2, respectivement.

Le déséquilibre de liaison entre paires de SNP a été étudié au sein du fragment 2. Un total de 1653 valeurs de r^2 entre paires de sites polymorphes a été estimé sur la base de 58 SNP retenus en fonction de la fréquence de leur allèle minoritaire (FAM> 15%). Sur ces 1653 valeurs, 127 montrent un écart significatif entre les associations alléliques observées et celles attendues sous l'hypothèse d'indépendance. La décroissance du DL au sein du gène *CCR* chez *E. camaldulensis* a été estimée sur la base de ces données (Figure 34). Cette étude montre comme pour *E. urophylla*, une décroissance rapide du DL avec une valeur moyenne de r^2 égale à 0,2 pour des sites SNP distants d'environ 250 pb.

2. Discussion

2.1. La méthode de séquençage utilisée

La méthode de séquençage de produits de réaction PCR clonés, est une approche fréquemment employée pour la description des haplotypes chez des individus hétérozygotes et diploïdes. Elle peut être employée de différentes manières et notamment en complément d'une approche de séquençage direct de produits PCR. Dans ce cas, l'étape de clonage est seulement réalisée pour l'identification des haplotypes chez les individus hétérozygotes. Cette dernière approche est généralement réservée aux espèces ou régions géniques peu variables présentant des niveaux d'hétérozygotie assez faible (Toomajian *et al*., 2002 ; Brock *et al*., 2007 ; Hanzawa *et al*., 2008). Dans le cas d'espèces très variables et hétérozygotes comme les arbres forestiers, l'utilisation de logiciels de reconstitution des haplotypes à partir de données génotypiques présente une alternative au clonage systématique des produits d'amplification par PCR souvent long et fastidieux. Parmi ces programmes, PHASE (Stephens *et al*., 2001 ; Stephens et Donnelly, 2003) est le plus utilisé. Plusieurs études basées sur des données empiriques et simulées ont montré l'efficacité de ce logiciel (Stephens et Donnelly, 2003 ; Sabbagh et Darlu, 2005). Une étude récente de Harrigan *et al*. (2008) montre que le logiciel

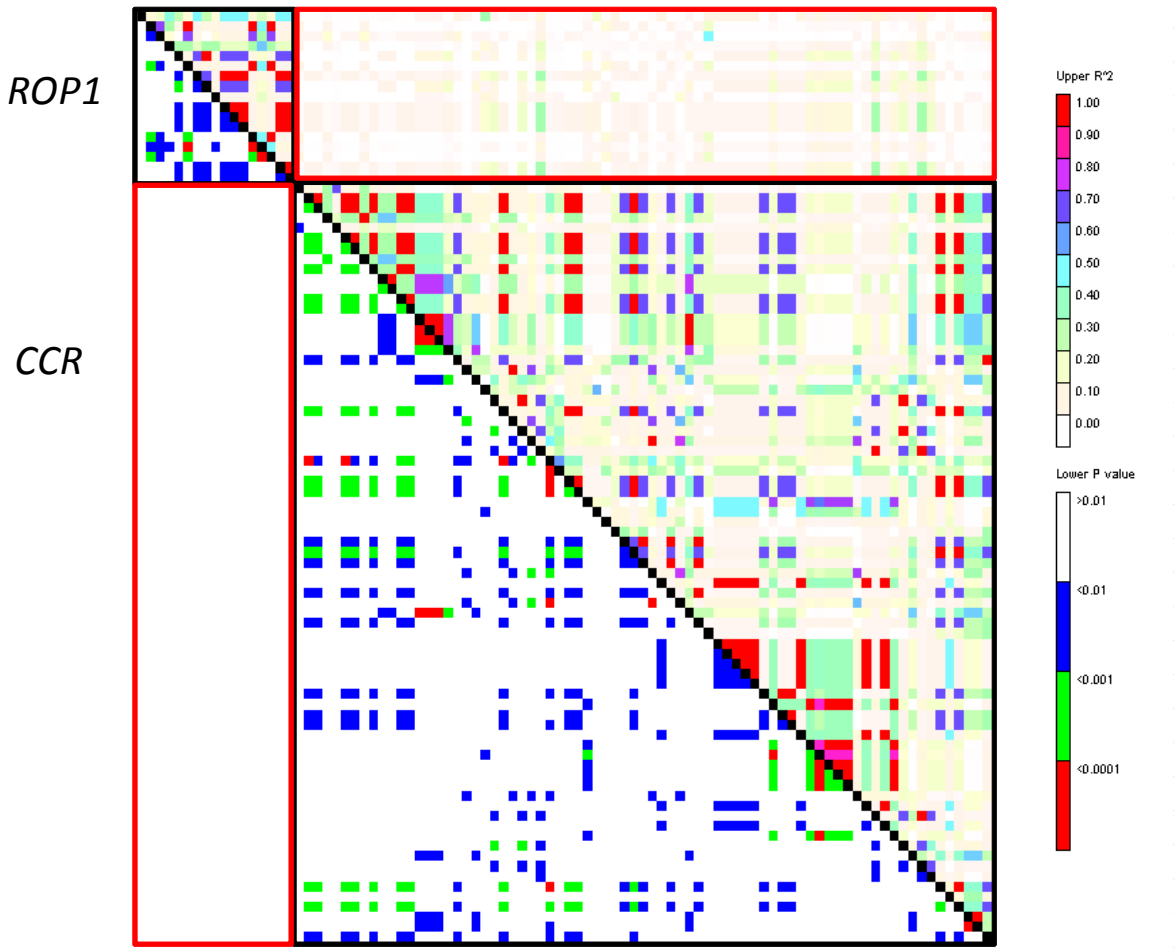


Figure 33: matrice du DL intra et inter-génique pour les gènes *ROP1* et *CCR* au sein d'un échantillon de 8 individus non apparentés de l'espèce *E. urophylla*. Les deux gènes sont situés à une distance génétique de 5 cM sur le groupe de liaison 6. La diagonale représente l'enchainement des SNP sélectionnés (MAF > 15%) dans l'ordre de leur position sur la séquence du gène de 5' en 3'. Au dessus de la diagonale, les valeurs de DL entre paires de sites SNP sont représentées par un carré dont la couleur dépend de la valeur de r^2 obtenue entre les sites étudiés. En dessous de la diagonale, et de la même manière, ce sont les P-values relatives au niveau de significativité du DL mesuré entre paires de sites SNP qui sont représentées. Les valeurs de r^2 et de P-value associées aux mesures de DL entre les deux gènes sont encadrées en rouge et montrent qu'aucune mesure du DL n'est significative au seuil de 1% entre polymorphismes des deux gènes étudiés.

PHASE était plus performant qu'une approche de clonage systématique, même dans le cas de régions géniques très variables et dans le cas d'échantillons de petites tailles. Cependant cette étude considère des fragments de petite taille (250 pb environ) et compare les données de PHASE avec une approche de clonage basée sur le séquençage d'un seul clone par individu. Dans le cas de certains individus hétérozygotes, l'identification des phases par clonage est perturbée par le séquençage d'allèles « chimères ». De tels allèles sont issus d'événements de recombinaison techniques qui ont lieu entre allèles maternels et paternels lors de l'amplification par PCR de portions génomiques au sein d'individus hétérozygotes (Brakenhoff, 1991). Dans le cadre notre étude de la variabilité du gène *CCR* chez *E. urophylla*, les études préliminaires à ce travail de thèse montraient une efficacité limitée du logiciel PHASE pour reconstruire les haplotypes de la longueur totale du gène sur la base de données génotypiques (Champurney, 2003). La longueur du gène (plus de 3kb) associée à une grande quantité de polymorphisme complexifiait les calculs de reconstitution des phases gamétiques. La longueur du gène peut expliquer la perte d'efficacité du logiciel par rapport à l'étude de Harrigan *et al*. (2008). Ces éléments ont motivé l'utilisation d'une approche lourde et fastidieuse de clonage systématique pour la détection de la variabilité nucléotidique des gènes candidats.

Comme l'avaient montré Harrigan *et al*. (2008) et comme nous avons pu le vérifier par le séquençage du gène *CCR* chez *E. urophylla*, l'approche de séquençage de produits PCR clonés implique la prise en compte de deux artéfacts techniques liés à la réaction PCR : les erreurs de réplication de la Taq polymérase et les événements de recombinaison techniques entre allèles d'individus hétérozygotes (« chimères »). Ces artéfacts techniques sont révélés tous deux lors du séquençage et induits par la séparation des fragments d'ADN produits par PCR lors de l'étape de clonage. Les erreurs de réplication de la Taq polymérase doivent pouvoir être identifiées pour ne pas être confondues avec les polymorphismes naturels des gènes. La production d'allèles chimères par la réaction PCR induit elle, une inversion de phase entre les polymorphismes présents au sein de la portion du génome ciblée entre l'allèle paternel et maternel chez les individus hétérozygotes. Les allèles chimères doivent donc être identifiés systématiquement car ils introduisent des événements de recombinaison techniques qui ne doivent pas être pris en compte dans l'identification des haplotypes.

Si les méthodes utilisées dans différentes études indiquent que ces artéfacts ont été pris en compte dans l'approche de séquençage mise en place (Toomajian *et al*., 2002 ; Brock *et al*., 2007 ; Hanzawa *et al*., 2008), aucune ne rapporte les taux d'erreurs détectés lors de l'étape

Tableau 14: comparaison de la variabilité du gène *CCR* au sein d'échantillons de populations d'*E. urophylla* et *E. camaldulensis*. * : 24 pb avant le codon d'initiation. ** : 108 pb après le codon stop. * : 3 SNP bialleliques et un SNP triallélique. [a] : taille de l'échantillon (nombre de copies indépendantes du gène). [b] : SNP trialléliques. TSP SNP : SNP communs aux deux espèces *E. urophylla* et *E. camaldulensis*.**

					Variabilité du gène *CCR chez E. urophylla* et *E. camaldulensis*													
	N [a]		Taille en pb		SNPs (total)		INDELs		poly A		SSR		SNPs tri [b]		Non Synonymes		TSP SNPs	Densité de SNPs 1/x pb	
Région génique	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*	*E. u*	*E. c*		*E. u*	*E. c*
Exon1*			156		1	7	-		-		-		-	-	-	1	1	156.0	22.3
Exon2			156		5	4	-		-		-		-	-	1	1	1	31.2	39.0
Exon3	32	16	183		8	4	-		-		-		-	-	1	-	2	22.9	45.8
Exon4			355		17	21	-		-		-		-	1	2	2	11	20.9	16.9
Exon5**			292		10	7	-		-		-		-	1	-	-	4	29.2	41.7
Total exons			1142		41	43	-		-		-		-	2	4	4	19	27.9	26.6
Intron1	32	16	117	135	2	7	2	1	1		-		-	-	-		1	58.5	19.3
Intron2			689	683	38	32	6	9	-		1		1	-	-		10	18.1	21.3
Intron3			166	166	11	12	1	-	-		-		1	2	-		3 + 1 ***	15.1	13.8
Intron4			1022	1004	64	51	8	9	-		1		2	1	-		31	16.0	19.7
Total introns			1994	1988	115	102	17	19	1		2		4	3	-		42	17.3	19.5
Total gene			3136	3130	156	145	17	19	1		2		4	5	4	4	65	20.1	21.6

de séquençage. Dans le cas du gène *CCR* chez *E. urophylla*, le nombre moyen d'erreur par séquence (0,88 erreurs par séquence en moyenne avec jusqu'à 7 erreurs identifiées sur une même séquence de 683 pb !) ainsi que le taux d'erreur par pb ($1{,}48.10^{-3}$) démontrent que les erreurs de réplication de la Taq polymérase peuvent induire des taux importants de faux variants si elles ne sont pas prise en compte. La fréquence des allèles chimères (7% des séquences obtenues pour le gène *CCR* chez *E. urophylla*) indique également que ces événements doivent être considérés avec attention lors de la mise en place d'une approche de séquençage de produits PCR clonés. Etant donné l'aspect aléatoire de ces deux types d'erreurs techniques, la multiplication des clones séquencés mise en place dans le cadre de cette étude permet de distinguer facilement les vrais événements de mutation et de recombinaison des erreurs techniques d'amplification PCR. La mise en place d'une telle approche, bien que coûteuse en temps de manipulations, reste le moyen le plus sûr d'obtenir des données haplotypiques fiables pour les études d'association ou de diversité génétique des populations.

Aujourd'hui, les outils de séquençage nouvelle-génération (NGS) et l'évolution rapide des techniques qui leur sont associées, permettent de mettre en évidence, en une seule expérimentation, la variabilité nucléotidique associée à un grand nombre de gènes, chez un grand nombre d'individus, même si leur taux d'erreur est de l'ordre de 10 fois supérieur à la technique Sanger. A titre d'exemple, Kulheim *et al.*, 2009 ont récemment utilisé cette méthode chez l'eucalyptus pour détecter la variabilité de 23 gènes impliqués dans la biosynthèse de métabolites secondaires, représentant une longueur de 50 kbp au sein de 1764 individus. Etant donné le nombre de paires de bases générées lors d'une seule expérimentation (plusieurs dizaines de millions), c'est le nombre de répétition obtenu pour chaque base identifiée (profondeur de séquençage) qui permet, comme dans le cas de l'approche de séquençage de produits PCR clonés, de détecter les SNP et de les distinguer des erreurs techniques. Ces méthode ne permettent de générer que de courtes séquences (de l'ordre de 200 pb) mais associées à des méthodes de marquage des individus, elles permettent d'accéder rapidement à l'information des haplotypes (pour peu que des SNP soient identifiées dans des régions chevauchantes) et de leur fréquence sur des régions génomiques de grande taille (il suffit pour cela de produire de nombreux produits PCR) et pour de grands échantillons (Meyer *et al.*, 2007).

Tableau 15: comparaison des niveaux de diversité nucléotidique et haplotypique au sein du gène *CCR* chez *E. urophylla* et *E. camaldulensis.* La diversité nucléotidique π a été estimée sur le total des sites.

	E. urophylla			*E. camaldulensis*		
Taille d'échantillon	32			16		
Portion génique étudiée	region 1	region 2	gène complet	region 1	region 2	gène complet
Taille de séquence en pb	626	2335	2942	627	2391	2927
Nombre de SNPs	18	138	156	30	115	145
Diversité nucléotidique (π par site)	0.0081	0.0144	0.0131	0.0125	0.0148	0.0143
Nombre d'haplotypes K	11	20	20	11	13	
H_d	0.804	0.958	0.958	0.925	0.975	
R_M	2	18	20	3	14	

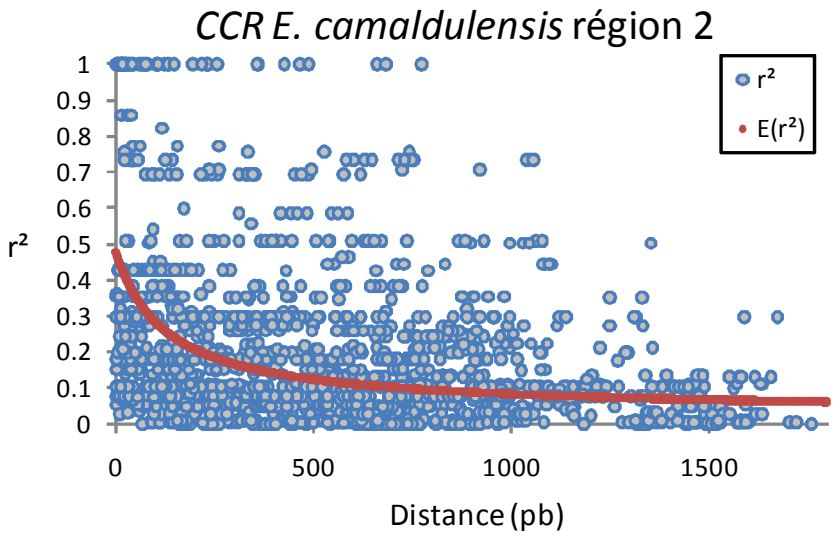


Figure 34: étendue du DL au sein de la région 2 du gène *CCR* chez *E. camaldulensis.*

2.2. Diversité nucléotidique, haplotypique et déséquilibre de liaison chez *E. urophylla*

2.2.1. Variabilité des gènes : densité de SNP et diversité nucléotidique

L'étude de la diversité nucléotidique des plantes a débuté dans les années 1990 avec pour principal objectif d'appréhender la variabilité de différentes portions du génome chez des espèces modèles comme Arabidopsis ou des espèces d'intérêt agronomique comme le maïs, le coton, l'orge, le blé, le sorgho (voir synthèse bibliographique par Wright et Gaut, 2005). Ces études montrent des niveaux de diversité nucléotidique variables entre les espèces mais globalement plus élevés que chez l'Homme. Les premières études menées chez les arbres forestiers sont plus tardives et datent du début des années 2000. Les données les plus importantes ont été obtenues chez le peuplier (Olson *et al.*, 2010) pour lequel un génome de référence est disponible (Tuskan *et al.*, 2006) et, différentes espèces de pin et d'épicéa pour lesquels des ressources génomiques (EST) ont été développées (Savolainen et Pyhäjärvi, 2007). Ces études montrent des niveaux de diversité nucléotidique intermédiaire en comparaison aux autres espèces de plantes et globalement plus faible que ceux détectés chez les plantes de grande culture comme le maïs ou le tournesol.

Très peu de données sont pour le moment disponibles chez l'eucalyptus. Les quelques études menées sur des gènes candidats (Poke *et al.*, 2003 ; Thumma *et al.*, 2005 ; Kulheim *et al.*, 2009 ; Thumma *et al.*, 2009) semblent indiquer des niveaux de variabilité globalement importants. Dans leur étude, Kulheim *et al.* (2009) rapportent des densités de SNPs moyennes (intron et exons confondus) comprises entre 1 SNP/33 pb chez *E. nitens* et 1 SNP/16 pb chez *E. camaldulensis* pour 23 gènes impliqués dans la biosynthèse de métabolites secondaires. Les résultats obtenus dans le cadre de ce travail de thèse pour l'ensemble des gènes candidats de la lignification chez *E. urophylla* (1SNP/42 pb en moyenne) indiquent des niveaux de variabilité moyens équivalents à ceux mis en évidence par Kulheim *et al.* (2009). Ces auteurs indiquent également des densités de SNPs 1,5 fois plus importantes en moyenne dans les régions introniques que dans les régions exoniques pour l'ensemble des espèces étudiées. Les données que nous avons recueillies chez *E. urophylla* indiquent des niveaux de variabilité 2 fois supérieurs au sein des introns et sont donc en accord avec ces résultats. L'ensemble de ces études suggère des niveaux de variabilité des gènes globalement élevés chez l'eucalyptus, malgré des différences entre les espèces, avec des densités de SNP de l'ordre de celles observées chez d'autres espèces forestières comme le chêne (Quang *et al.*, 2008) ou le

peuplier (Chu *et al.*, 2009). Les différences de variabilité observées entre régions exoniques et introniques sont compatibles avec l'effet de la sélection purifiante sur les régions géniques impliquées dans la structure primaire des protéines.

Les données d'estimation du paramètre θ ($4Ne\mu$) obtenues dans le cadre de ce travail de thèse constituent les premières données rapportées chez l'eucalyptus pour un ensemble de gènes répartis sur différentes régions du génome. Ces données confirment les résultats des études de diversité nucléotidique chez d'autres espèces. La Figure 29 montre la répartition en classes de valeurs de $\theta_{\pi\ total}$ (par site) de 211 estimations obtenues pour l'étude de différents gènes chez différentes espèces forestières. Cette figure indique également les classes auxquelles appartiennent les gènes étudiés chez *E. urophylla* dans le cadre de cette étude, la valeur moyenne pour ces gènes et la valeur moyenne pour l'ensemble des gènes inclus dans cette méta-analyse (voir références de la Figure 29). Cette figure montre que la valeur moyenne de $\theta_{\pi\ total}$ (par site) estimée chez *E. urophylla* est supérieure à celle obtenue globalement sur l'ensemble des gènes étudiés chez d'autres espèces d'arbres forestiers. Avec une valeur de $\theta_{\pi\ silent}$ (par site n'impactant pas la structure primaire de la protéine) de 0,0126 estimée pour 9 gènes candidats de la lignification, les niveaux de diversité nucléotidique détectés chez *E. urophylla* sont en effet parmi les plus élevés détectés chez les espèces d'arbres forestiers. Pour la plupart des conifères, les études rapportent des valeurs comprises entre 0,0038 (Kado *et al.*, 2003) et 0,0079 (Gonzalez-Martinez *et al.*, 2006). Chez *Populus tremula* les valeurs estimées sont proches de celles obtenues pour *E. urophylla* (0,016 – 0,013; Ingvarsson *et al.*, 2005, 2008) et suggèrent une taille efficace de population et un taux de mutation global important pour ces espèces (les sites silencieux étant supposés évoluer selon le MNSE considérant des populations non subdivisées de tailles finies au sein desquelles les individus se reproduisent au hasard et ont les mêmes chances de survivre et de se reproduire). Cependant, la variation du niveau moyen de diversité nucléotidique estimé chez différentes espèces peut être causée par différents facteurs tels que la stratégie d'échantillonnage des individus relatifs à l'étude et les locus étudiés. Dans notre cas, le nombre de locus étudié est du même ordre que pour les autres espèces comparées précédemment (cf synthèse bibliographique par Savolainen et Pyhäjärvi, 2007). Cependant, si ces valeurs ont été estimées dans la plupart des cas au sein d'échantillons représentatifs des espèces étudiées, cela n'est pas le cas pour l'échantillon d'individus constituant le panel de détection des SNP chez *E. urophylla*. Bien que Tripiana *et al.* (2007) et Payn *et al.* (2008) indiquent des faibles niveaux de différenciation entre populations chez cette espèce

d'eucalyptus, on peut penser que les niveaux de diversité nucléotidique mis en évidence chez *E. urophylla*, même s'ils sont importants, ont été sous-estimés. Enfin, les valeurs de θ_π estimées pour l'ensemble des sites silencieux (6338) et des sites non synonymes (4283) étudiés chez *E. urophylla* suggèrent, comme pour l'étude de la densité de SNP, une action de la sélection purifiante sur les sites impactant la structure primaire de la protéine.

Parmi les gènes étudiés dans le cadre de ce travail de thèse, les gènes *MYB1* et *MYB2* sont des facteurs de transcription (FT). Ces FT sont capables de se lier de manière spécifique aux régions promotrices de certains gènes et sont des éléments importants de la régulation de l'expression des gènes de la voie de biosynthèse des lignines chez l'eucalyptus (Goicoechea *et al.*, 2005 ; Legay *et al.*, 2008). Etant donné le nombre de gènes potentiellement régulés par ces FT, ces derniers pourraient jouer un rôle déterminant dans la variation de certaines propriétés du bois et notamment les propriétés chimiques relatives aux lignines. Cependant, étant donné leur fort niveau de conservation entre les espèces, et leur effet pléiotropique (les mêmes facteurs de transcriptions sont capables de réguler l'expression de plusieurs gènes), certaines études suggèrent que l'évolution de ces FT serait contrainte par l'effet d'une forte pression de sélection purifiante (Wagner et Lynch, 2008) et que se soient leurs sites de fixation (éléments cis de régulation) situés au sein des zones promotrices des gènes cibles qui portent la variabilité nucléotidique responsable de la variation phénotypique.

Nous n'avons pas lors de ce travail de thèse mené d'investigation sur la diversité des zones promotrices. Il est donc impossible de comparer le niveau de diversité des FT(*MYB1* et *MYB2*) à leurs cibles potentielles. Cependant, l'étude de ces deux FT est riche d'enseignement. Même si leur niveau de diversité nucléotidique globale reste inférieur à la moyenne de l'ensemble des gènes de structure étudiés (ce qui est attendu), la diversité des sites non synonymes indique des niveaux de variabilité supérieurs à la moyenne des gènes de structure. Cela suggère des effets potentiels directs des FT sur la variabilité des phénotypes. Nombreuses sont les études indiquant des coïncidences entre FT et QTL chez les plantes de grande culture. Chez les arbres forestiers, une étude de Kirst *et al.*, (2004) va également dans ce sens. Elle s'est intéressée à la variation de l'expression de gènes majeurs de la lignification. Les auteurs ont d'abord mis en évidence une corrélation entre la croissance et l'expression de gènes de la voie de biosynthèse des lignines au sein d'une famille obtenue par backcross entre un individu *E. urophylla* x *E. globulus* et un individu *E. globulus*. Ils ont ensuite montré que les patrons de variation d'expression des gènes de la lignification étaient corrélés avec la variation de la teneur et de la composition des lignines au sein de cette famille. Enfin, ils ont

positionné les QTLs d'expression pour ces gènes de la lignification et ont mis en évidence deux locus impliqués dans la régulation de leur expression et colocalisant avec des QTL de croissance. Les auteurs ont émis l'hypothèse de l'existence de locus communs impliqués dans la variation de la croissance et des caractères liés aux lignines. Ces locus ne colocalisant pas avec les gènes dont l'expression a été étudiée, les auteurs ont alors suggéré l'intervention de facteurs de régulation *trans* ayant des effets pleïotropiques et dont la variation pourrait impacter à la fois la croissance et les caractères liés aux lignines. Les FT de type *MYB*, étant connus pour interagir avec les gènes de la voie de biosynthèse des lignines (revue bibliographique par Zhong et Yé, 2009 et études de Goicoechea *et al.*, 2005 et Legay *et al.*, 2008 pour l'eucalyptus), les auteurs suggèrent qu'ils sont des bons gènes candidats pour ce type de régulation.

2.2.2. Ecart à la neutralité sélective et à l'équilibre démographique

Les résultats des tests de neutralité indiquent que pour la plupart des gènes étudiés, les patrons de diversité nucléotidique et haplotypique ne s'écartent pas de manière significative du MNSE. La diversité nucléotidique et haplotypique de la majeure partie des gènes étudiés ne semble donc pas montrer de signature de l'effet de la sélection naturelle au sein de l'échantillon étudié. Cependant, des valeurs de D de Tajima significativement différentes de 0 ont été observées pour les gènes *C3H* et *COMT2*. Une valeur positive est observée pour le gène *C3H*. Cette valeur est indicatrice d'un défaut de variants en faible fréquence par rapport à l'attendu sous le MNSE. Ce patron de diversité nucléotidique est caractéristique de ceux observés dans le cas de locus soumis à une sélection balancée favorisant la coexistence de groupes d'haplotypes en fréquences intermédiaires au sein des populations étudiées (Tajima, 1989). Dans le cas du gène *C3H*, seules des données de séquences diploïdes ont été obtenues et il n'est donc pas possible d'identifier les groupes d'haplotypes en présence. Une valeur de D de Tajima négative est observée pour le gène *COMT2* indiquant un excès de variants rares par rapport à l'entendu sous le MNSE. Ce patron de diversité nucléotidique est compatible avec l'effet de la sélection purifiante à ce locus. Cependant, le test du D de Tajima est sensible aux effets démographiques tels que des variations de tailles de population ou l'existence d'une structure génétique dans la population étudiée. Pour conclure à un effet de la sélection sur les gènes *C3H* et *COMT2*, il aurait fallu pouvoir exclure les hypothèses de changement de taille des populations. La différence majeure entre les effets de la sélection et les effets de la dérive génétique (liée à la taille des populations) réside dans le fait que la sélection agit de manière ciblée sur certaines zones du génome d'un organisme alors que les effets de la dérive

génétique se font ressentir sur la diversité nucléotidique de l'ensemble des locus du génome. Dans notre cas, seuls ces deux gènes montrent des écarts significatifs à la neutralité sélective avec des effets opposés. Dans le cas du gène *C3H*, l'hypothèse de l'existence d'une sélection balancée à ce locus est renforcée par le fait que la valeur de D de Tajima moyen estimée pour l'ensemble des gènes de la voie de biosynthèse des lignines est sensiblement négative. Ces résultats devraient faire néanmoins l'objet d'analyses plus poussées pour conclure à un effet de la sélection naturelle sur la diversité nucléotidique de ces 2 locus.

2.2.3. Diversité haplotypique et étendue du DL

Les indices de diversité haplotypique (nombre d'haplotypes et diversité haplotypique) estimés pour 7 des gènes étudiés chez *E. urophylla* présentent des valeurs modérées (*COMT2*, *CAD2*, *MYB2*) à fortes (*C4H*, *CCR*, *ROP1*) en comparaison aux valeurs estimées chez d'autres espèces d'arbres forestiers. Chez *Pseudotsuga mienzesii* (18 gènes de la qualité du bois), *Pinus teada* (18 gènes candidats impliqués dans la réponse aux stress hydriques), *Pinus pinaster* et *Pinus radiata* (8 gènes impliqués dans la formation du bois), *Picea abies* (22 gènes), *Populus nigra* (9 gènes), les valeurs de diversité haplotypiques moyennes estimées varient de 0,376 à 0,931 (Krutovsky et Neale, 2005 ; Gonzalez-Martinez *et al.*, 2006 ; Pot *et al*., 2005 ; Heuertz *et al.*, 2006 ; Chu *et al.*, 2009) pour des échantillons de taille équivalente à celui utilisé dans le cadre de notre travail (23 à 48 copies indépendantes des gènes généralement échantillonnés sur l'aire de répartition des espèces). La valeur moyenne estimée sur l'ensemble des gènes candidats de la lignification chez *E. urophylla* est de l'ordre de celle estimée chez *Pseudotsuga* par Krutovsky et Neale (2005). Cependant, pour un échantillon qui dans le cas de notre étude, n'est pas représentatif de l'aire de répartition naturelle de l'espèce on peut penser que la valeur moyenne estimée de la diversité haplotypique chez *E. urophylla* était sous-estimée par rapport aux valeurs obtenues chez les autres espèces.

Ces niveaux de diversité haplotypique élevés sont associés à une faible étendue du DL au sein des gènes candidats de la lignification chez *E. urophylla*. Les résultats obtenus constituent les premières estimations de l'étendue du DL chez l'eucalyptus et son cohérents d'une part, avec les niveaux élevés d'hétérozygotie détectés dans cette étude et d'autre part, avec le régime de reproduction principalement allogame de cette espèce (Tripiana *et al.*, 2007). Ces résultats sont également cohérents avec ceux obtenus chez d'autres espèces d'arbres forestiers qui montrent une décroissance généralement rapide du DL. Les données indiquent des valeurs de DL entre sites polymorphes très faibles ($r^2 < 0,2$) au-delà de 1000 pb

et parfois moins chez *Pinus taeda*, *Pinus sylvestris*, *Picea Abies* ou *Populus Tremula* (Neale et Savolainen, 2004 ; Brown *et al.*, 2004 ; Dvornyk *et al.*, 2002 ; Rafalski et Morgante, 2004, Heuertz *et al.*, 2006 ; Ingvarsson *et al.*, 2005 ; Savolainen et Pyhäjärvi, 2007). Ces patrons de DL contrastent avec ceux mis en évidence chez des espèces principalement autogames comme Arabidopsis au sein de laquelle le DL peut être maintenu sur des distances de 20 kb dans des échantillons représentatifs de la diversité de l'espèce (Remington *et al.*, 2001). Chez les arbres, l'estimation de la décroissance du DL a été étudiée principalement au sein de gènes ou portions de gènes. Ces données sont généralement estimées sur de faibles distances de l'ordre de quelques kb. Il n'existe pas d'estimation de l'étendue du DL sur de longues distances malgré le fait que ces études indiquent des valeurs ponctuelles pouvant être fortes au delà de quelques kb. Les résultats obtenus ici sur la décroissance du DL entre *CCR* et *ROP1* (situés sur le groupe de liaison 6 et distants de 11 cM) semblent indiquer (une généralisation nécessiterait une étude plus poussée) que le DL n'est généralement pas maintenu sur de longues distances entre paires de sites polymorphes. Au final, nous pouvons considérer que les estimations réalisées sur la base de plusieurs régions géniques de quelques kb sont représentatives du DL moyen observé chez cette espèce.

