HAL
open science

Développement de marqueurs moléculaires chez le pin maritime (Pinus pinaster Ait.) et cartographie génétique comparée des conifères

David Chagné

To cite this version:

David Chagné. Développement de marqueurs moléculaires chez le pin maritime (Pinus pinaster Ait.) et cartographie génétique comparée des conifères. Biologie végétale. Université Henri Poincaré - Nancy 1, 2004. Français. NNT : 2004NAN10008 . tel-01748676

HAL Id: tel-01748676

https://hal.univ-lorraine.fr/tel-01748676v1

Submitted on 29 Mar 2018

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.

UNIVERSITÉ DE LORRAINE

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

U.F.R. Sciences et Techniques Biologiques
Ecole Doctorale : Ingénieries des Ressources, Procédés, Produits et Environnement
Département de Formation Doctorale : Biologie Forestière

Thèse

Présentée pour l'obtention du diplôme de

Docteur de l'Université Henri Poincaré, Nancy I

Mention Biologie Forestière

Soutenue publiquement le 15 avril 2004 à Bordeaux

par David CHAGNE

Développement de marqueurs moléculaires chez le pin maritime (*Pinus pinaster* Ait.) et cartographie génétique comparée des conifères

Membres du jury :

Président :	J-M. FAVRE	Professeur, Université Henri Poincaré, Nancy I
Rapporteurs :	O. SAVOLAINEN	Professeur, Université de Oulu, Finlande
	P. LEROY	Ingénieur de recherche, INRA Clermont-Ferrand
Examinateurs :	M. CERVERA	Chercheur, INIA Madrid, Espagne
	C-E. DUREL	Chargé de recherche, INRA Angers
Directeur de thèse :	C. PLOMION	Directeur de recherche, INRA Bordeaux

UMR BIOGECO – INRA Equipe de génétique – 69 route d'Arcachon, 33612 Cestas cedex

Merci à tous ceux qui de près ou de loin ont participé ou influencé ce travail de thèse. La liste est longue, je vais essayer de rassembler tous ces gens, en espérant n'oublier personne…

Je remercie tout d'abord et surtout « the boss » Christophe Plomion pour tout le temps qu'il m'a consacré au cours de ces cinq (!!!) super années qui sont passées très vite. Merci pour la confiance qu'il a placé en moi, merci pour les corrections à domicile, merci pour ces moments passés ensemble devant l'ordinateur, merci pour ces « pfff ! laisse tomber, te prend pas la tête ! », les « OK, super, pas de problème !», les « hey guys !» etc… Merci de m'avoir envoyé à l'autre bout du monde parce que beaucoup de choses sont parties de « là-bas ». Merci aussi pour avoir partagé le même bureau pendant tout ce temps.

Je souhaite remercier Outi Savolainen (kiitos !) et Maite Cervera (gracias !) qui ont accepté de relire ce manuscrit en français. Je remercie aussi Philippe Leroy d'avoir accepter la tâche de rapporteur et Charles-Eric Durel et Jean-Michel Favre d'avoir accepté de participer à mon jury de thèse.

Je remercie les personnes qui ont participé à mes comités de thèse, à savoir Joël Gellin, Jean-Christophe Glaszmann et Jean-Marc Gion (« mais mon gars, ça sert à quoi la carto comparée ? »).

Je souhaite remercier particulièrement tous les gens avec qui j'ai travaillé plus ou moins directement. Céline Lalanne qui m'a appris à tenir une pipette et à faire des PCR. Et comment dissocier Céline de Delphine M. ? David pour ses QTL et ses SNP. Patricia pour ses questions existentielles, pour penser à moi quand la porte du frigo se décroche et pour toutes ces DGGE qui ne marchent pas toujours. Bon courage pour cette future double (!) maternité. Agnès pour la belle aventure des microsats et tous ces délires dans le labo et en dehors. Encore bravo pour ta réussite et bonne chance pour ce futur toulousain ! Philippe alias « MacGyver » pour son aide au tout début de ma thèse avec les microsats de radiata. Bon courage pour ta thèse et tes « puces ». Ivan pour cette magnifique carte qui tient encore tout un mur de la salle info, pour ces discussions sur les manières de refaire le monde et sur l'aide précieuse face au « tigre de papier ». Loick pour le temps passé devant l'ordinateur avec toutes ces belles photos de pruniers partout dans le bureau. JMF pour l'aide avec les scripts en Perl. Virginie pour les assemblages. Merci à Pauline pour toutes ces discussions, les bonbons au gingembre et ces moments passés ensemble, ainsi que pour la correction de l'article carto comparée, les discussions SNP et j'en oublie… Craig et Tom pour la collaboration Franco-NZ avec les microsats, pour les corrections de l'article cDNA-SSR et pour les bières à San Diego. Maite et Beppe et tous les membres du consortium pour les microsats de pin.

Cette thèse a été l'occasion pour moi de m'expatrier à l'autre bout du monde, dans la ville des Aggies, des vélos, de la DGGE, des soirées folles et de la Sudwerk's, c'est-à-dire Davis. Je voudrai donc remercier David Neale qui m'a permis de joindre son équipe par deux fois, lors de cette thèse et lors de mon DEA. Merci aussi de m'avoir prêté ses chaussons d'escalade, même si ça n'a as été une réussite…
Je remercie aussi Garth Brown qui a été mon supervisor pendant ces deux séjours. Merci pour son aide, son amitié et félicitations à Kim et à lui pour ce beau bébé !
Merci à Michela Troggio pour son amitié, sa bonne humeur et son célèbre « boiled chicken », merci aussi au Grand Kostya, Geoff et sa casquette, Elhan et son caractère d'or, Kathie, Robin, Dr Xeno…

Au cours de ces années à Pierroton j'ai eu la chance de côtoyer des gens extraordinaires, je les remercie pour tout ce qu'ils m'ont apporté aussi bien pour cette thèse mais aussi du point de vue personnel.

Je souhaite tout d'abord remercier Antoine Kremer pour m'avoir accueilli au sein de son équipe dans ce petit paradis perdu au milieu de la forêt. Merci de m'avoir permis de travailler dans ce laboratoire qui a vraiment quelque chose de spécial… et c'est sûrement dû à ce grand monsieur qu'est Antoine !

Une thèse c'est plein de moments partagés, à discuter avec les collègues, c'est pour cela je souhaite tout d'abord remercier tous les thésards du labo, et il y en a eu beaucoup… Je commencerai par saluer Paulo, pour qui j'ai une pensée et qui j'espère appréciera cette thèse d'où il est. J'ai eu la chance de le côtoyer

