HAL
open science

# Un modèle de variabilité fonctionnelle chez les arbres forestiers : le gène CCR d'eucalyptus

Jean-Marc J.-M. Gion, Fréderic Mortier, Eric Mandrou, Paulo Ricardo Hein Gherardi, Tristan Costecalde, Gilles Chaix, Marie Pierre Etienne, Pierre Sivadon, Jacqueline Grima Pettenati, Emilie Villar, et al.

## To cite this version:

Jean-Marc J.-M. Gion, Fréderic Mortier, Eric Mandrou, Paulo Ricardo Hein Gherardi, Tristan Costecalde, et al.. Un modèle de variabilité fonctionnelle chez les arbres forestiers : le gène CCR d'eucalyptus. 7. Colloque National: Ressources Génétiques, Oct 2008, Strasbourg, France. 17 p. hal-02821259

**HAL Id: hal-02821259**

**https://hal.inrae.fr/hal-02821259v1**

Submitted on 6 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Titre complet : Un modèle de variabilité fonctionnelle chez les arbres forestiers : le gène CCR d'eucalyptus
Titre courant : Variabilité fonctionnelle du gène CCR d'*Eucalyptus*

Jean-Marc GION [(1,2)*], Fréderic MORTIER [(1)], Eric MANDROU [(1,2,3)], Paulo Ricardo HEIN GHERARDI [(1)], Tristan COSTECALDE [(1,2)], Gilles CHAIX [(1)], Marie ETIENNE [(4)], Pierre SIVADON [(5)], Jacqueline GRIMA-PETTENATI [(5)], Emilie VILLAR [(6)], Aubin SAYA [(6)], Brigitte POLLET [(7)], Catherine LAPIERRE [(7)], Philippe VIGNERON [(1)]

[(1)] CIRAD UPR 39 Génétique Forestière Dpt BIOS Campus international de Baillarguet TA A/39 C 34398 Montpellier Cedex 5, France
[(2)] CIRAD UPR39 / UMR BIOGECO 1202 INRA, Equipe de Génétique 69 route d'Arcachon F-33612 CESTAS Cedex France
[(3)] Vallourec, 27 avenue du Général Leclerc 92100 Boulogne-Billancourt, France
[(4)] ENGREF –UMR 518. GRESE 19 Avenue du Maine 75732 Paris, France
[(5)] UMR UPS-CNRS 5546 « Surfaces Cellulaires et Signalisation Végétales » Pôle de Biotechnologie Végétale 24, chemin de Borde Rouge BP 42617 Auzeville, France
[(6)] Ur2pi, BP 1291 Pointe Noire, République du Congo
[(7)] UMR Chimie Biologique. INRA - Agro Paris Tech BP1 78850 Thiverval-Grignon, France
*Correspondance : gion@cirad.fr

Abstract: **CCR gene in *Eucalyptus*: a model of functional variability in forest trees.** Nucleotidic polymorphism of Cinnamoyl CoA Reductase (CCR) gene and its relation with lignin content is studied within a breeding population of *Eucalyptus urophylla* S.T. Blake ("Timor Mountain Gum"). The nearly full sequence (94%) are obtained for 15 parental trees. This gene (3220 bp) is highly polymorphic showing 131 single nucleotide polymorphism (SNP), 17 insertion-deletions (INDEL), 1 polyA sequence and a microsatellite site. Exons fragments encompass 10 non-synonymous SNPs, half of them within exon 5 (194 bp). Fifteen different haplotypes are reconstructed based on the polymorphism of exon 4 and intron 4. CCR promoting sequence (694 bp) including all the known regulatory sequences is described for the two alleles of one of the genitor trees displaying QTL and CCR gene colocalization in its genetic map. Five SNPs are present. Functional variability of the promoting sequence will be studied *in planta* through genetic modification of *Arabidopsis thaliana*. Lignin content was assessed within a sample of 35 full sib families (348 individuals) generated with the 15 parental trees, showing a high genetic additive control for this trait ($h^2$=0.76). A new algorithm based on Reversible-jump MCMC was developed in order to implement association studies. Half of the progeny trees (208) were genotyped using the microsatellite fragment. The results show that a significant part of the observed genetic variance of lignin content is due to the nucleotide polymorphism of the studied gene. Those preliminary results look promising in order to develop early gene assisted selection for eucalyptus clones used as raw material in charcoal and paper production.
Keywords : CCR gene, eucalyptus, variability, lignins, association study

Résumé : La variabilité nucléotidique du gène codant la Cinnamoyl CoA Reductase (CCR) et ses effets sur le taux de lignine est étudiée au sein d'une population d'*Eucalyptus urophylla* S.T. Blake. La presque totalité de la séquence (94%, 3220 paires de bases) est décrite pour 15 individus. Le gène est hautement polymorphe et présente 131 mutations ponctuelles (SNP) ainsi que divers autres types de mutations. Les fragments exoniques présentent 10 SNP non synonymes dont 5 dans l'exon 5. La séquence promotrice (694 pb) est décrite pour les deux allèles d'un des géniteurs. Elle regroupe 5 SNPs. La variabilité fonctionnelle de ce promoteur sera étudiée grâce à son expression dans *Arabidopsis thaliana*. L'analyse de la teneur en lignine de 348 arbres appartenant à 35 familles de pleins frères obtenues avec ces 15 géniteurs montre que ce caractère présente un fort contrôle génétique additif ($h^2$=0.76). Un nouvel algorithme type MCMC a été développé pour procéder aux études d'association sur 208 descendants génotypés grâce à un marqueur microsatellite présent dans le gène CCR. Les résultats montrent qu'une part importante de la variance du taux de lignine est due au polymorphisme du gène CCR. Ces résultats laissent envisager le développement d'une sélection précoce assistée par marqueurs.
Mots clefs : gène CCR, eucalyptus, variabilité, lignines, test d'association

# 1 INTRODUCTION

La conservation et l'exploitation durable des ressources génétiques forestières nécessitent une bonne connaissance de la variabilité naturelle existante. Cette variabilité a souvent été analysée grâce aux caractères phénotypiques eux-mêmes ou à l'aide de marqueurs moléculaires anonymes permettant de décrire la variabilité neutre au sein de populations forestières. L'analyse de la variabilité nucléotidique de gènes d'intérêt reste encore limitée chez les génomes forestiers même si le développement récent de programmes de séquençage fournit une quantité croissante de séquences exprimées (Gupta and Rustgi, 2004 [14]). Ceci permet d'envisager une caractérisation de la variabilité fonctionnelle afin d'élaborer des stratégies de conservation et de sélection assistées par marqueurs chez ces espèces peu domestiquées et à long cycle de révolution (Neale and Savolainen, 2004 [22]). Les premiers travaux sur la variabilité de populations forestières grâce à l'analyse de gènes d'intérêt ont été réalisés chez le pin et l'eucalyptus (Pot *et al.*, 2005 [30], Poke *et al.*, 2003 [27]). Ils semblent indiquer une variabilité nucléotidique plus importante ainsi qu'un déséquilibre de liaison intra-séquences plus faible chez ces populations forestières que chez l'homme, la drosophile ou encore le maïs. Une étude sur *E. nitens* indique que le déséquilibre de liaison au sein du gène CCR décroît rapidement avec la distance physique et montre une répartition hétérogène de la variabilité nucléotidique le long du gène (Thumma *et al.*, 2005 [34]). L'analyse de la diversité de gènes pour les espèces forestières concerne, en général, un fragment du gène (EST) et donc une partie limitée de sa variabilité totale. De plus, les régions promotrices, dont la variabilité est susceptible d'affecter l'expression du gène, sont rarement prises en compte (Neale and Savolainen, 2004 [22]). La conservation et l'exploitation de la variabilité fonctionnelle de gènes d'intérêt nécessitent un recensement des formes alléliques existantes et donc l'analyse préalable de séquences les plus complètes possibles.

Les propriétés du bois constituent un caractère majeur pour l'amélioration génétique des espèces forestières et leur adaptation au milieu. Ces propriétés dépendent en grande partie de la lignine, qui est déposée dans les parois secondaires des cellules du xylème. Ainsi, un bois avec une forte teneur en lignine nécessitera un traitement énergétique et chimique onéreux et polluant pour la production de pâte à papier alors qu'il sera adapté à la production d'énergie. La voie de biosynthèse spécifique des lignines est catalysée par deux enzymes, la cinnamoyl CoA reductase (CCR) et la cinnamyl alcool dehydrogenase (CAD). Elles utilisent comme substrat le cinnamoyl CoA produit par la voie des phenylpropanoïdes pour synthétiser les monolignols, qui sont finalement polymérisés en lignine grâce à un groupe de peroxydases et de laccases. Les connaissances approfondies de cette voie de biosynthèse, la disponibilité de séquences pour les gènes de lignification et l'impact reconnu des lignines sur la qualité du bois et son utilisation, en font un exemple de voie métabolique de choix pour mener à bien une analyse de la variabilité fonctionnelle de gène candidats. Un des gènes candidats évident pour ce type d'approche est le gène CCR. En effet, la modulation de l'expression de ce gène chez des plants de tabacs transgéniques entraîne une modification de la quantité et de la qualité des lignines, avec, par exemple, jusqu'à 50% de diminution du taux de lignines et un changement dans la composition en unités synapylique et coniférylique lorsque ce gène est réprimé (Grima-Pettenati and Goffner, 1999 [13]). De plus, ce gène est préférentiellement exprimé dans le xylème d'eucalyptus grâce à la présence de séquences promotrices spécifiques (Lacombe *et al.*, 1997 [17] ; Paux *et al.*, 2004 [25]). Enfin, le gène CCR a été cartographié chez *E. urophylla* (Gion *et al.*, 2000 [10]) et une colocalisation a été mise en évidence entre ce gène et un QTL de teneur en lignines (Gion, 2001 [11]).