2.3. Comparaison avec d'autres espèces d'Eucalyptus : cas du gène *CCR* chez *E. urophylla*, *E. camaldulensis* et *E. globulus*

Le séquençage complet du gène *CCR* réalisé chez *E. urophylla* et *E. camaldulensis* nous a permis de comparer les données obtenues chez ces espèces avec celles obtenues par Poke *et al.* (2003) pour *E. globulus*. L'ensemble de ces données font apparaitre des taux de variabilité assez importants au sein du gène *CCR* pour les 3 espèces considérées. Avec 1 SNP/48 pb au sein des exons et 1 SNP/33 bp au sein des introns, *E. globulus* est apparue comme l'espèce la moins variable des 3 étudiées. La densité de SNP la plus forte dans les régions exoniques a été détectée chez *E. camaldulensis* et dans les régions introniques chez *E. urophylla*. Cependant, si on considère la taille de l'échantillon utilisé pour ces trois espèces (46 copies pour *E. globulus*, 32 copies pour *E. urophylla* et 16 copies pour *E. camaldulensis*) on peut penser que les niveaux de variabilité du gène *CCR* chez *E. camaldulensis* sont largement sous-estimés par rapport à ceux des deux autres espèces. Une diversité (π_{total}) plus élevée chez *E. camaldulensis* est néanmoins confirmée malgré un échantillon de taille inférieure sur la base de données obtenues sur les autres gènes. Les données de diversité haplotypique et de

l'étendue du DL sont également en faveur d'une diversité plus forte chez *E. camaldulensis* comparé à *E. urophylla*.

Ces résultats sont cohérents avec ceux obtenus par Kulheim *et al.* (2009) qui indiquent également des taux de variabilité plus importants chez *E. camaldulensis* comparé aux espèces, *E. loxophera*, *E. nitens* et *E globulus* pour 23 gènes impliqués dans la synthèse des métabolites secondaires. Les auteurs expliquent que ces résultats sont en accord avec la taille des aires de répartition des espèces étudiées, celle d'*E. camaldulensis* étant la plus importante parmi toutes les espèces d'eucalyptus. Les auteurs indiquent également un ensemble de SNP communs pour les espèces étudiées avec des pourcentages de SNP partagés par plusieurs espèces compris entre 26% et 40% pour les exons et 21% et 43% pour les introns. Les auteurs suggèrent que la majeure partie de ces SNP pourraient avoir une origine ancestrale commune chez toutes ces espèces expliquant que dans ce cas, d'autres gènes de voies de biosynthèse différentes devraient présenter les mêmes tendances. Les données obtenues dans le cadre de notre étude indiquent bien l'existence d'un ensemble de SNP communs entre *E. urophylla* et *E. camaldulensis*. Sur la totalité du gène *CCR*, 65 SNP sont communs aux deux espèces, soit un total de 42% et 45% de SNP partagés chez ces 2 espèces respectivement. L'espèce *E. urophylla* étant une espèce endémique de l'archipel des îles de la Sonde, elle ne partage pas la même aire de répartition naturelle que les autres espèces d'eucalyptus. Les hybridations naturelles entre espèces d'eucalyptus (rapportées pour *E. camaldulensis* par exemple avec *E. tereticornis* (Butcher et al., 2002)) ne peuvent vraisemblablement pas être à la base de l'existence de ces polymorphismes partagés entre *E. urophylla* et *E. camaldulensis*. L'âge de séparation de la clade des *E. urophylla* de ses taxons frères d'Australie est estimé entre 2 et 5 millions d'années sur la base de données climatiques et tectoniques (Ladiges *et al.*, 2003) et entre 7 et 20 millions d'années sur la base de données moléculaires (Crisp *et al.*, 2004). La conservation de SNP ancestraux communs sur de telles périodes suggère un maintien de ces SNP par le fait de la sélection ou par des tailles de populations suffisamment importantes pour qu'ils soient conservés au cours du temps. La première hypothèse semble moins vraisemblable car peu de SNP non synonymes sont conservés au sein du gène *CCR* entre les espèces *E. camaldulensis* et *E. urophylla* (1 seul SNP non synonyme a été détecté en commun dans notre étude). Il se pourrait donc que ces polymorphismes communs aux différentes espèces d'eucalyptus puissent être présents dans une grande partie de leurs génomes. Ces polymorphismes pourraient constituer une source de marqueurs intéressants pour l'amélioration génétique des espèces hybrides. On pourrait par exemple envisager de

développer un ensemble de marqueurs universels qui permettent d'étudier le déterminisme génétique des caractères quantitatifs chez l'ensemble des espèces d'eucalyptus qui présentent un intérêt commercial.

2.4. Diversité génétique, DL et études d'association chez l'eucalyptus

L'ensemble de ces résultats indique des forts niveaux de variabilité nucléotidique et une faible étendue du DL pour *E. urophylla* et *E. camaldulensis*. La variabilité génétique étant la « matière première » de l'améliorateur, les résultats obtenus laissent espérer chez ces espèces faiblement domestiquées un fort potentiel d'amélioration génétique pour les caractères d'intérêt et principalement ceux dont la variation est importante et présente un contrôle génétique fort.

Les études d'association permettent d'établir des liens statistiques entre la variabilité nucléotidique des génomes et la variation des caractères quantitatifs en populations. Dans ce cadre, la faible étendue du DL observée chez *E. urophylla* et *E. camaldulensis* indique globalement une faible dépendance de ségrégation entre les polymorphismes en population. Cette propriété confère un avantage en termes de résolution des études d'association. Il sera possible de distinguer les variants causaux parmi un ensemble de polymorphismes physiquement liés situés à une distance relativement faible les uns des autres (au sein d'un même gène par exemple). Cependant, elle implique une description fine de la variabilité en présence et donc la nécessité de mener des efforts importants en termes de séquençage et génotypage pour l'identification de ces variants causaux. Pour ces raisons, la stratégie « gène candidat », visant à tester l'effet de la variation de quelques gènes, sélectionnés sur la base de données indiquant leur implication potentielle dans la variation des caractères d'intérêt, semble être la mieux adaptée à la mise en évidence des polymorphismes d'intérêt agronomique chez l'eucalyptus, avant que des stratégies plus globales ne soient entreprises grâce au séquençage d'un génome de référence (disponible début 2011) et au reséquençage de génotypes chez plusieurs espèces (en cours dans quelques laboratoires).

Chapitre 5 : Association entre variabilité des gènes de la lignification et variation de caractères d'intérêt agronomique chez *E. urophylla*

Faisant suite à l'étude du déterminisme génétique des caractères relatifs aux lignines (LK% et S/G) et à la description des patrons de diversité nucléotidique des gènes de la lignification (*4CL*, *C3H*, *C4H*, *F5H*, *COMT2*, *CAD2*, *CCR*, *MYB1*, *MYB2* et *ROP1*), la dernière partie de ce travail de thèse consistait à tester si la variabilité de ces gènes pouvait expliquer une partie de la variation des caractères quantitatifs étudiés (hauteur, circonférence à 1,30 m, densité du bois, teneur en lignines et rapport S/G).

Un premier objectif était de confirmer, dans une population à base génétique plus large, les co-localisations entre les gènes *CCR* et *ROP1* et des QTL de KL% et S/G (Gion *et al.*, 2001 ; Foucart *et al.*, 2009). Dans ce cadre, ce travail devait permettre de préciser les effets de ces deux gènes en discriminant, au sein des gènes, les polymorphismes liés statistiquement à la variation des caractères quantitatifs. Dans un deuxième temps, il s'agissait, de détecter de nouveaux gènes impliqués dans la variation de la teneur en lignines et du rapport S/G, et également de tester l'effet de l'ensemble des gènes sur d'autres caractères d'intérêt agronomique (croissance et densité du bois). Les tests d'association devant être réalisés dans des plans de croisement factoriels mis en place dans le cadre de programmes d'amélioration des eucalyptus, un troisième objectif de ce travail était d'évaluer la possibilité d'utiliser de tels dispositifs pour détecter des associations entre variabilité nucléotidique et variation de caractères quantitatifs chez l'eucalyptus.

Ce chapitre rapporte d'abord l'effet de la variabilité des gènes *4CL*, *C3H*, *C4H*, *F5H*, *COMT2*, *CAD2*, *CCR*, *MYB1*, *MYB2* et *ROP1* testé sur la variation des caractères de croissance (hauteur et circonférence à 1,30 m), de la densité du bois, de la teneur en lignines et du rapport S/G dans un plan de croisement factoriel *E. urophylla* x *E. urophylla*. Un accent particulier est mis sur le gène *CCR*, identifié sur la base de plusieurs études (génomique fonctionnelle, cartographie de QTL) comme le meilleur gène candidat pour le contrôle de la variation des caractères relatifs aux lignines. Ce chapitre rapporte également les difficultés qui

Tableau 16: séquence et taille des allèles mis en évidence pour le motif microsatellite de l'intron 4 du gène *CCR* chez les 16 génotypes parentaux du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Pour chacun des génotypes numérotés de 1 à 16, les 2 allèles du motif microsatellite détectés sont indiqués. Les génotypes parentaux 6 et 15 sont homozygotes pour l'ensemble de la région séquencée du gène *CCR* (indiqué par [1]).

Génotypes parentaux	Séquence du motif microsatellite	Taille du motif (bp)
1; 3; 7	(TT)2(CT)3(CA)1(CT)5	22
3; 4; 12; 10	(TT)2(CT)10	24
1; 5; 12; 7	(TT)2(CT)13	30
2; 14; 6^1; 6^1; 13	(TT)1(CT)14	30
11	(TT)2(CT)4(GT)1(CT)3(GT)1(CT)5	32
8; 13	(TT)2(CT)4(GT)1(CT)3(GT)1(CT)3(GT)1(CT)1	32
16	(TT)2(CT)15	34
14	(TT)2(CT)16	36
9	(TT)2(CT)4(GT)1(CT)12	38
10	(TT)2(CT)18	40
9	(TT)2(CT)4(GT)1(CT)14	42
16	(TT)2(CT)4(GT)1(CT)15	44
2; 11; 8; 15^1; 15^1	(TT)2(CT)4(GT)1(CT)16	46
5	(TT)2(CT)22(GT)1(CT)1	52
4	(TT)1(CT)26(GT)1(CT)1	58

ont été rencontrées pour l'étude de l'effet de la variabilité du gène *CCR* dans la descendance d'un plan de croisement interspécifique *E. camldulensis* x *E. urophylla*.

1. Résultats

1.1. Génotypage et sélection des SNP

1.1.1. Factoriel E. urophylla x E. urophylla

1.1.1.1.Cas du gène CCR

La population d'association était initialement composée de 328 individus répartis dans 33 familles de plein-frères. Ces familles sont issues des croisements de 16 parents *E. urophylla* selon un plan de croisement factoriel incomplet. La totalité de la séquence du gène *CCR* est disponible ainsi que l'information des haplotypes pour l'ensemble des 16 parents. Les résultats du séquençage et de la mise en évidence de la variabilité du gène sont exposés au chapitre précédent. Dans le cadre de ce travail nous avons mis en évidence un motif microsatellite au sein de l'intron 4 du gène *CCR* (motif di-nucléotide CT de 22 à 58 pb). La variabilité de ce SSR est suffisante pour distinguer chaque allèle du gène pour chacun des parents, malgré des allèles communs entre génotypes parentaux (Tableau 16). Au regard de sa variabilité nucléotidique totale, le gène *CCR* présente 20 allèles distincts pour les 16 parents étudiés. Un total de 15 allèles peuvent être différenciés grâce à la séquence du motif microsatellite et seulement 13 au regard de sa taille (nombre de répétition du motif). Au sein des descendances du plan de croisement factoriel *E. urophylla* x *E. urophylla*, le polymorphisme de taille du motif microsatellite est suffisant pour caractériser la ségrégation de tous les allèles parentaux au sein de chacune des 33 familles de plein-frères (toutes les classes génotypiques attendues dans les descendances peuvent être caractérisées). Ce motif microsatellite hypervariable a donc été utilisé pour caractériser la variabilité nucléotidique du gène *CCR* dans chaque famille de plein-frères en confrontant les données de génotypage microsatellite des descendants avec l'information de séquençage obtenue chez les parents. En effet, étant donnée la taille du gène *CCR* (environ 3 kbp) et le nombre d'individus dans chaque famille de plein-frères (10 individus par famille environ), aucun événement de recombinaison n'est attendu pour le gène *CCR* entre la variabilité des parents et celle observable au sein de la descendance.

Tableau 17: tests de ségrégation mendélienne des allèles du gène *CCR* dans la descendance de chacun des parents du plan de croisement factoriel *E. urophylla x E. urophylla*. N : effectifs totaux par parent. N(Allèle 1) et N(Allèle 2) : effectifs pour les Allèles 1 ou 2 transmis à la descendance. ns : différence non significative avec les proportions mendéliennes. [1] : génotypes parentaux homozygotes pour le gène *CCR*.

Génotypes parentaux		N	Allèle 1	Allèle 2	N(Allèle 1)	N(Allèle 2)	chi2 (1 ddl, α=5%)
Pères	1	36	h9	h8	20	16	0.22 ns
	2	37	h5	h1	21	16	0.34 ns
	3	27	h3	h9	18	9	1.50 ns
	4	36	h4	h15	19	17	0.06 ns
	5	39	h10	h20	20	19	0.01 ns
	6[1]	38	h5	.	38	.	
	7	47	h16	h19	24	23	0.01 ns
	8	48	h2	h1	30	18	1.50 ns
Mères	9	33	h6	h7	18	15	0.14 ns
	10	39	h4	h11	21	18	0.12 ns
	11	45	h1	h12	30	15	2.50 ns
	12	46	h16	h3	25	21	0.17 ns
	13	29	h2	h5	18	11	0.84 ns
	14	50	h13	h14	29	21	0.64 ns
	15[1]	28	h1	.	28	.	
	16	38	h18	h17	20	18	0.05 ns

Parmi les génotypes parentaux étudiés, 14 sont hétérozygotes pour le gène *CCR* et deux sont homozygotes, le génotype « père » numéro 6 et le génotype « mère » numéro 15. Le croisement entre ces deux génotypes n'étant pas présent dans le plan factoriel, seuls deux types de ségrégations mendélienne sont attendues : i/ dans le cas de deux parents hétérozygotes, quatre classes de descendants ségrégeant dans les proportions 1/4 :1/4 :1/4 :1/4, ii/ dans le cas d'un parent hétérozygote et d'un autre homozygote, deux classes de descendants dans les proportions 1/2 :1/2. Les ségrégations des allèles parentaux, ont été vérifiées pour chaque parent hétérozygote (en raison du faible effectif par famille). Le Tableau 17 présente pour chacun des parents, les effectifs observés au sein des différentes descendances pour les deux allèles, avec la valeur du $\chi 2$ associé à l'hypothèse de ségrégation mendélienne. Pour chaque parent, la ségrégation des allèles observée au sein de ces descendances est cohérente avec une ségrégation mendélienne. Au sein des familles de plein-frères, les allèles parentaux sont globalement transmis de manière homogène à la descendance. Cependant, dans une famille impliquant le père 8 et la mère 11, seul l'allèle h2 du père est transmis aux huit descendants.

Cette approche a permis de mettre en évidence la variabilité du gène *CCR* au sein des descendants des 33 familles de plein-frères du plan factoriel. Pour des problèmes liés à la qualité de l'ADN extrait, le génotype de 20 descendants répartis dans ces familles n'a pas pu être caractérisé. Au total, 308 individus ont été utilisés pour tester l'association entre variabilité du gène *CCR* et variation des caractères d'intérêt. Les données de génotypage du microsatellite ont donc permis de caractériser les génotypes des descendants pour 152 SNP bi-alléliques ainsi que 16 INDEL bi-alléliques. Parmi ces 168 polymorphismes, seuls ceux présentant une FAM supérieure ou égale à 5% dans la totalité de la descendance du plan de croisement factoriel ont été retenus pour l'analyse, soit 103 polymorphismes. Certains de ces SNP sont en DL complet ($r^2=1$) et dans ce cas, un seul SNP a été conservé. Finalement, 54 SNP ($r^2<1$) ont été utilisés pour tester l'effet de la variabilité nucléotidique du gène *CCR*. Ces SNP sont globalement bien répartis le long du gène.

1.1.1.2.Cas des autres gènes candidats de la lignification

Pour les 9 autres gènes de la lignification (*4CL*, *C4H*, *C3H*, *F5H*, *COMT2*, *CAD2*, *MYB1*, *MYB2*, *ROP1*), 231 SNP ont été identifiés après séquençage des 16 génotypes, à raison de 26 SNP en moyenne par gène (ces donnés de variabilité nucléotidique sont également exposées au chapitre précédent). Pour ces gènes, les données de séquençage n'ont été que

Tableau 18: part des SNP détectés par séquençage au sein de 9 gènes, utilisables pour les tests d'association au sein du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Les SNP génotypés satisfont aux critères imposés par la méthode de génotypage Sequenom MassARRAY iPLEX Gold. Parmi les SNP génotypés, les SNP utilisables sont de bonne qualité et polymorphes dans la population d'association.

Gènes Candidats	SNP détectés (1)	SNP génotypés (2)	$P_{(1)/(2)}$ (%)	SNP utilisables (3)	$P_{(1)/(3)}$ (%)
4CL	21	10	48	8	38
C3H	22	4	18	1	5
C4H	48	8	17	6	13
CAD2	30	10	33	4	13
COMT2	19	10	53	6	32
F5H	14	5	36	4	29
MYB2	14	6	43	4	29
MYB1	14	4	29	4	29
ROP1	49	11	22	9	18
Total	231	68	29	46	20

partielles (seul un fragment du gène a été séquencé) et n'ont pas été obtenues pour la totalité des génotypes parentaux. Une méthode de génotypage par spectrométrie de masse (Sequenom MassARRAY iPLEX Gold) a été utilisée dans ce cas pour identifier le génotype des 328 descendants du plan de croisement. Cette méthode n'a permis de caractériser qu'une faible proportion des 231 SNP identifiés, en effet 71% des SNP ne répondaient pas aux critères techniques imposés : i/ absence de variabilité dans les régions des gènes bordant les SNP à génotyper, ii/ compatibilité des amorces (multiplexage) pour l'amplification en mélange des SNP ciblés. Ainsi, seuls 68 SNP ont pu être pris en compte (Tableau 18). Parmi ces SNP, 22 sont de mauvaise qualité ou monomorphes dans la descendance du plan de croisement factoriel et n'ont pas été conservés pour l'analyse. Au final, 20% des SNP détectés par séquençage sont utilisables pour l'étude d'association.

Comme dans le cas du gène *CCR*, les SNP dont la FAM est inférieure à 5% dans le plan de croisement *E. urophylla* x *E. urophylla* ont été écartés pour l'analyse d'association. Au total, 36 SNP représentant la variabilité des 9 gènes candidats de la lignification ont été conservés. Au sein de chaque gène, le DL entre SNP n'a pas été testé dans la descendance cependant, pour les gènes *F5H*, *C4H*, *COMT2*, *CAD2*, *MYB2* et *ROP1*, les SNP sélectionnés ne sont pas en DL complet ($r^2=1$) au sein des 16 génotypes *E. urophylla* séquencés. L'information de génotypage au sein de la descendance du plan de croisement est très hétérogène en fonction des gènes : le gène *4CL* est le mieux représenté avec 7 SNP, *ROP1* avec 6 SNP, *C4H* avec 5 SNP, *F5H*, *MYB1* et *MYB2* avec 4 SNP, *CAD2* avec 3 SNP, *COMT2* avec 2 SNP et enfin *C3H* avec un seul SNP.

1.1.2. Factoriel E. camaldulensis x E. urophylla

Au sein de l'espèce *E. camaldulensis*, la variabilité nucléotidique n'a été étudiée que pour le gène *CCR* (ces données de variabilité sont décrites au chapitre 4). De la même manière que pour les gènes *4CL*, *C4H*, *C3H*, *F5H*, *COMT2*, *CAD2*, *MYB1*, *MYB2* et *ROP1* dans le dispositif *E. urophylla* x *E. urophylla*, les données de la variabilité nucléotidique du gène *CCR* ne sont pas disponibles pour la totalité des génotypes parentaux du plan de croisement factoriel *E. camaldulensis* x *E. urophylla*. La méthode Sequenom MassARRAY iPLEX Gold a donc été utilisée pour génotyper la variabilité du gène dans les descendances du plan de croisement. Le fait d'utiliser cette méthode de génotypage pour des individus hybrides de deux espèces présente quelques difficultés. Le séquençage du gène *CCR* chez *E. urophylla* et *E. camaldulensis* a révélé une variabilité nucléotidique importante de

Tableau 19: résultats de l'analyse de variance à deux facteurs pour le test des effets des haplotypes du gène *CCR* au sein des descendances de chacun des parents du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Seuls les tests montrant un effet « Haplotype » significatif sont présentés. ** : effet « Haplotype » significatif au seuil de 1%. * : effet « Haplotype » significatif au seuil de 0,1%. H favorable indique les haplotypes associés à une augmentation de la valeur du caractère et PVE est le pourcentage de variance expliquée par le facteur « Haplotype ».**

				Analyse de variance à deux facteurs						PVE	H favorable
Gène	Caractère	Géniteur	Génotype	Source	ddl	Som carrés	Moy carrés	F	Pr > F		
CCR	S/G	10	(H4;H11)	Modèle	4	1.249	0.312	5.24	0.002	12.9%	H4
				Famille	3	0.957	0.319	5.35	0.004		
				Haplotype	1	0.443	0.443	7.42	0.01 **		
				Erreur	34	2.028	0.060				
		5	(H10;H20)	Modèle	4	1.264	0.316	5.93	0.001	18.6%	H10
				Famille	3	0.665	0.222	4.16	0.01		
				Haplotype	1	0.564	0.564	10.59	0.003 **		
				Erreur	34	1.811	0.053				
		2	(H1;H5)	Modèle	4	1.322	0.331	6.34	0.0007	17.5%	H1
				Famille	3	1.047	0.349	6.69	0.001		
				Haplotype	1	0.576	0.576	11.05	0.002 **		
				Erreur	32	1.669	0.052				
	KL%	2	(H1;H5)	Modèle	4	12.047	3.012	7.60	$<10^{-4}$	13.2%	H5
				Famille	3	9.915	3.305	8.34	$<10^{-4}$		
				Haplotype	1	3.439	3.439	8.68	0.006 **		
				Erreur	32	12.683	0.396				

1 SNP/20 pb en moyenne chez ces deux espèces. Bien que certains SNP soient communs à *E. urophylla* et *E. camaldulensis*, une part importante de la variabilité nucléotidique est propre à chacune des espèces. Si elle est considérée en mélange, comme c'est le cas dans un contexte de populations d'hybrides interspécifiques, cette variabilité correspond à une densité de SNP de 1 SNP/13 pb. En raison des contraintes imposées par la méthode de génotypage Sequenom (faible variabilité dans les régions bordant les SNP), celle-ci n'est pas conçue pour génotyper des SNP dans des contextes de variabilité si forte. Aussi, très peu de SNP peuvent être génotypés. Pour diminuer les contraintes sur cette étape, les SNP propres aux deux espèces ont été génotypés de manière indépendante chez les descendants hybrides *E. urophylla* x *E. camaldulensis*. Les données de génotypage devaient permettre dans chacune des 16 familles de plein-frères, de reconstituer le génotype des descendants sur la base des allèles hérités du géniteur *E. camaldulensis* d'une part et *E. urophylla* d'autre part. Deux expériences de génotypage indépendantes ont donc été mises en place.

Pour *E. urophylla*, la variabilité nucléotidique du gène *CCR* n'étant pas connue chez les parents du dispositif expérimental, un total de 103 SNPs bi-allèliques a été soumis pour la conception de l'expérience de génotypage. Ces SNP correspondent à l'ensemble des 152 SNPs bi-allèliques détectés par séquençage du gène *CCR* complet chez 16 individus non apparentés auxquels les 49 singletons, considérés comme des événements de mutation rares, ont été retranchés. Au final, 21 SNP seulement ont pu être pris en compte par la méthode de génotypage. Dans le cas de l'espèce *E camaldulensis*, la totalité des 145 SNP détectés au sein du gène chez les parents du plan de croisement ont été soumis. Comme pour *E. urophylla*, seuls 21 SNP ont pu être pris en compte par la méthode. Ces deux sous-ensembles de SNP comprennent 5 SNP communs aux deux espèces.

Le génotypage a été réalisé sur l'ensemble des 182 descendants et des 12 géniteurs afin de pouvoir confirmer les génotypes obtenus pour les descendants par l'étude des parents. L'analyse des données révèle des incohérences dans les ségrégations de l'ensemble des SNP de chaque espèce entre les génotypes parentaux et les génotypes des descendants. De plus, beaucoup de données sont manquantes. Ces résultats peuvent avoir une origine multiple, parmi lesquelles une incompatibilité de la méthode avec le génotypage des hybrides ou une mauvaise information entre les parents et les descendances du dispositif expérimental, ce qui est monnaie courante dans un programme d'amélioration et enfin une possible fixation des amorces d'extension ou d'amplification sur des locus homologues comme par exemple le *CCR2* mis en évidence chez *E. urophylla* lors de l'étude de la diversité nucléotidique du gène

Tableau 20: résultats de l'analyse marqueur par marqueur selon le modèle mixte prenant en compte l'apparentement dans la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla*. Seuls les résultats des SNP montrant un effet significatif sur la variation de S/G sont présentés. Les mentions NS, S et UTR indiquent pour les SNPs détectés dans les exons s'ils sont non synonymes, synonymes ou dans la région non traduite du gène respectivement. R^2 est la part de variance phénotypique expliquée par le marqueur. SNP r^2=1 indique les SNP du gène *CCR* qui sont en DL complet avec le SNP testé et qui n'ont pas été inclus dans l'analyse.

Locus	Région		Allèles	p-value	Q-value	R^2	FAM	SNP r^2=1
SNP2	Intron 1		[T/G]	3.76E-02	7.11E-02			
SNP6	Exon 2	NS	[A/G]	6.80E-03	3.05E-02	1.18%	0,19	SNP9, SNP14, SNP20, SNP39, SNP63, SNP65, SNP86
SNP12	Intron 2		[A/G]	3.37E-02	7.11E-02			
SNP30	Intron 2		[A/G]	6.95E-04	8.32E-03	1.52%	0,13	
SNP35	Intron 2		[G/C]	6.40E-04	8.32E-03	1.61%	0,06	
SNP49	Exon 3	S	[T/C]	4.40E-03	2.63E-02	1.03%	0,06	SNP49
SNP50	Exon 3	S	[A/G]	1.60E-03	1.44E-02	1.39%	0,08	
SNP74	Exon 3	S	[T/C]	1.38E-02	4.51E-02	1.08%	0,08	
SNP87	Intron 4		[T/G]	8.80E-03	3.51E-02	1.17%	0,17	SNP95, SNP102
SNP94	Intron 4		[A/G]	1.61E-02	4.82E-02	1.19%	0,16	SNP107, SNP112
SNP117	Intron 4		[T/C]	1.00E-02	3.59E-02	1.22%	0,10	
SNP138	Intron 4		[A/G]	6.10E-04	8.32E-03	1.78%	0,38	
SNP147	Exon 5	UTR	[A/G]	3.20E-03	2.30E-02	1.66%	0,23	SNP31

	SNP30	SNP35	SNP49	SNP50	SNP74	SNP87	SNP94	SNP117	SNP138	SNP147
SNP6	0.61	0.25	0.25	0.37	0.39	0.85	0.05	0.01	0.38	0.07
SNP30		0.41	0.41	0.60	0.24	0.43	0.03	0.04	0.24	0.04
SNP35			0.30	0.68	0.18	0.30	0.01	0.16	0.10	0.02
SNP49				0.68	0.65	0.07	0.01	0.16	0.10	0.02
SNP50					0.43	0.18	0.02	0.09	0.14	0.03
SNP74						0.19	0.02	0.09	0.15	0.03
SNP87							0.04	0.02	0.32	0.06
SNP94								0.02	0.12	0.65
SNP117									0.17	0.01
SNP138										0.07

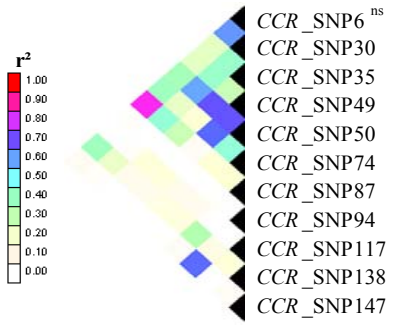


Figure 35: DL (r^2) entre les SNP du gène *CCR* associés à S/G dans la descendance du plan de croisement factoriel *E. urophylla* x *E. urophylla*.

CCR. De plus amples analyses permettraient de répondre à ces questions (génotypage microsatellite des descendants et des parents pour l'identification des relations de parenté, séquençage des parents *E. urophylla* pour l'étude complète de leur variabilité nucléotidique en ségrégation dans le plan de croisement). Pour des contraintes de temps ces études n'ont pas pu être réalisées. Ces données de génotypage ont été utilisées pour les 5 SNP communs aux deux espèces *E. urophylla* et *E. camaldulensis*, pour lesquels les résultats des deux expériences pouvaient être confrontés.

1.2. Tests d'association entre variabilité des gènes de la lignification et variation des caractères relatifs aux lignines

1.2.1. Cas du gène CCR dans le plan de croisement E. urophylla x E. urophylla

1.2.1.1.Effet des haplotypes parent par parent

Dans un premier temps l'effet des haplotypes de chaque parent a été testé grâce à une analyse de variance à deux facteurs (Famille et Haplotype) réalisée sur l'ensemble des familles d'un parent considéré. Cette analyse, basée sur l'information du génotypage microsatellite, a été réalisée pour chacun des 14 parents hétérozygotes du plan de croisement factoriel. Les résultats de l'analyse de variance pour les parents présentant un effet « Haplotype » significatif sur le rapport S/G (S/G) et la teneur en lignines (LK%) sont présentés au Tableau 19. Un effet « Haplotype » significatif sur la variation de S/G est observé pour trois des 14 parents (géniteurs 2, 5 et 10). Le géniteur 5 présente également un effet « Haplotype » significatif sur la variation de LK%. Cette analyse permet de mettre en évidence trois haplotypes favorables pour S/G : h4, h10 et h1 en combinaison avec les haplotypes h11, h20 et h5, respectivement. Pour le géniteur 2, l'haplotype h5 (défavorable pour S/G) est également favorable à LK%.

1.2.1.2.Effet de la variabilité nucléotidique

L'effet des 54 SNP sélectionnés le long du gène *CCR* a été testé sur la variation des caractères de hauteur, circonférence, densité du bois, LK% et S/G en utilisant le modèle mixte implémenté dans le logiciel TASSEL. L'information du pédigrée (relation entre les individus) à été incluse dans l'analyse afin de prendre en compte l'apparentement entre les descendants du plan de croisement factoriel pour contrôler le taux de faux positifs. Le Tableau 20 décrit

Géniteur	Haplotypes	SNP6	SNP30	SNP35	SNP49	SNP50	SNP74	SNP87	SNP94	SNP117	SNP138	SNP147	Effet S/G
10	h4	**G**	**G**	**G**	**C**	**G**	**C**	**T**	**A**	**T**	**G**	**G**	Favorable
	h11	**G**	**G**	**G**	**C**	**G**	**C**	**T**	**A**	**T**	**A**	**G**	Défavorable
5	h10	**A**	**A**	**C**	**T**	**A**	**T**	**G**	**A**	**C**	**G**	**G**	Favorable
	h20	**G**	**G**	**G**	**C**	**G**	**C**	**T**	**A**	**T**	**A**	**G**	Défavorable
2	h1	**G**	**G**	**G**	**C**	**G**	**C**	**T**	**A**	**T**	**A**	**G**	Favorable
	h5	**G**	**G**	**G**	**C**	**G**	**C**	**T**	**G**	**T**	**A**	**A**	Défavorable

Figure 36: répartition des allèles des 11 SNP significativement associés à la variation de S/G au sein des haplotypes des géniteurs 2, 5 et 10 pour lesquels un effet haplotype significatif a été détecté.

Tableau 21: répartition des haplotypes des géniteurs du plan de croisement en fonction des allèles des 11 SNP significativement associés à la variation de S/G. Les haplotypes des géniteurs 2, 5 et 10 sont indiqués en gras.