quelques mois à mes tous débuts et de pouvoir apprécier ce grand monsieur… avant de le regretter comme le regrettent tous ceux qui l'ont croisé. Je voudrais aussi remercier Stéphanie, avec qui tout a commencé. Merci pour tous ces moments inoubliables et merci de m'avoir donné goût aux microsats qui marchent pas. Désolé pour le coup de la carbo dans les tubes... n'empêche que c'était marrant. Merci à Stéphane pour ces bonnes idées et pour ce fameux week-end à San Francisco. Merci à Delphine et Christian pour avoir apprécié la blague de la carboglace (c'est d'ailleurs parti de Christian) et pour tous autres supers moments passés ensemble. Je promets de faire un saut à LA bientôt. Merci à Marie-France pour sa bonne humeur, son calme incroyable, mais surtout pour le bel exemple à suivre en matière de réussite mais surtout de bonheur. Elle est pas belle la vie, advienne que pourra inch hallah youpi ! Merci à Céline « HCL » avec ses hauts, ses bas, et enfin ses hauts. Bon courage à toi pour la suite. Merci à Grégoire pour ces petits coups de pouce de manips, pour ses blagues et les parties de pêche. Merci à Caroline pour toutes ces danses endiablées, pour avoir apprécié mes chemises et pour l'exemple à suivre en matière de densité d'évènements (colloque aux USA, soutenance, appendicite et mariage en 10 jours !!!). Merci à Jorge pour ces semaines passées ensemble au labo et à la maison, pour sa conduite légendaire et pour le vin portugais. Merci à Josquin, le plus français des australiens, pour cette personnalité unique, pour sa capacité à assimiler le rhum une veille de réunion et pour le domptage du megaspace. Merci à Jérémy pour son humour, sa foi en la tartiflette et surtout merci d'avoir rangé la salle radio après mes Dot Blot. Merci à Philippe pour ses astuces et Ramstein dans le bureau d'à côté, merci à Jérome (euh non, il ne le mérite pas. Si ! merci pour les impressions et les odeurs d'encens), à la petite Manue pour ses questions, à Laurent, à Li An, à Marta, à Iria de prendre la suite. Merci à ma génération de thésards, c'est à dire ceux qui ont commencé en même temps que moi et avec qui on a pu partager pleins de choses. Muchas gracias al professor Magni con quien la aventura empiezaba en Nancy, bravo pour cette belle fécondité (deux filles et une thèse). Bon courage pour la fin et surtout bon courage pour la suite au Chili en tant que grand Professeur de genetica forestal. Je viens quand tu veux pour donner des cours de ce que tu veux (pin-maritimologia ?), même en espagnol. Grazie mile a la bellissima Manuela pour cette habilité à ne jamais stresser et à tout prendre de manière calme (sauf les courses de F1). Ca fait vraiment du bien d'avoir quelqu'un comme ça dans le bureau d'à côté. Je promets de venir souvent dans les chambres d'hôtes d'Orvieto. Merci aussi pour les boudins au chocolat. Merci au « Grand David » pour tous ses moments passés ensemble, depuis la dure préparation des exposés d'Avignon jusqu'aux cours de kite blooob bloob surf. Merci d'avoir répondu à toutes ces questions et de m'avoir filé tous ces petits coups de main. Merci pour le « génome candidat ». C'est difficile d'écrire toute ma gratitude, en tout cas j'ai vraiment l'impression d'avoir bossé à côté d'un grand bonhomme pendant trois ans. D'où le terme « Grand David ». Et puis ça m'a mis une énorme claque de voir la 205 partir de Pierroton avec sa « vlanche à poils » sur le toit et de retrouver le conducteur et sa douce 10 jours plus tard dans Mon Davis ! Bonne chance en tout cas avec tes mouches.
Ouf ! Voilà pour les thésards.

Maintenant je veux remercier tous les autres gens de Pierroton et les gens qui auront le courage de lire ce paragraphe jusqu'au bout. D'abord merci à tous ceux qui vont rester à Pierroton à perpétuité. Merci à Fabrice et Jean-Marc pour ces petits moments passés ensemble au labo ou ailleurs. Merci surtout pour m'avoir aidé à dompter mon PC. Merci à Sophie pour ses corrections lors du rush final et pour les discussions autour d'un thé. Un grand merci à Annie pour la correction de l'orthographe. Merci à Guy d'avoir essayé de courir plus vite que moi. Merci à Mimie d'avoir couru derrière nous. Merci à Patrick pour les coups de main avec le megaspace. Merci à Henri pour son bel exemple de pugnacité contre l'administration. Merci à Christian pour la biblio sur la phylogénie et pour la superbe figure avec des boules vertes, jaunes et rouges. Merci à Jeannot pour les coups de fusil dans les branches et pour les blagues à table. Merci à Cyril pour ses microsats qui marchent encore moins. Merci à Thierry pour les pages web alors que c'est pas son boulot. Merci à Christophe O pour l'aide avec la base de données SSR. Merci aussi (dans le sens des aiguilles d'une montre qui marche à l'envers) à Rémy, Alexis, Marie-Pierre, Dominique, Catherine, Pierre-Yves, Annie, Florence et tous les autres d'à côté, comme Luc, Christelle, Sylvain, les footeux du mardi, les gens de la pépinière, d'entomo, du château et d'en face.
Je remercie aussi tous les stagiaires, contractuels, clandestins, intermittents de la science, visiteurs et autres stakhanovistes de la pipette pour ces supers moments passés à la paillasse. Merci à MH pour sa bonne humeur et ses soirées du jeudi, merci à Teresa pour les discussions sur la cartographie comparée,

sur les microsats qui marchent pas et sur la vie en général. Merci surtout pour les gâteaux au chocolat. Merci aux Laetitia, à Sabine la « mégère ménagère ». Merci à tous ceux qui ont quitté Pierroton comme par exemple Frank pour partir au pays des caribous, Guylaine au Moulon ou Corinne à Pessac. Merci à toute la nouvelle relève, au Jeune pour les photos, à Hakim le fils du forgeron pour les olives et les chips, à Olivier pour avoir supporté les mails qui arrivent et les autres bruits, au jeune basque Alex et à ses 8 000 extractions, à Muriel pour s'être inquiété de ma santé ces derniers temps, Fabiano pour avoir repris de la tête mon superbe tir qui a heurté le poteau droit, Franck, Caroline, Stéphane, Jean-François (bon courage avec mon ancien PC), la petite Michela, Julien le nancéen, et puis tous ceux que j'oublie.

Il n'y a pas que Pierroton dans la vie. Et d'ailleurs il y a beaucoup de gens dans le monde et dans le Lot-et-Garonne qui ne savent même pas où ça se trouve, ni ce que ça fait de rater une PCR. Tant mieux pour eux. Pourtant au cours des dernières années une partie de ces gens m'ont apporté un soutient crucial.

Je souhaite remercier tous les « petits frères marmandais », Franck, Cédric, Charlotte (et le petit Jules), Bidi, Sandrine, Max, Célia pour ces moments de délire qui font du bien à la tête. Merci surtout pour cette confiance en moi que vous m'avez donné, même quand j'étais à l'autre bout du monde. Merci à Aurélie et LN pour cette belle amitié née dans les Pyrénées. Merci à tous les arbitres de basket du sud-ouest pour avoir partagé avec moi le moment de défouloir du samedi soir. Merci à John Stockton pour cette carrière exceptionnelle. Merci aux gens qui attendent leur bus tous les matins à la même heure quand je passe en voiture pour aller au boulot. Est-ce que je vais leur manquer ? Et d'abord comment s'appellent-ils? Merci à toute la Familia de Davis, à Philou (go Kings !), Flo et à tous ceux que je reverrai sûrement à Davis ou ailleurs. Merci à Bono, Benabar et Tori Amos pour l'ambiance de fin de rédaction. Merci au professeur Thibault pour m'avoir attiré vers la science. Merci à ma prof de physique et chimie de 1ère S de m'avoir dit que je n'étais pas fait pour les sciences.

Merci à ma mère, mon père, mon frère, ma sœur, Laurent, ma grand-mère et mes grands-pères pour leur soutient de toujours.