Au sein des programmes d'amélioration des arbres, la sélection se fait classiquement par l'utilisation de modèles de génétique quantitative basés sur le phénotype. Ces modèles appartiennent à la classe des modèles à effets aléatoires où les effets aléatoires expriment les effets génétiques. L'utilisation de marqueurs moléculaires tels que les SNP permet d'avoir une information concernant les allèles des géniteurs et de la descendance sur l'ensemble du génome ou sur des zones d'intérêt comme des gènes candidats. Une question centrale est alors de pouvoir estimer les effets génétiques parentaux mais aussi les

associations entre les caractères phénotypiques et les SNP. L'objectif des études d'association est d'identifier les formes aux sites polymorphes qui varient systématiquement avec les individus porteurs de différents états d'une caractéristique phénotypique donnée, quelques soient leurs parents (Balding, 2006 [2]). Dans les études d'associations entre des caractères phénotypiques quantitatifs et un ensemble de SNP, les méthodes de régression sont classiquement utilisées. Il est néanmoins nécessaire de sélectionner les SNP significatifs parmi l'ensemble des variations nucléotidiques observées. Plusieurs méthodes ont été utilisées telles que les méthodes stepwise (Cordell and Clayton, 2002 [7]), les méthodes de bayésienne (shrinkage methods ; Carlson *et al.*, 2004 [3]) ou encore les méthodes SNP tagging (Wang *et al.*, 2005 [37]). Mais lorsque le nombre de SNP sélectionnés reste important il est difficile, voire impossible, de tenir compte des effets d'interactions. Une dernière approche alors employée pour diminuer la dimension du problème mais aussi pour étudier les phénomènes d'épistasie consiste à fonder son analyse sur les haplotypes (haplotype-based methods) (Clark, 2004 [6]; Chen and Rodriguez, 2007 [4]).

Dans le cadre de cette étude, nous proposons d'étudier l'effet de la variabilité du gène CCR d'*Eucalyptus urophylla* sur la teneur en lignine. Le genre *Eucalyptus* est un bon modèle car il représente un des feuillus les plus plantés au monde. Ses deux principales utilisations sont la production de bois d'énergie et de bois de trituration, toutes deux directement reliées à la teneur en lignine. L'ensemble des données actuelles traduit une très grande proximité des génomes entre les quelques espèces utilisées en plantation, pourtant adaptées à des conditions écologiques spécifiques (Gion *et al.*, 2005 [12]). Chez *E. urophylla*, des marqueurs microsatellites révèlent une importante variabilité dont la structuration ne reflète pas celle obtenue sur la base de caractères adaptatifs (Tripiana *et al.*, 2007 [35]). L'étude de la variabilité «utile» du gène CCR d'*E. urophylla* constituera un modèle de référence pour l'analyse de la diversité «fonctionnelle» chez les arbres forestiers. Nous proposons une méthode statistique qui permet simultanément, d'estimer les variances génétiques des géniteurs (additive et de dominance) et de sélectionner les SNP discriminants les allèles qui expliquent significativement une part de la variation phénotypique.

## 2 MATERIEL ET METHODES

### 2.1 Matériel végétal et conditions de croissance

Le matériel végétal est issu d'un plan de croisement factoriel incomplet de huit mères et huit pères *E. urophylla* non apparentés (graines d'origine de l'ile de Florès, Indonésie) ayant généré par croisement contrôlé 348 individus répartis en 35 familles de pleins frères de 9 ou 10 arbres chacune. Le test de descendances a été mis en place en janvier 1992 dans la station forestière de Kissoko, Pointe-Noire, République du Congo (4°45'S, 12°00'E, 50m d'altitude). La pluviométrie annuelle est de 1200mm avec 4 mois de saison sèche. La température annuelle moyenne est de 24°C. Les sols sont chimiquement pauvres (sable > 85%, CEC<0.5cmolc $kg^{-1}$), développés sur un matériau détritique épais (Jamet, 1975 [16]). 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 4x4 arbres plantés à 667 tiges / hectare. Les arbres ont été abattus à 14 ans, ayant atteint 22.5m de hauteur moyenne pour une circonférence à 1.3m de 53cm.
L'ADN génomique des 16 parents et des 348 descendants a été extrait à partir de jeunes feuilles séchées, suivant la méthode de Doyle et Doyle (1990, [9]). Les échantillons sont stockés à −20°C.

### 2.2 Caractères de croissance et analyse des lignines

La hauteur (Ht) et la circonférence (C) individuelles ont été mesurées lors de l'abattage. Un disque de 2cm d'épaisseur a été prélevé à 1.3m de hauteur sur chacun des 348 arbres. Un quartier de chaque disque a été découpé, puis broyé au broyeur à couteaux (poudre 4mm) et un aliquote de la poudre obtenue a été broyé avec un broyeur à fléau (poudre 0.5mm).

**2.2.1. Analyses NIRS :** Pour chaque échantillon, plusieurs spectres d'absorption dans le proche infrarouge ont été enregistrés sur les deux types de poudres. Les poudres sont conditionnées dans une salle régulée en température (20°C) et en humidité (humidité relative de 65%). L'humidité d'équilibre du bois est de l'ordre de 12%. Le matériel utilisé pour les données spectrales d'absorption est un spectrophotomètre proche infrarouge à transformée de Fourier équipé d'une sphère d'intégration (BRUKER, modèle Vector 22N-I). La résolution retenue du spectromètre est de 8cm$^{-1}$ et chaque spectre est la moyenne de 32 scans. La gamme spectrale est de 12,800-3,500cm$^{-1}$ (780-2,860nm). L'acquisition des spectres a été réalisée à partir d'une coupelle dont le fond en quartz laisse passer l'énergie lumineuse. Sur la base des données spectrales des 348 poudres et des prédictions de lignines obtenues au moyen d'équations de calibration existantes, 60 échantillons ont été sélectionnés pour être suffisamment représentatifs en termes de variabilité spectrale et en termes de gamme en lignine.

**2.2.2 Analyses chimiques de référence :** Pour les 60 échantillons sélectionnés, 5 grammes de poudre (0.5 mm) ont été soumis à extraction exhaustive dans un appareil de Soxhlet par les solvants toluène/éthanol (2/1, v/v), éthanol, puis eau. Le résidu pariétal (RP) ainsi obtenu, constitué essentiellement des parois, a été séché et pesé. Le dosage de la lignine acido-insoluble (dénommée Lignine Klason, LK) et de la lignine acido-soluble (LAS) a été réalisé à partir de 300 mg de RP et selon la méthode décrite dans la littérature (Dence, 1992 [8]). La structure des lignines a été étudiée par thioacidolyse, à partir de 20 mg de RP, et selon la méthode décrite antérieurement (Lapierre *et al.*, 1995 [19]). Les produits monomères de type guaïacyles (G) et syringyles (S) ont été analysés sous forme de leurs dérivés triméthylsilylés et par chromatographie en phase gazeuse couplée à la spectrométrie de masse (CPG-SM) (Lapierre *et al.*, 1995 [19]). Toutes ces analyses ont été réalisées en double.

## 2.3 Variabilité du CCR d'*E. urophylla*

**2.3.1 Amplification et séquençage :** Le gène CCR code la cinnamoyl CoA réductase. Cette enzyme catalyse la réduction de Cinnamoyl CoA en Aldéhyde Cinnamylique. La séquence du clone génomique caractérisé chez *E. gunnii* (Lacombe *et al.*, 1997 [17]; Acc : X97433) a été utilisé comme séquence de référence pour la définition d'amorces et l'analyse de séquence. Sept couples d'amorces spécifiques ont été définis grâce au logiciel OLIGO 4. Un couple d'amorces supplémentaires a été utilisé pour amplifier la partie promotrice la plus en aval du gène, située entre la position -628 pb en amont et +66 pb en aval du site +1 de l'initiation de la transcription.
L'amplification des 8 fragments du gène CCR a été réalisée sur un thermocycleur Applied Biosystems Gene Amp PCR system 2700 en utilisant 2µL de tampon 10X (Invitrogen), 0.8µL dNTPs (solution stock à 5mM), 0.8µL $MgCl_2$ (solution stock à 50mM), 0.4µL de chaque amorce (solutions stock à 10µM), 20ng d'ADN génomique (solutions stock à 20ng.µL$^{-1}$), 0.8 unités de Taq DNA polymérase native (Invitrogen) complété par une quantité suffisante d'eau ultrapure pour un volume final de 20µL. Les conditions de PCR impliquent une dénaturation initiale à 94°C pendant 5 minutes (min) puis 35 cycles de dénaturation à 94°C pendant 30 secondes (sec), hybridation au Tm des amorces pendant 1 min, élongation à 72°C pendant 1 min et enfin une étape d'élongation finale à 72°C pendant 10 min. La position des amorces sur la séquence référence et le Tm utilisé pour chaque couple sont donnés dans le tableau I.