SNP	Allèle 1	Allèle 2	Halpotypes allèle 1	Halpotypes allèle 2
SNP6	A	G	h7, h9, **h10**, h12, h15, h19	**h1,** h2, h3, **h4**, **h5**, h6, h8, **h11**, h13, h14, h16, h17, h18, **h20**
SNP30	A	G	h7, h9, **h10**, h12	**h1**, h2, h3, **h4**, **h5**, h6, h8, **h11**, h13, h14, h15, h16, h17, h18, h19, **h20**
SNP35	C	G	**h10**, h12	**h1**, h2, h3, **h4**, **h5**, h6, h7, h8, h9, **h11**, h13, h14, h15, h16, h17, h18, h19, **h20**
SNP49	T	C	h7, **h10**	**h1**, h2, h3, **h4**, **h5**, h6, h8, h9, **h11**, h12, h13, h14, h15, h16, h17, h18, h19, **h20**
SNP50	A	G	h7, **h10**, h12	**h1**, h2, h3, **h4**, **h5**, h6, h8, h9, **h11**, h13, h14, h15, h16, h17, h18, h19, **h20**
SNP74	T	C	h7, **h10**, h15	**h1**, h2, h3, **h4**, **h5**, h6, h8, h9, **h11**, h12, h13, h14, h16, h17, h18, h19, **h20**
SNP87	G	T	h9, **h10**, h12, h15, h19	**h1**, h2, h3, **h4**, **h5**, h6, h7, h8, **h11**, h13, h14, h16, h17, h18, **h20**
SNP94	A	G	**h1**, h2, h3, **h4**, h6, h7, h8, h9, **h10**, **h11**, h12, h14, h15, h16, h17, h18, h19, **h20**	**h5**, h13
SNP117	C	T	**h10**, h14, h17	**h1**, h2, h3, **h4**, **h5**, h6, h7, h8, h9, **h11**, h12, h13, h15, h16, h18, h19, **h20**
SNP138	G	A	h3, **h4**, h7, h9, **h10**, h12, h14, h15, h17, h19	**h1**, h2, **h5**, h6, h8, **h11**, h13, h16, h18, **h20**
SNP147	G	A	**h1**, h2, h3, **h4**, h6, h7, h8, h9, **h10**, **h11**, h12, h15, h16, h17, h19, **h20**	**h5**, h13, h14, h18

les résultats de l'étude d'association marqueur par marqueur entre les 54 SNP sélectionnés le long du gène *CCR* et les caractères de croissance, densité du bois, LK% et S/G.

Aucun des SNP testés n'est associé à la variation de la croissance, de la densité du bois ou de la teneur en lignines. Un total de 13 SNP est significativement associé à la variation de S/G (p-value<0,05). Après correction pour les tests multiples (FDR Q-value<0,05), 11 de ces SNP restent significativement associés à la variation du caractère. Ces SNP se répartissent de manière homogène le long du gène *CCR* et quatre d'entre eux sont détectés dans la partie codante du gène (exon 2 et exon 3). Les 11 SNP présentent des FAM comprises entre 6% et 38% dans la population d'association et expliquent entre 1% et 1,8% de la variation phénotypique du caractère. Parmi eux, les SNP30, SNP35 et SNP138 sont les plus significatifs et expliquent respectivement 1,5%, 1,6% et 1.8% de la variation du phénotype. Les 11 SNP associés ne sont globalement pas en fort DL les uns avec les autres (Figure 35). Cependant, une forte valeur de DL est observée entre le SNP30 et le SNP87 (r^2=0.85) et des valeurs de DL supérieures à 0.6 (r^2) sont observées pour 6 autres couples de SNP. De plus, les SNP6, SNP49, SNP87 et SNP94 sont en DL complet (r^2=1) avec d'autre SNP écartés de l'analyse. Le SNP6, est un SNP non synonyme. Il impacte l'acide aminée numéro 77 et provoque le changement d'une lysine (K) en acide glutamique (E). Ce SNP est en DL complet avec 8 SNP répartis le long du gène entre l'exon 2 et l'intron 4.

1.2.1.3.Relation entre variabilité haplotypique et variabilité nucléotidique

La Figure 36 montre la répartition des allèles des 11 SNP ayant un effet significatif sur la variation de S/G au sein des haplotypes des géniteurs 2, 5 et 10 pour lesquels des effets « Hapltoype » significatifs ont été mis en évidence au sein des familles de demi-frères. Pour le géniteur 5, 9 des 11 SNP sont à l'état hétérozygotes et ségrégent dans la descendance. Pour les géniteurs 2 et 10, deux SNP et un seul SNP respectivement sont à l'état hétérozygote. Les SNP à l'état hétérozygote pour le géniteur 2 (SNP94 et SNP147) sont à l'état homozygote pour les géniteurs 5 et 10 et discriminent l'haplotype h5 de tous les autres haplotypes détectés chez ces trois parents. Le SNP138 discrimine les haplotypes h4 et h10 de tous les autres. Les 8 autres SNP discriminent l'haplotype h4 de tous les autres. Le Tableau 21 montre la répartition des 20 haplotypes mis en évidence pour le gène *CCR* en fonction des allèles des 11 SNP détectés par l'analyse marqueur par marqueur. Pour les marqueurs SNP6, SNP30, SNP35, SNP49, SNP50, SNP74, SNP87 et SNP117, l'haplotype h10 est généralement inclus dans un petit groupe de 2 à 6 haplotypes. Pour les marqueurs SNP84 et SNP147, l'haplotype

h5 est également inclus dans un petit groupe comprenant 4 et 2 haplotypes respectivement. Le marqueur SNP138, discrimine deux groupes d'haplotypes de tailles équivalentes, dont un contient les haplotypes h4 et h10.

1.2.2. Cas des autres gènes dans le plan de croisement E. urophylla x E. urophylla

Au total, l'effet de 36 SNP représentant la variabilité des 9 autres gènes candidats de la lignification à été testé sur les caractères de croissance, la densité du bois, LK% et S/G. Pour ces SNP, 320 individus sont génotypés en moyenne au sein du plan de croisement factoriel. L'analyse marqueur par marquer ne révèle aucun SNP significativement associé à la variation des caractères de croissance, de la densité du bois ou de la teneur en lignines. Un seul SNP est significativement associé à S/G après correction pour les tests multiples (FDR Q-value<5%). Ce SNP est situé sur le gène *ROP1*, dans une région non codante. Dans l'échantillon de génotypes *E. urophylla* utilisé pour le séquençage du gène, ce SNP est en DL complet avec un autre SNP, situé lui aussi dans une région non codante de *ROP1* et qui n'a pas été pris en compte dans l'analyse. Le SNP21 présente une FAM de 16% dans la population d'association et explique 1.3% de la variance phénotypique du caractère.

1.2.3. Cas du gène CCR dans le factoriel E. camaldulensis x E. urophylla

Les données de génotypage SNP du gène *CCR* dans le plan de croisement *E. camaldulensis* x *E. urophylla* ne sont pas de bonne qualité. Les résultats des analyses présentées ici concernent 5 SNP communs aux deux espèces et sont donnés à titre indicatif. Ils doivent être considérés avec prudence étant donné le nombre de données manquantes et le nombre d'incohérences entre les génotypes parentaux et les génotypes des descendants au sein des familles de plein-frères du plan de croisement.

Parmi les 5 SNP étudiés, le SNP6, un SNP non synonyme situé dans l'exon 2 du gène *CCR*, est associé de manière très significative (p-value<10^{-3}) à la variation de la teneur en lignines dans le plan de croisement factoriel. Ces résultats ont été obtenus sur la base de l'étude de 137 descendants du plan de croisement (24% de données manquantes). La FAM de ce SNP dans la population des descendants est de 31%. Etant donné le nombre de biais possibles pour cette analyse ces résultats ne seront pas discutés et devront être vérifiés par d'autres analyses proposées précédemment.

2. Discussion

2.1. Variabilité fonctionnelle des gènes candidats de la lignification

Dans le cadre de cette étude, l'effet de la variabilité nucléotidique de 10 gènes candidats de la lignification sur la teneur en lignines et le rapport S/G a été testé dans la descendance d'un plan de croisement factoriel. L'utilisation d'un modèle mixte permettant de contrôler l'effet de l'apparentement entre individus à permis de mettre en évidence 12 SNP associés à la variation de la qualité des lignines (S/G). L'ensemble des SNP détectés montre un effet faible (entre 1% et 1,8% de la variation phénotypique) mais significatif sur le caractère. Parmi ces SNP, 11 sont situés sur le gène *CCR* et un sur le gène *ROP1*.

Le gène *CCR* code la cinnamoyl-CoA reductase. Cette enzyme catalyse la conversion des esters cinnamoyl-CoA en leur cinnamaldehydes correspondants. Cette étape représente la première étape spécifique de la biosynthèse des monomères de lignines (Pichon *et al.*, 1998). Cette enzyme est considérée comme une enzyme clé dans l'allocation du carbone vers les lignines et plusieurs études de transgénèse ont montré qu'une modulation d'expression du gène *CCR* pouvait entrainer une réduction importante du taux de lignines (Leplé *et al.*, 2007) ainsi qu'une variation du rapport S/G (Pichon *et al.*, 1998). Le gène *ROP1* code une protéine de régulation de type GTPase. Elle pourrait être impliquée dans les processus de différenciation précoce des cellules du xylème durant la mise en place du xylème secondaire (Foucart *et al.*, 2009). Des études de cartographie de QTL menées dans une famille de plein-frères *E. urophylla* x *E. grandis* ont mis en évidence des co-localizations entre ces deux gènes et des QTL majeurs de teneur en lignines (*CCR*) et rapport S/G (*CCR* et *ROP1*) (Gion *et al.*, 2001, Foucart *et al.*, 2009). Ces derniers résultats suggèrent un rôle de leur variabilité naturelle dans le contrôle de la variation des caractères en relation avec les lignines.

Les résultats obtenus dans le cadre de ce travail tendent vers la même conclusion. Ils semblent indiquer que les effets de la variabilité naturelle de ces gènes sont conservés à l'échelle de populations à base génétique plus large que celle qui sont traditionnellement utilisées dans le cadre des approches de cartographie de QTL. Cependant, comme dans la plupart des études d'association, les effets détectés pour les variants associés à la variation du caractère sont globalement faibles (entre 1% et 1,8% dans le cas de notre étude).

La majeure partie des SNP associés à la variation du S/G est située dans les régions introniques des gènes. Dans le cas du gène *CCR*, quatre SNP sont situés dans les exons et causent trois mutations synonymes (SNP49, SNP50 et SNP74), et une mutation non synonyme (SNP6). Plusieurs hypothèses peuvent être émises quand au rôle fonctionnel de ces SNP.

a- Dans le cas des mutations silencieuses situées dans les régions introniques du gène, celles-ci pourraient générer des variants d'épissage alternatif impliquant une modification de l'activité de la protéine. Chez *Capsella bursa-pastoris*, Slotte *et al.* (2009) ont montré que des variations dans des sites d'épissage du *locus FLC* induisaient la production de variants d'épissage alternatif et étaient associées avec une floraison précoce. Dans le cas de *CCR* et *ROP1*, ces mutations pourraient également toucher des sites de régulation introniques encore inconnu.

b- Dans le cas de mutations synonymes, celles-ci n'entrainent pas de modification de la structure primaire de la protéine mais impliquent un changement de codon. Plusieurs études menées sur le biais d'utilisation des codons semblent indiquer que l'utilisation des différents codons au sein des gènes ne se fait pas au hasard. Elle pourrait être influencée par le niveau d'expression du gène (Gouy et Gaultier, 1982 ; Sharp et Li, 1987), sa longueur (Moriyama *et al.*, 1998 ; Duret et Mourichoud, 1999) ou encore la structure secondaire de la protéine (Gupta *et al.*, 2000) et l'abondance des ARNt (Duret, 2000). L'étude de Duret (2000) indique que l'abondance des ARNt et l'usage des codons sont coadaptés chez *C. elegans* pour une traduction efficace des gènes fortement exprimés. Dès lors, des variations synonymes dans des gènes fortement exprimés, comme c'est le cas du gène *CCR* dans les tissus vasculaires en différentiation (Lacombe *et al.*, 1997), pourraient avoir un impact sur l'efficacité de la traduction et donc l'activité de la protéine *in vivo*. Ce type de mutation pourrait donc être impliqué dans la variation du rapport S/G.

c- Dans le cas de mutations synonymes, un changement dans la structure primaire de la protéine pourrait avoir un impact direct sur sa fonction *in planta*. Dans ce cas, un effet fort de la mutation sur la variation du caractère pourrait être attendu. Dans le cas du gène *CCR*, la mutation non synonyme détectée n'impacte pas un acide aminée très conservé de la protéine (Lacombe *et al.*, 1997). Ceci pourrait expliquer l'effet faible qui lui est associé (R^2=1,2%).

Enfin, il se pourrait que les polymorphismes identifiés ne soient pas les polymorphismes causaux de la variation du rapport S/G et qu'ils soient simplement en DL avec un ou plusieurs

polymorphismes non détectés dans la région des deux gènes *CCR* et *ROP1*. Ces polymorphismes pourraient être situés par exemple dans leurs régions promotrices. En effet, malgré la faible étendue du DL estimée au sein des gènes de la lignification chez cette espèce ($r^2<0,2$ au-delà de 1000 pb), un DL fort peut persister de manière ponctuelle entre sites distants de quelques milliers de paires de bases (l'étendue du DL chez *E. urophylla* est décrite dans le chapitre 2). Dans le cadre de ces travaux, les régions promotrice des gènes n'ont pas été étudiées. Celles-ci présentent néanmoins un intérêt majeur et pourraient faire l'objet de futurs travaux de caractérisation de la variabilité fonctionnelle.

Dans tous les cas, les effets des SNP associés à la variation du S/G restent faibles en comparaison de l'héritabilité au sens stricte du caractère. De nombreux locus et leurs interactions doivent donc être impliqués dans le contrôle génétique de la variation de ce carcatère en population.

2.2. L'effet des autres gènes candidats de la lignification

Au sein des gènes *4CL*, *C4H*, *C3H*, *F5H*, *COMT2*, *CAD2*, *MYB1* et *MYB2*, aucun SNP n'a pu être identifié comme significativement associé à la variation des caractères étudiés. Il est donc possible que ces gènes n'aient pas d'effet sur ces caractères. Cependant, malgré l'absence de données de co-localisation avec des QTL, notamment pour la teneur en lignines ou le rapport S/G, la plupart de ces gènes sont considérés comme de bons candidats pour le contrôle des caractères relatifs aux lignines en population. Pour l'ensemble de ces gènes, l'implication dans la biosynthèse des lignines est largement démontrée (Vanholme *et al.*, 2008). Des études d'association ont également révélé au sein de ces gènes des polymorphismes impliqués dans le contrôle de la variation des caractères en population (Gonzalez Martinez *et al.*, 2007 ; Wegrzyn *et al.*, 2010). Dans notre cas, un effet de la méthodologie employée pour la mise en évidence de la variabilité de ces gènes dans les descendances du plan de croisement factoriel peu être mise en cause.

En effet, pour le gène *CCR*, une étude exhaustive de la variabilité nucléotidique présente au sein des 16 lignées parentales a été réalisée. Le gène a été séquencé sur toute sa longueur, et les phases de liaison gamétiques entre les polymorphismes mis en évidence ont été obtenues. Cette étude, associée à une méthode simple de génotypage des différentes formes haplotypiques du gène dans la descendance du plan de croisement, a permis

d'identifier l'ensemble des SNPs en ségrégation dans la population d'association (l'équivalent de 104 SNP a été testé).

Dans le cas des 9 autres gènes de la lignification, la variabilité n'a pu être identifiée que pour une portion du gène, dans un échantillon n'incluant pas la totalité des 16 géniteurs du plan de croisement. Une vision plus partielle de la variabilité a donc été obtenue pour cet ensemble de gènes (152 SNPs identifiés au sein du seul gène *CCR* contre 231 SNPs pour l'ensemble des 9 autres gènes). La méthode de génotypage Sequenom MassARRAY iPLEX Gold, utilisée pour le génotypage de la variabilité de ces gènes dans la descendance du plan de croisement, n'a permis de génotyper que 20 % de cette variabilité (46 SNPs génotypés pour 231 détectés). De plus, dans le cadre de ce type d'approche, les SNP génotypés sont généralement sélectionnés pour représenter au mieux la variabilité des gènes étudiés (sélection de tag SNP). Cette approche a notamment été mise en places par Gonzalez Martinez *et al.* (2007) pour tester l'effet de la variabilité nucléotidique de 20 gènes candidas sur la variation des propriétés du bois dans une population de clones de *Pinus taeda*. Les auteurs ont utilisé des données de séquençage précédemment obtenues sur la base d'un échantillon (Brown *et al.*, 2004 ; Gonzalez Martinez *et al.*, 2006) pour sélectionner un sous ensemble de SNP à tester en population plus large. Les SNP ont été sélectionnés en fonction de leur rôle fonctionnel (SNP non synonymes), de la manière dont ils représentaient la variabilité totale obtenue par séquençage et de la fréquence des SNP détectés dans l'échantillon de séquençage. Un total de 58 SNP ont été génotypés pour l'étude des 20 gènes permettant de détecter plusieurs gènes associés à la variation de différentes propriétés du bois Dans notre cas, les SNP génotypés n'ont pas pu être choisis. Dès lors, il est possible que les polymorphismes associés avec la variation des caractères d'intérêt n'aient pas pu être étudiés car non détectés par séquençage, non génotypés dans la population d'association, ou mal représentés par l'ensemble des SNP étudiés.

2.3. Le dispositif expérimental utilisé

A ce jour il n'existe aucune étude (à notre connaissance) qui rapporte l'utilisation de plans de croisement factoriel chez des espèces forestières pour la détection des effets de la variabilité des gènes sur des caractères quantitatifs. Pourtant, ces dispositifs sont beaucoup utilisés dans le cadre de programmes d'amélioration chez les arbres. Ils sont particulièrement appréciés car ils permettent d'estimer facilement les différentes composantes de la variance pour des caractères en population. Le travail qui a été mené ici indique que ces dispositifs

expérimentaux peuvent être utilisés, à condition que l'apparentement entre les individus puisse être pris en compte, dans le cadre d'études d'association.

Cependant, la descendance d'un plan de croisement tel que celui qui a été utilisés ici ne correspond pas à une population naturelle (classiquement utilisée pour les études d'association). Elle implique une structure particulière du DL au sein des descendances issues de mêmes parents (plein-frères et demi-frères). L'impact de cette structure particulière de la population sur la résolution et la précision de la localisation des QTN est difficile à évaluer et il est difficile de déterminer si les polymorphismes mis en évidence sont les polymorphismes causaux de la variation du S/G. Par exemple, les SNP 35, 49, 50, 74 et 117 du gène *CCR* qui présentent un effet sur le rapport S/G sont en fréquence faible dans le plan de croisement (<10%) et sont portés par un haplotype parental qui présente un effet sur le même caractère au sein d'une descendance de demi-frères. Il est probable que tous ces SNP n'aient pas d'effet propre sur la variation du rapport S/G mais reflètent plutôt l'effet de l'haplotype h10 à l'échelle du plan de croisement. Pour tous ces SNP, il est alors probable que les associations détectées ne soient valables que dans le contexte étudié. Cependant, le fait d'identifier des effets pour plusisuers haplotypes pourrait indiquer que plusieurs variants causaux compris dans la région chromosomique de ces deux gènes pourraient être impliqués dans la variation du rapport S/G.

.Notre dispositif expérimental ne permet d'étudier que la variabilité de 16 individus non apparentés. Ces individus ne représentent pas la variabilité génétique de l'ensemble de la population d'*E. urophylla*. Les résultats de notre analyse indiquent qu'il est cependant possible dans des populations à base génétique plus large qu'un pédigrée de cartographie génétique d'identifier des liens entre la variabilité génétique et la variation phénotypique chez ces espèces. Cette approche révèle que l'effet d'une région chromosomique sur le rapport S/G, mis en évidence dans le cadre d'une analyse de cartographie de QTL (la région contenant les gènes *CCR* et *ROP1*), est conservé dans une population à base génétique plus large qu'un pédigrée de cartographie. Chez des espèces aussi variables que les eucalyptus (comme nous avons pu le démontrer au chapître 4 de ce travail de thèse), ces résultats semblent indiquer qu'il est encore possible de mettre en évidence des liens entre variabilité génétique et varaition phénotypique, pour peu que la variabilité haplotypique de la région chrommosomique étudiée soit bien carcatérisée (comme dans le cas des gènes *CCR* et *ROP1*).

Pour aller plus loin, une analyse QTL menée dans ce dispositif expérimental multi-parental pourrait permettre d'obtenir des informations complémentaires. Ce type d'approche est aujourd'hui possible puisque de nombreux marqueurs microsatellites transferables entre espèces d'eucalyptus sont maintenant disponibles et devraient permettre de suivre la ségrégation des blocs haplotypiques des 16 parents dans les descandances du plan de croisement (Brondani *et al.*, 2006, Faria *et al.*, 2010). Une étude d'association conduite dans une population constituée d'individus non apparentés permettrait peut être de confirmer les liens mis en évidence entre la variabilité des gènes étudiés et la variation du rapport S/G. Il s'agirait d'utiliser les polymorphismes mis en évidence dans notre étude pour réaliser une étude d'association gène candidat chez *E. urophylla* sur le modèle de celles qui ont été réalisées chez *E. nitens* par Thumma *et al.* (2005 ; 2009). Dans ce cadre, les résultats de ces travaux de thèse pourront également aider à l'ellaboration d'une stratégie efficace pour la mise en évidence de la variabilité nucléotidique des gènes candidats de la lignification dans la population d'étude.

Conclusion générale

L'objectif principal de cette thèse était d'étudier le déterminisme génétique de la quantité et de la qualité des lignines du bois d'eucalyptus et d'identifier des marqueurs moléculaires pour la sélection de variétés clonales.

Dans un premier temps, nous avons étudié le déterminisme génétique de caractères liés à la quantité et à la qualité des lignines, deux propriétés chimiques d'intérêt majeur pour l'industrie du charbon de bois et de la pâte à papier. Ces propriétés ayant été peu étudiées d'un point de vue génétique, la première partie de ce travail s'est attachée à quantifier leur variation et étudier les composantes de la variance à partir des données phénotypiques collectées dans des familles de plein-frères de plans de croisement factoriels. Dans un deuxième temps, nous avons décrit la diversité nucléotidique de gènes impliqués dans la biosynthèse des lignines au sein de deux espèces d'intérêt majeur en plantation : *E. urophylla* et *E. camaldulensis*. Enfin, la relation entre la variabilité de ces gènes et la variation des caractères relatifs aux lignines a été étudiée par une étude d'association. Les tests d'association ont été essentiellement menés dans la descendance d'un plan de croisement intraspécifique *E. urophylla* x *E. urophylla* avec pour objectif d'identifier des polymorphismes impliqués dans la variation de ces caractères.

1. Les principaux résultats

1.1. Déterminisme génétique de la quantité et de la qualité des lignines

L'étude des propriétés chimiques du bois est relativement récente. Dans le cadre de ce travail, la méthode SPIR s'est avérée très efficace pour prédire à moindre coût la teneur en lignines et le rapport S/G dans des descendances de plan de croisements factoriels impliquant 3 espèces d'eucalyptus *E. urophylla*, *E. camaldulensis* et *E. grandis*. Cependant, dans un objectif d'utiliser cette méthode de prédiction des caractères chimiques en routine dans le cadre des programmes d'amélioration génétique des eucalyptus en République du Congo et au Brésil, il apparait important de développer les échantillons de calibration pour espérer pouvoir prédire précisément ces caractères dans leur gamme de variation.

Quoi qu'il en soit, les résultats de ces premiers travaux indiquent des niveaux de variabilité plus importants (CV_P) pour le rapport S/G que pour la teneur en lignines. La teneur en lignines est globalement plus forte dans les descendances du plan de croisement impliquant l'espèce *E. camaludlensis*. La décomposition des valeurs phénotypiques des descendants de ces plans de croisements (en fonction de leur composante génétique et environnementale) indique des effets génétiques globalement forts et principalement additifs au sein des dispositifs expérimentaux étudiés, suggérant qu'une sélection est possible pour ces deux caractères. Les coefficients de variation génétiques additifs obtenus pour S/G montrent que des gains génétiques importants peuvent être obtenus, au moins chez *E. urophylla*. Chez cette espèce, et dans le contexte de croisements intraspécifique, nos travaux mettent également en évidence des corrélations génétiques additives entre les caractères relatifs aux lignines et des caractères classiquement pris en compte dans les programmes d'amélioration génétique : la hauteur et la densité du bois. Ces corrélations indiquent qu'une sélection orientée vers une amélioration de la croissance pourrait à terme diminuer la teneur en lignines et donc diminuer la qualité du bois dans un objectif de production de charbon. Elles indiquent également qu'une sélection pour l'augmentation de la densité du bois, comme ce pourrait être le cas pour la production de pâte à papier, pourrait diminuer fortement S/G et donc déprécier la qualité du bois pour cette application industrielle, une lignine plus riche en monomère S facilitant la délignification.

1.2. Diversité nucléotidique et étendue du déséquilibre de liaison au sein de gènes candidats de la lignification

Si de nombreuses séquences de gènes d'eucalyptus sont aujourd'hui disponibles dans les bases de données publiques (notamment des EST), les études consacrées à la description de la variabilité nucléotidique sont encore embryonnaires. De telles études permettent d'évaluer le niveau de la diversité nucléotidique des gènes ainsi que l'étendue du déséquilibre de liaison entre allèles, deux connaissances essentielles à la mise en place d'études d'association en génétique. Dans le cadre de notre travail, la diversité nucléotidique et le déséquilibre de liaison ont été étudiés au sein de 10 gènes candidats de la lignification chez *E. urophylla* avec une attention particulière portée au gène codant la *CCR* (cinnamoyl CoA reductase), dont la séquence totale a été considérée. Pour ce gène, l'étude de la variabilité nucléotidique a également été réalisée au sein d'un échantillon de l'espèce *E. camaldulensis* permettant de comparer la diversité nucléotidique observée et l'étendue du DL entre ces deux espèces d'eucalyptus. Les résultats montrent des niveaux de variabilité élevés pour la plupart

des gènes étudiés. Les gènes codant la CCR et la C3H montrent les niveaux de diversité génétique totale les plus élevés. Le nombre d'haplotypes le plus important (20) a été détecté pour le gène codant la CCR. A l'inverse, les gènes *F5H* et *COMT2* présentent les niveaux de diversité génétique plus faibles avec seulement 9 haplotypes détectés au sein du gène *COMT2*.

L'étendue du DL est globalement faible et de l'ordre de ce qui est observé chez d'autres espèces d'arbres forestiers comme le pin ou le peuplier. Nos résultats indiquent qu'il décroît rapidement avec la distance entre sites polymorphes et devient très faible au-delà de quelques centaines de paires de bases. Par ailleurs, les résultats obtenus pour le gène *CCR* chez *E. camaldulensis* indiquent des valeurs similaires à celles obtenues chez *E. urophylla* que ce soit en termes de diversité nucléotidique et d'étendue du DL. Cependant, l'étude de la variabilité du gène *CCR* dans un échantillon de taille plus limitée chez *E. camaldulensis* en comparaison à l'espèce *E. urophylla* suggère des niveaux de variabilité encore plus importants chez cette espèce. L'étude de la variabilité nucléotidique réalisée chez *E. urophylla* a permis de constituer une collection de 387 SNP, dont 116 semblent être des événements de mutation rares et 37 correspondent à des mutations non synonymes. Cette collection de SNP constitue le premier catalogue de marqueurs génétiques pour l'étude de la variabilité fonctionnelle de gènes candidats de la lignification chez *E. urophylla*. Au sein du seul gène *CCR*, 152 et 138 SNP bialléliques ont été détectés pour les espèces *E. urophylla* et *E. camaldulensis*, respectivement. Ces SNP représentent également un catalogue inédit de marqueurs génétiques pour l'étude de la variabilité fonctionnelle de ce gène chez les deux espèces pures *E. urophylla* et *E. camaldulensis* ou chez l'hybride *E. camaldulensis* x *E. urophylla*. L'ensemble de ces marqueurs génétiques pourraient également être utilisés pour étudier l'impact de la sélection naturelle sur les gènes de la voie de biosynthèse des lignines. Pour cela, il serait intéressant de disposer d'un échantillon d'individus représentatifs de l'aire de répartition des espèces.

Enfin, dans le cadre de cette étude, un deuxième locus homologue au gène *CCR* étudié (*CCR2*) a été mis en évidence chez *E. urophylla*. De plus amples investigations seront nécessaires pour déterminer si ce gène est fonctionnel ou s'il s'agit d'un pseudo-gène. Dans le premier cas, on pourrait envisager de prendre en compte ce deuxième gène dans le cadre des études d'association menées sur les caractères relatifs aux lignines.

1.3. Génotypage de la variabilité nucléotidique dans des descendances de plans de croisement factoriels

Des descendances de plein-frères ont été utilisées pour constituer les populations d'association destinées à mettre en évidence la variabilité fonctionnelles des gènes étudiés. Lorsque la variabilité nucléotidique est connue chez les parents utilisés pour générer les descendances, ce type de dispositif présente un avantage pour l'étape de génotypage de la variabilité des gènes dans les descendances. En effet, la comparaison entre les génotypes des parents et ceux des descendants permet déjà de vérifier l'exactitude des croisements. Par ailleurs, grâce à ces connaissances, des méthodes simples de génotypage peuvent être utilisées pour mettre en évidence de manière exhaustive la variabilité des gènes chez les descendants (cas du microsatellite du gène *CCR* pour le dispositif *E. urophylla* x *E. urophylla*). Lorsque la variabilité des parents n'est pas connue, le génotypage des individus nécessite l'utilisation de méthodes non ciblées. Parmi ces méthodes, celles qui sont basées sur le principe d'extension d'amorces ont été largement utilisées. Elles présentent cependant certaines contraintes, notamment chez les espèces dont le niveau de variabilité nucléotidique est important et pour lesquelles une faible étendue du DL est détectée (beaucoup de sites polymorphes doivent être génotypés pour représenter la variabilité totale et peu de SNP peuvent être pris en compte par la méthode). Cette méthode présente un autre inconvénient dans le cas où les polymorphismes étudiés ont été détectés dans des gènes appartenant à des familles multigéniques. En effet, la présence de régions conservées au sein de ces paralogues peut créer un bruit de fond important (présence d'allèles de paralogues différents). Les résultats de notre étude indiquent que chez l'eucalyptus, caractérisé par une très grande diversité nucléotidique, cette méthode de génotypage n'est pas très bien adaptée. Même si elle reste pour le moment moins chère que le séquençage direct, elle ne permet de prendre en compte qu'une faible partie de la variabilité nucléotidique. Dans le cas du génotypage du gène *CCR* au sein du plan de croisement *E. camaldulensis* x *E. urophylla*, les niveaux de variabilité très importants détectés chez les deux espèces et la présence d'un deuxième gène *CCR* très proche du gène ciblé sont certainement responsables de l'échec de cette méthode de génotypage au sein de ce dispositif.

1.4. Etude de la variabilité fonctionnelle de gènes de la lignification chez *E. urophylla*

L'étude d'association menée au sein des descendances du plan de croisement factoriel *E. urophylla* x *E. urophylla* montre des liaisons statistiques entre la variabilité nucléotidique

des gènes *CCR* et *ROP1* et la variation du rapport S/G. Ces résultats confirment les résultats d'analyses précédentes indiquant des co-localisations entre ces deux gènes et un QTL pour le rapport S/G dans une descendance hybride *E. urophylla* x *E. grandis*. Au total, 12 SNP (11 au sein du gène *CCR* et un au sein du gène *ROP1*) présentent un effet significatif mais faible sur la variation du caractère. Au sein du gène *CCR*, certains de ces SNP, en fréquence faible dans le plan de croisement, semblent fortement liés à un haplotype du gène dont l'effet est significatif dans la descendance d'un des parents du plan de croisement. Un autre SNP, en fréquence intermédiaire dans la descendance du plan de croisement, est porté par deux haplotypes ayant des effets sur la variation du caractère au sein des descendances de deux parents du plan de croisement. Ces résultats doivent être confirmés dans des populations plus importantes qui devraient rapidement être mises en place dans le cadre des actions de recherche menées au CIRAD.