Merci enfin à celle qui m'a permis de passer toutes les épreuves avec plaisir, de Davis à Bordeaux, celle qui a gommé mes doutes et qui m'a donné confiance en moi. Merci à celle avec qui je vais partager les épreuves à venir, mais surtout les joies, les rires et les plus grands évènements. Merci Karine pour cette énergie si contagieuse. Merci pour cet amour.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Remerciements

Table des matières 1

Préambule 5

Partie A : Introduction 8

A.1 Les conifères 9

A.1.1 Caractéristiques générales des conifères 9

A.1.2 A l'origine des conifères : des pins et des dinosaures 10

A.1.3 Phylogénie et classification des conifères 11

A.1.4 Le génome des conifères 12

A.1.5 Contexte "politico-scientifique" de la recherche en génomique chez les conifères 14

A.2 L'espèce étudiée : le pin maritime 15

A.2.1 Présentation générale 15

A.2.2 Diversité génétique du pin maritime au sein de son aire de répartition 16

A.2.3 Le programme d'amélioration génétique du pin maritime 17

A.2.4 Le programme de génomique du pin maritime 18

A.3 Situation au sein du contexte, objectifs et problématiques de la thèse 20

A.3.1 Construction d'une carte génétique du pin maritime 20

A.3.2 Cartographie génétique comparée chez les conifères 20

A.3.3 Développement de marqueurs 21

A.3.4 Problématique 22

Partie B : Méthodes mises en jeu 24

B.1 Développement de marqueurs moléculaires 24

B.1.1 Marqueurs anonymes révélés en masse : les AFLP 24

B.1.1.1 Description des marqueurs AFLP 24

B.1.1.2 Protocole 25

B.1.2 Marqueurs anonymes fondés sur les séquences répétées : les microsatellites ou SSR 25

B.1.2.1 Description des microsatellites 25

B.1.2.2 Les microsatellites chez les Pinaceae 26

B.1.2.3 Les méthodes de développement des microsatellites utilisées chez le pin Maritime 27

B.1.2.3.1 Transfert entre espèces proches 27
B.1.2.3.2 Recherche de microsatellites dans les banques d'EST 28
B.1.2.3.3 Développement de banques génomiques enrichies en microsatellites 29
B.1.2.4 Développement de bases de données pour les microsatellites de pin maritime 31
B.1.3 Marqueurs fondés sur les séquences codantes : les EST 31
B.1.3.1 Qu'est-ce qu'un EST ? 31
B.1.3.2 Méthodes de détection de polymorphisme sur les EST 32
B.1.3.2.1 Méthodes basées sur la longueur des fragments d'ADN 32
B.1.3.2.2 Méthodes basées sur la séquence d'ADN 33
B.1.3.2.2.1 Electrophorèse en gradient de dénaturation : la technique DGGE 34
B.1.3.2.2.2 Polymorphisme de conformation d'ADN simple brin : la technique SSCP 34
B.1.3.2.3 Méthode bioinformatique de détection de variation 35
B.2 Cartographie génétique 37
B.2.1 Principe de la cartographie génétique 37
B.2.2 Les stratégies de cartographie génétique utilisées chez les arbres forestiers 40
B.2.2.1 Stratégies utilisant le mégagamétophyte des conifères 40
B.2.2.2 Stratégie du double-pseudo backcross 40
B.2.3 Description des populations utilisées pour la construction des cartes génétiques de pin maritime et d'autres espèces de conifères 41
B.2.3.1 Population F2 41
B.2.3.2 Population G2 42
B.2.3.3 Les autres cartes utilisées dans le cadre de la cartographie comparée chez les Conifères 43
B.2.4 Méthodes et logiciels utilisés pour la construction des cartes génétiques de pin Maritime 43
B.2.4.1 Protocole utilisé pour la construction de la carte F2 (diploïde et haploïde) 43
B.2.4.2 Protocole utilisé pour la construction de la carte génétique de la G2 44
B.2.4.3 Protocole utilisé pour l'ajout de nouveaux marqueurs et pour tester l'ordre de ces marqueurs 44
B.2.5 Détection de QTL chez le pin maritime 45
B.3 Cartographie génétique comparée 47
B.3.1 Définition et exemples chez les plantes 47
B.3.2 Cartographie génétique comparée chez les arbres forestiers 48

Partie C : Résultats et discussion 50
C.1 Développement de marqueurs moléculaires chez le pin maritime 51
C.1.1 Marqueurs AFLP 51
C.1.1.1 Obtention de marqueurs AFLP et évaluation de leur utilité pour la cartographie génétique du pin maritime 51
C.1.1.2 Evaluation des AFLP pour la détection de QTL 52
C.1.1.3 Evaluation des AFLP pour la cartographie comparée chez les conifères 53
C.1.2 Marqueurs microsatellites 53
C.1.2.1 Transfert de marqueurs microsatellites entre espèces de pin 53
C.1.2.1.1 Quelques marqueurs polymorphes 53
C.1.2.1.2 Transférabilité et orthologie des marqueurs microsatellites : application en cartographie comparée chez les conifères 56
C.1.2.2 Détection *in silico* de marqueurs microsatellites 57
C.1.2.2.1 Quels types de motifs répétés se trouvent dans les régions codantes des conifères ? 57
C.1.2.2.2 Développement d'amorces PCR, transférabilité, polymorphisme et orthologie des microsatellites de régions codantes (SSR-ADNc) 59
C.1.2.3 Développement d'une banque enrichie en microsatellites 61
C.1.3 Développement de marqueurs codants 65
C.1.3.1 Des EST polymorphes révélés par DGGE et SSCP et EST-AFLP 65
C.1.3.2 Les polymorphismes ponctuels (*Single Nucleotide Polymorphism*, SNP) 67
C.1.3.2.1 Cartographie de SNP dans des gènes candidats liés à la formation du bois 67
C.1.3.2.2 Détection automatique de SNP dans les bases d'EST de pin maritime 68
C.2 Cartographie génétique chez le pin maritime 71
C.2.1 Etablissement d'une carte génétique à partir de la descendance G2 71
C.2.2 Vers une carte génétique consensus du pin maritime 71
C.2.2.1 Comparaison des cartes F2 et G2 71
C.2.2.2 Y a-t-il des QTL communs entre les cartes F2 et G2 ? 72
C.2.3 Qu'est-ce que la cartographie génétique peut nous apprendre sur la structure du génome des conifères ? 73
C.2.3.1 La relation entre tailles génétique et physique 74
C.2.3.2 Répartition de différents marqueurs moléculaires sur le génome de l'épicéa commun (*Picea abies* Karst.) 75
C.2.3.2.1 Avant-propos 75

C.2.3.2.2 Divers types de marqueurs moléculaires 76
C.2.3.2.3 Une répartition non aléatoire de différents types de marqueurs moléculaires 77
C.3 Cartographie comparée chez les Pinaceae 80
C.3.1 Comparaison des cartes génétiques de *Pinus taeda* et de *Pinus pinaster* 80
C.3.2 Synténie chez les conifères 81
C.3.3 Application de la cartographie génétique comparée entre espèces de pins 83

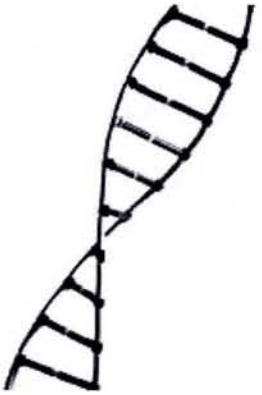
Partie D : Conclusion et perspectives 85
D.1 Des marqueurs moléculaires aux propriétés variables pour des applications diverses (cartographie génétique, cartographie comparée, approche gène candidat) 86
D.2 Des applications futures en génétique des populations 88
D.3 Une suite à la cartographie génétique comparée des conifères 89