Les clonages des produits de PCR ont été réalisés sans purification préalable en utilisant le kit de clonage TOPO-TA Cloning kit for sequencing (Invitrogen). Les transformations ont été réalisées en utilisant les cellules One Shot TOP 10 Chemically competent *E. coli* (Invitrogen). Pour chaque produit de clonage 12 colonies ont été prélevées et ont servi de matrice pour l'amplification PCR des fragments clonés (PCR sur colonie). Les amplifications ont été réalisées sur un thermocycleur Applied Biosystems Gene Amp PCR system 2700 en utilisant 2µL de tampon 10X (Invitrogen), 0.8µL dNTPs (solution stock à 5mM), 0.8µL $MgCl_2$ (solution stock à 50mM), 2µL de chaque amorce F13 et R13 (invotrogen) (solution stock à 2µM), 0.3 unités de Taq DNA polymérase native (Invitrogen) complété par une quantité suffisante d'eau ultrapure pour un volume final de 20µL. Les produits de l'amplification PCR sur colonie ont été

concervés pour séquençage. Les conditions de PCR sont une dénaturation de 10 min à 94°C, 40 cycles avec une dénaturation de 45 sec à 94°C, une hybridation à 55°C pendant 45 sec et une élongation à 72°C pendant 1 min et enfin une étape d'élongation finale à 72°C pendant 10 min.
Les produits d'amplification PCR sur colonie ont été purifiés par filtration sur membrane MultiScreen PCRµ96 Plate (Millipore). Entre 20 et 40ng d'ADN ont été prélevés et utilisés comme matrice pour la réaction de séquence. Le séquençage a été réalisé sur séquenceur ABI 3730 DNA Analyzer en utilisant le kit de séquençage Big Dye Terminator V1.1 cycle sequencing kit (Applied Biosystelms). Le séquençage a été effectué dans un seul sens pour les fragments 1, 2, 3, 4, 5 et 7 du gène CCR et le fragment 6 à été séquencé dans les deux sens. Les séquences ont été analysées en utilisant le logiciel CodonCode Aligner puis alignées dans l'éditeur de séquences BioEdit.

**2.3.2. Génotypage des descendants du factoriel :** Pour l'amplification du motif microsatellite une seconde PCR est réalisée avec 1.5µl de produits d'amplification, 0.3µl de chaque amorce à 2µM (amorce M13 marquée à l'IRD700 ou IRD800 et amorce antisens), 0.08 µl de Taq (5 U), 1.5 µl Tampon 10X « Stand Taq Buffer » Biolabs, 0.6 µl de dNTPs (5mM) et 10.72 µl d'$H_2O$. Les conditions d'amplification sont une dénaturation de 4 min à 94°C, suivie de 34 cycles avec dénaturation de 30 sec à 94°C, hybridation de 1 min à 53°C (Tm spécifique de l'IRD700 et 800), une élongation de 1 min à 72°C et enfin une élongation finale de 10 min à 72°C. Pour la révélation sur LI-COR4300 (NE-USA) 1µl de produits PCR est ajouté à 10µl de bleu (78ml formamide, 10ml de xylène cyanol 1%, 10ml de bleu de bromophénol 1%, 2ml d'EDTA [0.5 M] pH8) et 9µl d'H2O. Les échantillons sont dénaturés 3 min à 94°C puis aussitôt refroidis dans la glace. Ils sont ensuite déposés sur un gel d'acrylamide non dénaturant (960µl de TBE 10X , 10g d'Urée, 5g d'Acrylamide 19/1 et qsp 26g avec de l'eau milliQ, 16µl de Temed et 160µl d'APS à 10%). La migration est réalisée à 1500V, 40mA, 35W et 35°C pendant 6 heures. La révélation de la fluorescence se fait par excitation laser à une longueur d'onde de 700 ou 800 nm.

## 2.4. Analyses statistiques

**2.4.1. Estimation des paramètres génétiques :** L'analyse des données est faite pour chaque caractère (procedure Proc Mixed de SAS) selon le modèle $X_{ijk} = \mu_x + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$ (1) où $\mu_x$ est la moyenne de la population, $\alpha_i$ effet aléatoire de la mère i $N(0,\sigma^2m)$, $\beta_j$ effet aléatoire du père j $N(0,\sigma^2p)$, $\delta_{ij}$ effet aléatoire d'interaction mère–père $N(0,\sigma^2mp)$, $\varepsilon_{ijk}$ erreur $N(0,\sigma^2r)$. Les variances associées aux différents effets sont déduites de l'espérance des carrés moyens (méthode REML). Les variances génétiques sont calculées comme suit : la variance additive mère $\sigma^2A_m=4\sigma^2m$ (2), la variance additive père $\sigma^2A_p=4\sigma^2p$ (3). Ces deux variances sont deux estimations de la même variance additive de la population parentale, les mères et pères étant issus d'une population unique. Les différences d'estimations sont attribuées aux effets d'échantillonnage. La variance génétique additive dans la population de descendants est ainsi $\sigma^2A=2(\sigma^2m+\sigma^2p)$ (4), la variance de dominance $\sigma^2D=4\sigma^2mp$ (5), la variance génétique totale en absence d'épistasie $\sigma^2G=\sigma^2A+\sigma^2D$ (6) et la variance phénotypique totale dans la population de descendants $\sigma^2P=\sigma^2m+\sigma^2p+\sigma^2mp+\sigma^2r$ (7). Ces variances permettent de calculer l'héritabilité au sens strict: $h^2=\sigma^2A/\sigma^2P$ (8) et l'héritabilité au sens large de la valeur familiale $H^2=\sigma^2G/\sigma^2P$ (9). L'analyse multiple de variance (Proc GLM MANOVA SAS) permet de calculer les variances associées aux différents effets pour chacun des caractères à expliquer ainsi que les covariances $COVm_{xy}$, $COVp_{xy}$, $COVmp_{xy}$, $COVr_{xy}$ associées aux différents couples de variables X et Y. Ces covariances sont décomposées en Covariance additive, Covariance de dominance et Covariance résiduelle. Les variances et covariances sont utilisées pour le calcul des corrélations entre caractères. La corrélation phénotypique entre caractères $\rho_P(x,y)$ peut avoir diverses origines, génétique quand les ensembles de gènes contrôlant les deux caractères sont identiques ou partiellement identiques (quelques gènes communs et des gènes spécifiques) et environnementale quand les variations de l'environnement affectent conjointement les deux caractères. Les corrélations génétiques additives sont estimées par $\rho_A(x,y)=CovA(x,y)/\sqrt{(\sigma^2A_x\ \sigma^2A_y)}$ (10), les corrélations génétiques de dominance par $\rho_D(x,y)=CovD(x,y)/\sqrt{(\sigma^2D_x\ \sigma^2D_y)}$ (11), et enfin la corrélation d'origine purement environnementale est déduite de la relation algébrique entre les trois types, génétique $\rho_A$, phénotypique $\rho_P$

et environnementale $\rho_E$ et sous l'hypothèse d'une corrélation de dominance nulle, est donnée par $\rho_E=(\rho_P - h_x h_y \rho_A)/\sqrt{(1-h^2_x)(1-h^2_y)}$ (12) où $h^2_x$ et $h^2_y$ sont les héritabilités au sens strict des caractères x et y (8).

**2.4.2. Sélection des SNP et test d'associations :** La méthodologie développée (Mortier *et al.*, 2008 [21]) peut s'expliquer de la façon suivante : si K est le nombre total de SNP observés au sein du gène CCR, $2^K$ formes alléliques sont théoriquement observables. L'objectif est de déterminer un nombre $k \leq K$ de SNP ainsi que leur effet sur la variation du taux de lignine. On notera ($\alpha_1$, …, $\alpha_k$) ces effets théoriques. En pratique, on observe un nombre p d'allèles dans la population d'études, associés aux K SNP ($p<2^K$). Par exemple, si un seul SNP est inclus dans le modèle (on dira qu'il est « allumé ») alors seul p=2 allèles sont observables. Dans le cas où 3 SNP sont allumés, $2^3=8$ allèles sont alors au plus observables. Le modèle peut donc être formulé, dans un cadre bayésien, de la façon suivante : soit k un ensemble de SNP allumés, soit y le taux de lignine observé sur n individus issus d'un plan de croisement de sorte que

$$y|m, p, \underline{\alpha}, k, \sigma^2 \sim N(\alpha_1 + \alpha_2 + m + p, \sigma^2)$$

où m et p désignent les effets mères et pères. Nous considérons les effets de dominance négligeables.
$\alpha_i$, i=1,2 sont les effets des 2 allèles (individus diploïdes) dépendant des SNP inclus dans le modèle.
Dans le cadre bayésien, pour définir le modèle il est nécessaire de stipuler les lois *a priori*. Ainsi, on suppose que les effets mères et pères sont distribués selon une loi gaussienne d'espérance nulle et de variance $\sigma^2_m$ et $\sigma^2_p$ respectivement. On suppose que connaissant les SNP "allumés" les effets des allèles suivent une loi gaussienne centrée de variance $\sigma^2_\alpha$. La distribution du nombre de SNP est uniforme sur [0,K]. Finalement, nous supposons que $\sigma^2_m$, $\sigma^2_p$, $\sigma^2_\alpha$ et $\sigma^2$ sont indépendantes et suivent une loi inverse gamma $IG(10^{-3}, 10^{-3})$. Les lois *a priori* ont donc été choisies comme non informatives. L'inférence bayésienne repose sur l'étude de la loi *a posteriori* des paramètres sachant les observations. Comme le nombre k de SNP n'est pas connu, on est conduit à changer l'espace à parcourir. On propose donc un algorithme Reversible-Jump MCMC définit de la manière suivante (Richardson and Green, 1997 [31]):

1/ Initialiser les paramètres $\theta$ et une configuration $\gamma$ de SNP « allumés ». $\gamma$ est un vecteur qui contient des 1 et des 0 (1 pour les SNP « allumés » 0 sinon).