En effet, à l'aube de la mise à disposition des données du séquençage du génome d'*E. grandis*, et avec le développement des technologies de séquençage et de génotypage, la production de données moléculaires ne devrait plus être limitante. En revanche, la mise à disposition de dispositifs expérimentaux *ad hoc* est un enjeu majeur et notamment pour la détection des locus impliqués dans la variation des caractères quantitatifs. Dans ce cadre, il serait intéressant de disposer d'une population d'individus non apparentés, avec des copies clonales. Comme le montrent les résultats de notre étude, les locus détectés auront probablement, pour la majorité d'entre eux, des effets faibles. Cette population d'association devra donc être suffisamment importante pour permettre de détecter de tels locus. Etant donnée la bonne aptitude au bouturage des eucalyptus, on pourrait également imaginer planter cette population dans des environnements différents afin de valider la stabilité des associations détectées dans différents contextes de plantation.

2. A la recherche des polymorphismes qui contrôlent la variation de la teneur en lignines

Dans le cadre de la production de charbon de bois pour des applications industrielles, la teneur en lignines est un caractère majeur de la qualité du bois. Les résultats de cette étude n'ont pas permis de mettre en évidence des effets de la variabilité nucléotidique des gènes étudiés sur la variation de la teneur en lignines. Cependant, même si l'effet d'une grande part de la variabilité du gène *CCR* mise en évidence chez *E. urophylla* a été testé sur la variation

de la teneur en lignines, l'approche mise en place n'a permis de prendre en compte que peu d'individus (le dispositif expérimental utilisé ne comporte que 16 parents *E. urophylla*), peu de gènes (10 gènes au total) et surtout peu de polymorphismes parmi ceux détectés au sein de la plupart de ces gènes. Comme dans le cas du rapport S/G, la prise en compte de plus de marqueurs dans les analyses devrait pouvoir permettre de détecter des polymorphismes d'intérêt impliqués dans le contrôle de la variation de la teneur en lignines. Dans ce cadre, la possibilité d'étudier une part plus importante (voire exhaustive) de la variabilité des gènes candidats doit maintenant être un objectif prioritaire, moins de 20% des SNP détectés au sein des gènes ayant pu être étudiés. Cela passera par le séquençage des gènes complets (y compris des zones promotrices) dans les descendances des plans de croisement utilisés, ce qui est envisageable en considérant la disponibilité de la séquence du génome de l'eucalyptus, ainsi que les possibilités ouvertes par les nouvelles technologies de séquençage à haut débit couplé à l'étiquetage des fragments amplifiés.

3. La sélection génomique : une nouvelle approche en cours d'évaluation chez les arbres

Depuis peu, des approches dites de sélection génomique sont envisagées chez l'eucalyptus. La sélection génomique propose la sélection d'individus sur la base de valeurs génétiques additives prédites à partir de données de marqueurs génétiques dispersés sur tout le génome et permettant de capturer un maximum de l'information des QTL pour un caractère (Grattapaglia et Resende, 2010). Ce type d'approche est déjà utilisé chez certaines espèces animales et présente un réel intérêt chez les arbres forestiers pour lesquels la sélection fait face à des problématiques communes avec ces espèces (des temps de génération longs et des caractères d'intérêt dont l'évaluation se fait souvent à un âge avancé). Cette approche a été proposée pour la première fois en 2001 par Meuwissen *et al.* (2001), et initiée par le développement des méthodes de génotypage à haut débit chez les bovins. Contrairement à la génétique d'association, la sélection génomique ne cherche plus à identifier des marqueurs impliqués dans la variation des caractères, mais à prédire la valeur génétique des individus en utilisant un ensemble de marqueurs sur le génome sans s'intéresser à ceux qui sont en cause dans la variation des caractères. La possibilité de mettre en place ce type d'approche nécessite le génotypage d'un grand nombre de marqueurs génétiques et la mesure des phénotypes pour les caractères d'intérêt dans un échantillon d'individus de grande taille (population de référence). Un modèle de prédiction est alors construit sur la base de ces données. Il permettra

de prédire des valeurs génétiques additives pour les caractères chez des individus dont seul le génotype aux marqueurs est obtenu. Dans ce type d'approche, la précision de la prédiction dépend de l'étendue du DL (déterminant pour le nombre de marqueurs génétiques nécessaires à la prédiction), du nombre d'individus pris en compte pour la mise au point du modèle de prédiction et de la précision des valeurs génétiques additives estimées pour ces individus, de l'héritabilité du caractère ciblé, du nombre de QTL impliqués dans la variation du caractère et des effets de ces QTL. Chez les arbres, la faible étendue du DL qui caractérise la plupart des espèces d'intérêt agronomique est a priori très limitante car elle implique qu'un grand nombre de marqueurs soit pris en compte pour la prédiction. Cependant, des niveaux de DL plus important peuvent être obtenus en considérant des populations élites de quelques dizaines d'individus. Ces approches sont en cours d'évaluation chez les arbres en utilisant la simulation (Grattapaglia et Resende, 2010). Pour le moment, les études d'association, qu'elles soient basées sur un faible nombre de gènes candidats ou sur un balayage plus dense du génome restent très peu pratiquées chez les arbres. Nul doute que ces approches sont amenées à se développer grâce à l'émergence des nouvelles technologies de séquençage et de génotypage, des capacités de calcul grandissant, des nouveaux algorithmes mathématiques, sans oublier les dispositifs mis en place depuis des décennies par les améliorateurs, qui verront leurs dispositifs revisités pour améliorer l'efficacité de la sélection.

Bibliographie

Abbott, R.J., and Gomes, M.F. (1989). Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. Heredity **62,** 411-418.

Abdi, H. (2007). The Bonferonni and Sidak Corrections for Multiple Comparisons.

Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y.M., Lench, N.J., Carey, A., Cardon, L.R., Moffatt, M.F., and Cookson, W.O.C. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. American Journal of Human Genetics **68,** 191-197.

Agarwal, U.P. (2006). Raman imaging to investigate ultrastructure and composition of plant cell walls: distribution of lignin and cellulose in black spruce wood (Picea mariana). Planta **224,** 1141-1153.

Agrama, H.A., George, T.L., and Salah, S.F. (2002). Construction of genome map for Eucalyptus camaldulensis DEHN. Silvae Genetica **51,** 201-206.

Ahuja, M.R. (2001). Recent advances in molecular genetics of forest trees. Euphytica **121,** 173-195.

Altshuler, D., Brooks, L.D., *et al.* (2005). A haplotype map of the human genome. Nature **437,** 1299-1320.

Alves, A., Schwanninger, M., Pereira, H., and Rodrigues, J. (2006). Calibration of NIR to assess lignin composition (H/G ratio) in maritime pine wood using analytical pyrolysis as the reference method. Holzforschung **60,** 29-31.

Amin, N., van Duijn, C.M., and Aulchenko, Y.S. (2007). A Genomic Background Based Method for Association Analysis in Related Individuals. Plos One **2**.

Antal, M.J., Allen, S.G., Dai, X.F., Shimizu, B., Tam, M.S., and Gronli, M. (2000). Attainment of the theoretical yield of carbon from biomass. Industrial & Engineering Chemistry Research **39,** 4024-4031.

Anterola, A.M., and Lewis, N.G. (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry **61,** 221-294.

Anterola, A.M., Jeon, J.H., Davin, L.B., and Lewis, N.G. (2002). Transcriptional control of monolignol biosynthesis in Pinus taeda - Factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. Journal of Biological Chemistry **277,** 18272-18280.

Aoyagi, S., Sugiyama, M., and Fukuda, H. (1998). BEN1 and ZEN1 cDNAs encoding S1-type DNases that are associated with programmed cell death in plants. Febs Letters **429,** 134-138.

Apiolaza, L.A., Raymond, C.A., and Yeo, B.J. (2005). Genetic variation of physical and chemical wood properties of Eucalyptus globulus. Silvae Genetica **54,** 160-166.

Apiolaza, L.A. (2009). Very early selection for solid wood quality: screening for early winners. Annals of Forest Science **66,** 601p601-601p610.

Aranzana, M.J., Kim, S., Zhao, K.Y., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C.L., Toomajian, C., Traw, B., Zheng, H.G., Bergelson, J., Dean, C., Marjoram, P., and Nordborg, M. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. Plos Genetics **1,** 531-539.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics **3,** 299-309.

Atanassova, R., Favet, N., Martz, F., Chabbert, B., Tollier, M.T., Monties, B., Fritig, B., and Legrand, M. (1995). Altered Lignin Composition in Transgenic Tobacco Expressing O-Methyltransferase Sequences in Sense and Antisense Orientation. Plant Journal **8,** 465-477.

Atwell, S., Huang, Y.S., *et al.* (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465,** 627-631.

Bailleres, H., Davrieus, F., and Pichavant, F.H. (2002). Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. Annals of Forest Science **59,** 479-490.

Balasaravanan, T., Chezhian, P., Kamalakannan, R., Yasodha, R., Varghese, M., Gurumurthi, K., and Ghosh, M. (2006). Identification of species-diagnostic ISSR markers for six Eucalyptus species. Silvae Genetica **55,** 119-122.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. Nature Reviews Genetics **7,** 781-791.

Bao, F.C., Jiang, Z.H., Jiang, X.M., Lu, X.X., Luo, X.Q., and Zhang, S.Y. (2001). Differences in wood properties between juvenile wood and mature wood in 10 species grown in China. Wood Science and Technology **35,** 363-375.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S. (2007). SNP discovery via 454 transcriptome sequencing. Plant Journal **51,** 910-918.

Baril, C.P., Verhaegen, D., Vigneron, P., Bouvet, J.M., and Kremer, A. (1997). Structure of the specific combining ability between two species of Eucalyptus .1. RAPD data. Theoretical and Applied Genetics **94,** 796-803.

Baril, C.P., Verhaegen, D., Vigneron, P., Bouvet, J.M., and Kremer, A. (1997). Structure of the specific combining ability between two species of Eucalyptus .2. A clustering approach and a multiplicative model. Theoretical and Applied Genetics **94,** 804-809.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., and Edwards, D. (2003). Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiology **132,** 84-91.

Baucher, M., Halpin, C., Petit-Conil, M., and Boerjan, W. (2003). Lignin: Genetic engineering and impact on pulping. Critical Reviews in Biochemistry and Molecular Biology **38,** 305-350.

Beers, E.P. (1997). Programmed cell death during plant growth and development. Cell Death and Differentiation **4,** 649-661.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological **57,** 289-300.

Bergelson, J., Stahl, E., Dudek, S., and Kreitman, M. (1998). Genetic variation within and among populations of Arabidopsis thaliana. Genetics **148,** 1311-1323.

Boerjan, W., Ralph, J., and Baucher, M. (2003). Lignin biosynthesis. Annual Review of Plant Biology **54,** 519-546.

Bogunic, F., Muratovic, E., Brown, S.C., and Siljak-Yakovlev, S. (2003). Genome size and base composition of five Pinus species from the Balkan region. Plant Cell Reports **22,** 59-63.

Botstein, D. (1980). A theory of modular evolution for bacteriophages. Annals of the New York Academy of Sciences **354,** 484-491.

Bouffier, L., Raffin, A., Rozenberg, P., Meredieu, C., and Kremer, A. (2009). What are the consequences of growth selection on wood density in the French maritime pine breeding programme? Tree Genetics & Genomes **5,** 11-25.

Bourquin, V., Nishikubo, N., Abe, H., Brumer, H., Denman, S., Eklund, M., Christiernin, M., Teeri, T.T., Sundberg, B., and Mellerowicz, E.J. (2002).

Xyloglucan endotransglycosylases have a function during the formation of secondary cell walls of vascular tissues. Plant Cell **14,** 3073-3088.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23,** 2633-2635.

Brakenhoff, R.H., Schoenmakers, J.G.G., and Lubsen, N.H. (1991). Chimeric Cdna Clones - a Novel Pcr Artifact. Nucleic Acids Research **19,** 1949-1949.

Brett, C.T. (2000). Cellulose microfibrils in plants: Biosynthesis, deposition, and integration into the cell wall. International Review of Cytology - a Survey of Cell Biology, Vol 199 **199,** 161-199.

Brock, M.T., Tiffin, P., and Weinig, C. (2007). Sequence diversity and haplotype associations with phenotypic responses to crowding: GIGANTEA affects fruit set in Arabidopsis thaliana. Molecular Ecology **16,** 3050-3062.

Brondani, R.P.V., Brondani, C., Tarchini, R., and Grattapaglia, D. (1998). Development, characterization and mapping of microsatellite markers in Eucalyptus grandis and E-urophylla. Theoretical and Applied Genetics **97,** 816-827.

Brondani, R.P.V., Brondani, C., and Grattapaglia, D. (2002). Towards a genus-wide reference linkage map for Eucalyptus based exclusively on highly informative microsatellite markers. Molecular Genetics and Genomics **267,** 338-347.

Brondani, R.P.V., Williams, E.R., Brondani, C., and Grattapaglia, D. (2006). A microsatellite-based consensus linkage map for species of Eucalyptus and a novel set of 230 microsatellite markers for the genus. Bmc Plant Biology **6**.

Brooker, M.I.H. (2000). A new classification of the genus Eucalyptus L'Her. (Myrtaceae). Australian Systematic Botany **13,** 79-148.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., and Neale, D.B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America **101,** 15255-15260.

Browning, B.L. (1967). In Methods of wood chemistry, I. John Wiley & Sons, ed (New York, USA), pp. 791-792.

Buckler, E.S., and Thornsberry, J.M. (2002). Plant molecular diversity and applications to genomics. Current Opinion in Plant Biology **5,** 107-111.

Bundock, P.C., Hayden, M., and Vaillancourt, R.E. (2000). Linkage maps of Eucalyptus globulus using RAPD and microsatellite markers. Silvae Genetica **49,** 223-232.

Bundock, P.C., Potts, B.M., and Vaillancourt, R.E. (2008). Detection and stability of quantitative trait loci (QTL) in Eucalyptus globulus. Tree Genetics & Genomes **4,** 85-95.

Butcher, P.A., and Williams, E.R. (2002). Variation in outcrossing rates and growth in Eucalyptus camaldulensis from the Petford Region, Queensland; Evidence of outbreeding depression. Silvae Genetica **51,** 6-12.

Butcher, P.A., Otero, A., McDonald, M.W., and Moran, G.F. (2002). Nuclear RFLP variation in Eucalyptus camaldulensis Dehnh. from northern Australia. Heredity **88,** 402-412.

Butcher, P.A., McDonald, M.W., and Bell, J.C. (2009). Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, Eucalyptus camaldulensis. Tree Genetics & Genomes **5,** 189-210.

Byrne, M., Murrell, J.C., Allen, B., and Moran, G.F. (1995). An Integrated Genetic-Linkage Map for Eucalypts Using Rflp, Rapd and Isozyme Markers. Theoretical and Applied Genetics **91,** 869-875.

Byrne, M., Parrish, T.L., and Moran, G.F. (1998). Nuclear RFLP diversity in Eucalyptus nitens. Heredity **81,** 225-233.

Campbell, M.M., and Sederoff, R.R. (1996). Variation in lignin content and composition - Mechanism of control and implications for the genetic improvement of plants. Plant Physiology **110,** 3-13.

Camus-Kulandaivelu, L., Veyrieras, J.B., Madur, D., Combes, V., Fourmann, M., Barraud, S., Dubreuil, P., Gouesnard, B., Manicacci, D., and Charcosset, A. (2006). Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the Dwarf8 gene. Genetics **172,** 2449-2463.

Carle, J., and Holmgren, P. (2008). Wood from Planted Forests. Forest Products Journal **58,** 6-18.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L., and Nickerson, D.A. (2003). Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nature Genetics **33,** 518-521.

Chaffey, N. (2002). Why is there so little research into the cell biology of the secondary vascular system of trees? New Phytologist **153,** 213-223.

Champurney, N. (2003). Mise en évidence de polymorphisme au sein du gène CCR chez Eucalyptus urophylla. In Mémoire de DEA Ressources Phytogénétiques et Interactions Biologiques (Montpellier: Ecole Doctorale Biologie Intégrative), pp. 1-20.

Chang, H.M., and Sarkanen, K.V. (1973). Species variation in lignin. Effect of species on the rate of kraft delignification. Tappi **56,** 132-134.

Chiang, V.L., and Funaoka, M. (1988). The Formation and Quantity of Diphenylmethane Type Structures in Residual Lignin During Kraft Delignification of Douglas-Fir. Holzforschung **42,** 385-391.

Chiang, V.L. (2006). Monolignol biosynthesis and genetic engineering of lignin in trees, a review. Environmental Chemistry Letters **4,** 143-146.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., and Rafalski, A.J. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. Bmc Genetics **3**.

Christensen, J.H., Baucher, M., O'Connell, A.P., Van Montagu, M., and Boerjan, W. (2000). Control of lignin biosynthesis. In Molecular Biology of Woody Plants, S.M. SM Jain, eds, ed (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 227–237.

Chu, Y.G., Su, X.H., Huang, Q.J., and Zhang, X.H. (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.

Cosgrove, D.J. (1997). Assembly and enlargement of the primary cell wall in plants. Annual Review of Cell and Developmental Biology **13,** 171-201.

Cosgrove, D.J. (1999). Enzymes and other agents that enhance cell wall extensibility. Annual Review of Plant Physiology and Plant Molecular Biology **50,** 391-417.

Costa, M.A., Collins, R.E., Anterola, A.M., Cochrane, F.C., Davin, L.B., and Lewis, N.G. (2003). An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in Arabidopsis thaliana and limitations thereof. Phytochemistry **64,** 1097-1112.

Costa e Silva, J., Borralho, N.M.G., Araujo, J.A., Vaillancourt, R.E., and Potts, B.M. (2009). Genetic parameters for growth, wood density and pulp yield in Eucalyptus globulus. Tree Genetics & Genomes **5**.

Cown, D.J. (1978). Comparison of the Pilodyn and Torsiometer Methods for the Rapid Assessment of Wood Density in Living Trees. New Zealand Journal of Forestry Science **8,** 384-391.

Cramer, S., Kretschmann, D., Lakes, R., and Schmidt, T. (2005). Earlywood and latewood elastic properties in loblolly pine. Holzforschung **59,** 531-538.

Crisp, M., Cook, L., and Steane, D. (2004). Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present-day communities? Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **359,** 1551-1571.

Cronn, R., Cedroni, M., Haselkorn, T., Grover, C., and Wendel, J.F. (2002). PCR-mediated recombination in amplification products derived from polyploid cotton. Theoretical and Applied Genetics **104,** 482-489.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J.M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics **182,** 375-385.

Dejardin, A., Laurans, F., Arnaud, D., Breton, C., Pilate, G., and Leple, J.C. (2010). Wood formation in Angiosperms. Comptes Rendus Biologies **333,** 325-334.

del Rio, J.C., Gutierrez, A., Hernando, M., Landin, P., Romero, J., and Martinez, A.T. (2005). Determining the influence of eucalypt lignin composition in paper pulp yield using Py-GC/MS. Journal of Analytical and Applied Pyrolysis **74,** 110-115.

Dence, C.W. (1992). The determination of lignin. In Methods in Lignin Chemistry, S.Y. Lin and C.W. Dence, eds (Berlin: Springer Verlag), pp. 33-61.

Deutsch, S. (2001). The case for large-size mutations. Ieee Transactions on Biomedical Engineering **48,** 124-127.

Doblin, M.S., Kurek, I., Jacob-Wilk, D., and Delmer, D.P. (2002). Cellulose biosynthesis in plants: from genes to rosettes. Plant and Cell Physiology **43,** 1407-1420.

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics **3,** 43-52.

Dooner, H.K., and MartinezFerez, I.M. (1997). Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell **9,** 1633-1646.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **96,** 4482-4487.

Duret, L. (2000). tRNA gene number and codon usage in the C-elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics **16,** 287-289.

Dvornyk, V., Sirvio, A., Mikkonen, M., and Savolainen, O. (2002). Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. Molecular Biology and Evolution **19,** 179-188.

Dwivedi, U.N., Campbell, W.H., Yu, J., Datla, R.S.S., Bugos, R.C., Chiang, V.L., and Podila, G.K. (1994). Modification of Lignin Biosynthesis in Transgenic Nicotiana through Expression of an Antisense O-Methyltransferase Gene from Populus. Plant Molecular Biology **26,** 61-71.

Eckert, A.J., Bower, A.D., Wegrzyn, J.L., Pande, B., Jermstad, K.D., Krutovsky, K.V., Clair, J.B.S., and Neale, D.B. (2009). Asssociation Genetics of Coastal Douglas Fir (Pseudotsuga menziesu var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits. Genetics **182,** 1289-1302.

Edenberg, H.J., and Liu, Y. (2009). Laboratory Methods for High-Throughput Genotyping. Cold Spring Harb Protoc **2009,** pdb.top62-.

Edwards, D., Forster, J.W., Chagne, D., and Batley, J. (2007). What are SNPs? Association mapping in plants**,** 41-52.

Edwards, D., Forster, J.W., Cogan, N.O.I., Batley, J., and Chagne, D. (2007). Single nucleotide polymorphism discovery. Association mapping in plants**,** 53-76.

Eldridge, K.G., Davidson, J., Harwood, C.E., and van Wyk, G. (1993). Eucalypt domestication and breeding. (Oxford, United Kingdom: Clarendon Press).

Ericsson, T., and Fries, A. (2004). Genetic analysis of fibre size in a full-sib Pinus sylvestris L. progeny test. Scandinavian Journal of Forest Research **19,** 7-13.

Ewens, W.J., and Spielman, R.S. (1995). The Transmission Disequilibrium Test - History, Subdivision and Admixture. American Journal of Human Genetics **57,** 455-464.

FAO. (2007). States of the world's forests 2007. (Rome: FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS).

FAO. (2007). ITEM 5 : Global wood and wood products flow. In Advisory committee on paper and wood products (Shanghai, China: FAO).

FAO. (2009). State of the world's forests 2009. (Rome).

Feuillet, C., Boudet, A.M., and Grimapettenati, J. (1993). Nucleotide-Sequence of a Cdna-Encoding Cinnamyl Alcohol-Dehydrogenase from Eucalyptus. Plant Physiology **103,** 1447-1447.

Flint-Garcia, S.A., Thornsberry, J.M., and Buckler, E.S. (2003). Structure of linkage disequilibrium in plants. Annual Review of Plant Biology **54,** 357-374.

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

Frazer, K.A., Ballinger, D.G., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature **449,** 851-U853.

Freeman, J.S., Potts, B.M., Shepherd, M., and Vaillancourt, R.E. (2006). Parental and consensus linkage maps of Eucalyptus globulus using AFLP and microsatellite markers. Silvae Genetica **55,** 202-217.

Freeman, J.S., Whittock, S.P., Potts, B.M., and Vaillancourt, R.E. (2009). QTL influencing growth and wood properties in Eucalyptus globulus. Tree Genetics & Genomes **5,** 713-722.

Fries, A., and Ericsson, T. (2009). Genetic parameters for earlywood and latewood densities and development with increasing age in Scots pine. Annals of Forest Science **66**.

Frisse, L.M., Bartoszewicz, A., Wall, J.D., Hudson, R.R., and Di Rienzo, A. (2001). Sequence variation and linkage disequilibrium in the human genome. American Journal of Human Genetics **69,** 1383.

Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147,** 915-925.

Fujimoto, T., Kurata, Y., Matsumoto, K., and Tsuchikawa, S. (2008). Application of near infrared spectroscopy for estimating wood mechanical properties of small clear and full length lumber specimens. Journal of near Infrared Spectroscopy **16,** 529-537.

Fukuda, H. (1996). Xylogenesis: Initiation, progression, and cell death. Annual Review of Plant Physiology and Plant Molecular Biology **47,** 299-325.

Fukuda, H. (2000). Programmed cell death of tracheary elements as a paradigm in plants. Plant Molecular Biology **44,** 245-253.

Funada, R., Furusawa, O., Shibagaki, M., Miura, H., Miura, T., Abe, H., and Ohtani, J. (2000). The role of cytoskeleton in secondary xylem differentiation in conifers. Cell and Molecular Biology of Wood Formation**,** 255-264.

Gabriel, S.B., Schaffner, S.F., *et al.* (2002). The structure of haplotype blocks in the human genome. Science **296,** 2225-2229.

Ganal, M.W., Altmann, T., and Roder, M.S. (2009). SNP identification in crop plants. Current Opinion in Plant Biology **12,** 211-217.

Gapare, W.J., Wu, H.X., and Abarquez, A. (2006). Genetic control of the time of transition from juvenile to mature wood in Pinus radiata D. Don. Annals of Forest Science **63,** 871-878.

Garg, K., Green, P., and Nickerson, D.A. (1999). Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Research **9,** 1087-1092.

Gaspar, M.J., Louzada, J.L., Silva, M.E., Aguiar, A., and Almeida, M.H. (2008). Age trends in genetic parameters of wood density components in 46 half-sibling families of Pinus pinaster. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **38,** 1470-1477.

Gaspar, M.J., Louzada, J.L., Aguiar, A., and Almeida, M.H. (2008). Genetic correlations between wood quality traits of Pinus pinaster Ait. Annals of Forest Science **65**.

Gibbs, R.A., Belmont, J.W., *et al.* (2003). The International HapMap Project. Nature **426,** 789-796.

Gilchrist, E.J., Haughn, G.W., Ying, C.C., Otto, S.P., Zhuang, J., Cheung, D., Hamberger, B., Aboutorabi, F., Kalynyak, T., Johnson, L., Bohlmann, J., Ellis, B.E., Douglas, C.J., and Cronk, Q.C.B. (2006). Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of Populus trichocarpa. Molecular Ecology **15,** 1367-1378.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. (1996). Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics **144,** 1883-1891.

Gilmour, A.R., Thompson, R., Cullis, B.R., and Welham, S.J. (2002). ASReml estimates variance matrices from multivariate data using the animal model. Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. Session 28**,** 1-2.

Gion, J.M., Rech, P., Grima-Pettenati, J., Verhaegen, D., and Plomion, C. (2000). Mapping candidate genes in Eucalyptus with emphasis on lignification genes. Molecular Breeding **6,** 441-449.

Gion J.M. , Boudet C., Grima-Pettenati J., Ham Pichavant F., Plomion C., Baillères H., Verhaegen D. (2001). A candidate genes approach identifies CCR, PAL and C4H as loci for Syringyl / Guaiacyl ratio in a interspecific hybrid between *E.urophylla* and *E. grandis* . Proceedding IUFRO Desarrolando el eucalipto del futuro 10-15 septembre,Chili.

Gion, J.M., Carouché, A., Dewaere, S., Boudet, C., Bedon, F., Pichavant, F., Charpentier, J.P., Bailleres, H., Rozenberg, P., Carocha, V., Ognouabi, N., Verhaegen, D., Grima-Pettenati, J., Vigneron, P., and Plomion, C. Genetic architecture of wood properties in Eucalyptus. soumis.

Goff, S.A., Ricke, D., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science **296,** 92-100.

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

Gonzalez-Martinez, S.C., Ersoz, E., Brown, G.R., Wheeler, N.C., and Neale, D.B. (2006). DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in Pinus taeda L. Genetics **172,** 1915-1926.

Gonzalez-Martinez, S.C., Wheeler, N.C., Ersoz, E., Nelson, C.D., and Neale, D.B. (2007). Association genetics in Pinus taeda L. I. Wood property traits. Genetics **175,** 399-409.

Gonzalez-Martinez, S.C., Huber, D., Ersoz, E., Davis, J.M., and Neale, D.B. (2008). Association genetics in Pinus taeda L. II. Carbon isotope discrimination. Heredity **101,** 19-26.

Gou, J.Y., Park, S., Yu, X.H., Miller, L.M., and Liu, C.J. (2008). Compositional characterization and imaging of "wall-bound" acylesters of Populus trichocarpa reveal differential accumulation of acyl molecules in normal and reactive woods. Planta **229,** 15-24.

Gough, G., and Barnes, R.D. (1984). A comparison of three methods of wood density assessment in a Pinus elliottii progeny test. South African Forestry Journal**,** 22-25.

Goujon, T., Sibout, R., Eudes, A., MacKay, J., and Joulanin, L. (2003). Genes involved in the biosynthesis of lignin precursors in Arabidopsis thaliana. Plant Physiology and Biochemistry **41,** 677-687.

Gouy, M., and Gautier, C. (1982). Codon Usage in Bacteria - Correlation with Gene Expressivity. Nucleic Acids Research **10,** 7055-7074.

Grans, D., Hannrup, B., Isik, F., Lundqvist, S.O., and McKeand, S. (2009). Genetic variation and relationships to growth traits for microfibril angle, wood density and modulus of elasticity in a Picea abies clonal trial in southern Sweden. Scandinavian Journal of Forest Research **24,** 494-503.

Grattapaglia, D., and Bradshaw, H.D. (1994). Nuclear-DNA Content of Commercially Important Eucalyptus Species and Hybrids. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **24,** 1074-1078.

Grattapaglia, D., and Sederoff, R. (1994). Genetic-Linkage Maps of Eucalyptus-Grandis and Eucalyptus-Urophylla Using a Pseudo-Testcross - Mapping Strategy and Rapd Markers. Genetics **137,** 1121-1137.

Grattapaglia, D., Bertolucci, F.L.G., Penchel, R., and Sederoff, R.R. (1996). Genetic mapping of quantitative trait loci controlling growth and wood quality traits in Eucalyptus grandis using a maternal half-sib family and RAPD markers. Genetics **144,** 1205-1214.

Grattapaglia, D., and Kirst, M. (2008). Eucalyptus applied genomics: from gene sequences to breeding tools. New Phytologist **179,** 911-929.

Grattapaglia, D., and Resende, M. (2010). Genomic selection in forest tree breeding. Tree Genetics & Genomes**,** 1-15.

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

Griffin, A.R., Burgess, I.P., and Wolf, L. (1988). Patterns of Natural and Manipulated Hybridization in the Genus Eucalyptus Lherit - a Review. Australian Journal of Botany **36,** 41-66.

Grimapettenati, J., Feuillet, C., Goffner, D., Borderies, G., and Boudet, A.M. (1993). Molecular-Cloning and Expression of a Eucalyptus-Gunnii Cdna Clone Encoding Cinnamyl Alcohol-Dehydrogenase. Plant Molecular Biology **21,** 1085-1095.

Grimmig, B., and Matern, U. (1997). Structure of the parsley caffeoyl-CoA O-methyltransferase gene, harbouring a novel elicitor responsive cis-acting element. Plant Molecular Biology **33,** 323-341.

Groover, A., and Jones, A.M. (1999). Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. Plant Physiology **119,** 375-384.

Guillaumie, S., Mzid, R., Mechin, V., Leon, C., Hichri, I., Destrac-Irvine, A., Trossat-Magnin, C., Delrot, S., and Lauvergeat, V. (2010). The grapevine transcription factor WRKY2 influences the lignin pathway and xylem development in tobacco. Plant Molecular Biology **72,** 215-234.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., and Ghosh, T.C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochemical and Biophysical Research Communications **269,** 692-696.

Haigler, C.H., Ivanova-Datcheva, M., Hogan, P.S., Salnikov, V.V., Hwang, S., Martin, K., and Delmer, D.P. (2001). Carbon partitioning to cellulose synthesis. Plant Molecular Biology **47,** 29-51.

Hallingback, H.R., Jansson, G., and Hannrup, B. (2010). Genetic correlations between spiral grain and growth and quality traits in Picea abies. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **40,** 173-183.

Halpin, C., Knight, M.E., Foxon, G.A., Campbell, M.M., Boudet, A.M., Boon, J.J., Chabbert, B., Tollier, M.T., and Schuch, W. (1994). Manipulation of Lignin Quality by Down-Regulation of Cinnamyl Alcohol-Dehydrogenase. Plant Journal **6,** 339-350.

Hamberger, B., Ellis, M., Friedmann, M., Souza, C.D.A., Barbazuk, B., and Douglas, C.J. (2007). Genome-wide analyses of phenylpropanoid-related genes in Populus trichocarpa, Arabidopsis thaliana, and Oryza sativa: the Populus lignin toolbox and conservation and diversification of angiosperm gene families. Canadian Journal of Botany-Revue Canadienne De Botanique **85,** 1182-1201.