Bibliographie 92

Annexes 114
Annexe I : A high density genetic map of maritime pine based on AFLPs 116
Annexe II : Microsatellite markers for *Pinus pinaster* Ait. 122
Annexe III : Comparative genome and QTL mapping between maritime and loblolly pines. 125
Annexe IV : Cross species transferability and mapping of genomic and cDNA SSRs in pines. 132
Annexe V : Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. 144
Annexe VI : Analysis of the distribution of marker classes in a genetic linkage map of Norway spruce (*Picea abies* Karst.). 154

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

PREAMBULE



U... NANCY 1
BIBLIOTHEQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

L'essor actuel de la **génomique**, science qui étudie la structure et le fonctionnement du **génome**, a été initié par la découverte de la structure de l'ADN par Watson et Crick (1953). Cette discipline s'est rapidement développée ces dernières années et a été ponctuée de dates clefs comme par exemple, la construction d'une **carte génétique** de l'Homme (Botstein et al. 1980), le séquençage complet des génomes du nématode *Caenorhabditis elegans* (Hodgkin et al. 1998), de l'Homme (International Human Genome Sequencing Consortium 2001), d'*Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), et récemment d'un arbre forestier, le peuplier (Brunner et al. 2004). La génomique a bénéficié des progrès techniques cruciaux de la **biologie moléculaire** comme le clonage, le séquençage, la transformation génétique ou l'amplification de l'ADN par PCR, et s'est alliée récemment à l'informatique pour créer une autre discipline émergente : la **bioinformatique**. La génomique a permis de compléter les recherches effectuées sur la compréhension des mécanismes génétiques contrôlant des caractères comme par exemple les maladies héréditaires humaines ou la productivité d'espèces agronomiques animales et végétales. Ces recherches utilisaient auparavant des approches classiques comme la génétique mendélienne ou la génétique quantitative. Ces dernières bénéficient maintenant des apports de la génomique, comme par exemple l'identification des gènes et des mécanismes moléculaires sous-jacents aux caractères d'intérêt et agronomique.

La recherche en génétique forestière, et en particulier celle concernant les **conifères**, a elle aussi bénéficié des apports de la génomique. La construction de cartes génétiques et le **développement de marqueurs moléculaires** pour ces espèces ont été réalisés pour bon nombre d'espèces, avec pour objectifs une meilleure connaissance de leur génome, une utilisation de ces outils dans des programmes d'**amélioration génétique**, ou une description plus fine de leur **diversité génétique**.

Cette thèse applique des méthodes de la génomique à une espèce forestière de première importance en France : le **pin maritime** (*Pinus pinaster* Ait.), dans le but :

1/ d'établir une carte génétique saturée du pin maritime;

2/ de placer des marqueurs orthologues pour comparer la structure de son génome à d'autres espèces de conifères (conservation de la synténie et de l'ordre des gènes);

3/ de transférer des informations génétiques (QTL, gènes) du pin taeda (espèce de référence pour la cartographie comparée des Pinaceae) au pin maritime;

4/ enfin, de fournir des outils moléculaires (marqueurs microsatellites et polymorphismes ponctuels) pour aider les sélectionneurs à mieux exploiter et gérer leur

ressources génétiques dans le cadre de programme de sélection assistée par marqueurs ou de programmes de conservation.

Dans un premier temps, nous décrirons le contexte de la thèse en présentant l'origine et la biologie des conifères et en particulier celle du pin maritime. Nous verrons ensuite le matériel végétal et les méthodes mises en jeu tant pour le développement de marqueurs que pour la construction d'une carte génétique et la comparaison avec des cartes obtenues chez d'autres conifères. Les résultats seront présentés et discutés en se référant aux articles placés en annexe.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

INTRODUCTION

Partie A : Introduction

A.1 Les conifères

Les conifères sont des organismes remarquables à de nombreux points de vue. Ils comptent parmi les organismes les plus grands et les plus longévifs de notre planète. Ils sont apparus il y a beaucoup plus longtemps que les autres espèces de plantes vivant actuellement sur Terre et sont adaptés à une grande variété d'écosystèmes. Les pins, en particulier, sont des espèces d'importance écologique majeure pour les forêts tempérées. Les paragraphes suivants porteront sur différents aspects des ces organismes : leur biologie, leur origine, leur classification et les connaissances actuelles sur leur génome.

A.1.1 Caractéristiques générales des conifères

Les conifères sont les représentants actuels des gymnospermes. Ce sont toujours des arbres, ou au minimum, des arbustes, également connus sous le nom de résineux. Ils s'opposent aux autres arbres (angiospermes) par le fait qu'ils portent des aiguilles et que les canaux sécréteurs présents dans tous leurs organes contiennent de la résine. Ces espèces dans leur ensemble montrent des capacités d'adaptation incroyables à de nombreux environnements, allant des régions montagneuses tibétaines, aux plateaux désertiques de l'ouest des Etats-Unis, en passant par les régions marécageuses de Floride ou les dunes de la côte landaise.

L'importance économique des conifères est énorme. Leur bois reste leur principale production; il sert de matière première dans la construction (bois d'œuvre) et dans l'industrie (pâtes à papier et panneaux de particules). Il peut être aussi utilisé comme bois de feu et représente la principale source d'énergie pour la plupart des pays en voie de développement. La résine est également une production économiquement importante, ou plutôt les résines, car leur qualité varie beaucoup avec les espèces, depuis la résine dont on extrait l'essence de térébenthine jusqu'aux produits tels que le baume du Canada, produit par un sapin (*Abies balsamea*). Enfin, la valeur ornementale et protectrice de nombreuses essences est indéniable et fait même l'objet actuellement d'un certain engouement.

Les pins (genre *Pinus*) appartiennent à la famille des Pinaceae. Il s'agit d'arbres dont le développement et le port sont très variés, mais qui sont tous caractérisés par des aiguilles

Figure A1 : Echelle géologique présentant les trois périodes du Mésozoïque (Trias, Jurassique et Crétacé).

Era.syst	Série	Etage	Ma
Mésozoïque – Crétacé	Crétacé supérieur	Maastrichtien	72
		Campanien	83
		Santonien	87
		Coniacien	88
		Turonien	91
		Cénomanien	96
	Crétacé inférieur	Albien	108
		Aptien	114
		Barrémien	116
		Hauterivien	122
		Valanginien	130
		Berriasien	135
Mésozoïque – Jurassique	Jurassique supérieur (Malm)	Tithonien	141
		Kimméridgien	146
		Oxfordien	154
	Jurassique moyen (Dogger)	Callovien	160
		Bathonien	167
		Bajocien	176
		Aalénien	180
	Jurassique inférieur (Lias)	Toarcien	187
		Pleinsbachien	194
		Sinémurien	201
		Hettangien	205
Mésozoïque – Trias	Trias supérieur	(Rhétien)	
		Norien	220
		Carnien	230
	Trias moyen	Ladinien	235
		Anisien	240
	Trias inférieur	Scythien	245