2/ Pour i=1 à N itérations

3/ Choisir si on augmente ou on diminue le nombre de SNP « allumés ».

4/ Choisir un SNP parmi les SNP allumés ($\gamma$==1), on le transforme en $0 \rightarrow \gamma^*$.

5/ Déterminer les nouveaux allèles efficaces, proposer de nouvelles valeurs des paramètres selon la loi *a posteriori* sachant les SNP choisis. $\rightarrow \theta^*$

6/ Calculer, le ratio de Métropolis-Hastings en tenant compte du changement de dimension.

7/ On accepte ou on rejette $\theta^*$ et $\gamma^*$ proportionnellement au ratio calculé.

# 3 RESULTATS

## 3.1. Analyse des caractères quantitatifs

Le tableau 1 présente la moyenne, le coefficient de variation phénotypique ainsi que les différents paramètres génétiques associés aux variables analysées. Les différents effets du modèle (1) sont toujours très hautement significatifs (significatif pour l'effet père sur C). La variabilité observée est globalement bien structurée par le modèle : la réalisation d'une telle structure sous l'hypothèse $H_o$ (les variables explicatives introduites, mère, père et interaction, n'ont pas d'effet) a une probabilité généralement inférieure à 0.0001.
L'origine parentale a un effet majeur sur les performances des descendants, ceci quel que soit le caractère considéré. Le contrôle de la variabilité phénotypique observée entre familles est essentiellement d'origine génétique pour le taux de lignine et le rapport S/G ($H^2$= 0.85 à 0.65). Les effets environnementaux sont plus sensibles pour les caractères de croissance ($H^2$= 0.41 et 0.21). La variance de dominance représente la moitié de la variance génétique totale pour la circonférence. Pour les autres caractères, le contrôle

génétique est essentiellement additif. Il s'en suit que les héritabilités au sens strict, $h^2 = \sigma^2A/\sigma^2P$, pour les caractères chimiques sont très élevées (0.65 et 0.76), confirmant ainsi le fait que les caractères « bois » sont plus héritables que la croissance.
L'analyse multiple de variance est utilisée pour estimer les covariances entre caractères pris deux à deux et en dériver les corrélations génétiques et environnementales (formules 10, 11, 12). Les résultats sont présentés dans le tableau 2. Les corrélations génétiques additives sont toutes significativement différentes de zéro sauf entre C et LK. Les corrélations environnementales montrent qu'une meilleure disponibilité des ressources a un effet positif sur la croissance en hauteur et en circonférence. Elle favorise de la même façon une augmentation du taux de lignine et est plutôt favorable à la synthèse du monomère G. L'ensemble de ces résultats démontre à l'évidence l'importance de la distinction entre corrélations phénotypiques, génétiques et environnementales, particulièrement claire dans le cas du couple croissance/taux de lignines.

## 3.2. Variabilité du gène CCR chez *E. urophylla*

**3.2.1. Séquençage du gène CCR des géniteurs du factoriel :** Les 7 fragments ciblés du gène ont été générés pour 15 des 16 génotypes *E. urophylla* étudiés. Le gène n'a pas pu être amplifié chez un géniteur du fait d'une mauvaise qualité de l'ADN génomique. Au total, la description de la variabilité de CCR est possible sur 94% du gène avec en moyenne 116 séquences par fragment (120 attendues avec 8 clones par parent). La totalité des introns 1, 2, 3 et 4 sont disponibles. Pour les régions codantes, on dispose d'information de séquence pour 100% des exons 2, 3, 4, 48% et 88% respectivement pour les exons 1 et 5. Ces régions correspondent aux extrémités 5'UTR et 3'UTR non traduite. Des polymorphismes de type SNP, INDEL, poly A et microsatellite ont pu être mis en évidence au sein du gène. Le tableau 3 donne la répartition des sites polymorphes au sein des régions codantes et non codantes. Le type et le nombre de sites variables sont différents entre les introns et les exons. Les SNP sont en moyenne plus fréquents dans les introns (1/23) que dans les exons (1/30). Dans les exons, les sites de mutations sont de plus en plus fréquents au sein de l'exon 4 (1/22). Les fréquences des SNP sont de 1/78 pour l'exon 1 (sur 48% de l'exon 1), 1/31 pour l'exon 2, 1/37 pour l'exon 3 et 1/28 pour l'exon 5. L'ensemble des INDEL est mis en évidence dans les zones introniques. Enfin, le ratio entre le nombre de transitions et le nombre de transversions est plus élevé dans les exons (2.7) que dans les introns (1.6). Ceci montre un biais pour le type de mutation entre régions codantes et non codantes. Sur les 131 sites SNP étudiés, on met en évidence 10 mutations non synonymes dont 5 sont portées par le dernier exon.

**3.2.2. Haplotypes parentaux :** Les haplotypes ont été reconstruits pour 1161pb, soit 36% du gène CCR, en considérant pour chacun des 15 géniteurs séquencés les fragments correspondant à l'exon 4 et l'intron 4. La sélection de cette région du gène a été motivée par son niveau de variabilité plus important (1 SNP pour 21pb en moyenne) et par la présence d'un microsatellite permettant de génotyper les descendants. Un total de 15 haplotypes a été mis en évidence à partir des 30 allèles portés par les 15 parents séquencés. L'un d'entre eux n'est reconstruit que partiellement avec plus de 50% de l'information de séquence disponible (géniteur 14-132).
L'arbre des distances entre les 14 haplotypes complets d'*E. urophylla* et les régions correspondantes des clones génomiques CCR de trois autres espèces du même sous genre (*Symphyomyrtus*), *E. gunnii* (X97433), *E. globulus* (AY656821) et *E. saligna* (AF297877) a été obtenu en utilisant la méthode des plus proches voisins (Neighbor-Joining ; Saitou and Nei, 1997 [32]). Il met en évidence deux groupes d'haplotypes correspondant aux deux sections *Maidenaria* (*E. globulus* et *E. gunnii*) d'une part et *Latoangulatae* (*E. saligna* et *E. urophylla*) d'autre part.

**3.2.3. Génotypage des descendants du factoriel :** Le génotypage du microsatellite a été réalisé sur 208 individus issus de 21 familles (2 à 3 familles pour chacune des huit mères du dispositif). Le tableau 4 indique la ségrégation des 15 haplotypes séquencés ainsi que celle des deux haplotypes du géniteur 14-147 ($H_a$ et $H_b$ identifiés grâce au locus microsatellite). On dispose en moyenne de 9.5 individus par famille.

L'homozygotie ou l'hétérozygotie supposées des géniteurs (données du séquençage) est confirmée par les ségrégations observées de type ½ ½ ou ¼ ¼ ¼ ¼. La figure 3 illustre la ségrégation des haplotypes H9, H4, H7, H3 et H11 dans deux familles issues de la mère 14-138.

**3.2.4. Variabilité de la partie promotrice chez le géniteur 14-144 :** Un fragment de 694 pb correspondant à la partie promotrice la plus en aval du gène (-628 pb/+66 pb par rapport au site +1 de l'initiation de la transcription) a été généré à partir du génotype *E. urophylla* 14-144. Ce fragment contient toutes les séquences connues pour intervenir dans la régulation de l'expression du gène CCR chez *Eucalyptus* (Lacombe *et al*., 1997 [17]). Les deux formes allèliques présentes différent par 5 SNPs, dont 4 portent le même nucléotide que la séquence d'*E. gunnii.* Ils diffèrent de cette dernière par 20 SNPS et 1 INDEL de 14pb (Figure 3).

## 3.3 Variabilité fonctionnelle du gène CCR

Les tests d'associations ont été réalisés uniquement pour le caractère LK et les haplotypes reconstruits à partir des fragments 5 et 6. Le programme ne prenant pas encore en charge les données manquantes, les haplotypes présentant des données manquantes ne sont pas considérés dans l'analyse. Au total, 68 SNP sont observés soit un nombre potentiel d'allèles de $2^{68} \approx 3\ 10^{20}$. En pratique on observe uniquement 11 haplotypes sur les 15 mis en évidence en 3.2.2.
Dans le cadre de nos algorithmes d'estimation type MCMC, nous avons lancé 20 chaînes avec des points de départs différents sur 100000 itérations avec un burning de 10000. Systématiquement, 3 SNP sont sélectionnés (SNP N°3, 10 et 50) avec des probabilités *a posteriori* pour l'une des chaînes égales respectivement à 0.68, 0.90 et 0.61. Les résultats pour les autres chaînes sont du même ordre. Ces SNP correspondent à deux SNP présents sur l'exon 4 et un sur l'intron 4. Le site SNP N°3 sélectionné correspond à un des deux SNP NS de l'exon 4, dont la variabilité T/G observée implique un changement d'acide aminé amidé (Asparagine) en acide aminé dibasique (Lysine).
Quatre groupes d'allèles sont reconstruits à partir des 3 SNP sélectionnés. La variance des effets alléliques est estimée à 0.39 avec une variance d'estimation de 0.082. Les variances des effets mères et pères sont égales à 0.236 et 0.274 avec des variances d'estimation égales à 0.025 et 0.055 respectivement. Par analogie à l'expression de l'héritabilité au sens strict utilisée dans les modèles classiques de génétique quantitative (cf section 2.4.1), nous avons calculé l'héritabilité du gène CCR (identifié aux fragments 5 et 6), l'héritabilité parentale et globale. Ces héritabilités sont respectivement estimées à 0.37, 0.48 et 0.85. L'héritabilité globale est la même que celle obtenue dans le cas classique, ce qui tend à valider les résultats.
On peut confirmer ce résultat à l'aide d'une analyse de variance qui montre des effets significatifs pour les groupes d'allèles (données non montrées). Lors des décompositions de type I, l'ordre d'introduction des facteurs dans le modèle est important dans le cas déséquilibré (Scheffé, 1959 [33]). Quelque soit l'ordre d'introduction, l'effet « groupe d'allèles » reste significatif, même si la significativité des effets est plus faible dans le deuxième cas. Néanmoins, ces résultats doivent être pris avec précaution car le plan d'expérience est fortement déséquilibré avec 132 individus pour le groupe 1, 23 pour le groupe 2, 6 le groupe 6 et enfin 5 pour le dernier groupe. Ce dernier groupe est constitué d'une seule forme haplotypique (H8) portée par un seul géniteur.