Hamilton, M.G., and Potts, B.M. (2008). Eucalyptus nitens genetic parameters. New Zealand Journal of Forestry Science **38,** 102-119.

Hamrick, J.L., Godt, M.J.W., and Sherman-Broyles, S.L. (1992). Factors influencing levels of genetic diversity in woody plant species. New Forests **6,** 95-124.

Hannrup, B., Cahalan, C., Chantre, G., Grabner, M., Karlsson, B., Le Bayon, I., Jones, G.L., Muller, U., Pereira, H., Rodrigues, J.C., Rosner, S., Rozenberg, P., Wilhelmsson, L., and Wimmer, R. (2004). Genetic parameters of growth and wood quality traits in Picea abies. Scandinavian Journal of Forest Research **19,** 14-29.

Hanzawa, Y., Sasaki, T., Mizugaki, M., Ishikawa, M., and Hiratsuka, M. (2008). Genetic polymorphisms and haplotype structures of the human CYP2W1 gene in a Japanese population. Drug Metabolism and Disposition **36,** 349-352.

Harakava, R. (2005). Genes encoding enzymes of the lignin biosynthesis pathway in Eucalyptus. Genetics and Molecular Biology **28,** 601-607.

Harrigan, R.J., Mazza, M.E., and Sorenson, M.D. (2008). Computation vs. cloning: evaluation of two methods for haplotype determination. Molecular Ecology Resources **8,** 1239-1248.

Hatfield, R., and Fukushima, R.S. (2005). Can lignin be accurately measured? Crop Science **45,** 832-839.

Hattersley, A.T., and McCarthy, M.I. (2005). Genetic Epidemiology 5 - What makes a good genetic association study? Lancet **366,** 1315-1323.

Hatton, D., Sablowski, R., Yung, M.H., Smith, C., Schuch, W., and Bevan, M. (1995). 2 Classes of Cis Sequences Contribute to Tissue-Specific Expression of a Pal2 Promoter in Transgenic Tobacco. Plant Journal **7,** 859-876.

Hauffe, K.D., Lee, S.P., Subramaniam, R., and Douglas, C.J. (1993). Combinatorial Interactions between Positive and Negative Cis-Acting Elements Control Spatial Patterns of 4cl-1 Expression in Transgenic Tobacco. Plant Journal **4,** 235-253.

Hayashi, K., Hashimoto, N., Daigen, M., and Ashikawa, I. (2004). Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. Theoretical and Applied Genetics **108,** 1212-1220.

Hebenbrock, K., Williams, P.M., and Karger, B.L. (1995). Single-Strand Conformational Polymorphism Using Capillary Electrophoresis with 2-Dye Laser-Induced Fluorescence Detection. Electrophoresis **16,** 1429-1436.

Hein, P.R.G., Campos, A.C.M., Trugilho, P.F., Lima, J.T., and Chaix, G. (2009). Near infrared spectroscopy for estimating wood basic density in Eucalyptus urophylla and Eucalyptus grandis. Cerne **15,** 133-141.

Hein, P.R.G., Lima, J.T., and Chaix, G. (2010). Effects of sample preparation on NIR spectroscopic estimation of chemical properties of Eucalyptus urophylla ST Blake wood. Holzforschung **64,** 45-54.

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

Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics **38,** 226-231.

Hill, W.G., and Weir, B.S. (1988). Variances and Covariances of Squared Linkage Disequilibria in Finite Populations. Theoretical Population Biology **33,** 54-78.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science **307,** 1072-1079.

Hirschhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genetics in Medicine **4,** 45-61.

Hodge, G.R., and Woodbridge, W.C. (2004). Use of near infrared spectroscopy to predict lignin content in tropical and sub-tropical pines. Journal of near Infrared Spectroscopy **12,** 381-390.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.V., Skovmand, B., Taba, S., and Warburton, M. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? Proceedings of the National Academy of Sciences of the United States of America **96,** 5937-5943.

Holman, J.E., Hughes, J.M., and Fensham, R.J. (2003). A morphological cline in Eucalyptus: a genetic perspective. Molecular Ecology **12,** 3013-3025.

Hosokawa, M., Suzuki, S., Umezawa, T., and Sato, Y. (2001). Progress of lignification mediated by intercellular transportation of monolignols during tracheary element differentiation of isolated Zinnia mesophyll cells. Plant and Cell Physiology **42,** 959-968.

Hsia, A.P., Wen, T.J., Chen, H.D., Liu, Z.W., Yandeau-Nelson, M.D., Wei, Y.L., Guo, L., and Schnable, P.S. (2005). Temperature gradient capillary electrophoresis (TGCE) - a tool for the high-throughput discovery and mapping of SNPs and IDPs. Theoretical and Applied Genetics **111,** 218-225.

Hudson, L. (1985). Experimental Studies on Trypanosoma-Cruzi. Annales De La Societe Belge De Medecine Tropicale **65,** 71-77.

Huttley, G.A., Smith, M.W., Carrington, M., and O'Brien, S.J. (1999). A scan for linkage disequilibrium across the human genome. Genetics **152,** 1711-1722.

Hylen, G. (1999). Age trends in genetic parameters of wood density in young Norway spruce. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **29,** 135-143.

Iglesias, G., and Wiltermann, D. (2008). Eucalyptus universalis: Global cultivated eucalypt forests map. Version 1.0. In Eucalyptologics information resources on eucalypt cultivation worldwide, www.git-forestry.com, ed.

Im, K.H., Cosgrove, D.J., and Jones, A.M. (2000). Subcellular localization of expansin mRNA in xylem cells. Plant Physiology **123,** 463-470.

Ingvarsson, P.K. (2005). Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (Populus tremula L., Salicaceae). Genetics **169,** 945-953.

Ingvarsson, P.K., Garcia, M.V., Luquez, V., Hall, D., and Jansson, S. (2008). Nucleotide polymoirphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (Populus tremula, Salicaceae). Genetics **178,** 2217-2226.

Isik, F., and Li, B.L. (2003). Rapid assessment of wood density of live trees using the Resistograph for selection in tree improvement programs. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **33,** 2426-2435.

Isik, F., Li, B.L., Goldfarb, B., and McKeand, S. (2008). Prediction of wood density breeding values of Pinus taeda elite parents from unbalanced data: a method for adjustment of site and age effects using common checklots. Annals of Forest Science **65,** 406p401-406p412.

ITTO. (2006). Status of Tropical Forest Management 2005. (Yokohama. Japan).

Ivkovich, M., Namkoong, G., and Koshy, M. (2002). Genetic variation in wood properties of interior spruce. I. Growth, latewood percentage, and wood density. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **32,** 2116-2127.

Jeffreys, A.J., Holloway, K., Kauppi, L., May, C., Neumann, R., Slingsby, T., and Taylor, T. (2002). Patterns of human meoitic recombination. European Journal of Human Genetics **10,** 53-53.

Jones, R.C., Steane, D.A., Potts, B.M., and Vaillancourt, R.E. (2002). Microsatellite and morphological analysis of Eucalyptus globulus populations. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **32,** 59-66.

Jovanovic, T., and Booth, T.H. (2002). Improved species climatic profiles. (Kingston: RIRDC Publications), pp. 30-31, 46-47.

Kado, T., Yoshimaru, H., Tsumura, Y., and Tachida, H. (2003). DNA Variation in a Conifer, Cryptomeria japonica (Cupressaceae sensu lato). Genetics **164,** 1547-1559.

Kauppi, L., Sajantila, A., and Jeffreys, A.J. (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Human Molecular Genetics **12,** 33-40.

Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K., and Ebinuma, H. (2000). Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. Plant Journal **22,** 289-301.

Kelley, S.S., Rowell, R.M., Davis, M., Jurich, C.K., and Ibach, R. (2004). Rapid analysis of the chemical composition of agricultural fibers using near infrared spectroscopy and pyrolysis molecular beam mass spectrometry. Biomass & Bioenergy **27,** 77-88.

Kermicle, J.L., Eggleston, W.B., and Alleman, M. (1995). Organization of Paramutagenicity in R-Stippled Maize. Genetics **141,** 361-372.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. Nature Genetics **39,** 1151-1155.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61,** 893-903.

Kingsmore, S.F., Lindquist, I.E., Mudge, J., Gessler, D.D., and Beavis, W.D. (2008). Genome-wide association studies: progress and potential for drug discovery and development. Nature Reviews Drug Discovery **7,** 221-230.

Kirst, M., Myburg, A.A., De Leon, J.P.G., Kirst, M.E., Scott, J., and Sederoff, R. (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiology **135,** 2368-2378.

Kirst, M., Cordeiro, C.M., Rezende, G., and Grattapaglia, D. (2005). Power of microsatellite markers for fingerprinting and parentage analysis in Eucalyptus grandis breeding populations. Journal of Heredity **96,** 161-166.

Knowler, W.C., Williams, R.C., Pettitt, D.J., and Steinberg, A.G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture.

Konieczny, A., and Ausubel, F.M. (1993). A Procedure for Mapping Arabidopsis Mutations Using Codominant Ecotype-Specific Pcr-Based Markers. Plant Journal **4,** 403-410.

Kota, R., Wolf, M., Michalek, W., and Graner, A. (2001). Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (Hordeum vulgare L.). Genome **44,** 523-528.

Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K., and Graner, A. (2003). Snipping polymorphisms from large EST collections in barley (Hordeum vulgare L.). Molecular Genetics and Genomics **270,** 24-33.

Koutaniemi, S., Warinowski, T., Karkonen, A., Alatalo, E., Fossdal, C.G., Saranpaa, P., Laakso, T., Fagerstedt, K.V., Simola, L.K., Paulin, L., Rudd, S., and Teeri, T.H. (2007). Expression profiling of the lignin biosynthetic pathway in Norway spruce using EST sequencing and real-time RT-PCR. Plant Molecular Biology **65,** 311-328.

Kretschmann, D. (2003). Natural materials: Velcro mechanics in wood. Nat Mater **2,** 775-776.

Kruglyak, L. (2008). The road to genome-wide association studies. Nature Reviews Genetics **9,** 314-318.

Krutovsky, K.V., and Neale, D.B. (2005). Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics **171,** 2029-2041.

Krypuy, M., Newnham, G.M., Thomas, D.M., Conron, M., and Dobrovic, A. (2006). High resolution melting analysis for the rapid and sensitive detection of mutations in clinical samples: KRAS codon 12 and 13 mutations in non-small cell lung cancer. Bmc Cancer **6**.

Kube, P.D., Raymond, C.A., and Banham, P.W. (2001). Genetic parameters for diameter, basic density, cellulose content and fibre properties for Eucalyptus nitens. Forest Genetics **8,** 285-294.

Kube, P.D., and Raymond, C.A. (2002). Prediction of whole-tree basic density and pulp yield using wood core samples in Eucalyptus nitens. Appita Journal **55,** 43-48.

Kuhn, D.N., Borrone, J., Meerow, A.W., Motamayor, J.C., Brown, J.S., and Schnell, R.J. (2005). Single-strand conformation polymorphism analysis of candidate genes for reliable identification of alleles by capillary array electrophoresis. Electrophoresis **26,** 112-125.

Kulheim, C., Yeoh, S.H., Maintz, J., Foley, W.J., and Moran, G.F. (2009). Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. Bmc Genomics **10**.

Kumar, S. (2004). Genetic parameter estimates for wood stiffness, strength, internal checking, and resin bleeding for radiata pine. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **34,** 2601-2610.

Kuriyama, H. (1999). Loss of tonoplast integrity programmed in tracheary element differentiation. Plant Physiology **121,** 763-774.

Lacombe, E., Hawkins, S., VanDoorsselaere, J., Piquemal, J., Goffner, D., Poeydomenge, O., Boudet, A.M., and GrimaPettenati, J. (1997). Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: Cloning, expression and phylogenetic relationships. Plant Journal **11,** 429-441.

Lacombe, E., Van Doorsselaere, J., Boerjan, W., Boudet, A.M., and Grima-Pettenati, J. (2000). Characterization of cis-elements required for vascular expression of the Cinnamoyl CoA Reductase gene and for protein-DNA complex formation. Plant Journal **23,** 663-676.

Ladiges, P.Y., Udovicic, F., and Nelson, G. (2003). Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. Journal of Biogeography **30,** 989-998.

Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nature Reviews Genetics **7,** 385-394.

Lam, E. (2004). Controlled cell death, plant survival and development. Nature Reviews Molecular Cell Biology **5,** 305-315.

Lander, E.S., Linton, L.M., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature **409,** 860-921.

Lapierre, C., Pollet, B., and Rolando, C. (1995). New Insights into the Molecular Architecture of Hardwood Lignins by Chemical Degradative Methods. Research on Chemical Intermediates **21,** 397-412.

Lauvergeat, V., Rech, P., Jauneau, A., Guez, C., Coutos-Thevenot, P., and Grima-Pettenati, J. (2002). The vascular expression pattern directed by the Eucalyptus gunnii cinnamyl alcohol dehydrogenase EgCAD2 promoter is conserved among woody and herbaceous plant species. Plant Molecular Biology **50,** 497-509.

Le Dantec, L., Chagne, D., Pot, D., Cantin, O., Garnier-Gere, P., Bedon, F., Frigerio, J.M., Chaumeil, P., Leger, P., Garcia, V., Laigret, F., de Daruvar, A., and Plomion, C. (2004). Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Molecular Biology **54,** 461-470.

Legay, S., Lacombe, E., Goicoechea, M., Briere, C., Seguin, A., Mackay, J., and Grima-Pettenati, J. (2007). Molecular characterization of EgMYB1, a putative transcriptional repressor of the lignin biosynthetic pathway. Plant Science **173,** 542-549.

Lenz, P., Cloutier, A., MacKay, J., and Beaulieu, J. (2010). Genetic control of wood properties in Picea glauca - an analysis of trends with cambial age. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **40,** 703-715.

Leple, J.C., Dauwe, R., *et al.* (2007). Downregulation of cinnamoyl-coenzyme a reductase in poplar: Multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. Plant Cell **19,** 3669-3691.

Lepoittevin, C., Frigerio, J.M., Garnier-Gere, P., Salin, F., Cervera, M.T., Vornam, B., Harvengt, L., and Plomion, C. (2010). In Vitro vs In Silico Detected SNPs for the Development of a Genotyping Array: What Can We Learn from a Non-Model Species? Plos One **5**.

Lewontin, R.C., and Kojima, K.-i. (1960). The Evolutionary Dynamics of Complex Polymorphisms. Evolution **14,** 458-472.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.

Li, B., McKeand, S., and Weir, R. (1999). Tree improvement and sustainable forestry - impact of two cycles of loblolly pine breeding in the U.S.A. Forest Genetics **6,** 229-234.

Li, L.G., Cheng, X.F., Leshkevich, J., Umezawa, T., Harding, S.A., and Chiang, V.L. (2001). The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. Plant Cell **13,** 1567-1585.

Li, L., and Wu, H.X. (2005). Efficiency of early selection for rotation-aged growth and wood density traits in Pinus radiata. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **35,** 2019-2029.

Libby, W.J., and Palmberg-Lerche, C. (2002). Forest plantation productivity. (Rome: FAO).

Liu, J.S., Sabatti, C., Teng, J., Keats, B.J.B., and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Research **11,** 1716-1724.

Long, A.D., and Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research **9,** 720-731.

Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., and Verdier, V. (2005). Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (Manihot esculenta Crantz). Theoretical and Applied Genetics **110,** 425-431.

Ma, X.F., Szmidt, A.E., and Wang, X.R. (2006). Genetic structure and evolutionary history of a diploid hybrid pine Pinus densata inferred from the nucleotide variation at seven gene loci. Molecular Biology and Evolution **23,** 807-816.

MacDonald, A.C., Borralho, N.M.G., and Potts, B.M. (1997). Genetic variation for growth and wood density in Eucalyptus globulus ssp. globulus in Tasmania (Australia). Silvae Genetica **46,** 236-241.

Manolio, T.A., Collins, F.S., *et al.* (2009). Finding the missing heritability of complex diseases. Nature **461,** 747-753.

Mardis, E.R. (2007). ChIP-seq: welcome to the new frontier. Nature Methods **4,** 613-614.

Markussen, T., Fladung, M., Achere, V., Favre, J.M., Faivre-Rampant, P., Aragones, A., Perez, D.D., Harvengt, L., Espinel, S., and Ritter, E. (2003). Identification of QTLs controlling growth, chemical and physical wood property traits in Pinus pinaster (Ait.). Silvae Genetica **52,** 8-15.

Marques, C.M., Araujo, J.A., Ferreira, J.G., Whetten, R., O'Malley, D.M., Liu, B.H., and Sederoff, R. (1998). AFLP genetic maps of Eucalyptus globulus and E-tereticornis. Theoretical and Applied Genetics **96,** 727-737.

Marques, C.M., Brondani, R.P.V., Grattapaglia, D., and Sederoff, R. (2002). Conservation and synteny of SSR loci and QTLs for vegetative propagation in four Eucalyptus species. Theoretical and Applied Genetics **105,** 474-478.

Martin, B., and Cossalter, C. (1976). The Eucalypts of the Sunda Islands [Part 3]. Bois et Forets des Tropiques**,** 3-20.

Martin, B., and Cossalter, C. (1976). The Eucalypts of the Sunda Islands [Part 4]. Bois et Forets des Tropiques**,** 3-22.

Martin, B., and Cossalter, C. (1976). The Eucalypts of the Sunda Islands [Part 5]. Bois et Forets des Tropiques**,** 3-24.

Martinez-Arias, R., Mateu, E., Bertranpetit, J., and Calafell, F. (2001). Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene. Human Genetics **109,** 7-10.

Martinez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andres, A., and Bertranpetit, J. (2001). Sequence variability of a human pseudogene. Genome Research **11,** 1071-1085.

McGowen, M.H., Wiltshire, R.J.E., Potts, B.M., and Vaillancourt, R.E. (2001). The origin of Eucalyptus vernicosa, a unique shrub eucalypt. Biological Journal of the Linnean Society **74,** 397-405.

McMillin, C.W. (1968). Morphological characteristics of loblolly pine wood as related to specific gravity, growth rate and distance from pith. Wood Science and Technology **2,** 166-176.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science **304,** 581-584.

Meder, R., Gallagher, S., Mackie, K.L., Bohler, H., and Meglen, R.R. (1999). Rapid determination of the chemical composition and density of Pinus radiata by PLS modelling of transmission and diffuse reflectance FTIR spectra. Holzforschung **53,** 261-266.

Megraw, R.A. (1985). Wood quality factors in loblolly pine. The influence of tree age, position in tree, and cultural practice on wood specific gravity, fiber length, and fibril angle. Wood quality factors in loblolly pine. The influence of tree age, position in tree, and cultural practice on wood specific gravity, fiber length, and fibril angle.**,** xii + 88pp.

Mellerowicz, E.J., Baucher, M., Sundberg, B., and Boerjan, W. (2001). Unravelling cell wall formation in the woody dicot stem. Plant Molecular Biology **47,** 239-274.

Meuwissen, T.H.E., and Goddard, M.E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155,** 421-430.

Meuwissen, T.H.E., and Goddard, M.E. (2001). Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution **33,** 605-634.

Meyer, M., Stenzel, U., Myles, S., Prufer, K., and Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Research **35**.

Miranda, I., Almeida, M.H., and Pereira, H. (2001). Influence of provenance, subspecies, and site on wood density in Eucalyptus globulus Labill. Wood and Fiber Science **33,** 9-15.

Mizutani, M., Ohta, D., and Sato, R. (1997). Isolation of a cDNA and a genomic clone encoding cinnamate 4-hydroxylase from Arabidopsis and its expression manner in planta. Plant Physiology **113,** 755-763.

Molitor, J., Marjoram, P., and Thomas, D. (2003). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. American Journal of Human Genetics **73,** 1368-1384.

Moriyama, E.N., and Powell, J.R. (1998). Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Research **26,** 3188-3193.

Mott, L., Groom, L., and Shaler, S. (2002). Mechanical properties of individual southern pine fibers. Part II. Comparison of earlywood and latewood fibers with respect to tree height and juvenility. Wood and Fiber Science **34,** 221-237.

Moura, V.P.G., Barnes, R.D., and Birks, J.S. (1987). A Comparison of 3 Methods of Assessing Wood Density in Provenances of Eucalyptus-Camaldulensis Dehnh and Other Eucalyptus Species in Brazil. Australian Forest Research **17,** 83-90.

Moura, J., Bonine, C.A.V., Viana, J.D.F., Dornelas, M.C., and Mazzafera, P. (2010). Abiotic and Biotic Stresses and Changes in the Lignin Content and Composition in Plants. Journal of Integrative Plant Biology **52,** 360-376.

Myburg, A.A., Potts, B.M., Marques, C.M., Kirst, M., Gion, J.M., Grattapaglia, D., and Grima-Pettenatti, J. (2007). Eucalypts. In Genome Mapping and Molecular Breeding in Plants, Forest Trees, C. Kole, ed (Berlin Heidelberg, Deutschland: Springer-Verlag), pp. 115-160.

Myers, R.M., Sheffield, V.C., and Cox, D.R. (1988). Detection of single base changes in DNA: ribonuclease cleavage and denaturing gradient gel electrophoresis. In Genome Analysis: A Practical Approach, K.E. Davies, Editor, ed (Oxford (UK): IRL Press), pp. 95–139.

Nakashima, J., Chen, F., Jackson, L., Shadle, G., and Dixon, R.A. (2008). Multi-site genetic modification of monolignol biosynthesis in alfalfa (Medicago sativa): effects on lignin composition in specific cell types. New Phytologist **179,** 738-750.

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

Neale, D.B. (2007). Genomics to tree breeding and forest health. Current Opinion in Genetics & Development **17,** 539-544.

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

Ni, W.T., Paiva, N.L., and Dixon, R.A. (1994). Reduced Lignin in Transgenic Plants Containing a Caffeic Acid O-Methyltransferase Antisense Gene. Transgenic Research **3,** 120-126.

Nicholls, J.W.P. (1985). A New Method for Determining Wood Density in the Standing Tree. Australian Forest Research **15,** 195-206.

Nielsen, D.M., and Zaykin, D. (2001). Association mapping: where weve been, where were going. Expert Review of Molecular Diagnostics **1,** 334-342.

Nordborg, M., and Tavare, S. (2002). Linkage disequilibrium: what history has to tell us. Trends in Genetics **18,** 83-90.

Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R., and Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. Bmc Genomics **9**.

Novaes, E., Osorio, L., Drost, D.R., Miles, B.L., Boaventura-Novaes, C.R.D., Benedict, C., Dervinis, C., Yu, Q., Sykes, R., Davis, M., Martin, T.A., Peter, G.F., and Kirst, M. (2009). Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels. New Phytologist **182,** 878-890.

Obara, K., Kuriyama, H., and Fukuda, H. (2001). Direct evidence of active and rapid nuclear degradation triggered by vacuole rupture during programmed cell death in Zinnia. Plant Physiology **125,** 615-626.

Oda, Y., and Hasezawa, S. (2006). Cytoskeletal organization during xylem cell differentiation. Journal of Plant Research **119,** 167-177.

Okagaki, R.J., and Weil, C.F. (1997). Analysis of recombination sites within the maize waxy locus. Genetics **147,** 815-821.

Olson, M.S., Robertson, A.L., Takebayashi, N., Silim, S., Schroeder, W.R., and Tiffin, P. (2010). Nucleotide diversity and linkage disequilibrium in balsam poplar (Populus balsamifera). New Phytologist **186,** 526-536.

Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., Silva, H.N., Edwards, D., Forster, J.W., Chagné, D., and Batley, J. (2007). What Are SNPs? In Association Mapping in Plants (Springer New York), pp. 41-52.

Orita, M., Suzuki, Y., Sekiya, T., and Hayashi, K. (1989). Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics **5,** 874-879.

Pallett, R.N., and Sale, G. (2004). The relative contributions of tree improvement and cultural practice toward productivity gains in Eucalyptus pulpwood stands. Forest Ecology and Management **193,** 33-43.

Palme, A.E., Wright, M., and Savolainen, O. (2008). Patterns of Divergence among Conifer ESTs and Polymorphism in Pinus sylvestris Identify Putative Selective Sweeps. Molecular Biology and Evolution **25,** 2567-2577.

Palumbi, S.R., and Baker, C.S. (1994). Contrasting Population-Structure from Nuclear Intron Sequences and Mtdna of Humpback Whales. Molecular Biology and Evolution **11,** 426-435.

Passialis, C., and Kiriazakos, A. (2004). Juvenile and mature wood properties of naturally-grown fir trees. Holz Als Roh-Und Werkstoff **62,** 476-478.

Patzlaff, A., McInnis, S., Courtenay, A., Surman, C., Newman, L.J., Smith, C., Bevan, M.W., Mansfield, S., Whetten, R.W., Sederoff, R.R., and Campbell, M.M. (2003). Characterisation of a pine MYB that regulates lignification. Plant Journal **36,** 743-754.

Paux, E., Tamasloukht, M., Ladouce, N., Sivadon, P., and Grima-Pettenati, J. (2004). Identification of genes preferentially expressed during wood formation in Eucalyptus. Plant Molecular Biology **55,** 263-280.

Pavy, N., Parsons, L.S., Paule, C., MacKay, J., and Bousquet, J. (2006). Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. Bmc Genomics **7**.

Payn, K.G., Dvorak, W.S., Janse, B.J.H., and Myburg, A.A. (2008). Microsatellite diversity and genetic structure of the commercially important tropical tree species Eucalyptus urophylla, endemic to seven islands in eastern Indonesia. Tree Genetics & Genomes **4,** 519-530.

Pichon, M., Courbou, I., Beckert, M., Boudet, A.M., and Grima-Pettenati, J. (1998). Cloning and characterization of two maize cDNAs encoding Cinnamoyl-CoA Reductase (CCR) and differential expression of the corresponding genes. Plant Molecular Biology **38,** 671-676.

Plomion, C., Leprovost, G., and Stokes, A. (2001). Wood formation in trees. Plant Physiology **127,** 1513-1523.

Poke, F.S., Vaillancourt, R.E., Elliott, R.C., and Reid, J.B. (2003). Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). Molecular Breeding **12,** 107-118.

Poke, F.S., Vaillancourt, R.E., Potts, B.M., and Reid, J.B. (2005). Genomic research in Eucalyptus. Genetica **125,** 79-101.

Poke, F.S., Potts, B.M., Vaillancourt, R.E., and Raymond, C.A. (2006)a. Genetic parameters for lignin, extractives and decay in Eucalyptus globulus. Annals of Forest Science **63,** 813-821.

Poke, F.S., and Raymond, C.A. (2006)b. Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in Eucalyptus globulus. Journal of Wood Chemistry and Technology **26,** 187-199.

Pot, D., Chantre, G., Rozenberg, P., Rodrigues, J.C., Jones, G.L., Pereira, H., Hannrup, B., Cahalan, C., and Plomion, C. (2002). Genetic control of pulp and timber properties in maritime pine (Pinus pinaster Ait.). Annals of Forest Science **59,** 563-575.

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

Pot, D., Rodrigues, J.C., Rozenberg, P., Chantre, G., Tibbits, J., Cahalan, C., Pichavant, F., and Plomion, C. (2006). QTLs and candidate genes for wood properties in maritime pine (Pinus pinaster Ait.). Tree Genetics & Genomes **2,** 10-24.

Potts, B.M., and Wiltshire, R.J.E. (1997). Eucalypt genetics and genecology. In Eucalypt Ecology: Individuals to Ecosystems, J.a.W. Williams, J., ed (Cambridge: Cambridge University Press), pp. 56–91.

Potts, B.M., and Dungey, H.S. (2004). Interspecific hybridization of Eucalyptus: key issues for breeders and geneticists. New Forests **27,** 115-138.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38,** 904-909.

Pritchard, J.K., and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics **65,** 220-228.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P. (2000). Association mapping in structured populations. American Journal of Human Genetics **67,** 170-181.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics **155,** 945-959.

Quang, N.D., Ikeda, S., and Harada, K. (2008). Nucleotide variation in Quercus crispula Blume. Heredity **101,** 166-174.

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing, R.D.C. Team, ed (Vienna, Austria).

Raes, J., Rohde, A., Christensen, J.H., Van de Peer, Y., and Boerjan, W. (2003). Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiology **133,** 1051-1071.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology **5,** 94-100.

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

Ralph, J., Lundquist, K., Brunow, G., Lu, F., Kim, H., Schatz, P.F., Marita, J.M., Hatfield, R.D., Ralph, S.A., Christensen, J.H., and Boerjan, W. (2004). Lignins: Natural polymers from oxidative coupling of 4-hydroxyphenyl- propanoids. Phytochemistry Reviews **3,** 29-60.

Raymond, C.A., and Greaves, B.L. (1997). Developing breeding objectives for kraft and cold caustic soak (CCS) pulping of eucalypts. Timber management toward wood quality and end-product value. Proceedings of the CTIA/IUFRO International Wood Quality Workshop, Quebec City, Canada, August 18-22, 1997**.,** IV.27-IV.34.

Raymond, C.A., and Muneri, A. (2000). Effect of fertilizer on wood properties of Eucalyptus globulus. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **30,** 136-144.

Raymond, C.A. (2002). Genetics of Eucalyptus wood properties. Annals of Forest Science **59,** 525-531.

Raymond, C.A., and Apiolaza, L.A. (2004). Incorporating wood quality and deployment traits in Eucalyptus globulus and Eucalyptus nitens. Plantation forest biotechnology for the 21st century**,** 87-99.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends in Genetics **17,** 502-510.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., and Lander, E.S. (2001). Linkage disequilibrium in the human genome. Nature **411,** 199-204.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doeblay, J., Kresovich, S., Goodman, M.M., and Buckler, E.S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences of the United States of America **98,** 11479-11484.

Robinson, A.R., and Mansfield, S.D. (2009). Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and near-infrared reflectance-based prediction modeling. Plant Journal **58,** 706-714.

Rocha, R.B., Barros, E.G.a., Cruz, C.D.o., Rosado, A.n.M., and AraÃºjo, E.F.d. (2007). Mapping of QTLs related with wood quality and developmental characteristics in hybrids (Eucalyptus grandis x Eucalyptus urophylla). Revista Ã□rvore **31,** 13-24.

Rodrigues, J., Meier, D., Faix, O., and Pereira, H. (1999). Determination of tree to tree variation in syringyl/guaiacyl ratio of Eucalyptus globulus wood lignin by analytical pyrolysis. Journal of Analytical and Applied Pyrolysis **48,** 121-128.

Rogers, L.A., and Campbell, M.M.C. (2004). The genetic control of lignin deposition during plant growth and development. New Phytologist **164,** 30.

Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P.R., Roose, M.L., Svensson, J.T., Stein, N., Varshney, R.K., Marshall, D.F., Grainer, A., Close, T.J., and Waugh, R. (2006). Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proceedings of the National Academy of Sciences of the United States of America **103,** 18656-18661.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19,** 2496-2497.

Ruel, K., Berrio-Sierra, J., Derikvand, M.M., Pollet, B., Thevenin, J., Lapierre, C., Jouanin, L., and Joseleau, J.P. (2009). Impact of CCR1 silencing on the assembly of lignified secondary walls in Arabidopsis thaliana. New Phytologist **184,** 99-113.

Sabbagh, A., and Darlu, P. (2005). Inferring haplotypes at the NAT2 locus: the computational approach. Bmc Genetics **6**.

Sahana, G., Guldbrandtsen, B., Janss, L., and Lund, M.S. (2010). Comparison of Association Mapping Methods in a Complex Pedigreed Population. Genetic Epidemiology **34,** 455-462.

Samuels, A.L., Kaneda, M., and Rensing, K.H. (2006). The cell biology of wood formation: from cambial divisions to mature secondary xylem. Canadian Journal of Botany-Revue Canadienne De Botanique **84,** 631-639.

Sano, A., and Tachida, H. (2005). Gene genealogy and properties of test statistics of neutrality under population growth. Genetics **169,** 1687-1697.

Savolainen, O., and Pyhajarvi, T. (2007). Genomic diversity in forest trees. Current Opinion in Plant Biology **10,** 162-167.

Schaid, D.J., Sinnwell, J.P., and Thibodeau, S.N. (2005). U-statistics for testing the association of genotype similarity with trait similarity: Methods for quantitative and censored traits. Genetic Epidemiology **29,** 142.

Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T., and Weisshaar, B. (2003). Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Research **13,** 1250-1257.

Schmidt, E.A. (2005). A practical model relating kraft pulping costs to hardwood chemical properties and morphology. Appita Journal **58,** 218-224.

Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M., and Sandberg, G. (2004). A high-resolution transcript profile across the wood-forming

meristem of poplar identifies potential regulators of cambial stem cell identity. Plant Cell **16,** 2278-2292.

Sharp, P.M., and Li, W.H. (1987). The Codon Adaptation Index - a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications. Nucleic Acids Research **15,** 1281-1295.

Shepherd, M., Sexton, T.R., Thomas, D., Henson, M., and Henry, R.J. (2010). Geographical and historical determinants of microsatellite variation in Eucalyptus pilularis. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **40,** 1051-1063.

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., and Lubberstedt, T. (2007). Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint * Flint maize recombinant inbred line population. BMC Genomics **8,** (18 January 2007).

Shi, R., Sun, Y.H., Li, Q.Z., Heber, S., Sederoff, R., and Chiang, V.L. (2010). Towards a Systems Approach for Lignin Biosynthesis in Populus trichocarpa: Transcript Abundance and Specificity of the Monolignol Biosynthetic Genes. Plant and Cell Physiology **51,** 144-163.

Sibout, R., Eudes, A., Pollet, B., Goujon, T., Mila, I., Granier, F., Seguin, A., Lapierre, C., and Jouanin, L. (2003). Expression pattern of two paralogs encoding cinnamyl alcohol dehydrogenases in Arabidopsis. Isolation and characterization of the corresponding mutants. Plant Physiology **132,** 848-860.

Šidàk, Z. (1967). Rectangular confidence region for themeans of multivariate normal distributions. Journal of the American Statistical Association **62,** 626-633.

Slotte, T., Huang, H.-R., Holm, K., Ceplitis, A., Onge, K.S., Chen, J., Lagercrantz, U., and Lascoux, M. (2009). Splicing Variation at a FLOWERING LOCUS C Homeolog Is Associated With Flowering Time Variation in the Tetraploid Capsella bursa-pastoris. Genetics **183,** 337-345.

Smith, S., Hughes, J., and Wardell-Johnson, G. (2003). High population differentiation and extensive clonality in a rare mallee eucalypt: Eucalyptus curtisii - Conservation genetics of a rare mallee eucalypt. Conservation Genetics **4,** 289-300.

Somerville, C. (2006). Cellulose synthesis in higher plants. Annual Review of Cell and Developmental Biology **22,** 53-78.

Spielman, R.S., and Ewens, W.J. (1993). Transmission Disequilibrium Test (Tdt) for Linkage and Linkage Disequilibrium between Disease and Marker. American Journal of Human Genetics **53,** 863-863.

Stackpole, D.J., Vaillancourt, R.E., Aguigar, M.d., and Potts, B.M. (2010). Age trends in genetic parameters for growth and wood density in Eucalyptus globulus. Tree Genetics and Genomes **6,** 179-193.

Stackpole, D.J., Vaillancourt, R.E., Downes, G.M., Harwood, C.E., and Potts, B.M. (2010). Genetic control of kraft pulp yield in Eucalyptus globulus. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **40,** 917-927.

Steane, D.A., Byrne, M., Vaillancourt, R.E., and Potts, B.M. (1998). Chloroplast DNA polymorphism signals complex interspecific interactions in Eucalyptus (Myrtaceae). Australian Systematic Botany **11,** 25-40.

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics **68,** 978-989.

Stephens, J.C., Schneider, J.A., *et al.* (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science **293,** 489-493.

Stephens, M., and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics **73,** 1162-1169.

Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nature Reviews Genetics **10,** 681-690.

Stewart, D., Yahiaoui, N., McDougall, G.J., Myton, K., Marque, C., Boudet, A.M., and Haigh, J. (1997). Fourier-transform infrared and Raman spectroscopic evidence for the incorporation of cinnamaldehydes into the lignin of transgenic tobacco (Nicotiana tabacum L) plants with reduced expression of cinnamyl alcohol dehydrogenase. Planta **201,** 311-318.

Stich, B., Melchinger, A.E., Heckenberger, M., Mohring, J., Schechert, A., and Piepho, H.P. (2008). Association mapping in multiple segregating populations of sugar beet (Beta vulgaris L.). Theoretical and Applied Genetics **117,** 1167-1179.

Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B-Statistical Methodology **64,** 479-498.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America **100,** 9440-9445.

Sykes, R., Li, B.L., Isik, F., Kadla, J., and Chang, H.M. (2006). Genetic variation and genotype by environment interactions of juvenile wood chemical properties in Pinus taeda L. Annals of Forest Science **63,** 897-904.

Syvanen, A.C. (1999). From gels to chips: "Minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. Human Mutation **13,** 1-10.

Syvanen, A.C. (2001). Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nature Reviews Genetics **2,** 930-942.

Syvänen, A.-C., Aalto-Setälä, K., Harju, L., Kontula, K., and Söderlund, H. (1990). A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. Genomics **8,** 684-692.

Tabor, H.K., Risch, N.J., and Myers, R.M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics **3,** 391-A396.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics **123,** 585-595.

Taylor, F.W. (1981). Rapid-Determination of Southern Pine Specific-Gravity with a Pilodyn Tester. Forest Science **27,** 59-61.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F., and Gaut, B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp mays L.). Proceedings of the National Academy of Sciences of the United States of America **98,** 9161-9166.

Thamarus, K.A., Groom, K., Murrell, J., Byrne, M., and Moran, G.F. (2002). A genetic linkage map for Eucalyptus globulas with candidate loci for wood, fibre, and floral traits. Theoretical and Applied Genetics **104,** 379-387.

Thamarus, K., Groom, K., Bradley, A., Raymond, C.A., Schimleck, L.R., Williams, E.R., and Moran, G.F. (2004). Identification of quantitative trait loci for wood and fibre properties in two full-sib pedigrees of Eucalyptus globulus. Theoretical and Applied Genetics **109,** 856-864.

The Arabidopsis genome initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408,** 796-815.

Thelen, M.P., and Northcote, D.H. (1989). Identification and Purification of a Nuclease from Zinnia-Elegans L - a Potential Molecular Marker for Xylogenesis. Planta **179,** 181-195.

Thomas, S., Porteous, D., and Visscher, P.M. (2004). Power of direct vs. indirect haplotyping in association studies. Genetic Epidemiology **26,** 116-124.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E.S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics **28,** 286-289.

Thumma, B.R., Nolan, M.R., Evans, R., and Moran, G.F. (2005). Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. Genetics **171,** 1257-1265.

Thumma, B.R., Matheson, B.A., Zhang, D.Q., Meeske, C., Meder, R., Downes, G.M., and Southerton, S.G. (2009). Identification of a Cis-Acting Regulatory Polymorphism in a Eucalypt COBRA-Like Gene Affecting Cellulose Content. Genetics **183,** 1153-1164.

Thumma, B.R., Southerton, S.G., Bell, J.C., Owen, J.V., Henery, M.L., and Moran, G.F. (2010). Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. Tree Genetics & Genomes **6,** 305-317.

Timell, T.E. (1969). The chemical composition of tension wood. Svensk Papperstidning **72**.

Todokoro, S., Terauchi, R., and Kawano, S. (1995). Microsatellite Polymorphisms in Natural-Populations of Arabidopsis-Thaliana in Japan. Japanese Journal of Genetics **70,** 543-554.

Toomajian, C., and Kreitman, M. (2002). Sequence variation and haplotype structure at the human HFE locus. Genetics **161,** 1609-1623.

Tripiana, V., Bourgeois, M., Verhaegen, D., Vigneron, P., and Bouvet, J.M. (2007). Combining microsatellites, growth, and adaptive traits for managing in situ genetic resources of Eucalyptus urophylla. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **37,** 773-785.

Tsuchikawa, S., Hayashi, K., and Tsutsumi, S. (1996). Near-infrared spectroscopy. Applied Spectroscopy **50,** 1117-1124.

Tsuchikawa, S. (2007). A review of recent near infrared research for wood and paper. Applied Spectroscopy Reviews **42,** 43-71.

Tuskan, G.A., DiFazio, S., *et al.* (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313,** 1596-1604.

Ukrainetz, N.K., Kang, K.Y., Aitken, S.N., Stoehr, M., and Mansfield, S.D. (2008). Heritability and phenotypic and genetic correlations of coastal Douglas-fir (Pseudotsuga menziesii) wood quality traits. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere **38,** 1536-1546.

Van Der Nest, M.A., Steenkamp, E.T., Wingfield, B.D., and Wingfield, M.J. (2000). Development of simple sequence repeat (SSR) markers in Eucalyptus from amplified inter-simple sequence repeats (ISSR). Plant Breeding **119,** 433-436.

van Raemdonck, D., Pesquet, E., Cloquet, S., Beeckman, H., Boerjan, W., Goffner, D., El Jaziri, M., and Baucher, M. (2005). Molecular changes associated with the setting up of secondary growth in aspen. Journal of Experimental Botany **56,** 2211-2227.

Vanholme, R., Morreel, K., Ralph, J., and Boerjan, W. (2008). Lignin engineering. Current Opinion in Plant Biology **11,** 278-285.

Vanholme, R., Demedts, B., Morreel, K., Ralph, J., and Boerjan, W. (2010). Lignin Biosynthesis and Structure. Plant Physiology **153,** 895-905.

Venter, J.C., Adams, M.D., *et al.* (2001). The sequence of the human genome. Science **291,** 1304-+.

Verhaegen, D., and Plomion, C. (1996). Genetic mapping in Eucalyptus urophylla and Eucalyptus grandis using RAPD markers. Genome **39,** 1051-1061.

Verhaegen, D., Plomion, C., Gion, J.M., Poitel, M., Costa, P., and Kremer, A. (1997). Quantitative trait dissection analysis in Eucalyptus using RAPD markers .1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. Theoretical and Applied Genetics **95,** 597-608.

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution **34,** 275-305.

Vigneron, P. (1995). Production and improvement of Eucalyptus varietal hybrids in the Congo. Traitements statistiques des essais de selection: strategies d'amelioration des plantes perennes. Actes du seminaire de biometrie et genetique quantitative, 12-14 septembre 1994, Montpellier, France.**,** 259-273.

Vigneron, P., and Bouvet, J.M. (2000). Eucalyptus. In Tropical plant breeding, A. Charrier, M. Jacquot, and D. Nicolas, eds (Collection repère du cirad), pp. 223-245.

Volker, P.W., Potts, B.M., and Borralho, N.M.G. (2008). Genetic parameters of intra- and inter-specific hybrids of Eucalyptus globulus and E-nitens. Tree Genetics & Genomes **4,** 445-460.

Wachowiak, W., Balk, P.A., and Savolainen, O. (2009). Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (Pinus sylvestris L.). Tree Genetics & Genomes **5,** 117-132.

Wagner, G.P., and Lynch, V.J. (2008). The gene regulatory logic of transcription factor evolution. Trends in Ecology & Evolution **23,** 377-385.

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. American Journal of Human Genetics **71,** 1227-1234.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: Theoretical and practical concerns. Nature Reviews Genetics **6,** 109-118.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7,** 256-276.

Weber, A.L., Briggs, W.H., Rucker, J., Baltazar, B.M., Sanchez-Gonzalez, J.D., Feng, P., Buckler, E.S., and Doebley, J. (2008). The Genetic Architecture of Complex Traits in Teosinte (Zea mays ssp parviglumis): New Evidence From Association Mapping. Genetics **180,** 1221-1232.

Wegrzyn, J.L., Eckert, A.J., Choi, M., Lee, J.M., Stanton, B.J., Sykes, R., Davis, M.F., Tsai, C.J., and Neale, D.B. (2010). Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem. New Phytologist **188,** 515-532.

Wei, X., and Borralho, N.M.G. (1997). Genetic control of wood basic density and bark thickness and their relationships with growth traits of Eucalyptus urophylla in south east China. Silvae Genetica **46,** 245-250.

Williams, C.G., and Savolainen, O. (1996). Inbreeding depression in conifers: Implications for breeding strategy. Forest Science **42,** 102-117.

Wittwer, C.T., Reed, G.H., Gundry, C.N., Vandersteen, J.G., and Pryor, R.J. (2003). High-resolution genotyping by amplicon melting analysis using LCGreen. Clinical Chemistry **49,** 853-860.

Wright, S. (1931). Evolution in mendelian populations. Genetics **16,** 97-159.

Wright, S.I., and Gaut, B.S. (2005). Molecular population genetics and the search for adaptive evolution in plants. Molecular Biology and Evolution **22,** 506-519.

Xu, X.J., Hsia, A.P., Zhang, L., Nikolau, B.J., and Schnable, P.S. (1995). Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. Plant Cell **7,** 2151-2161.

Xu, Z.Y., Zhang, D.D., Hu, J., Zhou, X., Ye, X., Reichel, K.L., Stewart, N.R., Syrenne, R.D., Yang, X.H., Gao, P., Shi, W.B., Doeppke, C., Sykes, R.W., Burris, J.N., Bozell, J.J., Cheng, Z.M., Hayes, D.G., Labbe, N., Davis, M., Stewart, C.N., and Yuan, J.S. (2009). Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. Bmc Bioinformatics **10**.

Yahiaoui, N., Marque, C., Myton, K.E., Negrel, J., and Boudet, A.M. (1998). Impact of different levels of cinnamyl alcohol dehydrogenase down-regulation on lignins of transgenic tobacco plants. Planta **204,** 8-15.

Ye, Z.H., and Varner, J.E. (1996). Induction of cysteine and serine proteases during xylogenesis in Zinnia elegans. Plant Molecular Biology **30,** 1233-1246.

Ye, Z.H. (2002). Vascular tissue differentiation and pattern formation in plants. Annual Review of Plant Biology **53,** 183-202.

Yeh, T.F., Chang, H.M., and Kadla, J.F. (2004). Rapid prediction of solid wood lignin content using transmittance near-infrared spectroscopy. Journal of Agricultural and Food Chemistry **52,** 1435-1439.

Yeh, T.F., Goldfarb, B., Chang, H.M., Peszlen, I., Braun, J.L., and Kadla, J.F. (2005). Comparison of morphological and chemical properties between juvenile wood and compression wood of loblolly pine. Holzforschung **59,** 669-674.

Yu, J., Hu, S.N., Wang, J., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp indica). Science **296,** 79-92.

Yu, J.M., and Buckler, E.S. (2006). Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology **17,** 155-160.

Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., and Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38,** 203-208.

Zamudio, F., Rozenbergb, P., Baettig, R., Vergara, A., Yanez, M., and Gantz, C. (2005). Genetic variation of wood density components in a radiata pine progeny test located in the south of Chile. Annals of Forest Science **62,** 105-114.

Zeggini, E., and Ioannidis, J.P.A. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics **10,** 191-201.

Zhang, K., Akey, J.M., Wang, N., Xiong, M., Chakraborty, R., and Jin, L. (2003). Randomly distributed crossovers may generate block-like pattern's of linkage disequilibrium: an act of genetic drift. Human Genetics **113,** 51-59.

Zhao, Z.M., Fu, Y.X., Hewett-Emmett, D., and Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene **312,** 207-213.

Zhao, K.Y., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C.L., Toomajian, C., Zheng, H.G., Dean, C., Marjoram, P., and Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. Plos Genetics **3**.

Zhong, R.Q., Ripperger, A., and Ye, Z.H. (2000). Ectopic deposition of lignin in the pith of stems of two Arabidopsis mutants. Plant Physiology **123,** 59-69.

Zhong, R.Q., and Ye, Z.H. (2007). Regulation of cell wall biosynthesis. Current Opinion in Plant Biology **10,** 564-572.

Zhong, R., and Ye, Z.-H. (2009). Transcriptional regulation of lignin biosynthesis. Plant Signal Behav **4,** 1028-1034.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., and Cregan, P.B. (2003). Single-nucleotide polymorphisms in soybean. Genetics **163,** 1123-1134.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and Prospects of Association Mapping in Plants. Plant Gen. **1,** 5-20.

Zollner, S., and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. Genetics **169,** 1071-1092.

Zwick, M.E., Cutler, D.J., Yohn, C.T., Tobin, K.P., Kashuk, C.S., Shah, N.A., Warrington, J.A., Eichler, E.E., and Chakravarti, A. (2000). Characterizing human genomic variation and linkage disequilibrium in multiple 100kb genomic segments using large-scale, microarray-based SNP detection. American Journal of Human Genetics **67,** 70.

ANNEXE 1: A candidate gene for lignin composition in *Eucalyptus*: *Cinnamoyl-CoA Reductase* (*CCR*)

L'Annexe 1 contient le manuscript de l'article soumis le 11 Septembre 2010 au journal Heredity : A candidate gene for lignin composition in *Eucalyptus*: *Cinnamoyl-CoA Reductase* (*CCR*)

A candidate gene for lignin composition in *Eucalyptus*: *Cinnamoyl-CoA Reductase* (*CCR*)

Eric Mandrou [1, 2, 3], Paulo Ricardo Gherardi Hein [4], Emilie Villar [2, 3, 5], Philippe Vigneron [5], Christophe Plomion [3], Jean-Marc Gion [2, 3,*]

[1] Centre de recherche Vallourec, CEV, Route de Leval BP 17, Aulnoye Aymeries, France;

[2] CIRAD, UMR AGAP, Equipe Génétique et amélioration des espèces pérennes : modèles Forêt & Palmier, Campus international de Baillarguet TA A-39/C, Montpellier cedex 5, France;

[3] INRA, UMR 1202 BIOGECO, 69 route d'Arcachon, Cestas Pierroton, France;

[4] CIRAD, UPR40 Production and Processing of Tropical Woods, 73 rue J-F Breton TA B-40/16, Montpellier, France;

[5] CRDPI, BP 1291, Pointe-Noire, République du Congo.

***Corresponding author**

Jean-Marc Gion

INRA, UMR 1202 BIOGECO

69 route d'Arcachon, 33612 Cestas-Pierroton, France

Tel: +33 5 57 12 28 93

Fax: +33 5 57 12 28 81

gion@cirad.fr

Abstract

Lignin content and composition are considered as mandatory traits of eucalyptus breeding programs, especially for pulp, paper and bioenergy production. In this article we used 33 *Eucalyptus urophylla* full-sib families of a 8x8 factorial design to provide estimates of genetic parameters for lignin and growth related traits. Secondly, from the sequencing of the 16 unrelated founders, we described the nucleotide and haplotype variability of *CCR*, a candidate gene for lignin related traits encoding the Cinnamoyl-CoA Reductase. Finally, we tested the association between *CCR* polymorphisms and trait variation using a mixed linear model. A high value of narrow sense heritability was obtained for lignin content (h^2=0.85) and S/G ratio (h^2=0.62) indicating that these traits are under strong genetic control. High levels of nucleotide (θ_π=0.0131) and haplotype (Hd=0.958) diversity were detected for *CCR*. From an initial set of 152 biallelic SNPs, a subset of 65 non redundant loci was selected. Three intronic SNPs were found to be associated to the variation of S/G ratio after multiple testing correction. In the line of what has been obtained in forest trees, these SNPs explained between 2.45% and 2.87% of the genetic variance of the trait. This study demonstrates the interest of the candidate gene approach for quantitative trait nucleotide detection in *Eucalyptus* and paves the way to gene assisted selection of lignin composition in *E. urophylla*.

Keywords

Eucalyptus, association study, Cinnamoyl-CoA reductase (CCR), lignin. S/G ratio

Thomas Drudge; email: tgg@ucdavis.edu

Introduction

Lignin content and composition are becoming important targets for forest tree breeding. In papermaking industry, lignins are undesirable compounds, and are eliminated by delignification during chemical pulping. This elimination involves energy intensive and natural impacting processes depending on both lignin quantity and quality (Bose *et al.* 2009). To a certain extent, poorly lignified wood is also preferred for bioethanol production as it facilitates the hydrolysis of cellulose and hemicelluloses to fermentable sugars (Alvira *et al.* 2009). Conversely, highly lignified wood is required in charcoal production, since more heat from combustion energy is produced from lignins than an equal mass unit of carbohydrates (Savidge 2000).

Lignins, one of the most abundant biopolymers on earth, accumulate in the secondary cell wall of xylem cells. They are believed to be involved in the transition of plants from aquatic to terrestrial life as they confer stiffness and hydrophobicity to the cell wall, two essential properties for straightness as well as water and nutrient transport. Lignins are known to vary both quantitatively and qualitatively among taxa (Boerjan *et al.* 2003), cell types (tracheids or xylem fiber cells) (Peter and Neale 2004) and cell wall layers. Furthermore, they are influenced by developmental factors and environmental cues (Campbell and Sederoff 1996). In angiosperms, lignins are principally produced by polymerization of monolignols Syringyl (S) and Guaiacyl (G). Typically, two traits are used to characterize lignin quantity and quality: lignin content and S/G ratio, respectively. Lignin content displays a pattern of continuous phenotypic distribution. It has been shown to be under moderate to high genetic control, with heritability ranging from 0.40 to 0.60 (Pot *et al.* 2002; Hannrup *et al.* 2004; Sykes *et al.* 2006; Poke *et al.* 2006) and relatively low coefficient of variation ranging from 3 to 5% (above mentioned references). Genetic parameters for S/G ratio have been much less studied. Baillères *et al.* (2002) reported a phenotypic variation coefficient of 13.4% in a full-sib family of *Eucalyptus urophylla* x

Thomas Drudge; email: tgg@ucdavis.edu

E. grandis hybrids. This indicates that lignin quality presents more variation than lignin quantity, but to our best knowledge no result has yet been published in respect to genetic variability estimates.

The lignin biosynthesis pathway is well known in plants. Actually, most of the structural enzymes catalyzing the chemical transformations from phenylalanine to monolignols have been identified and described in model plants species such as *Arabidopsis thaliana* (reviewed in Humphreys and Chapple 2002 and Boerjan *et al.* 2003). Abnormal lignin phenotypes, obtained from mutants or genetic modifications, were reported for these genes (reviewed in Anterola and Lewis 2002) providing useful information regarding their effect *in planta* and their involvement in the molecular regulation on both lignin quantity and quality. In eucalyptus most of these genes have been partially or completely sequenced (Hawkins and Boudet 1994; Lacombe *et al.* 1997; Feuillet *et al.* 1993; De Melis *et al.* 1999; Poeydomenge *et al.* 1994; Harakava 2005; Rengel *et al.* 2009). For some of them, expression data are available in different tissues and experimental conditions (Grima-Pettenati *et al.* 1993; Lacombe *et al.* 2000; Kirst *et al.* 2004; Paux *et al.* 2005).

The cinnamoyl-CoA reductase (CCR) is a key enzyme in lignin biosynthesis as it is the first player of the monolignols-specific pathway. It is considered as a control point for the allocation of carbon toward lignins. The expression of the gene encoding CCR was shown to be associated to areas undergoing lignification and to be controlled, at least in part, by the activity of its promoter (Lacombe *et al.* 2000). In transgenic tobacco, a significant down regulation of *CCR* expression was shown to impact both lignin content and S/G ratio (Piquemal *et al.* 1998; O'Connell *et al.* 2002). These results suggest *CCR* as a relevant functional candidate gene to control part of variation of lignin quantity and quality. Forward genetic experiments have also demonstrated that *CCR* could be considered as a relevant positional candidate gene, since its map position was shown to coincide with lignin content and S/G ratio QTLs in *E. urophylla* and *E. globulus* (Freeman *et al.* 2009; Gion *et*

Thomas Drudge; email: tgg@ucdavis.edu

al. unpublished). However, because of large confidence intervals of QTLs, such gene-trait association needs to be confirmed with more accurate methods and in broader genetic backgrounds before being used as diagnostic tools in breeding. Association mapping is probably the method of choice to deal with these pitfalls. It could improve the resolution for identification of the polymorphisms underlying phenotypic variation, especially in forest trees generally characterized by a rapid decay of linkage disequilibrium within population (Neale and Savolainen 2004). In *Eucalyptus*, Thumma *et al.* (2005; 2009) demonstrated the efficiency of the candidate gene association mapping approach by detecting associations between cellulose related traits (microfibril angle and cellulose content) and SNPs in *CCR* and a *COBRA*-like gene. These studies confirmed the potential role of lignification and cellulose biosynthesis related genes in controlling part of the variability of wood properties.

The present study is part of the genetic improvement programs of *Eucalyptus* in the Republic of Congo (CRDPI institute) and Brazil (Vallourec & Mannesmann do Brasil), which objective is to provide high quality wood material for papermaking and charcoal industries. Identifying early indirect selection criteria of lignin related traits is of major interest for the breeders. Consequently, our study had three aims: first of all, to provide for the first time estimates of genetic parameters for lignin quality (S/G ratio) in trees, then to describe the nucleotide diversity of *CCR* in a widely used species in forest plantations (*E. urophylla* ST Blake), and finally to test the association between *CCR* polymorphisms and variability of growth and lignin related traits using a factorial mating design of *E. urophylla*.

Material and methods

Plant material

Based on the *E. urophylla* seed collection made by Martin and Cossalter in the 1970's (1976 a; 1976 b; 1976 c), we selected 16 unrelated genotypes of the Flores Island in the

Sunda archipelago (122°-127°E, 8°-10°S). These trees belonged to the eucalyptus breeding program managed by CRDPI (Pointe Noire, Republic of Congo). They were conserved in a seed orchard in Kissoko station. Leaves were sampled for each tree and then dried in Silica Gel. DNA was extracted according to Doyle and Doyle (1990) and stored at -20°C before being used.

These 16 trees were crossed according to an incomplete factorial mating design (8 females x 8 males, Online Resource 1). A total of 328 offsprings in 33 full-sib families (8-10 in each FS) were phenotyped for genetic parameter estimation and genotyped for association mapping. While FS families were randomly allocated in the trial, offsprings within families were not. Therefore, it was not possible to separate dominance from micro-environmental effects. This progeny test was established in 1992 in Kissoko (Pointe Noire, Republic of Congo), and was harvested at 14 years old. 2 cm wood discs were sampled at 1.3 m height. Leaves were sampled for each offspring and DNA extracted as previously described.

Phenotyping

Growth: total height and circumference at 1.3 m were measured, after harvesting at age 14.

Lignin related traits: Klason Lignin (KL) and S/G ratio values were assessed by Near Infrared (NIR) Spectroscopy. These NIR-based models were previously presented in Hein *et al.* (2010). Briefly, 60 trees out of the 328 were selected as representative as possible of the range of variation of KL and S/G ratio for classic chemical measurements in order to develop the NIR calibrations. NIR spectra were obtained from the grounded wood using a Bruker spectrophotometer (model Vector 22/N, Bruker Optik GmbH, Ettlingen, Germany) in the diffuse reflectance mode. NIRS scans were repeated three times by sample, and the triplicate NIR spectra were averaged into a single NIR spectrum per tree. Therefore, Partial Least Square (PLS) regression analyses were applied to describe the relationship between

the NIR spectra and the reference chemical analysis (Klason method for lignin content and thioacidolysis for lignin composition). The coefficient of determination between the NIR-predicted and the LAB-measured values was 0.86 for both KL and S/G ratio. The standard errors of these PLS-R predictions were calculated as 0.48% for KL and 0.12% for S/G ratio (Hein *et al.* 2010).

Sequencing

Sequencing data were obtained from cloned PCR products generated from each of the 16 parental trees. Overlapping fragments were amplified using 7 primer pairs designed from a genomic clone of *Eucalyptus gunnii CCR* gene (3,202 bp) previously described by Lacombe *et al.* (1997; accession number X97433). For each amplicon, the sequences of the forward and reverse primers are described in Online Resource 2. Amplifications were performed using a Tetrad 2 PTC-0240 Thermo Cycler (MJ Research, Whaltam, MA, USA) in 20 µl final reaction volume composed by 2 µl of 10X 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, 0.8 unit of native Taq polymerase (Invitrogen) and dH_2O to complete the final reaction volume. A PCR cycle in 3 steps was used involving an initial denaturation of 4 min at 94 °C, 35 cycles of 30 s at 94 °C, 1 min at primers Tm, 1 min at 72 °C followed by a final extension of 10 min at 72 °C. Each PCR product was cloned independently using the TOPO-TA cloning kit for sequencing (Invitrogen) according to the manufacturer's protocol. A total of 16 positive transformed clones were collected. *CCR* inserts were sequenced (6 to 12 positive clones by cloning product) using Big Dye Terminator V 1.1 cycle sequencing kit (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. Electrophoreses were run on an ABI 3730 DNA Analyzer (Applied Biosystems). Sequencing reactions were performed in one direction (forward or reverse) by using universal primers T3 or T7, except for PCR fragments 2 and 6 that were sequenced in both

directions because of the presence of poly A and Simple Sequence Repeat motifs (SSR). For each of the 7 overlapping PCR fragments and for each of the 16 genotypes, sequences were aligned (112 independent alignments), checked for base calling errors by comparing electrophoregrams and called bases using CodonCode Aligner software version 1.6.1 (CodonCode, Deadham, MA, USA). Overlapping fragments were then assembled resulting into 16 contigs (one for each genotype), including 2 alleles of the complete *CCR* gene each. Contigs were exported in BioEdit sequence editor (Hall 1999) and manually checked to identify true polymorphic sites (SNPs and INDELs), discard amplification errors and chimera PCR products (due to PCR mediated recombination) and finally reconstruct complete phased haplotypes. Haplotype phases were identified by assessing allele concordances at polymorphic sites detected in the overlapping areas of contiguous PCR fragments. The partition of polymorphic sites in these overlapping areas allowed obtaining the two expected complete phased haplotypes of the gene in each individual. SNPs, INDELs, as well as poly A and SSR polymorphisms were described from alignment of the final complete phased haplotypes. INDELs, poly A and SSR were excluded from diversity and association analysis.

Haplotype genotyping

The segregation of the parental full-length haplotypes was followed in each FS using the SSR motif detected in intron #4 (Online Resource 3). This SSR enabled the detection of the two haplotypic forms of *CCR* in 308 of the 328 offsprings. The SSR motif was amplified for each parent and offsprings using the primer pair described in Table 1. PCR amplifications were performed in a Tetrad 2 PTC-0240 Thermo Cycler (MJ Research) using a 15 µl final reaction volume containing 25 ng of genomic DNA with 2.5 µl 1X PCR reaction Buffer (Invitrogen), 0.2 mM dNTPs, 2 mM $MgCl_2$, 0.10 µM of forward primer, 0.06 µM of reverse primer, 0.10 µM of the infrared dye IRdye M13/700 or M13/800, and 0.13 U/µl Taq DNA polymerase (Invitrogen). A PCR cycle in three steps was used

Thomas Drudge; email: tgg@ucdavis.edu

involving an initial denaturation of 4 min at 94 °C; 30 cycles of 30 s at 94 °C, 45 s at 51 °C and 45 s at 72 °C, followed by a final extension of 5 min at 72 °C. The reverse primer was probed with a 17 nucleotide extension at the 5' tail end with the sequence 5' GTAAAACGACGGCCAGT 3'. This sequence, which is complementary to the IRdye M13/700 and M13/800 included in the PCR reaction, allowed the IR-labeling of the PCR products. Electrophoreses of IR-labeled PCR products were run on an IR DNA analyzer (Li-Cor, Inc., Lincoln, NE, USA), which can detect labeled products at the wave lengths of 700 nm and 800 nm. The SSR alleles were identified according to their sizes, and their segregation was followed in each FS.

Nucleotide variability and linkage disequilibrium analysis

Analysis of sequencing data were performed using DNAsp v. 5.0 (Rozas *et al.*, 2003). Nucleotide and haplotype diversities were estimated by Waterson's θ_w and θ_π (the average number of pairwise nucleotide differences among sequences in a sample) reported as per site values and Hd, respectively. Linkage disequilibrium (LD) values between pairs of biallelic SNPs were estimated using r^2 (the squared correlation in allelic states between pairs of SNPs) in the 16 parents of the factorial mating design using the TASSEL software package (http://www.maizegenetics.net). This LD measurement was used to select a subset of the SNPs to be tested for association with growth and lignin related traits, i. e. one SNP was selected to represent all the SNPs that were completely linked ($r^2 = 1$).