Figure A2 : Des dinosaures et des conifères

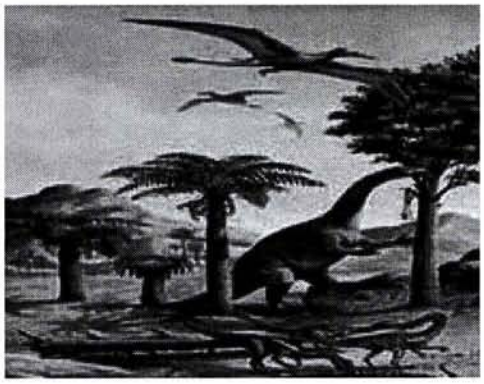
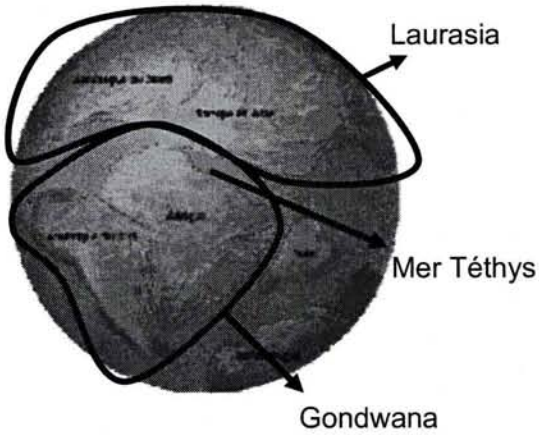


Figure A3 : La Terre lors de l'apparition des premiers pins durant la période Jurassique (il y a environ 140 Ma).



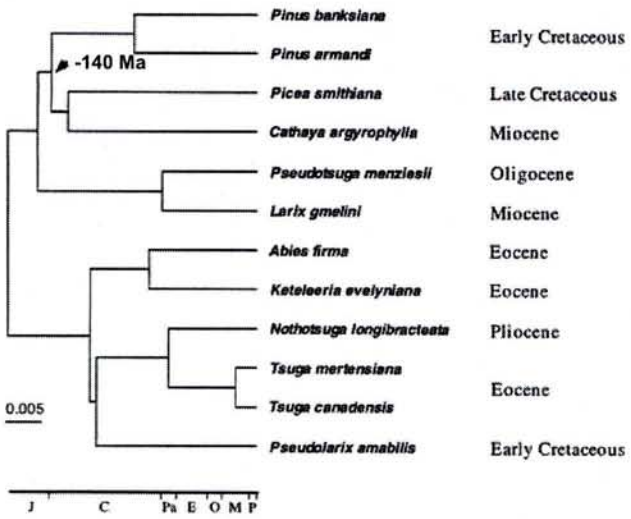


Figure A4 : Phylogénie moléculaire des Pinaceae (tiré de Wang et al. 2000) basée sur la séquence du gène chloroplastique *matK*.
J, Jurassique; C, Crétacé; Pa, Paléocène; E, Eocène; O, Oligocène; M, Miocène; P, Pliocène.

pointues, longues ou courtes, réunies en groupes de 2, 3, 4 ou 5. Les pins sont des plantes monoïques. L'inflorescence femelle ou cône, une fois la fécondation accomplie, mûrit en deux (rarement trois) ans. Après la formation des graines et l'ouverture des écailles, le cône peut tomber ou rester sur l'arbre. Les graines sont souvent ailées, ce qui facilite leur dissémination par le vent et l'extension de leur aire de distribution.

A.1.2 A l'origine des conifères : des pins et des dinosaures

Des restes fossiles datent l'apparition des premiers conifères à la fin de l'ère Primaire (Permien). Néanmoins le développement des gymnospermes fut important au cours du Mésozoïque (ère Secondaire, figure A1). Cette période a été marquée par la dominance d'une végétation composée de fougères à graines et de Gymnospermes (Ginkgos, Cycas et Conifères). Grâce à leurs graines, ces espèces se sont très bien adaptées à la colonisation des écosystèmes terrestres. Ainsi, les premières espèces de pin peuvent avoir constitué, au cours du Jurassique, le "pin quotidien" des Diplodocus et autres dinosaures herbivores (figure A2). Au point de vue géologique, le Mésozoïque est marqué par la séparation de la Pangée en deux sous-continents : Gondwana au Sud et Laurasia au Nord, séparés par un espace océanique (Téthys, figure A3). Les fossiles de pins ont été retrouvés à des périodes postérieures à cette séparation en deux continents (Jurassique, Mirov 1967), ce qui peut expliquer que l'on trouve actuellement des pins à l'état naturel uniquement dans l'hémisphère Nord. L'origine géographique précise des pins n'est pas connue, néanmoins on peut penser que ces espèces pionnières ont colonisé les terres émergées par le Nord (origine en Asie et/ou Amérique du Nord ?). L'étude de la colonisation des pins au sein de leurs aires de distribution actuelles est en particulier rendue difficile par la succession de périodes glaciaires et interglaciaires au cours du Quaternaire où les espèces ont tantôt colonisé de nouveaux territoires et tantôt se sont retrouvées dans des zones refuges.

Les données paléobotaniques peuvent être comparées avec les données de phylogénie moléculaire. Wang et al. (2000), ont étudié les séquences de gènes appartenant aux trois génomes (chloroplastique, mitochondrial et nucléaire) chez plusieurs espèces de la famille des Pinaceae. Ces auteurs se sont servis d'une horloge moléculaire en prenant 140 millions d'années (Ma) comme date de divergence du genre *Pinus* du reste des Pinaceae (figure A4). Ces résultats montrent que les espèces pins ont commencés à divergé vers le début du Crétacé, ce qui est en accord avec les données paléobotaniques qui indiquent que les deux sous-genres du genre *Pinus* étaient déjà différenciés lors de cette période (Mirov 1967). La

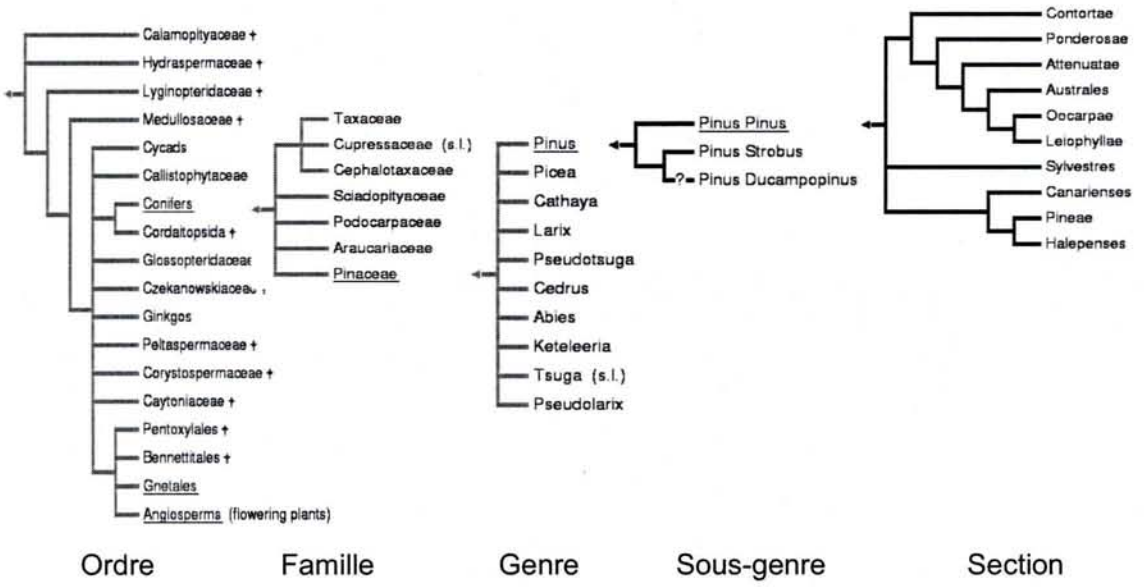


Figure A5 : Phylogénie simplifiée des conifères situant les différents niveaux de classification utilisés pour classer les espèces de pins (source : http://tolweb.org/tree/).