# 4 DISCUSSION

Différents auteurs ont mis en évidence l'importance du contrôle génétique de la qualité du bois d'eucalyptus (Poke *et al.*, 2006 [28] ; Vigneron *et al.*, 2003 [36]). Ces caractères présentent généralement une héritabilité supérieure à celle des caractères de croissance, comme cela a pu être montré chez quelques espèces résineuses (Pot *et al.*, 2002 [29] ; Hannrup *et al.*, 2004 [15]). Les résultats obtenus dans cette étude vont dans le même sens. L'analyse des caractères quantitatifs indique l'importance des effets pléiotropiques des gènes (un gène peut avoir un effet sur deux caractères différents). Le taux de lignine est

négativement corrélé aux autres caractères. Pour un niveau de ressources environnementales donné (eau, éléments minéraux, lumière…), les individus qui poussent rapidement semblent favoriser la synthèse de cellulose ou d'hémicellulose au détriment de celle de la lignine. Dans le même temps, ces individus favorisent la synthèse de S au détriment du G. Le matériel végétal choisi se prête particulièrement bien à l'objectif général de l'étude, notamment en ce qui concerne ses caractéristiques phénotypiques. En effet, l'observation d'une forte variabilité des caractères phénotypiques est un pré requis à l'analyse de leur association à la variabilité moléculaire des gènes supposés en être à l'origine. L'analyse montre l'importance du contrôle génétique de la variabilité des différents caractères étudiés. L'héritabilité au sens large du taux de lignine, caractère majeur de cette étude, est supérieure à 0.8. La variance génétique additive représente une large part du contrôle génétique, conduisant à une très forte héritabilité au sens strict (0.76 pour le taux de lignine). Les gains génétiques directs attendus sont donc élevés, pouvant atteindre 2.5 points pour le taux de lignine, ceci avec un taux de sélection de 1%. Le dispositif utilisé permet de faire la part des effets génétiques et environnementaux dans la corrélation entre caractères. L'absence de corrélation phénotypique entre caractères de croissance et taux de lignine résulte des effets antagonistes des corrélations génétiques et environnementales. La réduction du taux de lignine induit par la sélection sur la croissance peut être corrigée, au moins en partie, par une amélioration des pratiques culturales.

L'amplification de 94% du gène CCR chez *E. urophylla* à partir de la séquence d'un clone génomique d'*E. gunnii* traduit un niveau de conservation important du gène chez ces deux espèces d'*Eucalyptus* appartenant au sous genre *Symphyomyrtus*. Les travaux réalisés sur ce gène chez *E. globulus* (Poke *et al.*, 2003 [27]) vont aussi dans ce sens. S'il existe une conservation relative entre espèces pour ce gène, la variabilité du gène CCR mise en évidence chez *E. urophylla* apparaît plus élevée que celle décrite antérieurement chez *E. globulus* (Poke *et al.*, 2003 [27]). En effet, dans les introns on observe en moyenne un SNP pour 23 pb chez *E. urophylla* et 1/33 pb chez *E. globulus*. Cette tendance se retrouve aussi au niveau des exons avec un SNP pour 30 pb chez *E. urophylla* et 1/49 pb chez *E. globulus*. Trois hypothèses peuvent expliquer ces différences : i/ le pool de variabilité présent au sein des échantillons utilisés pour décrire la variabilité du gène est différent, ii/ la stratégie de séquençage, produits clonés dans notre étude *vs* produits PCR chez *E. globulus*, a un effet sur la variabilité mise en évidence, iii/ enfin, un niveau de variabilité plus important de CCR chez *E. urophylla* par rapport à *E. globulus*. Lorsqu'on s'intéresse aux fréquences observées chez des espèces plus éloignées et sur des gènes différents on constate aussi des variations sensibles. Pour 18 gènes, le maïs présente en moyenne 1 SNP pour 124 pb au sein des zones codantes et 1 SNP pour 31 pb au sein des zones non codantes (Ching *et al.*, 2002 [5]). Les auteurs ont travaillé sur un ensemble de variétés représentant la diversité génétique des variétés améliorées aux Etats-Unis. On constate que pour une espèce ayant subit une forte pression de sélection le niveau de variabilité des introns, ne codant pas la protéine, reste proche de celui observé chez *Eucalyptus*, espèce encore peu sélectionnée. Pour les régions codantes, le niveau de variabilité au sein du gène CCR d'*E. urophylla* et d'*E. globulus* semble bien supérieur à celui observé pour d'autres gènes chez des espèces annuelles fortement sélectionnées ou chez d'autres espèces forestières (Pot *et al.*, 2005 [30]). L'impact de la variabilité nucléotidique sur la structure protéique de la CCR est plus fort chez *E. globulus* que chez *E. urophylla* malgré un niveau de variabilité global (exon et intron) moins important. Lacombe *et al.* (1997 [17]) ont décrit chez *E. gunnii* plusieurs sites structuraux différents pouvant avoir un rôle fonctionnel chez la protéine CCR. Un site de huit acides aminés (KNWYCYGK) est proposé comme le site catalytique de la protéine (exon 4). Ce site catalytique, fortement conservé chez de nombreuses espèces végétales (Pichon *et al.*, 1998 [26]), n'est pas variable dans notre échantillon *E. urophylla*. Les SNPs non synonymes mis en évidence peuvent cependant avoir des effets non négligeables sur la structure tertiaire de la protéine et affecter ainsi ses propriétés physicochimiques.

Les résultats préliminaires obtenus sur le promoteur de CCR pour un des géniteurs *E. urophylla* semblent indiquer qu'aucun des sites polymorphes n'affecte les sites consensus de type MYB présents sur le promoteur *CCR* (Lacombe *et al.*, 2000 [18]) et conduisant à une expression spécifique de ce gène dans le

xylème. Malgré tout, on peut envisager que des variations de séquences dans le promoteur puissent influencer l'expression du gène. En effet, d'une part les complexes protéiques régulateurs de la transcription associent plusieurs protéines qui peuvent se fixer en plusieurs endroits sur la partie promotrice. D'autre part, leur fixation peut être potentiellement influencée par l'environnement nucléotidique. Il a donc été décidé de poursuivre ce travail *in planta* en exprimant un gène rapporteur GUS sous le contrôle de ces différentes formes alléliques. Un comparaison des résultats observés entre les deux allèles du géniteur *E. urophylla* 14-144 et les résultats décrits pour le promoteur de *E. gunnii* (Baghdady *et al*., 2006 [1]) permettront de voir si l'expression du gène est modulée entre les génotypes.

L'analyse de la variabilité fonctionnelle du gène CCR réalisée sur 36% de la séquence comprenant l'exon 4 et l'intron 4, zones les plus variable dans notre échantillon et codant le site catalytique présumé, met en évidence que le polymorphisme entre 11 haplotypes testés explique environ 40% de la variance génétique du taux de lignine dans notre échantillon. Sur quatre groupes d'haplotypes différenciés par 3 SNP, un des groupes composé de l'allèle H8 présente des teneurs en lignines significativement supérieures aux trois autres groupes d'allèles. Un effet de la variabilité du gène CCR sur des propriétés du bois a déjà été mis en évidence chez *E. nitens*. Thumma *et al.* (2005 [34]) ont mis en évidence un effet significatif de deux haplotypes sur l'angle des micro-fibrilles. Le résultat du test d'association présenté dans cette étude demande à être confirmé en considérant à la fois le nombre total d'haplotypes, la séquence entière du gène, et l'ensemble des descendances du plan de croisement factoriel *E. urophylla* x *E. urophylla*.
D'un point de vue statistique, il reste à étudier l'influence des lois *a priori* pour les effets « groupes alléliques » : i/ L'équivalent bayésien d'une analyse de variance avec effet fixe pour les effets groupe d'allèles consiste à choisir une loi normale de variance fixée et grande comme prior pour ces même effets. Dans ce cas on s'intéresse à l'effet spécifique de chaque groupe d'allèles sur le caractère lignine. ii/ Le choix d'une loi normale avec une variance inconnue que l'on cherche à estimer et sur laquelle on pose un prior Inverse-Gamma est la version bayésienne d'un modèle à effets aléatoires pour les groupes d'allèles. Cette situation propose de rendre compte de l'effet de la variabilité (qui se résume aux groupes d'allèles considérés) du gène CCR sur le caractère lignine. Enfin, il est nécessaire pour valider la méthode de la comparer avec les méthodes d'association classiquement utilisées (Balding 2006 [2]).
Par la suite, on souhaite développer une extension de ce modèle qui prendrait en compte les données manquantes et une dépendance entre les effets des groupes d'allèles. On se demande aussi si le partage d'une même histoire évolutive conduit à des effets similaires. L'utilisation d'une similarité calculée à partir d'une distance de Nei (1972 [23]) ou d'une distance fondée sur le graphe des coalescents (Nordborg, 2001 [24]) pourrait être envisagée pour valider cette hypothèse.