Statistical methods for genetic parameter estimation

Quantitative genetic analysis was performed with ASReml version 3.0 (Gilmour *et al.* 2002). Growth and lignin related traits were all normally distributed. They were first analyzed independently (univariate analysis) to estimate the variance components by using an individual model. The following mixed linear model was considered:

$$y = X.b + Z.a + e$$

where, y is the vector of observations, b is the vector of fixed effects, a is the vector of genetic effects, e is the vector of residuals and X and Z are the incidence matrices linking observations to the effects.

The random effect fits a normal distribution whose parameters were

$$E\begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \quad Var\begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

The variance-covariance matrices were defined as follows:

$$G = A.\sigma^2_A \quad \text{and} \quad R = I.\sigma^2_e$$

where A is the additive genetic relationship matrix computed from a pedigree file that takes into account all the relationships between the individuals, I is the identity matrix, σ^2_A the additive genetic variance and σ^2_e the residual variance. The variances associated to random effects were estimated by restricted maximum likelihood (REML method) using ASReml (Gilmour *et al.* 2002). As the variances are assumed to be independent, the total phenotypic variance σ^2_P was calculated as follows:

$$\sigma^2_P = \sigma^2_A + \sigma^2_e$$

Because, FSs were not replicated in the experimental design, dominance and micro-environmental effects were confounded, thus family effect was not considered in the model.

The variation of each trait as expressed by its phenotypic variation coefficients and narrow sense heritabilities were calculated as follows:

$$CV = \frac{\sigma^2_P}{\overline{X}} \quad \text{and} \quad h^2 = \frac{\sigma^2_A}{\sigma^2_P}$$

Thomas Drudge; email: tgg@ucdavis.edu

To estimate phenotypic and genetic additive correlations (r_P and r_A respectively), we considered a bivariate analysis using the same individual model as for univariate analysis. r_P and r_A were estimated as follows:

$$r_P = \frac{Cov_p(x, y)}{\sqrt{\sigma^2_{Px}.\sigma^2_{Py}}} \quad \text{and} \quad r_A = \frac{Cov_A(x, y)}{\sqrt{\sigma^2_{Ax}.\sigma^2_{Ay}}}.$$

Standard errors of h^2, σ^2_A, σ^2_P, r_P and r_A were calculated with ASReml using a standard Taylor series approximation (Gilmour *et al.* 2002).

Statistical method for association study

A marker-by-marker approach was made using a mixed linear model (MLM), fitted independently for each marker and trait, and implemented in TASSEL version 2.1. This method allows taking into account structure and relatedness between individuals of the association mapping population and is thus suited to the analysis of a factorial matting design involving full-sibs, half-sibs and unrelated individuals (Yu *et al.* 2006). Structure was not implemented in the model as a very low genetic structure was reported in populations of *E. urophylla* throughout its entire species range (F_{ST}=0.03-0.04 in Tripiana *et al.* 2007 and Payn *et al.* 2008) and particularly on Flores island (Fst=0.012-0.014 in Payn *et al.* 2008), from where the 16 founders of our study were sampled. Pedigree information was used to build the kinship matrix (K) implemented in the model to control for genetic covariation between relatives (in our case full-sibs and half-sibs). The statistical model is as follows:

$$y = X\beta + Zu + e,$$

where, y is the vector of observations (phenotypic values of tested traits), β is the vector of fixed SNP effects (encoded as genotype), u is the vector of random polygenic effect of individual genotypes, e is the vector of residuals and X and Z are the incidence matrices

Thomas Drudge; email: tgg@ucdavis.edu

linking observations to the effects. The random effect fits a normal distribution whose parameters are:

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad Var\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} K\sigma^2{}_a & 0 \\ 0 & I\sigma^2{}_e \end{bmatrix},$$

where, K is the kinship matrix (Identity By Descent) between individuals of the association mapping population, I is the identity matrix, σ^2_a is the additive genetic variance (random polygene variance) and σ^2_e is the residual variance. The REML estimates of σ^2_a and σ^2_e were obtained in TASSEL through the expectation and maximization algorithm.

In order to avoid spurious associations, we used the very stringent Bonferroni method to correct the level of significance for multiple testing. Experimental wise error rate was set to 5%.

Results

Genetic control of growth and lignin related traits

The quantitative genetic analysis allowed estimating the phenotypic variance, the genetic additive variance and the narrow sense heritability for each of the 4 studied traits. For each trait, genetic parameters are shown in Table 1. The phenotypic coefficients of variation for growth (17% for height and 21% for circumference) were higher than that for lignin related traits (4% for lignin content and 12% for S/G ratio). However, the estimates of narrow sense heritabilities indicated higher additive genetic control for lignin related traits (0.85 for KL and 0.62 for S/G ratio) than for growth (0.30 for height and 0.12 for circumference), a complex trait obviously subjected to greater environmental factors.

Estimated phenotypic and genetic additive correlations between traits are shown in Table 2. Although their standard deviations were quite high, these estimates suggested on the one

Thomas Drudge; email: tgg@ucdavis.edu

hand, a high correlation at both phenotypic and genetic levels between height and circumference (0.77 to 0.68, respectively), and on the other hand, lower negative correlations at both phenotypic and genetic levels between height and KL (-0.1 and -0.53, respectively) and similar trends between KL and S/G ratio (-0.23 and -0.25, respectively).

Sequencing data

A total of 1 291 clones were sequenced corresponding to an average of 9 clones per PCR fragment for each of the 16 parents. This sequencing data represented a total of 1 200 kb. It helped to avoid two sources of error: first of all, Taq polymerase replication errors, which occurred in our study at the frequency of 1.3 errors every kb, and finally, chimeric alleles that correspond to PCR mediated recombining alleles which can occur when highly similar matrices are melted in a PCR. These errors were removed prior SNP identification and linkage phase determination.

The sequenced region of *CCR*, from position 151 to position 3 194 of the sequence described by Lacombe *et al.* (1997) in *E. gunnii* (accession number X97433), represented 2 992 bp in *E. urophylla*, excluding all alignment gaps. As a result, the nucleotide variability of 94% of the gene was finely described including 100% of the coding sequence (CDS).

DNA polymorphisms and haplotype structure

Different types of polymorphisms were detected including substitutions, insertion deletion (INDELs), simple sequence repeats (SSR) and poly A length variation. Nature, number and gene location of all detected polymorphisms are given in Table 3. A total of 156 SNPs (152 biallelic and 4 triallelic) were detected corresponding to an average of 1 SNP every 20 bp for the entire *CCR* gene. In intronic regions, 115 SNPs were detected leading to an average of 1 SNP every 17 bp. In exonic regions, SNP density was lower with an average of 1 SNP every 28 bp (41 SNPs detected). A total of 4 SNPs corresponded to non

Thomas Drudge; email: tgg@ucdavis.edu

synonymous (NS) variations impacting amino acids #77, #123, #177 and #241 (Table 4). The change in amino acid #123 did not alter the physico-chemical properties of the protein, while the three others involved differences in terms of electric charge, size, polarity, hydrophobicity or functional group. A total of 17 INDELs were detected, exclusively in intronic regions. These INDELs were variable in size ranging from 1 to 35 bp. INDELs of one bp were the most abundant. They represented 65% (11 INDELs) of the detected INDELs. Intron #1 contained one poly A repeat. Finally, a dinucleotide SSR was detected in intron #4. A total of 15 or 13 alleles could be identified in terms of either sequence or size (from 11-29 repeats) variation (Online Resource 3).

The distribution of the SNPs according to their minor allele frequencies (MAF) is presented in Figure 1 for all detected biallelic SNPs. A total of 42% of the 152 biallelic SNPs were rare with MAFs lower than 10%. This included 48 singletons with MAF of 3.2% (1 occurrence of the minor allele among the 32 sampled copies of the gene). Two of the NS mutations (affecting amino acids #123 and #241) were detected as singletons in our sequencing panel. The two others NS mutations (impacting amino acids #77 and #177) exhibited MAFs of 22% and 12.5%, respectively. Based on these data, nucleotide diversity was estimated to 0.0131 and 0.0135 for θ_π and θ_w respectively. A total of 20 full-length gene haplotypes were detected with an average number of differences of 38.5 bp (ranging from 1 to 67 bp). Haplotype diversity (Hd) was estimated to 0.958.

Association between traits and SNP variation

The full-length *CCR* gene haplotypes obtained from the parents were identified in the progenies of the factorial mating design by genotyping the SSR motif detected in intron #4. The high level of variability of this SSR made it possible to follow the segregation of the parental haplotypes in each FS. Genotypes for each SNP were thus inferred for 308 of the

Thomas Drudge; email: tgg@ucdavis.edu

328 offsprings. Finally, a total of 65 biallelic SNPs were retained for the association study after discarding redundancy (complete LD) between markers.

No SNP was significantly associated with height, circumference or Klason lignin at the experimental wise error risk of 5%. Three SNPs were found to be strongly associated with S/G ratio. They were located in intronic regions of the gene. Table 5 indicates for each SNP harbouring an effect on S/G ratio: its location, its MAF in the association population, the number of genotypic classes compared, the P-value of the associated F statistic, and the effect of the SNP expressed as the part of phenotypic and genetic variance explained (R^2 marker). SNP#35 was rare in the association population with a MAF of 6% while SNP#30 and SNP#138 were more frequent with MAFs of 13% and 38% respectively. Linkage disequilibrium analysis indicated that these 3 SNPs were not highly correlated. The greatest r^2 value (0.41) was obtained between SNP#30 and SNP#35 (Table 6). In addition, they were found to be unlinked with the SNPs that were discarded from the analysis. Each of these 3 SNPs explained between 1.5 to 1.8% of the phenotypic variation of S/G ratio. Given the heritability of the trait, their contribution to the genetic variation ranged from 2.4 to 2.9%. The distributions of P-values related to the 65 association tests performed for each trait of interest are given in Online Resource 4.

Discussion

Genetic determinism of growth and lignin related traits in *E. urophylla*

Our experimental design revealed elevated narrow sense heritability for KL (0.85) and S/G ratio (0.62) compared to growth, indicating a strong genetic effect with a mode of inheritance that was mainly additive. To our knowledge we report here on the first estimate of lignin quality heritability, although an estimate of repeatability (estimated with clonally replicated full-sibs), was recently published in poplar (0.378 in Novaes *et al.* 2009). In respect to the level of heritability (or repeatability) this result is consistent with the general

Thomas Drudge; email: tgg@ucdavis.edu

trends reported in the literature. Different studies reported that wood properties, especially chemical characteristics, are generally less sensitive to environmental variations compared to growth performance (Raymond 2002; Freeman *et al.* 2009; Novaes *et al.* 2009). Growth is a highly complex process in which many inputs such as photosynthesis, water and nutrient availability are integrated and translated into developmental programs that orchestrate biomass production. It thus involves the interaction of many genes and their products. Conversely, chemical properties (amount and composition of polymeric compounds) are much less complex, often involving a single biosynthesis pathway. Therefore they present less opportunity of interaction with environmental factors than traits affected by many different physiological processes. Finally, our estimate for KL also agrees to that for the clonal heritability (0.83) made by Gominho *et al.* (1997) in *E. globulus*. In respect to the mode of inheritance, different patterns were observed in the few studies published so far. In *Picea abies*, Hannrup *et al.* (2004) reported a moderated value of broad sense heritability of 0.54 with a narrow sense heritability value of 0.10, suggesting a genetic effect principally due to dominance. Conversely, and in agreement with our findings, Pot *et al.* (2002) reported narrow sense heritability value of 0.47 in *Pinus pinaster*, with no dominance effect showing that lignin content was predominantly additively inherited.

In terms of genetic correlation, KL and S/G ratio were moderately negatively correlated at both phenotypic and genetic levels (-0.23). Even if the standard deviations associated with correlation coefficients were large, this result provides interesting insights about the genetic basis of their covariation due to either the effect of the same genes or the linkage disequilibrium between genes affecting both traits independently. Pleiotropy and physical linkage are indeed supported by the colocalization of QTLs for KL and S/G ratio in an *E. urophylla* x *E. grandis* full-sib family (Gion *et al.* unpublished). Pleiotropy is also well supported by transgenesis experiments. Piquemal *et al.* (1998) reported both a decrease in

Thomas Drudge; email: tgg@ucdavis.edu

lignin content and an increase in S/G ratio in tobacco plants down regulated for the expression of *CCR*. The same result was reported by O'Connell *et al.* (2002) for one severely *CCR*-suppressed tobacco line showing a decrease in lignin content (58% less lignins) coupled to an increase in S/G ratio (two-fold increase). In both studies the increase in S/G ratio was mostly explained by a decrease in the incorporation of G units in the polymer.

Growth and KL were also negatively correlated. The molecular basis of this correlation was nicely depicted by Kirst *et al.* (2004). By analysing together QTLs for growth and transcript abundance (eQTLs) they showed two interesting facts. On the one hand, fast growing trees presented a down-regulation of genes involved in lignin biosynthesis and associated methylation pathways. On the other hand, eQTLs for these genes colocalized with QTLs for growth diameter. In addition, they found a clear negative phenotypic correlation between growth and KL, and growth and S/G ratio. These results lead these authors to hypothesize that high lignin levels, as one of the main sinks for carbon in the xylem, could limit availability of carbon for cell division and growth. This hypothesis of carbon flow competition is also well supported by transgenesis experiments in which down-regulation of lignification genes led to a significant increase in growth (Hu *et al.* 1999; Li *et al.* 2003; Wu *et al.* 1999).

Nucleotide diversity in the *CCR* gene

The collected sequencing data provided useful information to describe with confidence and accuracy the nucleotide variability of 94% of the full-length *CCR* gene in a sample of 16 unrelated trees. A high level of nucleotide variability was detected regarding the observed value of θ_π (0.0131). This value is high compared to what has been obtained in other forest tree species so far for different sets of genes and for samples of individuals representative of species ranges (Online Resource 5). In terms of SNP densities, the level of

Thomas Drudge; email: tgg@ucdavis.edu

detected variability was, as expected, higher in introns (1 SNP/17 bp) than in exons (1 SNP/28 bp) with 60% more polymorphic sites in introns. This suggests the action of purifying selection on exonic regions directly linked to the primary structure of the protein. These values are higher but in the range of what was reported by Poke *et al.* (2003) in the same gene (1 SNP/48 bp in exons and 1 SNP/33 bp in introns), in a sample of 23 *E. globulus* from 2 open pollinated families selected for their differences in wood density. In our dataset, 4 non-synonymous (NS) mutations were detected, representing 10% of the exonic SNPs. The hypothetical catalytic site KNWYCYGK and the putative cofactor binding site, both reported by Lacombe *et al.* (1997) in *CCR* of *E. gunnii* remained invariant. In contrast, Poke *et al.* (2003) reported a higher number of NS mutations (12 mutations) representing 57% of the exonic mutations, from which 6 affected highly conserved amino acids with a high level of non conservative changes in terms of physico-chemical properties (5 out of the 6 detected). Many SNPs reported in the *E. globulus* study, were detected in *E. urophylla* with 50% common SNPs in introns (16 SNPs common to both species over 32 reported in *E. globulus*) and 13% in exons (3 common SNPs over 22 in *E. globulus*). These two species belong to the same sub-genus (*Symphyomyrtus*) but different sections (*Latoangulatae* for *E. urophylla* and *Maidenaria* for *E. globulus*). They occupy non-overlapping areas (*E. globulus* being present in South East Australia and Tasmania) making natural hybridizations unlikely. Besides, the proposed separation age between the *E. urophylla* clade of the Sunda islands and sister Taxons from Australia ranges from 2-5 mya according to Ladiges *et al.* (2003) based on climatic and tectonic events and 7-20 mya based on molecular dating (Crisp *et al.* 2004). So many SNPs in common between these two species would suggest the persistence of common ancestral polymorphisms through selection over a long period. However, this seems unlikely since none of the NS mutations reported in *E. globulus* were detected in *E. urophylla.* Alternatively; this could be due to the conservation of ancestral polymorphisms of recently derived species favored by large population sizes. If evolutionary time from speciation is

the dominant factor, genes from other biosynthetic pathways should show similar patterns as do terpenoid and flavonoid related genes between *E. globulus* and *E. nitens* of the *Globulares* series (28% shared polymorphisms reported by Külheim *et al.* (2009)).

Effect of *CCR* polymorphism on S/G ratio

Given their biological and demographical characteristics, association mapping shows great promise for the detection of quantitative trait nucleotides in forest trees (Neale and Savolainen 2004). While most association studies published so far in these long lived species have been based on populations with low or absence of relatedness (Online Resource 6), this study has aimed to test the efficiency of association mapping in a complex pedigreed population using a factorial design. This type of design is routinely used by forest tree geneticists in breeding programs, and offers a great opportunity to detect genes for marker-assisted breeding, providing that they can be associated with trait variation. In addition, results from association mapping within such elite germplasm could be used directly in breeding, because allelic variation present in the studied germplasm is investigated. Using a mixed model and considering relative kinship based on pedigree information (Yu *et al.* 2006), we detected significant associations between SNPs and lignin composition. Consequently, we demonstrated that this approach is efficient for identifying alleles associated with quantitative traits in trees. The way that such alleles could be used in breeding has still to be investigated.

Our finding is consistent with other experimental results in forward genetics. A collocation between *CCR* and a major QTL region accounting for 37% of S/G ratio (Gion *et al.* unpublished) was detected in a mapping population of *E. urophylla* x *E. grandis*. In *E. nitens*, a polymorphism in *CCR* was also shown to be associated with cellulose microfibril angle (MFA), and the gene was found to be located at the QTL peak for MFA (Thumma *et al.* 2005; Thumma *et al.*, 2010). In *E. globulus,* Thamarus *et al.* (2004) also

Thomas Drudge; email: tgg@ucdavis.edu

reported a QTL for pulp yield and cellulose, while Freeman *et al.* (2009) identified a QTL for density and lignin content in the same chromosomal region. Our results also agree with genetic manipulation of *CCR*. In the model plant tobacco and poplar, genetic engineering revealed a link between the expression level of the gene and the variation of lignin content and S/G ratio (Ralph *et al.* 1998; Piquemal *et al.* 1998; O'connell *et al.* 2002; Leplé *et al.* 2007). All together these results make out a strong case for *CCR* as a relevant candidate gene involved in the genetic control of wood quality traits in *Eucalyptus*.

The three SNPs found to be linked with trait variation were all localized in intronic regions of the gene. As they do not impact the primary structure of the protein, it's not clear how such mutations can impact S/G ratio. One hypothesis is that intronic SNPs could generate alternative splicing variants leading to less efficiency in protein activity. Hull *et al.* (2004) detected associations between common intronic SNPs and alternative splicing phenotypes at different loci in Human. The authors detected a dose dependent effect for most of concerned loci with homozygous genotypes exhibiting larger effects on the relative abundance of two transcript isoforms due to an alternative splicing event. In plant, Slotte *et al.* (2009) also reported that splice site polymorphisms in the *FLOWERING LOCUS* (*FLC*) were associated with early flowering. In eucalyptus, Thumma *et al.* (2005) identified such splicing variants in *CCR* but failed to demonstrate that it was the cause of the link detected between *CCR* polymorphisms and variation in MFA. An alternative hypothesis is that the detected SNPs are in LD with a causal mutation located in the surrounding region of *CCR*. In both *E. globulus* (Poke *et al.* 2005) and *E. urophylla* (this study, Online Resource 7) the decay of LD was rapid but some distant pairs of SNPs remained linked suggesting that this hypothesis is plausible. Testing this hypothesis would require further investigations as for example in the promoter region, in particular in the *cis*-regulating sites (*Myb* binding sites) identified by Lacombe *et al.* (2000), or in the transcription factors binding the *CCR* promoter (Goicoechea *et al.* 2005; Legay *et al.* 2007). Variations in these cis- and trans-

regulatory elements could indeed influence the level of allelic expression of the gene *in planta* and thus the resulting phenotype. Statistical power to detect single nucleotide polymorphism (SNP)/phenotype associations largely depends on the experimental design used, the number of individuals or families examined, the number of SNPs genotyped, and the explained proportion of variance in the trait imputable to a polymorphism. It also depends on specific characteristics of the studied traits, such as its heritability, level of variability and the accuracy of the measurement, especially when indirect measurements are used, as in the case of lignin related traits. In our study, the rather high values of narrow sense heritability (0.62) and phenotypic variability (CV=12.28%) for S/G ratio associated with a high quality of the NIRS calibration obtained to predict lignin composition (R^2=0.86), were considered as favourable characteristics to detect significant associations. The performance of the NIRS prediction was similar for KL and although the heritability estimate was even higher than for S/G ratio (0.85), the quite low variability (CV=4.33%) of this trait could have reduced the detection power. This could explain the absence of effect of *CCR* polymorphisms on the variation of KL in this study. The effects detected for each SNP independently on S/G ratio were rather small (between 2.4 to 2.9% of the additive variance and 1.5 to 1.8% of the phenotypic variance explained). Such small effects were reported in all association studies conducted so far in forest trees (Online Resource 6). In *Eucalyptus globulus*, the two SNPs in *CCR* associated with MFA (Thumma *et al.* 2005), explained 4.6% of the phenotypic variance. In *Pinus taeda*, Gonzalez-Martinez *et al.* (2007) identified 6 SNPs associated with wood properties, explaining between 2.2 and 3.6% of the phenotypic variance. In our case, the low level of LD detected between the three associated SNPs could indicate that each one explains a given part of the variance of S/G ratio. These results are encouraging and will have to be confirmed at the scale of the whole breeding population in order to determine the interest of such molecular tools in breeding. Finally, it should be recognized that given the high level of heritability of S/G ratio and the relatively low level of variation explained by the polymorphisms of the *CCR* gene, many other genes

Thomas Drudge; email: tgg@ucdavis.edu

(and their interactions) are likely involved in the genetic determinism of lignin quality in *Eucalyptus*.

Acknowledgements

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 N°25), from Agence Nationale de la Recherche, Plateformes Technologiques du Vivant (BOOST-SNP project, 07PFTV002), the Aquitaine Region (20061201004PFM, "ABIOGEN" FEDER project) and CIRAD. The field experiments were carried out at the CRDPI station (Pointe-Noire, Republic of Congo). The founders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alvira P, Tomas-Pejo E, Ballesteros M, Negro MJ (2009). Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. *Bioresour Technol* **101**(13)**:** 4851-4861.

Anterola AM, Lewis NG (2002). Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* **61**(3)**:** 221-294.

Boerjan W, Ralph J, Baucher M (2003). Lignin biosynthesis. *Annual Review of Plant Biology* **54:** 519-546.

Bose SK, Francis RC, Govender M, Bush T, Spark A (2009). Lignin content versus syringyl to guaiacyl ratio amongst poplars. *Bioresource Technology* **100**(4)**:** 1628-1633.

538 Campbell MM, Sederoff RR (1996). Variation in Lignin Content and Composition
539 (Mechanisms of Control and Implications for the Genetic Improvement of Plants). *Plant*
540 *Physiol* **110**(1)**:** 3-13.
541 Crisp M, Cook L, Steane D (2004). Radiation of the Australian flora: what can comparisons
542 of molecular phylogenies across multiple taxa tell us about the evolution of diversity in
543 present-day communities? *Philosophical Transactions of the Royal Society of London*
544 *Series B-Biological Sciences* **359**(1450)**:** 1551-1571.
545 De Melis LE, Whiteman PH, Stevenson TW (1999). Isolation and characterisation of a
546 cDNA clone encoding cinnamyl alcohol dehydrogenase in Eucalyptus globulus Labill.
547 *Plant Science* **143**(2)**:** 173-182.
548 Doyle JJ, and Doyle JL (1990). Isolation of plant DNA from fresh tissue. *Focus* **12:** 13–15.
549 Feuillet C, Boudet AM, Grimapettenati J (1993). Nucleotide Sequence of a cDNA encoding
550 Cinnamyl Alcohol Dehydrogenase from *Eucalyptus*. *Plant Physiol* **103**(4)**:** 1447-1447.
551 Freeman JS, Whittock SP, Potts BM, Vaillancourt RE (2009). QTL influencing growth and
552 wood properties in Eucalyptus globulus. *Tree Genetics and Genomes* **5**(4)**:** 713-722.
553 Gilmour A R, Gogel B J, Cullis B R, Welham S J, Thompson R (2002). ASReml User
554 Guide Release 1.0. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.*
555 Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A *et al.* (2005).
556 EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary
557 cell wall formation and lignin biosynthesis. Plant J 43(4): 553-567.
558 Gominho J, Rodrigues J, Almeida M H, Leal A, Cotterill P P, Pereira H (1997).
559 Assessment of pulp yield and lignin content in a first-generation clonal testing of
560 Eucalyptus globulus in Portugal. *Proceedings of the IUFRO Conference on Silviculture*
561 *and Improvement of Eucalypts, Salvador, Brazil*, **August 24–29**: 84–89.
562 Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007). Association
563 genetics in Pinus taeda L. I. Wood property traits. *Genetics* **175**(1)**:** 399-409.

Thomas Drudge; email: tgg@ucdavis.edu

564 Grima-Pettenati J, Feuillet C, Goffner D, Borderies G, Boudet AM (1993). Molecular
565 cloning and expression of a *Eucalyptus gunnii* cDNA clone encoding Cinnamyl Alcohol
566 Dehydrogenase. *Plant MolBiol* **21**(6)**:** 1085-1095.
567 Hall T A (1999). BioEdit: a user-friendly biological sequence alignment editor and
568 analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **41:**95-98.
569 Hannrup B, Cahalan C, Chantre G, Grabner M, Karlsson B, Le Bayon I *et al.* (2004).
570 Genetic parameters of growth and wood quality traits in Picea abies. *Scandinavian*
571 *Journal of Forest Research* **19**(1)**:** 14-29.
572 Harakava R (2005). Genes encoding enzymes of the lignin biosynthesis pathway in
573 Eucalyptus. *Genetics and Molecular Biology* **28**(3)**:** 601-607.
574 Hawkins S, Goffner D, Boudet AM (1994). Cinnamyl alcohol dehydrogenase
575 polymorphism and its potential role in the control of lignin heterogeneity. *Acta*
576 *Horticulturae* (381)**:** 280-286.
577 Hein P R G, Lima J T, Chaix G (2010). Effects of sample preparation on NIR spectroscopic
578 estimation of chemical properties of Eucalyptus urophylla S.T. Blake wood.
579 *Holzforschung* 64 (1): 45-54.
580 Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M *et al.* (2006).
581 Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic
582 history of Norway spruce [Picea abies (L.) Karst]. *Genetics* **174**(4)**:** 2095-2105.
583 Hu WJ, Harding SA, Lung J, Popko JL, Ralph J, Stokke DD *et al.* (1999). Repression of
584 lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees.
585 *Nature Biotechnology* **17**(8)**:** 808-812.
586 Hull J, Campino S, Rowlands K, Chan MS, Copley RR, Taylor MS *et al.* (2007).
587 Identification of common genetic variation that modulates alternative splicing. *Plos*
588 *Genetics* **3**(6)**:** 1009-1018.
589 Humphreys JM, Chapple C (2002). Rewriting the lignin roadmap. *Curr Opin Plant Biol*
590 **5**(3)**:** 224-229.

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol* **135**(4)**:** 2368-2378.

Kulheim 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**.

Lacombe E, Hawkins S, VanDoorsselaere J, Piquemal J, Goffner D, Poeydomenge O *et al.* (1997). Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: Cloning, expression and phylogenetic relationships. *Plant J* **11**(3)**:** 429-441.

Lacombe E, Van Doorsselaere J, Boerjan W, Boudet AM, Grima-Pettenati J (2000). Characterization of cis-elements required for vascular expression of the Cinnamoyl CoA Reductase gene and for protein-DNA complex formation. Plant J 23: 663-676.

Ladiges P Y, Udovicic F, Nelson G (2003). Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. Journal of Biogeography 30: 989-998.

Legay S, Lacombe E, Goicoechea M, Briere C, Seguin A, Mackay J et al. (2007). Molecular characterization of EgMYB1, a putative transcriptional repressor of the lignin biosynthetic pathway. Plant Science 173: 542-549.

Leplé JC, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B et al. (2007). Downregulation of cinnamoyl-coenzyme a reductase in poplar: Multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. Plant Cell **19**(11)**:** 3669-3691.

Li L, Zhou YH, Cheng XF, Sun JY, Marita JM, Ralph J *et al.* (2003). Combinatorial modification of multiple lignin traits in trees through multigene cotransformation.

Thomas Drudge; email: tgg@ucdavis.edu

Proceedings of the National Academy of Sciences of the United States of America **100**(8)**:** 4939-4944.

Martin B and Cossalter C (1976) a. Les Eucalyptus des îles de la Sonde. Partie 1. *Bois For. Trop.* **165:** 3-20

Martin B and Cossalter C (1976) b. Les Eucalyptus des îles de la Sonde. Partie 2. *Bois For. Trop.* **166:** 3-22

Martin B and Cossalter C (1976) c. Les Eucalyptus des îles de la Sonde. Partie 3. *Bois For. Trop.* **167:** 3-24

Neale DB, Savolainen O (2004). Association genetics of complex traits in conifers. *Trends in Plant Science* **9**(7)**:** 325-330.

Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C *et al.* (2009). Quantitative genetic analysis of biomass and wood chemistry of Populus under different nitrogen levels. *New Phytologist* **182**(4)**:** 878-890.

O'Connell A, Holt K, Piquemal J, Grima-Pettenati J, Boudet A, Pollet B *et al.* (2002). Improved paper pulp from plants with suppressed cinnamoyl-CoA reductase or cinnamyl alcohol dehydrogenase. *Transgenic Research* **11**(5)**:** 495-503.

Paux E, Carocha V, Marques C, de Sousa AM, Borralho N, Sivadon P *et al.* (2005). Transcript profiling of Eucalyptus xylem genes during tension wood formation. *New Phytologist* **167**(1)**:** 89-100.

Payn KG, Dvorak WS, Janse BJH, Myburg AA (2008). Microsatellite diversity and genetic structure of the commercially important tropical tree species Eucalyptus urophylla, endemic to seven islands in eastern Indonesia. *Tree Genetics & Genomes* **4**(3)**:** 519-530.

Peter G, Neale D (2004). Molecular basis for the evolution of xylem lignification. *Curr Opin Plant Biol* **7**(6)**:** 737-742.

Piquemal J, Lapierre C, Myton K, O'Connell A, Schuch W, Grima-Pettenati J *et al.* (1998). Down-regulation of cinnamoyl-CoA reductase induces significant changes of lignin profiles in transgenic tobacco plants. *Plant J* **13**(1)**:** 71-83.

Thomas Drudge; email: tgg@ucdavis.edu

644 Poeydomenge O, Boudet AM, Grimapettenati J (1994). A cDNA Encoding S-Adenosyl-L-
645 Methionine: Caffeic Acid 3-O-Methyltransferase from *Eucalyptus*. *Plant Physiol*
646 **105**(2)**:** 749-750.
647 Poke FS, Potts BM, Vaillancourt RE, Raymond CA (2006). Genetic parameters for lignin,
648 extractives and decay in Eucalyptus globulus. *Annals of Forest Science* **63**(8)**:** 813-821.
649 Poke FS, Vaillancourt RE, Elliott RC, Reid JB (2003). Sequence variation in two lignin
650 biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol
651 dehydrogenase 2 (CAD2). *Molecular Breeding* **12**(2)**:** 107-118.
652 Pot D, Chantre G, Rozenberg P, Rodrigues JC, Jones GL, Pereira H *et al.* (2002). Genetic
653 control of pulp and timber properties in maritime pine (Pinus pinaster Ait.). *Annals of*
654 *Forest Science* **59**(5-6)**:** 563-575.
655 Ralph J, Hatfield RD, Piquemal J, Yahiaoui N, Pean M, Lapierre C *et al.* (1998). NMR
656 characterization of altered lignins extracted from tobacco plants down-regulated for
657 lignification enzymes cinnamyl-alcohol dehydrogenase and cinnamoyl-CoA reductase.
658 *Proceedings of the National Academy of Sciences of the United States of America*
659 **95**(22)**:** 12803-12808.
660 Raymond CA (2002). Genetics of Eucalyptus wood properties. *Annals of Forest Science*
661 **59**(5-6)**:** 525-531.
662 Rengel D, San Clemente H, Servant F, Ladouce N, Paux E, Wincker P *et al.* (2009). A new
663 genomic resource dedicated to wood formation in Eucalyptus. *BMC Plant Biology*
664 **9**(36)**:** (27 March 2009).
665 Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA
666 polymorphism analyses by the coalescent and other methods. *Bioinformatics (Oxford,*
667 *England)* **19**(18)**:** 2496-2497.
668 Slotte T, Huang HR, Holm K, Ceplitis A, Onge KS, Chen J *et al.* (2009). Splicing variation
669 at a FLOWERING LOCUS C homeolog is associated with flowering time variation in
670 the tetraploid Capsella bursa-pastoris. Genetics 183(1): 337-345.