date de divergence du genre *Pinus* (140 Ma) peut être comparée avec celles d'autres genres de plantes. Ainsi, il semble que les pins sont des espèces beaucoup plus anciennes que le reste des plantes actuelles. Par exemple, chez les Fagaceae, une famille d'arbres forestiers feuillus, les données paléobotaniques (Xu 2004) montrent que les chênes ont divergé des châtaigniers et des hêtres au cours de l'Eocène (ère Tertiaire, -35 à -40 Ma). Ces dates de divergence correspondent également à peu près aux dates où ont divergé les différents genres de céréales : à l'intérieur du complexe *Triticeae / Aegilops*, le genre *Lolium* aurait divergé il y a 35 Ma, le genre *Hordeum* il y a 11 Ma et le genre *Secale* il y a 7 Ma (Huang et al. 2002).

A.1.3 Phylogénie et classification des conifères

En dehors des formes fossiles, les conifères sont regroupés en un seul ordre, celui des Coniférales, généralement divisé en sept familles (figure A5) : les Pinaceae (mélèzes, sapins, pins, cèdres, épicéas), les Taxodiaceae (séquoias), les Cupressaceae (cyprès, thuyas, genévriers), les Taxaceae (ifs), les Céphalotaxaceae (*Cephalotaxus*), les Araucariaceae (araucarias) et les Podocarpaceae (*Podocarpus*). La famille des Pinaceae est généralement divisée en dix genres : *Pinus*, *Cathaya*, *Picea*, *Larix*, *Pseudotsuga*, *Abies*, *Cedrus*, *Keteeleria*, *Tsuga* et *Pseudolarix*.

Le genre *Pinus* comporte un peu plus d'une centaine d'espèces répandues dans l'hémisphère Nord, des régions arctiques et sub-arctiques nord américaines et européennes, aux régions tropicales et sub-tropicales d'Amérique Centrale et d'Asie. Des représentations détaillées des aires de répartition des différentes espèces de pins ont été présentées par Mirov (1967) et par Farjon (1984). La classification du genre *Pinus* est assez complexe et a été soumise à de nombreux remaniements selon les caractères pris en compte. Elle est rendue en particulier complexe à cause de l'aptitude de certaines espèces à s'hybrider entre elles ou à former des continuum. Actuellement, il est encore difficile d'avoir une vue d'ensemble précise de la phylogénie du genre *Pinus*. La classification de Mirov (1967) est remarquable du fait qu'elle fut la première à prendre en compte plusieurs approches (morphologique, géographique, génétique, physiologique, chimique et paléobotanique) et qu'elle explore la quasi-totalité des espèces de pin (sa classification tient compte de 105 espèces au total). Néanmoins, cette classification a montré ses limites et des incertitudes. Les frontières entre les différents groupes de pins ont été redéfinies plus récemment (Little et Chritchfield 1969, Van der Burgh 1973), en particulier grâce aux données obtenues à partir des séquences d'ADN chloroplastique et nucléaire (Krupkin et al. 1996, Price et al. 1998, Liston et al. 1999, Wang et

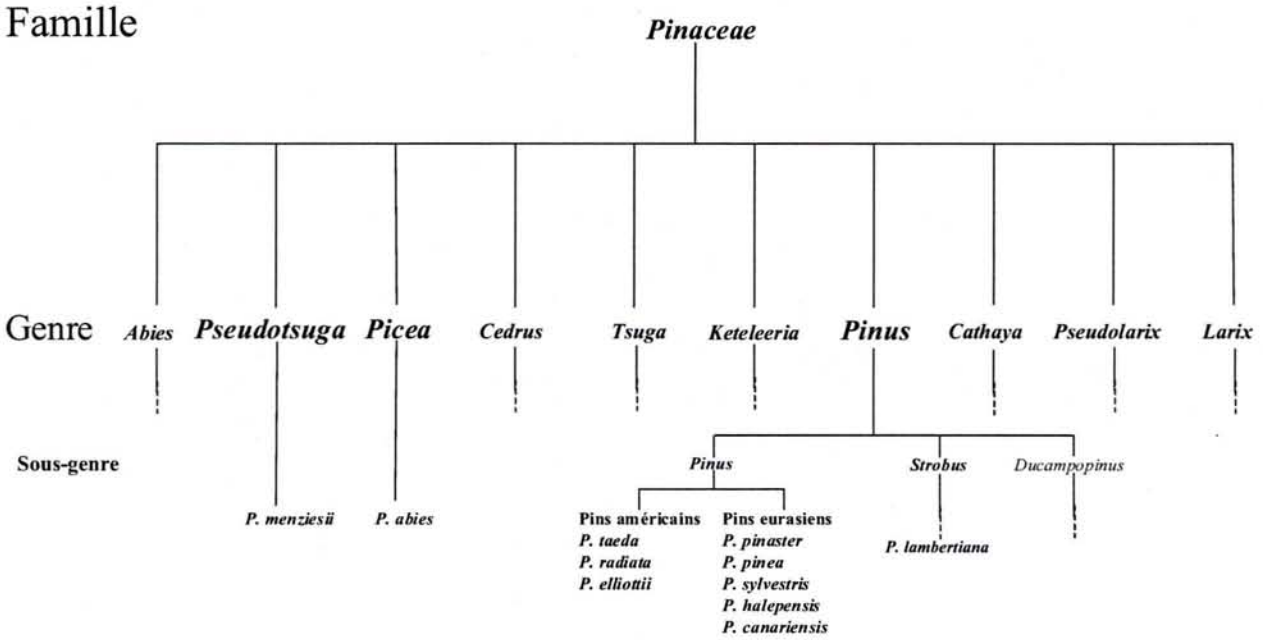


Figure A6 : Phylogénie simplifiée des Pinaceae : positionnement des espèces impliquées dans le projet de cartographie génétique comparée chez les conifères et dans le projet de développement de marqueurs microsatellites.