Ces résultats préliminaires sur l'effet de la variabilité du gène CCR sur la teneur en lignines chez *E. urophylla* sont très prometteurs pour envisager une sélection assistée par marqueurs pour la teneur en lignines au sein de programme d'amélioration des eucalyptus au Congo.

REMERCIEMENTS

Ce travail a été financé en partie par le Bureau des Ressources Génétiques et par une bourse CIFRE entre l'ANRT, le groupe Vallourec et le CIRAD (N° 213/2007)

REFERENCES

[1] Baghdady A., Blervacq A.-S., Jouanin L., Grima-Pettenati J., Sivadon P., Hawkins S., 2006. - *Eucalyptus gunnii* CCR and CAD2 promoter activities are coordinated in lignifying cells during primary and secondary xylem formation in *Arabidopsis thaliana*. Plant Physiology and Biochemistry, 44: 674-683.

[2] Balding D.J., 2006. - A tutorial on statistical methods for population association studies. Nature Reviews; Genetics, 7: 781-791.

[3] Carlson C.S., Eberle M.A., Rieder M.J., Yi O., Kruglyak L., Nickerson D.A., 2004. - Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet., 74: 106-120.

[4] Chen J. and Rodriguez C., 2007. - Conditional likelihood methods for haplotype-based association analysis using matched case-control data. Biometrics, 63: 1099-1107.

[5] Ching A., Caldwell K.S., Juing 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: 19.

[6] Clark A.G., 2004. - The role of haplotypes in candidate gene studies. Genetic Epidemiology, 27: 321-333.

[7] Cordell H.J. and Clayton D.G., 2002. - A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to hla in type 1 diabetes. Am. J. Hum. Genet., 70: 124-141.

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

[9] Doyle J.J., and Doyle J.L., 1990. - Isolation of DNA from fresh plant tissue. Focus, 12: 13-15.

[10] Gion J.M., Rech P., Grima Pettenati J., Verhaegen D., Plomion C., 2000. - Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Mol. Breed. 6: 441-449.

[11] Gion J.M., 2001. - Etude de l'architecture génétique des caractères complexes chez l'eucalyptus: des marqueurs anonymes aux gènes candidats. Thèse de doctorat, Université de Rennes I

[12] Gion J.M., Cabannes E., Poitel M., Vigneron Ph. 2005. - Conservation of xylem preferentially Expressed Sequence Tags in several subgenera of *Eucalyptus* L'Hérit. (Abstract) Plant and Animal Genome conference XIII, San-Diego, January 2005

[13] Grima-Pettenati J., Goffner D., 1999. - Lignin genetic engineering revisited. Plant Science, 145: 51–65.

[14] Gupta P.K., Rustgi S., 2004. - Molecular markers from the transcribed/expressed region of the genome in higher plants. Funct. Integr. Genomics, 4: 139-162.

[15] Hannrup B., Cahalan C., Chantre G., Grabner M., Karlsson B., Bayon I. le, Jones G. L., Müller U., Pereira H., Rodrigues J.C., Rosner, S., Rozenberg P., Wilhelmsson L., Wimmer R., 2004. - Genetic parameters of growth and wood quality traits in *Picea abies*. Scandinavian Journal of Forest Research, Vol. 19 (1): 14-29.

[16] Jamet R., 1975. - Internal report. ORSTOM, Cote MC 189, Centre de Brazzaville, 35 p.

[17] Lacombe E., Hawkins S., Van Doorsselaere J., Piquemal J., Goffner D., Poeydomenge O., Boudet A.M., and Grima-Pettenati J., 1997. - Cinnamoyl CoA Reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships. The Plant Journal 11 : 429-441.

[18] Lacombe E., Van Doorsselaere J., Boerjan W., Boudet A.M., Grima-Pettenati J., 2000. - Characterization of cis-elements required for vascular expression of the Cinnamoyl CoA reductase gene and for protein-DNA complex formation. The Plant Journal, 23: 663-676.

[19] Lapierre C., Pollet B., Rolando C., 1995. - New insights into the molecular architecture of hardwood lignins by chemical degradative methods. Res. Chem. Intermed., 21: 397-412.

*[20] Mandrou E., Vigneron P., Plomion C., Gion J.M., 2008 – Functional variability of the CCR gene in Eucalyptus urophylla. *in prep.*

*[21] Mortier F., Etienne M.P., Gion J.M., 2008. - SNP selection and association study for related individuals : a reversible jump approach.. *in prep.*

[22] Neale D.B., Savolainen O., 2004. - Association studies of complex traits in conifers. Trends in Plant Science 9 (7): 325-330.

[23] Nei M., 1972. - Genetic distance between populations. Amer. Natur.: 106-283

[24] Nordborg M., 2001. - Coalescent Theory. In Handbook of Statistical Genetics, Edited Balding D.J., Bishop M., Cannings C. John Wiley & Sons: 179-212.

[25] Paux E., Tamasloukht M., Ladouce N., Sivadon P., Grima-Pettenati J., 2004. - Identification of genes preferentially expressed during wood formation in *Eucalyptus*. Plant Mol Biol. 55: 263-280

[26] Pichon M., Courbou I., Beckert M., Boudet A.M., and Grima-Pettenati. J., 1998. - Cloning and characterization of two maize cDNAs encoding Cynnamoyl-CoA Reductase (CCR) and differential expression of the corresponding genes. Plant Molecular biology, 38: 671-676.

[27] Poke F.S., Vaillancourt R.E., Elliott R., Reid J.B., 2003. - Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). Mol. Breed., 12: 107-118.

[28] Poke F.S., Potts B.M., Vaillancourt R.E., Raymond C.A., 2006. - Genetic parameters for lignin, extractives and decay in *Eucalyptus globulus*. Ann. for. sci., 63, 8: 813-821.

[29] Pot D., Chantre G., Rozenberg Ph., Rodrigues J.C., Jonese 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.). Ann. For. Sci. 59: 563-575.

[30] Pot D., McMillan L., Echt C., Le Provost G., Garnier-Géré P., Catoand S., Plomion C., 2005. - Nucleotide variation in genes involved in wood formation in two pine species. New Phytol. 167 (1): 101-112.

[31] Richardson S. and Green P.J., 1997. - On bayesian analysis of mixtures with an unknown number of components. (with discussion). J. R. Stat. Soc., Ser. B, 59 (4): 731-792.

[32] Saitou N., & Nei M., 1987. - The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

[33] Scheffé H., 1959. - The Analysis of Variance. John Wiley & Sons

[34] Thumma B.R., Nolan M.F., 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.

[35] Tripiana V., Bourgeois M., Verhaegen D., Vigneron P., and Bouvet J.M., 2007. - Combining Microssatellites, growth, and adaptative traits for managing in situ genetic resources of *Eucalyptus urophylla*. Can. J. For. Res., 37: 773-785.

[36] Vigneron Ph., Giordanengo Th., Ognouabi N., Gion J.M., Chaix G and Baillères H., 2003. - Genetic components of wood quality traits in *Eucalyptus urophylla* x *grandis* full sib families. IUFRO Congress unit 2.08.03 “Eucalyptus in a Changing World”, 11 a 15 de Outubro de 2004, Aveiro, Portugal.

[37] Wang H., Zhang Y.M., Li X., Masinde G.L., Mohan S., Baylink D.J., and Xu S., 2005. - Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics, 170: 465-480.

Tableau I : Caractéristiques des amorces utilisées pour l'amplification des 7 fragments du gène CCR (d'après Mandrou *et al.*, 2008 [20]). La séquence de chaque amorce, leur position sur la séquence de référence *E. gunnii* (start et end) et la température d'hybridation sont données.

| Fragment | | Amorce 5'→3' | Start (pb) | End(pb) | Tm (°C) |
|---|---|---|---|---|---|
| CCR1 | Forward | CACCTCCTGAACCCCTCT | 151 | 168 | 63 |
| | Reverse | CGCACCCCTTGATGGCTTCT | 555 | 528 | |
| CCR2 | Forward | GCGAGGAACCGTCAGGAAC | 302 | 320 | 58 |
| | Reverse | TTTCCTCCCCAATCGTCTG | 920 | 902 | |
| CCR3 | Forward | AAGAATGTGCGATGGCGAACC | 840 | 860 | 70 |
| | Reverse | GTCCCGATCACCGCTGGCT | 1283 | 1265 | |
| CCR4 | Forward | ACGTAAGAAAGAGGGACCG | 1086 | 1104 | 66 |
| | Reverse | ACTTGAGGATGTGGATGATG | 1757 | 1738 | |
| CCR5 | Forward | GCTACGGCAAGGCAGTGG | 1614 | 1631 | 66 |
| | Reverse | AACCGACAACCCACACCTG | 2244 | 2226 | |
| CCR6 | Forward | CTTAGATAGATAGTCCCGC | 2047 | 2065 | 56 |
| | Reverse | CAAAGGGATTCAAGACAGG | 2695 | 2677 | |
| CCR7 | Forward | CGTCATCATCGTTCTCTCT | 2496 | 2514 | 56 |
| | Reverse | TGACAACTTCCATTCCAA | 3194 | 3117 | |

Tableau II : Moyenne (Moy), coefficient de variation (CV), variances (génétique additive mère $\sigma^2$Am, génétique additive père $\sigma^2$Ap, génétique additive totale $\sigma^2$A, dominance $\sigma^2$D, génétique totale $\sigma^2$G, phénotypique $\sigma^2$P) et héritabilités au sens strict, $h^2$, et large de la valeur familiale $H^2$.