671 Savidge R A (2000). Biochemistry of seasonal cambial growth and wood formation – an
672 overview of the challenges. Savidge R A, Barnett J R, Napier R (eds.). Cell and
673 molecular biology of wood formation. BIOS Scientific publishers Ltd, Oxford, UK : 1-
674 30.
675 Sykes R, Li BL, Isik F, Kadla J, Chang HM (2006). Genetic variation and genotype by
676 environment interactions of juvenile wood chemical properties in Pinus taeda L. Annals
677 of Forest Science 63(8): 897-904.
678 Thumma BR, Matheson BA, Zhang DQ, Meeske C, Meder R, Downes GM et al. (2009).
679 Identification of a Cis-Acting Regulatory Polymorphism in a Eucalypt COBRA-Like
680 Gene Affecting Cellulose Content. *Genetics* **183**(3)**:** 1153-1164.
681 Thumma BR, Nolan MR, Evans R, Moran GF (2005). Polymorphisms in cinnamoyl CoA
682 reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp.
683 *Genetics* **171**(3)**:** 1257-1265.
684 Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF (2010).
685 Quantitative trait locus (QTL) analysis of wood quality traits in Eucalyptus nitens. *Tree*
686 *Genetics and Genomes* **6**(2)**:** 305-317.
687 Tripiana V, Bourgeois M, Verhaegen D, Vigneron P, Bouvet JM (2007). Combining
688 microsatellites, growth, and adaptive traits for managing in situ genetic resources of
689 Eucalyptus urophylla. *Canadian Journal of Forest Research-Revue Canadienne De*
690 *Recherche Forestiere* **37**(4)**:** 773-785.
691 Wu RL, Remington DL, MacKay JJ, McKeand SE, O'Malley DM (1999). Average effect of
692 a mutation in lignin biosynthesis in loblolly pine. *Theoretical and Applied Genetics*
693 **99**(3-4)**:** 705-710.
694 Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF *et al.* (2006). A unified
695 mixed-model method for association mapping that accounts for multiple levels of
696 relatedness. *Nature Genetics* **38**(2)**:** 203-208.

Table 1: Genetic parameter estimates for growth (circumference and height) and lignin related traits (KL (%) and S/G ratio) in *E. urophylla*.

	Mean	CV (%) [1]	(σ^2_A) [2]	(σ^2_P) [3]	(h^2) [4]
Height (m)	21.15	17.31	4.08 ± *2.24*	13.66 ± *1.41*	0.30 ± *0.14*
Circumference (cm)	52.77	21.20	18.47 ± *11.69*	125.95 ± *10.87*	0.15 ± *0.09*
KL (%)	28.15	4.33	1.38 ± *0.57*	1.63 ± *0.29*	0.85 ± *0.21*
S/G ratio	2.42	12.28	0.06 ± *0.03*	0.09 ± *0.01*	0.62 ± *0.19*

[1] CV (%) is the phenotypic variation coefficient given in percent.

[2] σ^2_A is the genetic additive variance. Standard deviations are indicated in italic.

[3] σ^2_P is the total phenotypic variance. Standard deviations are indicated in italic.

[4] h^2 is the narrow sense heritability. Standard deviations are indicated in italic.

Thomas Drudge; email: tgg@ucdavis.edu

706 Table 2: Correlation matrix between growth (height and circumference) and lignin related traits (KL (%) and S/G ratio).

	Phenotypic correlations[1]			
	Height (m)	Circumference (cm)	KL (%)	S/G ratio
Height (m)	/	0.766 ± *0.029*	-0.103 ± *0.094*	0.024 ± *0.086*
Circumference (cm)	0.679 ± *0.213*	/	0.086 ± *0.080*	-0.050 ± *0.074*
KL (%)	-0.533 ± *0.241*	-0.276 ± *0.351*	/	-0.234 ± *0.107*
S/G ratio	0.312 ± *0.305*	0.222 ± *0.367*	-0.246 ± *0.281*	/
	Genetic additive correlations[1]			

707 [1] Phenotypic correlations and corresponding standard deviations are shown on the upper side of the diagonal and genetic additive correlations, and

708 corresponding standard deviations are shown on the lower side.

Thomas Drudge; email: tgg@ucdavis.edu

Table 3: Summary of the detected polymorphic sites in the 5 exons and 4 introns of *CCR* of *E. urophylla*.

Gene region	Length (bp)	Number of polymorphic sites: biallelic SNPs	triallelic SNPs	INDELs	poly A	SSR	NS SNPs[3]	SNP density (bp/SNP)
Exon #1[1]	156	1						1/156
Exon #2	156	5					1	1/31
Exon #3	183	8					1	1/23
Exon #4	355	17					2	1/21
Exon #5[2]	292	10						1/29
Total exons	1142	41					4	1/28
Intron #1	117	2		2	1			1/59
Intron #2	689	38	1	6		1		1/18
Intron #3	166	11	1	1				1/15
Intron #4	1022	64	2	8		1		1/16
Total introns	1994	115	4	17	1	2		1/17
Total gene	3136	156	4	17	1	2	4	1/20

[1] 24 bp before the start codon.

[2] 108 bp after the stop codon.

[3] The NS SNP column indicates the number of non-synonymous SNPs detected in each exon.

Table 4: Nature of the 4 non-synonymous mutations detected in *CCR*.

Gene region	Codon Change	Amino Acid change	Amino Acid Nb
Exon #2	AAA / GAA	Lys / Glu	77
Exon #3	GTG / TTG	Val / Leu	123
Exon #4	AAG / AAT	Lys / Asn	177
Exon #4	ACC / GCC	Thr / Ala	241

Table 5: Significant associations for the marker-by-marker approach after correction for multiple testing by the Bonferroni method.

					R^2 marker [3]	
Locus	Gene location	MAF[1]	Nb of genotypes	P-value [2]	Phenotypic	Genetic
SNP#30	Intron #2	0.13	3	0.00069	1.52	2.45
SNP#35	Intron #2	0.06	3	0.00064	1.61	2.6
SNP#138	Intron #4	0.38	3	0.00061	1.78	2.87

[1] Minor Allele Frequency

[2] P-value of the Fisher's exact test.

[3] Part of phenotypic and genetic variance explained by the marker (%). The part of genetic variance explained was obtained dividing the part of phenotypic variance explained by the heritability of the trait ($h^2 = 0.62$).

Thomas Drudge; email: tgg@ucdavis.edu

Table 6: Matrix of pairwise linkage disequilibrium (r^2) between the 3 SNPs significantly associated with S/G ratio.

	SNP#30	SNP#35	SNP#138
SNP#30	/	0.41	0.24
SNP#35	/	/	0.10

LD values are highly significant (P-value < 0.001).

Fig. 1: Distribution of biallelic SNPs according to their minor allele frequency (MAF)

Supplementary Tables and Figures:

Online Resource 1: Factorial experimental design.

x		Fathers							
		1	2	3	4	5	6	7	8
Mothers	9		FS	FS	FS		FS		
	10	FS	FS				FS	FS	
	11	FS		FS	FS	FS			FS
	12	FS	FS		FS			FS	FS
	13						FS	FS	FS
	14		FS	FS		FS	FS	FS	
	15				FS	FS			FS
	16	FS				FS		FS	FS

16 founders consisting of 8 fathers *E. urophylla* (from 1 to 8) and 8 mothers *E. urophylla* (from 9 to 16) collected in Flores island and 33 full-sib families (FS) obtained by controlled crosses.

Thomas Drudge; email: tgg@ucdavis.edu

Online Resource 2: Primer pairs designed for the PCR amplification of 7 overlapping fragments from the *CCR* gene of *E. urophylla*.

PCR fragment		PCR primer (5' → 3')[1]	Tm (°C)[2]	Product size[3]
F1	forward	CACCTCCTGAACCCCTCT	63	397 bp
	reverse	CGCACCCCTTGATGGCTTCT		
F2	forward	GCGAGGAACCGTCAGGAAC	58	619 bp
	reverse	TTTCCTCCCCAATCGTCTG		
F3	forward	AAGAATGTGCGATGGCGAACC	70	474 bp
	reverse	GTCCCGATCACCGCTGGCT		
F4	forward	ACGTAAGAAAGAGGGACCG	66	672 bp
	reverse	ACTTGAGGATGTGGATGATG		
F5	forward	GCTACGGCAAGGCAGTGG	66	631 bp
	reverse	AACCGACAACCCACACCTG		
F6	forward	CTTAGATAGATAGTCCCGC	56	649 bp
	reverse	CAAAGGGATTCAAGACAGG		
F7	forward	CGTCATCATCGTTCTCTCT	56	695 bp
	reverse	TGACAACTTCCATTCCAA		
SSR	forward	AGGTGTGGGTTGTCG	56	112 bp
	reverse	ATTTCCTCCCTTTTGCCC		

[1] All primer sequences were based on the *E. gunnii* gene (X97433).

[2] Annealing temperatures are given for each primer pair.

[3] Expected lengths are given for each amplicon based on the *E. gunnii* gene (X97433).

Thomas Drudge; email: tgg@ucdavis.edu

Online Resource 3: Allele sequences and allele sizes for the SSR detected in intron #4 of *CCR* among the 16 founders of *E. urophylla*.

Founder	Allele	Size (bp)
1; 3; 7	$(TT)_2(CT)_3(CA)_1(CT)_5$	22
3; 4; 12; 10	$(TT)_2(CT)_{10}$	24
1; 5; 12; 7	$(TT)_2(CT)_{13}$	30
2; 14; 6[1]; 6[1]; 13	$(TT)_1(CT)_{14}$	30
11	$(TT)_2(CT)_4(GT)_1(CT)_3(GT)_1(CT)_5$	32
8; 13	$(TT)_2(CT)_4(GT)_1(CT)_3(GT)_1(CT)_3(GT)_1(CT)_1$	32
16	$(TT)_2(CT)_{15}$	34
14	$(TT)_2(CT)_{16}$	36
9	$(TT)_2(CT)_4(GT)_1(CT)_{12}$	38
10	$(TT)_2(CT)_{18}$	40
9	$(TT)_2(CT)_4(GT)_1(CT)_{14}$	42
16	$(TT)_2(CT)_4(GT)_1(CT)_{15}$	44
2; 11; 8; 15[1]; 15[1]	$(TT)_2(CT)_4(GT)_1(CT)_{16}$	46
5	$(TT)_2(CT)_{22}(GT)_1(CT)_1$	52
4	$(TT)_1(CT)_{26}(GT)_1(CT)_1$	58

[1] Homozygous line for *CCR*.

Thomas Drudge; email: tgg@ucdavis.edu

Online Resource 4: P-value distribution of the association tests

Density of P-values obtained from the association tests between 65 SNPs of *CCR* and S/G ratio, Klason lignin, height and circumference. The grey area under the curve obtained for S/G ratio represents the proportion of the 65 independent tests considered as significant after Bonferroni correction at the experimental wise level of 5% (-log(P-value)=3.114).

Online Resource 5: Nucleotide diversity in forest tree species.

This histogram represents the distribution (histogram) and the density (line) of genes according to their estimated nucleotide diversity θ_π. This includes results obtained from different genes or portions of genes, for different populations of different forest tree species. This figure is not exhaustive and is based on the studies of Kado *et al.* (2003), Fujimoto *et al.* (2008), Pot *et al.* (2005), Brown *et al.* (2004), Gonzales Martinez *et al.* (2006), Ma *et al.* (2006), Palmé *et al.* (2008), Wachowiak *et al.* (2009), Krutovsky and Neale (2005), Heuertz *et al.* (2006), Ingvarson *et al.* (2005), Breen *et al.* (2009), Quang *et al.* (2008).

References of Online Resource 5:

Breen AL, Glenn E, Yeager A, Olson MS (2009). Nucleotide diversity among natural populations of a North American poplar (Populus balsamifera, Salicaceae). *New Phytologist* **182**(3)**:** 763-773.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America* **101**(42)**:** 15255-15260.

Thomas Drudge; email: tgg@ucdavis.edu

Fujimoto A, Kado T, Yoshimaru H, Tsumura Y, Tachida H (2008). Adaptive and slightly deleterious evolution in a conifer, Cryptomeria japonica. *Journal of Molecular Evolution* **67**(2)**:** 201-210.

Gonzalez-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**(3)**:** 1915-1926.

Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M *et al.* (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. *Genetics* **174**(4)**:** 2095-2105.

Ingvarsson PK (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (Populus tremula L., Salicaceae). *Genetics* **169**(2)**:** 945-953.

Kado T, Yoshimaru H, Tsumura Y, Tachida H (2003). DNA Variation in a Conifer, Cryptomeria japonica (Cupressaceae sensu lato). *Genetics* **164**(4)**:** 1547-1559.

Krutovsky KV, Neale DB (2005). Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. *Genetics* **171**(4)**:** 2029-2041.

Ma XF, Szmidt AE, Wang XR (2006). Genetic structure and evolutionary history of a diploid hybrid pine Pinus densata inferred from the nucleotide variation at seven gene loci. *Molecular Biology and Evolution* **23**(4)**:** 807-816.

Palme AE, Wright M, Savolainen O (2008). Patterns of divergence among conifer ESTs and polymorphism in Pinus sylvestris identify putative selective sweeps. *Molecular Biology and Evolution* **25**(12)**:** 2567-2577.

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Gere P, Cato S *et al.* (2005). Nucleotide variation in genes involved in wood formation in two pine species. *New Phytologist* **167**(1)**:** 101-112.

Thomas Drudge; email: tgg@ucdavis.edu

812 Quang ND, Ikeda S, Harada K (2008). Nucleotide variation in Quercus crispula Blume.
813 *Heredity* **101**(2)**:** 166-174.
814 Wachowiak W, Balk PA, Savolainen O (2009). Search for nucleotide diversity patterns of
815 local adaptation in dehydrins and other cold-related candidate genes in Scots pine
816 (Pinus sylvestris L.). *Tree Genetics & Genomes* **5**(1)**:** 117-132.
817

Online Resource 6: A review of association studies in forest trees.

Species	Traits	Associated Locus	PEV [1]	Method	Multiple testing correction method	Population	Reference
Pseudotsuga menziesii	Traits related to emergence, growth and ressource partitionning, phenology and cold tolerance	15 SNPs from 12 candidate genes	1.9% to 3.6%	GLM with traits as dependant variables and genotypic values corrected for population structure as explanatory variables	FDR Q-values (threshold 0.1)	700 unrelated individuals	Eckert *et al.* 2009
Populus tremula	Timing of bud set	2 SNPs from one candidate gene	1.4 to 5.9% [2]	Mixed Linear Model (MLM) corrected for population structure and relatedness	FDR Q-values (threshold 0.05)	120 clones almost unrelated	Ingvarson *et al.* 2008
Populus Trichocarpa	Sugar and lignin related traits	27 SNPs from 40 candidate genes	1.1 to 3.8%	GLM with traits as dependant variables and genotypic values corrected for population structure as explanatory variables	FDR Q-values (threshold 0.1)	1189 trees (448 replicated clones) from 101 provenances of Cascade mountains	Wegrzyn *et al.* 2010
Pinus taeda	Wood quality traits	4 SNPs from 4 candidate genes	2.2% to 3.6%	Mixed Linear Model (MLM) corrected for population structure and relatedness	FDR Q-values (threshold 0.1 and 0.05)	435 related clones from first and second generation of *Pinus taeda* breeding program	Gonzalez-Martinez *et al.* 2007
Pinus taeda	Carbon isotope discrimination	4 SNPs from 4 candidate genes	0.5% to 3.4%	Transmission Disequilibrium Test for Quantitative traits	Permutation tests (1000) with experiment wise level of 0.05, Bonferroni correction with experiment wise level of 0.05	961 clones from 61 families of *Pinus taeda* breeding program (partial diallele)	Gonzalez-Martinez *et al.* 2008
Pinus taeda	Four aridity indexes	5 SNPs from 5 unique ESTs	1.5% to 4%	GLM with environment as dependant variables and corrected genotypic values as explanatory variables	FDR Q-values (threshold 0.05), Bonferroni corrections with experiment wise level of 0.05	622 trees, largely unrelated, from *Pinus taeda* breeding programs	Eckert *et al.* 2010
Pinus radiata	Wood quality traits	10 SNPs form 9 candidate genes	2% to 3.5%	GLM with traits as dependant variables and genotypic values corrected for population structure as explanatory variables	FDR Q-values (threshold 0.1), Bayes factor with priors for allele frequency, P-value and population size	447 unrelated trees from 3 provenances of the natural distribution in California, 458 individuals of second generation progeny trial (from 229 families)	Dillion *et al.* 2010
Eucalyptus nitens	Microfibril angle	2 SNPs from one gene	4.60%	ANOVA with traits as dependant variables and genotypic values as explanatory variables, haplotype analysis (three markers sliding wiondows)	Bonferroni correction, permutation tests	290 trees from different open pollinated families	Thumma *et al.* 2005
Eucalyptus nitens	Cellulose content and kraft pulp yield	1 SNP from one candidate gene	Less than 1%	GLM with traits as dependant variables and genotypic values as explanatory variables, haplotype analysis (three markers sliding wiondows)	Permutation tests (1000) with experiment wise level of 0.05	300 trees from different open pollinated families	Thumma *et al.* 2009

[1] Percentage of phenotypic variance explained

[2] Corrected using method of moment approach

References of Online Resource 6:

Dillon SK, Nolan M, Li W, Bell C, Wu HX, Southerton SG (2010). Allelic variation in cell wall candidate genes affecting solid wood properties in natural populations and land races of Pinus radiata. *Genetics* **185**(4)**:** 1477-1487.

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV *et al.* (2009). Association genetics of coastal Douglas fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-hardiness related traits. *Genetics* **182**(4)**:** 1289-1302.

Eckert AJ, van Heervaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC *et al.* (2010). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). Genetics.

Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008). Association genetics in Pinus taeda L. II. Carbon isotope discrimination. *Heredity* **101**(1)**:** 19-26.

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007). Association genetics in Pinus taeda L. I. Wood property traits. *Genetics* **175**(1)**:** 399-409.

Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S (2008). Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (Populus tremula, Salicaceae). *Genetics* **178**(4)**:** 2217-2226.

Thumma BR, Matheson BA, Zhang DQ, Meeske C, Meder R, Downes GM *et al.* (2009). Identification of a Cis-Acting Regulatory Polymorphism in a Eucalypt COBRA-Like Gene Affecting Cellulose Content. *Genetics* **183**(3)**:** 1153-1164.

Thumma BR, Nolan MR, Evans R, Moran GF (2005). Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. *Genetics* **171**(3)**:** 1257-1265.

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R et al (2010). Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem. New Phytologist 188(2): 515-532.

Thomas Drudge; email: tgg@ucdavis.edu

Online Resource 7: Scatter plot of the squared correlation of allele frequencies (r^2), according to the distance between pairs of polymorphic sites in bp for *CCR* in *E. urophylla.*

This plot is based on biallelic SNPs with MAF superior or equal to 0.15 in the sequenced parental lines (16 unrelated *E. urophylla*). Linkage disequilibrium values (grey circles) were estimated using the software package TASSEL (http://www.maizegenetics.net). The decay of linkage disequilibrium with physical distance [E(r^2), black line] was estimated by nonlinear regression (Remington *et al.* 2001). The expected value of r^2 between pairs of adjacent sites was calculated from the formula:

$$E(r^2) = \left[\frac{10+C}{(2+C)(11+C)}\right]\left[1+\frac{(3+C)(12+12C+C^2)}{n(2+C)(11+C)}\right],$$

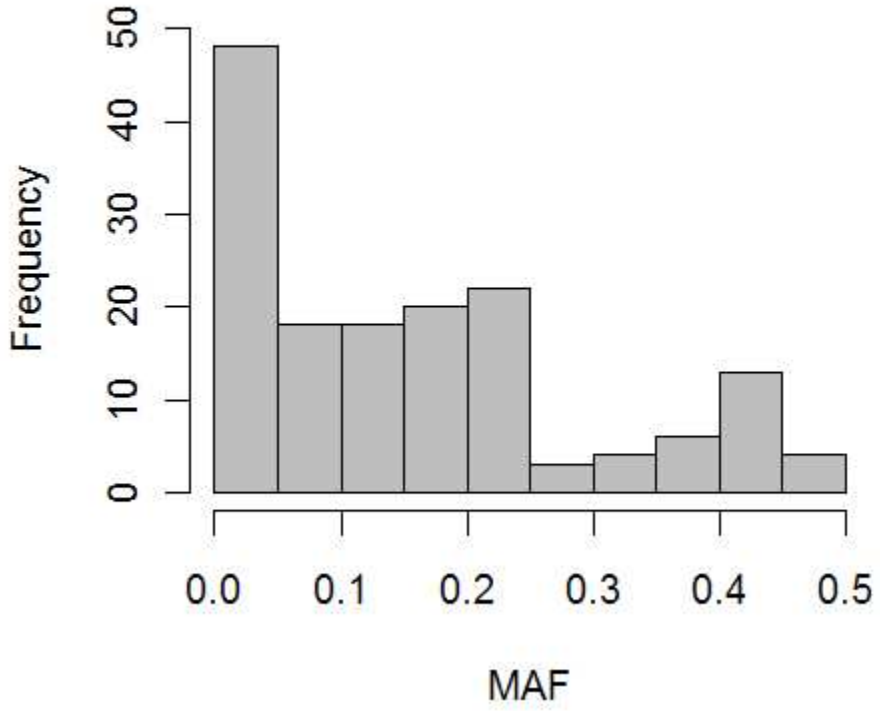
under drift recombination equilibrium, low mutation rate and with an adjustment for sample size (Weir and Hill 1988) and where *C* is the population recombination parameter ($\rho=4Ner$ where *Ne* is the effective size and *r* is the recombination rate per site per generation) and n the sample size (n=32). To fit this formula to our data, a nonlinear regression was used (nls function) in the R software (R Development Core Team 2005) replacing C by C x distance between pairs of sites in base pairs.

References of Online Resource 7:

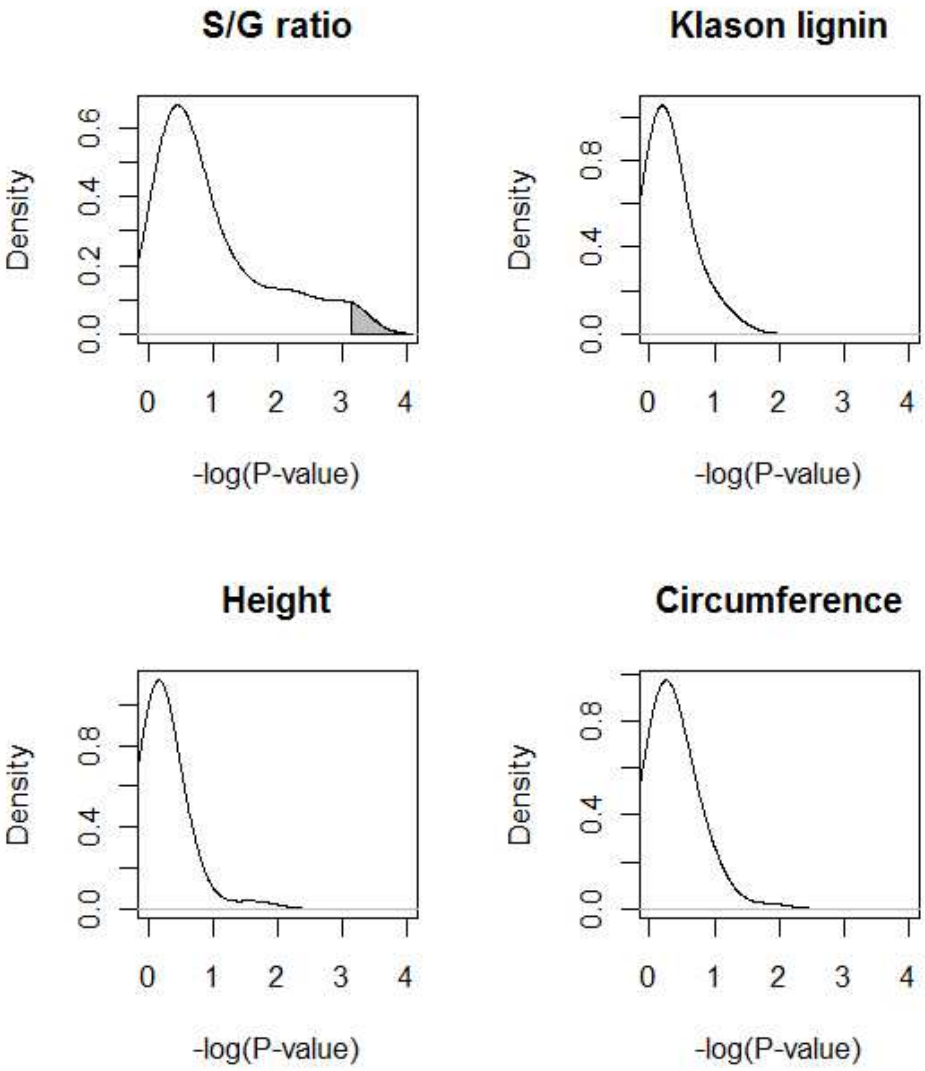
Remington D L, Thornsberry J M, Matsuoka Y, Wilson L M, Whitt S R *et al.* (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* **98:** 11479–11484.

Hill W G, and Weir B S (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33:** 54–78.

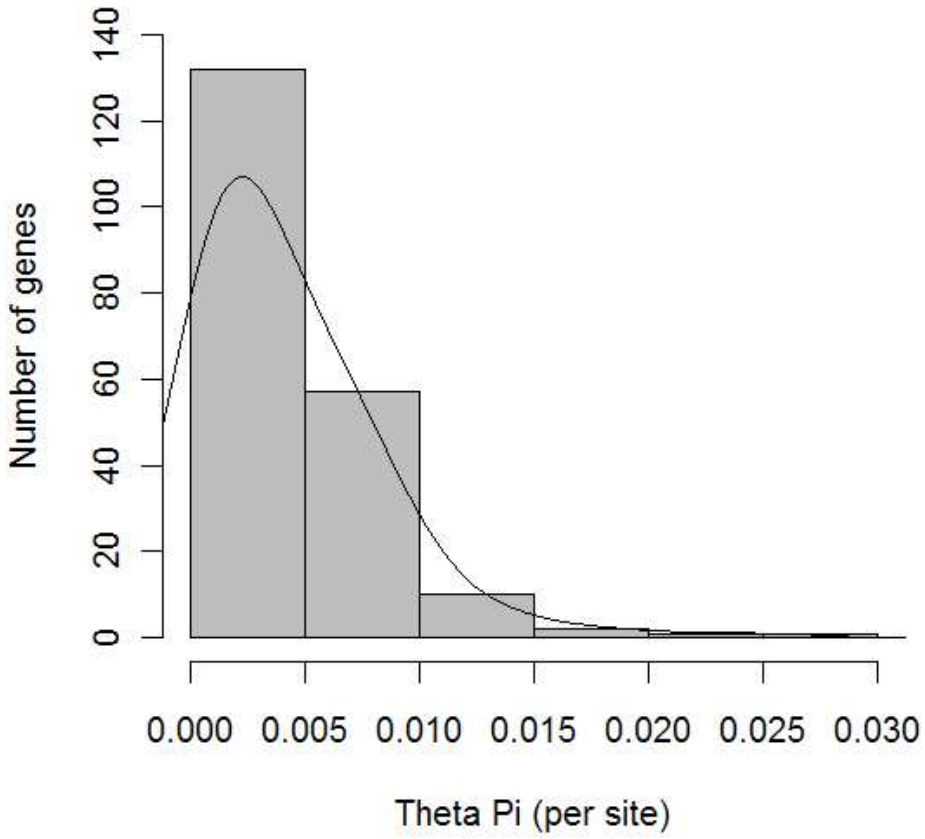
Thomas Drudge; email: tgg@ucdavis.edu



Distribution of biallelic SNPs according to their minor allele frequency (MAF)
172x166mm (72 x 72 DPI)



197x226mm (72 x 72 DPI)



203x210mm (72 x 72 DPI)

Résumé :

Les lignines représentent 25% de la biomasse des végétaux terrestre. Leur quantité (teneur en lignines) et leur qualité (rapport des monomères S/G) sont variables au niveau intraspécifique et constituent des cibles majeures de l'amélioration génétique des eucalyptus. L'étude de ces deux caractères est relativement récente et leur déterminisme génétique reste encore mal connu chez les arbres forestiers. La mise en évidence de marqueurs génétiques liés à la variation de ces caractères représente un autre enjeu majeur pour leur prise en compte dans les programmes d'amélioration génétique. Elle permettrait de disposer d'outils de diagnostic moléculaire pour la sélection précoce d'idéotypes ciblés pour des applications spécifiques (pâte à papier, charbon de bois).

Dans un premier temps, nous avons mené une étude sur le déterminisme génétique de ces caractères sur la base de descendances de plein-frères de plans de croisements factoriels impliquant les espèces *E. urophylla*, *E. camaldulensis* et *E. grandis*. Dans un second temps nous nous sommes attachés à décrire chez *E. urophylla* la diversité nucléotidique et haplotypique de 10 gènes impliqués dans la biosynthèse des lignines. Parmi ces gènes une attention particulière a été est portée sur le gène de structure codant la Cinnamoyl CoA Reductase (*CCR*) dont la variabilité est décrite chez *E. urophylla* et *E. camaldulensis*. Enfin, nous avons recherché des liens statistiques entre la variabilité des gènes et la variation des caractères relatifs aux lignines par une étude d'association au sein de deux plans de croisement factoriels.

Cette étude a permis d'identifier des mutations ponctuelles (SNP), situés au sein des gènes codant pour la *CCR* et une protéine impliquée dans la signalisation cellulaire (*ROP1*), associés à la variation du rapport S/G. Ces polymorphismes expliquent une faible part de la variation du caractère mais confirment les données de cartographie de QTL indiquant des co-localisations entre ces deux gènes et des QTL du rapport S/G dans un croisement interspécifique *E. urophylla* x *E. grandis*. Une étude menée à l'échelle d'une population à base génétique plus large permettrait de confirmer l'effet de ces gènes sur la variation du caractère et si les effets sont confirmés, d'utiliser les polymorphismes détectés comme critère de sélection précoce pour la qualité des lignines.

Abstract :

Lignins account for 25% of the terrestrial plant biomass. Their quantity (lignin content) and their quality (S/G ratio) are variable at the species level and constitute major targets of eucalyptus breeding programs. The study of these traits is still in its infancy and their genetic determinism remains poorly described in forest trees. The identification of molecular markers associated with the phenotypic variation is another challenge that would allow to put in place breeding strategies based on molecular tools for the early selection of ideotypes with improved properties for the pulp and paper as well as the charcoal industry.

We first studied the genetic determinism of these two traits using full-sib progenies of factorial designs involving *E. urophylla*, *E. camaldulensis* and *E. grandis*. Second, we described in *E. urophylla* the nucleotide and haplotype diversity of 10 genes involved in lignin biosynthesis. Emphasis was given to the gene encoding a CCR for which the variability was described in *E. urophylla* and *E. camaldulensis*. Finally, the statistical association between nucleotide variability and phenotypic variation was tested based on the information gathered in two factorial designs.

This study allows to identify single nucleotide polymorphisms (SNPs) in CCR and ROP1 (a gene encoding for a protein involved in cellular signaling) accounting for a small part of the variation of S/G ratio. It is noteworthy to mention that these two genes co-localised with S/G ratio-QTLs in an inter-specific cross between *E. urophylla* x *E. grandis*. Further studies using a broader genetic background should be carried out to validate our findings. If confirmed, the identified polymorphisms could be used as early selection criteria.