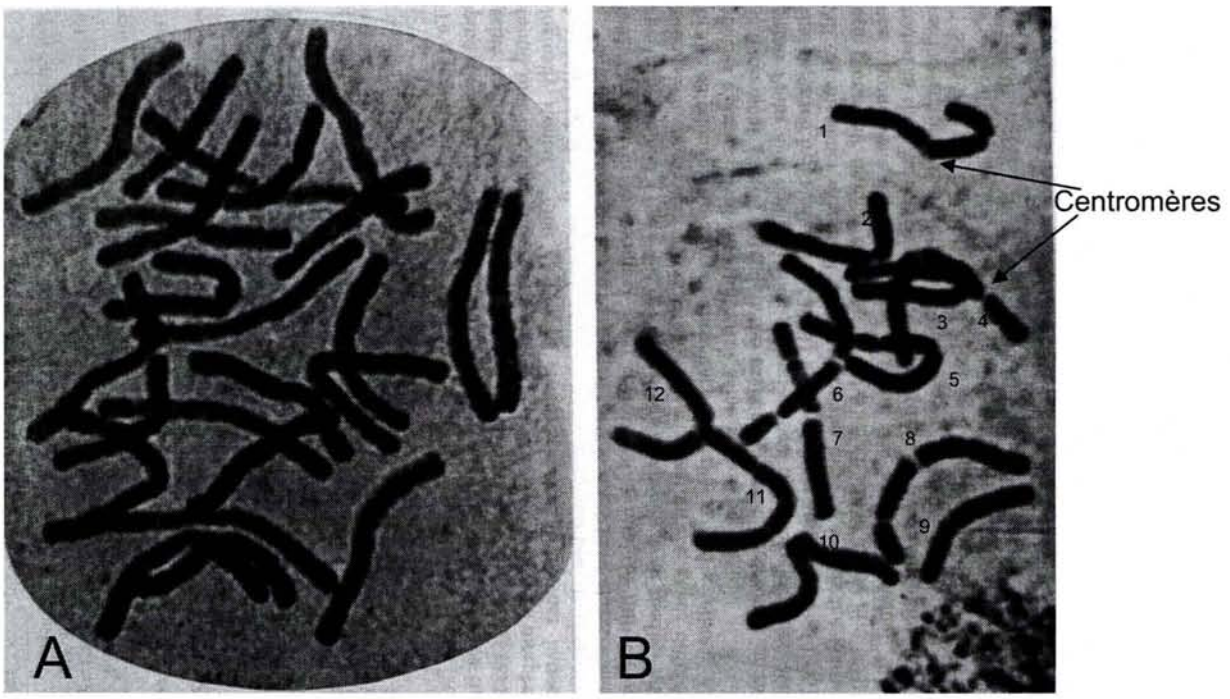


Figure A7 : Karyotypes du genre Pinus (planches tirées de Mirov 1967).
A/ *Pinus gerardiana* (2n = 24);
B/ *Pinus densiflora* (Endosperme, n = 12).
Noter la position médiane ou submédiane des centromères sur la plupart des chromosomes.

al. 1999, 2000). Ces classifications présentent des différences quant aux relations phylogénétiques des espèces entre elles. Il serait nécessaire dans le futur d'intégrer les différentes approches morphologiques et moléculaires afin d'obtenir une classification la plus précise et exacte possible. Beaucoup d'études restent donc à réaliser en matière de phylogénie des pins. Tous les auteurs cités s'accordent néanmoins pour dire que le genre *Pinus* est subdivisé en deux sous-genres : le sous-genre *Strobus* (ou haploxylon), appelé "*Soft pines*" en anglais, dont la plupart des espèces possèdent cinq aiguilles, et le sous-genre *Pinus* (ou diploxylon), "*Hard pines*", dont les espèces possèdent deux ou trois aiguilles. Les espèces de ces deux sous-genres ont la particularité de ne pas s'hybrider entre elles, ce qui souligne l'isolement génétique de ces deux groupes. Les sous-genres sont divisés en sections, subdivisées elles-mêmes en sous-sections. Pour simplifier les choses dans le cadre de cette thèse, je prendrai en compte un type de classification s'inspirant de la classification de Krupkin et al. (1996) et de Price et al. (1998), qui, à l'intérieur du sous-genre *Pinus*, distinguent les pins américains (Amérique du Nord et Mexique) des pins eurasiens (méditerranéens et asiatiques). Cette classification (figure A6), même si elle n'est pas parfaite, a un intérêt pratique et prend en compte des données moléculaires récentes.

A.1.4 Le génome des conifères

Dans les paragraphes qui suivent, comme tout au long de ce manuscrit, je limiterai le terme "génome" à celui du génome nucléaire, les autres génomes (chloroplastiques et mitochondriaux) n'étant pas en relation directe avec le sujet de cette thèse.

Les conifères sont caractérisés par un génome de grande taille, un nombre de chromosomes constant au sein des différentes familles et un karyotype très conservé entre espèces. De plus, très peu de conifères sont polyploïdes, ce qui semble montrer, au premier abord, que la structure du génome des gymnospermes est très simple. Néanmoins, les études portant sur le génome des conifères sont peu nombreuses comparées à celles réalisées chez les plantes de grande culture ou les animaux, et beaucoup d'information restent à obtenir.

Ferguson (1901) fut un précurseur dans l'étude du génome des pins car il a été le premier à observer un "comportement normal" du noyau cellulaire de *Pinus strobus* lors de la méiose. Sax et Sax (1933) ont par la suite étudié le karyotype d'espèces de pins et ont observé qu'ils

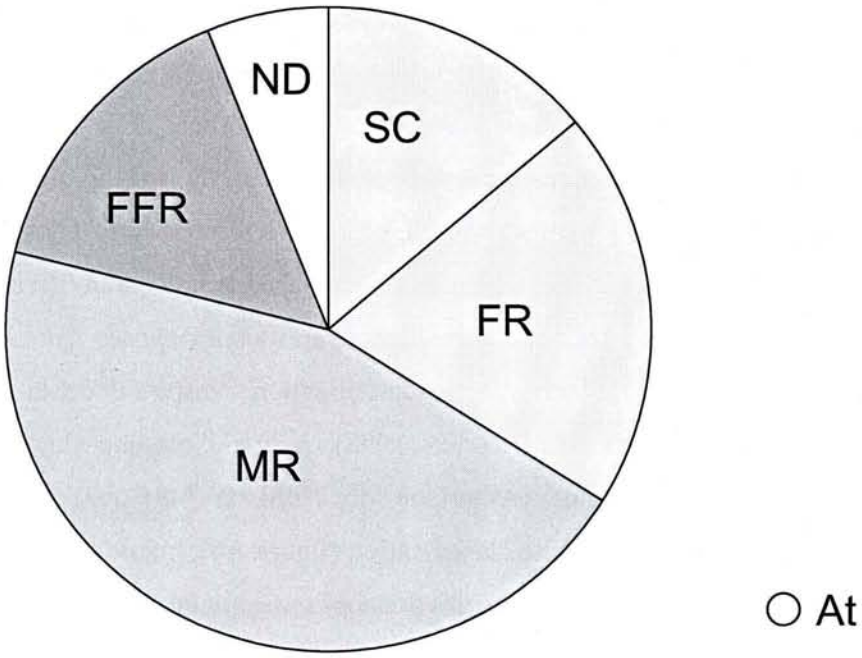


Figure A8 : Composition du génome des pins (d'après Elsik et Williams 2000) :

SC : fraction simple-copie (environ 4000 Mb)
FR : fraction faiblement répétée (moins de 100 copies)
MR : fraction moyennement répétée (entre 100 et 10 000 copies)
FFR : fraction fortement répétée (plus de 10 000 copies)
ND : non-défini (pas de réassociation de cette fraction)

At : génome d'*Arabidopsis thaliana* (taille proportionnelle)