| | Moy | CV(%) | $\sigma^2$Am | $\sigma^2$Ap | $\sigma^2$A | $\sigma^2$D | $\sigma^2$G | $\sigma^2$P | $\sigma^2$A/$\sigma^2$G | $h^2$ | $H^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ht (m) | 21.2 | 17.2 | 5.06 | 3.20 | 4.13 | 1.40 | 5.53 | 13.59 | 0.75 | 0.30 | 0.41 |
| C (cm) | 53.0 | 21.2 | 26.20 | 2.18 | 14.19 | 11.90 | 26.10 | 127.11 | 0.54 | 0.11 | 0.21 |
| LK (%) | 28.1 | 4.3 | 1.48 | 0.71 | 1.10 | 0.12 | 1.22 | 1.442 | 0.90 | 0.76 | 0.85 |
| S/G | 2.42 | 12.4 | 0.092 | 0.035 | 0.064 | 0.000 | 0.064 | 0.098 | 1.00 | 0.65 | 0.65 |

Tableau III : Corrélations génétiques additives (gras) et environnementales (italique) entre les 4 variables. La valeur en grisé n'est pas significativement différente de zéro.

| | LK | S/G | Ht | C |
|---|---|---|---|---|
| LK | 1 | **-0.15** | **-0.55** | **-0.03** |
| S/G | *-0.39* | 1 | **0.27** | **0.37** |
| Ht | *0.44* | *-0.17* | 1 | **0.64** |
| C | *0.27* | *-0.22* | *0.82* | 1 |

Tableau IV : Répartition et caractéristiques des variations en fonction des zones codantes ou non codantes du gène CCR. Les données suivantes sont présentées : la taille des fragments en paire de base (pb) chez *E. gunnii* (*E.g.*) et *E. urophylla* (*E.u.*), le nombre de SNP (SNP, Bi-SNP et Tri-SNP), de SNP non synonymes (SNP NS), d'insertions/délétions (INDEL), de motif polyA et microsatellite. La fréquence des SNP par paire de bases (F SNP), la fréquence totale des variations (F Tot), le nombre de transversions (Nb V), le nombre de transitions (Nb S) et le rapport transition sur transversion (S/V) (d'après Mandrou *et al.*, 2008 [20]).

| Nature de la séquence | Taille (pb) | | Caractéristiques des sites polymorphes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *E. g* | *E. u* | SNP | SNP NS | Bi-SNP | tri-SNP | INDEL | Poly A | Microsat | F SNP | F Tot | Nb V | NB S | S/V |
| Exon 1 | 324 | 156 | 1 | 1 | - | - | - | - | - | 1/78 | 1/78 | 1 | 1 | 1 |
| Exon 2 | 156 | 156 | 4 | 1 | - | - | - | - | - | 1/31 | 1/31 | 2 | 3 | 1.5 |
| Exon 3 | 185 | 186 | 4 | 1 | - | - | - | - | - | 1/37 | 1/37 | 2 | 3 | 1.5 |
| Exon 4 | 355 | 355 | 14 | 2 | - | - | - | - | - | 1/22 | 1/22 | 2 | 14 | 7 |
| Exon 5 | 220 | 194 | 1 | 5 | - | 1 | - | - | - | 1/28 | 1/28 | 3 | 6 | 2 |
| Total exons | 1240 | 1047 | 24 | 10 | - | 1 | - | - | - | 1/30 | 1/30 | 10 | 27 | 2.7 |
| Intron 1 | 109 | 117 | 3 | - | - | - | 2 | 1 | - | 1/39 | 1/20 | 2 | 1 | 0.5 |
| Intron 2 | 663 | 748 | 29 | - | 1 | - | 6 | - | - | 1/26 | 1/21 | 12 | 19 | 1.6 |
| Intron 3 | 166 | 166 | 6 | - | - | - | 1 | - | - | 1/28 | 1/24 | 1 | 4 | 4 |
| Intron 4 | 1024 | 1142 | 52 | - | 5 | - | 8 | - | 1 | 1/20 | 1/17 | 23 | 37 | 1.6 |
| Total introns | 1962 | 2173 | 90 | - | 6 | - | 17 | 1 | 1 | 1/23 | 1/19 | 38 | 61 | 1.6 |
| Total gène | 3202 | 3220 | 114 | 10 | 6 | 1 | 17 | 1 | 1 | 1/25 | 1/22 | 48 | 88 | 1.8 |

Tableau V : Haplotypes parentaux mis en évidence à partir des fragments 5 et 6 du gène CCR. La répartition des haplotypes au sein de chacune des familles est déterminée grâce au génotypage microsatellite des 208 descendants étudiés (d'après Mandrou *et al.*, 2008 [20]).

| Mères | Pères | Effectif par famille | Génotype parentaux mère/père | Génotypes des descendants | Effectif pour chaque génotype |
|---|---|---|---|---|---|
| 14-138 | 14-137 | 10 | (H4;H7) / (H6;H11) | (H7;H6) / (H4;H6) / (H7;H11) / (H4;H11) | 2 / 7 / 1 / 0 |
| | 14-130 | 10 | (H4;H7) / (H3;H11) | (H4;H11) / (H7;H11) / (H4;H3) / (H7;H3) | 6 / 3 / 1 / 0 |
| | 14-142 | 10 | (H4;H7) / (H9;H6) | (H4;H9) / (H7;H9) / (H7;H6) / (H4;H6) | 3 / 2 / 2 / 3 |
| 14-144 | 14-135 | 10 | (H3;H6) / (H4;H5) | (H3;H4) / (H6;H5) / (H3;H5) / (H6;H4) | 4 / 2 / 2 / 2 |
| | 14-130 | 9 | (H3;H6) / (H3;H11) | (H3;H11) / (H3;H3) / (H6;H11) / (H6;H3) | 2 / 4 / 2 / 1 |
| | 14-142 | 9 | (H3;H6) / (H9;H6) | (H3;H6) / (H6;H9) / (H6;H6) / (H3;H9) | 2 / 1 / 3 / 3 |
| 14-128 | 14-137 | 8 | (H2;H1) / (H6;H11) | (H2;H6) / (H1;H11) / (H2;H11) / (H1;H6) | 4 / 1 / 3 / 0 |
| | 14-135 | 10 | (H2;H1) / (H4;H5) | (H2;H5) / (H1;H5) / (H1;H4) / (H2;H4) | 4 / 2 / 2 / 2 |
| | 14-142 | 7 | (H2;H1) / (H9;H6) | (H2;H6) / (H1;H6) / (H1;H9) / (H2;H9) | 1 / 3 / 2 / 1 |
| 14-149 | 14-132 | 10 | (H4;H4) / (H12;H3) | (H4;H12) / (H4;H3) | 4 / 6 |
| | 14-148 | 10 | (H4;H4) / (H4;H7) | (H4;H7) / (H4;H4) | 6 / 4 |
| 14-133 | 14-135 | 10 | (H13;H6) / (H4;H5) | (H6;H5) / (H13;H4) / (H6;H4) / (H13;H5) | 4 / 1 / 1 / 4 |
| | 14-130 | 10 | (H13;H6) / (H3;H11) | (H13;H11) / (H13;H3) / (H6;H11) / (H6;H3) | 2 / 4 / 2 / 2 |
| 14-136 | 14-146 | 10 | (H7;H5) / (H5;H5) | (H5;H5) / (H7;H5) | 4 / 6 |
| | 14-148 | 10 | (H7;H5) / (H4;H7) | (H7;H7) / (H4;H7) / (H5;H4) / (H5;H7) | 2 / 4 / 2 / 2 |
| | 14-147 | 9 | (H7;H5) / (Ha;Hb) | (H5;Hb) / (H7;Hb) / (H7;Ha) / (H5;Ha) | 2 / 4 / 2 / 1 |
| 14-140 | 14-132 | 9 | (H8;H15) / (H12;H3) | (H8;H12) / (H15;H3) / (H15;H12) / (H8;H3) | 2 / 1 / 4 / 2 |
| | 14-146 | 10 | (H8;H15) / (H5;H5) | (H8;H5) / (H15;H5) | 2 / 8 |
| | 14-147 | 10 | (H8;H15) / (Ha;Hb) | (H8;Ha) / (H15;Hb) / (H8;Hb) / (H15;Ha) | 4 / 2 / 2 / 2 |
| 14-152 | 14-132 | 10 | (H14;H10) / (H12;H3) | (H10;H12) / (H10;H3) / (H14;H3) / (H14;H12) | 4 / 1 / 4 / 1 |
| | 14-148 | 8 | (H14;H10) / (H4;H7) | (H10;H7) / (H14;H7) / (H10;H4) / (H14;H4) | 3 / 3 / 2 / 2 |
| | 14-147 | 9 | (H14;H10) / (Ha;Hb) | (H14;Hb) / (H10;Hb) / (H10;Ha) / (H14;Ha) | 3 / 2 / 2 / 2 |

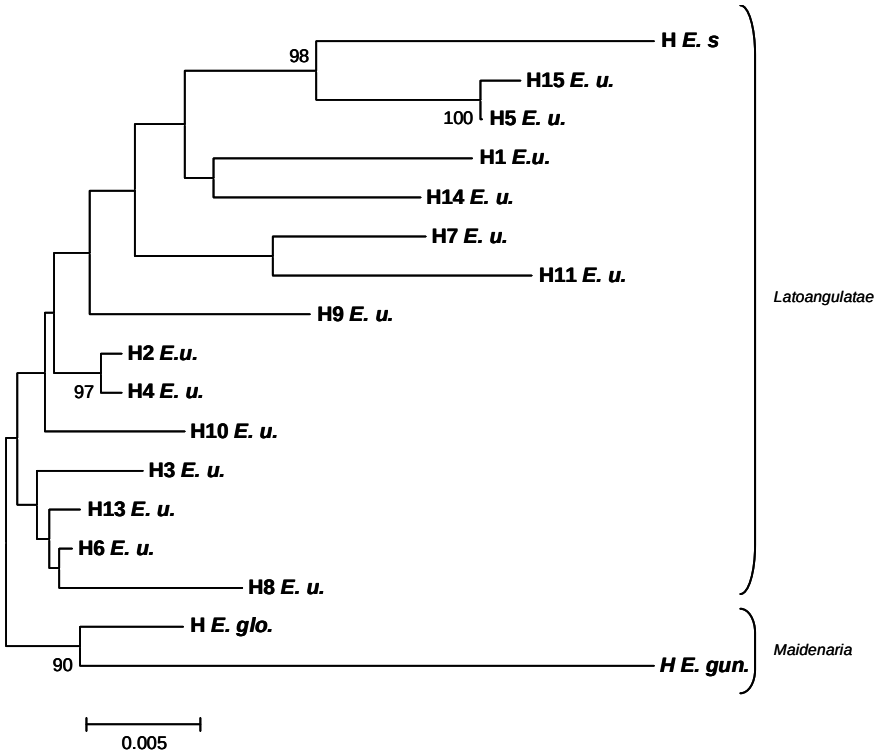


Figure 1: Arbre des distances entre les 14 halpotypes d'*E. urophylla* (H *E. u.*) et des haplotypes pour *E. gunnii* (H *E. gun.*), *E. globulus* (H *E. glo.*) et *E. saligna* (H *E. s.*). Le calcul des distance est réalisée grâce à la méthode du plus proche voisin, pour une partie du gène CCR regroupant l'exon 4 et l'intron 4. Les résultats d'un test de bootstrap (1000 répétitions) supérieurs à 80% sont indiqués. L'unité de longueur de branche correspond au nombre de nucléotides différents entre deux séquences pondéré par le nombre de sites comparés. Tous les sites d'insertion délétion ont été conservés.

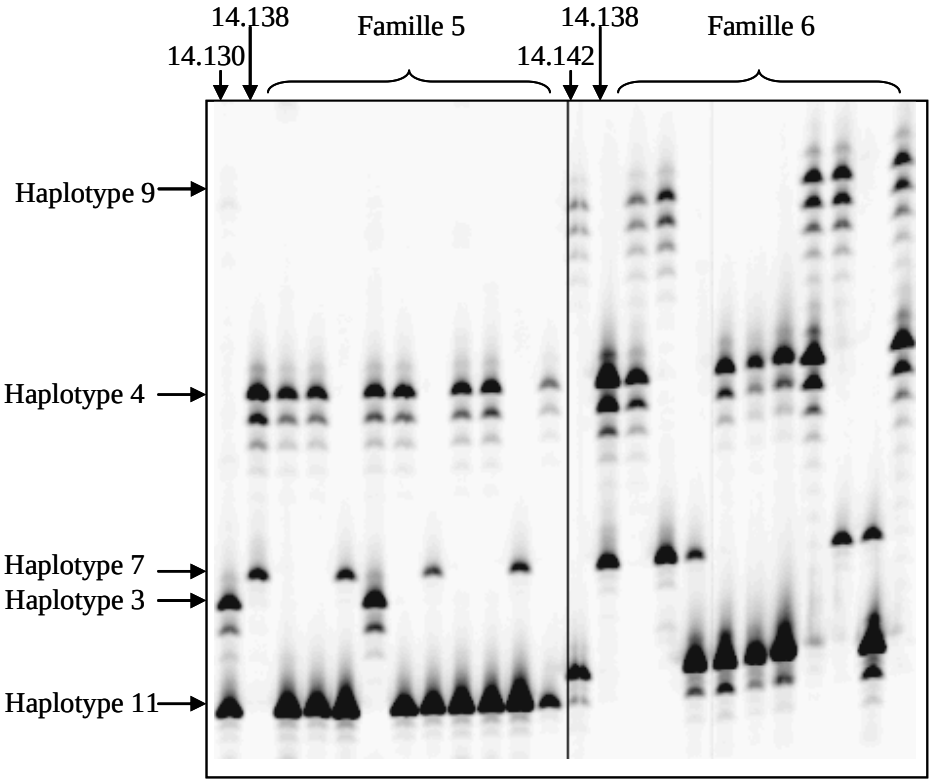


Figure 2: Profils microsatellite obtenus pour deux familles intra-spécifique d'*E. urophylla* issues de la mère 14-138 et des pères 14-130 et 14-142. La ségrégation des 5 haplotypes portés par ces trois géniteurs (H3, H4, H7, H9 et H11) est donnée par le locus microsatellite.

```
        -628
          |
E. g.     CTCCTCCTCCAACTGACCTAACAAATGGGTATGATTTAACTTACACGGCATGGATCACACGCAAAA
E. u. a   CTCCTCCTCCAACTGACCTAACAAACGGGTATGATTTAACTTACACGGCATGGATCACACGCAAAA
E. u. b   CTCCTCCTCCAACTGACCTAACAAACGGGTATGATTTAACTTACACGGCATGGATCATACGCAAAA

AAGAAAAAG-----TCGATCCTCAAGGGTTATGACCATTTCTCATTTGGCTGCCGAACTCCTCCAGTAAGGACAT
AAGAAAAAGGAAAGTCGATCCTCAAGAGTTATGACCATTTCTCATTTGGCTGCCGAACTCCTCCAGTAAGGACAT
AAGAAAAAGGAAAGTCGATCCTCAAGGGTTATGACCATTTCTCATTTGGCTGCCGAACTCCTCCAGTAAGGACAT

TTGATAAAAGATACATAGTACGTTCAT--------------ATGATTTTTTTAAAAAACACTTCACTAAAAGAAA
TTAATAAAAGATCCATAGTACGTTCATTTTCATGAAAAAGTAATGATTTAAAAAAAAAAACACTTCACTAAAAGAAA
TTAATAAAAGATCCATAGTACGTTCAT--------------ATGATTTTTTTCAAAAACACTTCACTAAAAGAAG

TTATTTACATTATTCAGAAAAATTCATCAATGAATTTTTCATTTAACAATTACAATTGGCGCCTTACCATTGTCG
TTATTTACATTATTCAGAAAAATTAATCAATGAATTTTTCATTTAACAATTACAATTGGTGCCTTACCATTATTG
TTATTTACATTATTCAGAAAAATTGATCAATGAATTTTTCATTTAACAATTACAATTGGTGCCTTACCATTATTG

TTCCAAACTAAATAAACAATCATTTCTTGAAAACGCTTCATAATTTTTTGTGAAGTAAACACACCCATTAACACT
TTCCCAACTAAATAAACAATCATTTCTTGGAAACGCTTTATAATTTTTTGTGAAGTAAACACACCCATTAACACT
TTCCCAACTAAATAAACAATCATTTCTTGGAAACGCTTTATAATTTTTTGTGAAGTAAACACACCCATTAACACT

AATATATTTTATAGATATCTCCAAGTACTAGAATCTCATCCTTGAACCATGGAAAATAAGGCTTATCCAAAAAAA
AATATATTTTATAGGTATCTCCAAGTACTACAATCTCATCCTTGAACCATGGAAAATAAGGCTTATCCAAAAAAA
AATATATTTTATAGGTATTTCCAAGTACTACAATCTCATCCTTGAACCATGGAAAATAAGGCTTATCCAAAAAAA

AAAA---CATGGAAAATAAGGGCAAATGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTTAGAGAAAGGAGTGGTC
AAAAA----TGGAAAATAAGGGCAAATGCTCTCTCTCTCTCTCTCTCTCTCTCT--------TAGAGAGAGGAGTGGTC
AAAAAAACATGGAAAATAAGGGCAAATGCTCTCTCTCTCTCTCTTTCTCT--------TAGAGAGAGGAGTGGTC

CTTATAGGGGAGCGGGTCATTTCATGCGGTAGGTGGGTCTTGGTAAACATTGTCTTTTTCCCTTATATATATATA
CTTATAGGGGAGCGGGTCATTTCGTGCGGTAGGTGAGTCTTGGTAAACATTGTCTTTTTCCCTTATATATATATA
CTTATAGGGGAGCGGGTCATTTCGTGCGGTAGGTGAGTCTTGGTAAACATTGTCTTTTTCCCTTATATATATAT~

                                        +1
                                         |
TATATATATATATATATATATATATATGTCTTGAGTGCTAGCCACCTTCGAATCAAGCCGTCAAAAAGCGTCCAA
T-------------------------------------------------------------------------
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CTTGAGTGCTAGCCACCTTCGACTTAAGCCGTCCACAAGCGTCCAC
```

Figure 3: Comparaison des deux haplotypes du géniteur *E. urophylla* 14-144 (*E. u.* a et *E. u.* b) et du clone *E. gunnii* (*E. g.*, Acc. EGU132750) pour la partie promotrice du gène CCR (-628 pb/+66 pb par rapport au site +1 de l'initiation de la transcription). Les sites variables entre *E. gunnii* et *E. urophylla* sont surlignés en gris, ceux qui varient entre haplotypes *E. urophylla* sont surlignés en noir.