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

avaient tous un nombre de chromosomes de 2n = 2x = 24. Saylor (1961) a montré qu'ils avaient la même morphologie (position médiane ou submédiane des centromères, figure A7). A ce niveau, aucune différence n'étant décelable entre les différentes espèces considérées, les auteurs suivants se sont employés à mesurer plus précisément la taille des chromosomes. Ainsi, Miksche (1967) a montré que les gymnospermes contenaient une grande quantité d'ADN en comparaison avec d'autres organismes, ce qui a été rediscuté plus tard par Ohri et Khoshoo (1986). Wakayima et al. (1993) ont observé des corrélations significatives entre la taille du génome des conifères et des caractères liés à l'adaptation de ces espèces à leur habitat (température, précipitations). Chez le pin maritime, par exemple, la taille physique est de 25,5 pg/C (soit environ 25 000 mégabases, Mb, annexe I), ce qui est bien supérieur à la taille du génome humain (3 pg/C) et à celle d'*Arabidopsis thaliana* (0,15 pg/C). Des études de cinétique de réassociation de l'ADN réalisées par Rake et al. (1980) et Kriebel (1985) ont montré qu'à l'intérieur de ce "mégagénome", il y avait une faible portion d'ADN codant (0,1%, à comparer aux 3% chez l'homme) et une grande proportion (76%) de séquences hautement répétées. Ces données ont été remesurées à l'aide de la même méthode sur *Pinus strobus* par Elsik et Williams (2000) et les résultats, divisés en 4 classes selon le nombre de copies, sont présentés à la figure A8. Ces auteurs ont de plus montré que la partie peu répétée du génome des pins, qui à elle seule est supérieure à la taille physique du génome d'Arabidopsis, contient de nombreux rétrotransposons[1].

Seuls quelques auteurs ont caractérisé les régions répétées d'espèces de conifères. Kamm et al. (1996) ont démontré la présence de rétrotransposons de la famille des *TPE*1 (éléments de type *Ty*1-*copia*) sur les 12 chromosomes de *Pinus elliottii* en hybridant des sondes marquées de manière fluorescente. De même, Brown et al. (1993, 1997, 1998) et Lubaretz et al. (1996) ont montré la présence de familles d'ADN répété (ADN ribosomaux et *Sau*3A) chez deux espèces d'épicéa. Murray et al. (1998), Kossack et Kinlaw (1999) et Schmidt et al. (2000) ont aussi montré la présence de grandes quantités de rétroéléments (LINE, *gypsy*, *copia*) dans les régions répétées du génome des gymnospermes.

Certains auteurs ont également souligné l'importance des familles multigéniques dans la partie codante du génome des conifères (Perry et Furnier 1996, Kinlaw et Neale 1997). Ces derniers ont montré que les régions codantes sont elles-mêmes répétées et que la plupart des gènes étaient présents en plusieurs copies dans le génome. Le développement des marqueurs

[1] Rétrotransposon : un transposon créé par rétrotranscription d'une molécule d'ARN.

moléculaires a permis de cartographier des marqueurs de type RFLP[2] sur les chromosomes de pin (Neale et Williams 1991), et de réaliser des cartes génétiques pour ces espèces (Devey et al. 1991). Le développement de nombreux types de marqueurs moléculaires couvrant la totalité du génome a permis de construire une première carte génétique saturée pour une espèce de pin (Plomion et al. 1995a). D'autres cartes saturées de conifères ont alors suivi celle-ci (voir chapitre B.2).

Outre les quelques références citées ci-dessus, on connaît très peu de choses sur le génome des conifères. En particulier aucun projet de séquençage complet et de cartographie physique n'a été envisagé du fait de la taille de leur génome. Des banques BAC très partielles de *Picea abies* (Morgante, communication personnelle), *Pinus taeda* (C. Kinlaw et D. Neale, communication personnelle) et *Pinus pinaster* (F. Canovas, communication personnelle) ont cependant été initiées. Ce contexte montre l'utilité de la cartographie génétique comparée pour la poursuite de la compréhension de la structure du génome des conifères.

A.1.5 Contexte "politico-scientifique" de la recherche en génomique chez les conifères

Contrairement au cas de l'homme, des animaux d'élevage ou des plantes de grandes cultures, la communauté scientifique s'intéressant à l'étude du génome d'espèces forestières, et en particulier à celui des conifères, est réduite. De plus, cette communauté est caractérisée par un nombre important d'espèces étudiées : la règle étant souvent qu'un laboratoire s'intéresse préférentiellement aux espèces qui ont une importance économique et écologique dans la région où il se trouve. Cette diffusion des forces pourrait être néanmoins comblée par la cartographie génétique comparée. En effet, cette discipline a pour but de rassembler les données génétiques obtenues chez des espèces apparentées pour pouvoir transférer des informations (position de gènes et de QTL[3]) entre ces espèces. De plus, dans le cas des conifères, elle peut pallier au manque d'espèce de référence en permettant de considérer ces espèces comme un système génétique unique. De plus, les études antérieures sur le génome des conifères présentées dans le chapitre précédent montre que la cartographie comparée peut être efficace chez les conifères, du fait par exemple de la conservation du nombre de chromosomes chez les pins.

---

[2] RFLP : *Restriction Fragment Length Polymorphism*. Technique basée sur les différences au niveau d'un site de restriction et révélée sur gel d'électrophorèse.

3 QTL : *Quantitative Trait Locus*. Région chromosomique contrôlant un caractère quantitatif.

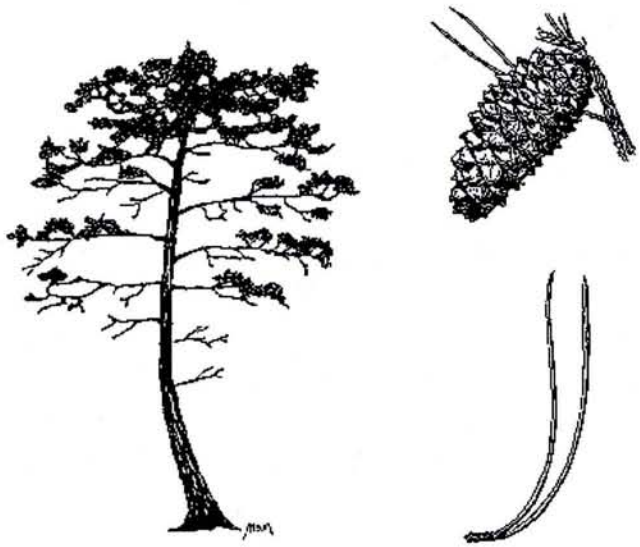


Figure A9 : le pin maritime (*Pinus pinaster* Ait.) : port général, aiguilles et cône.

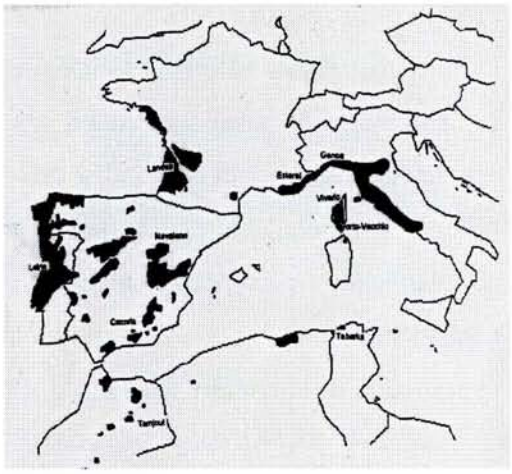


Figure A10 : Aire de distribution du pin maritime

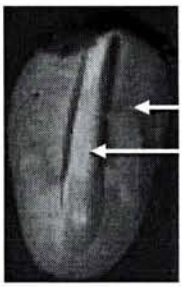


Figure A11 : Coupe d 'une graine de pin maritime
1: Endosperme
2: Embryon

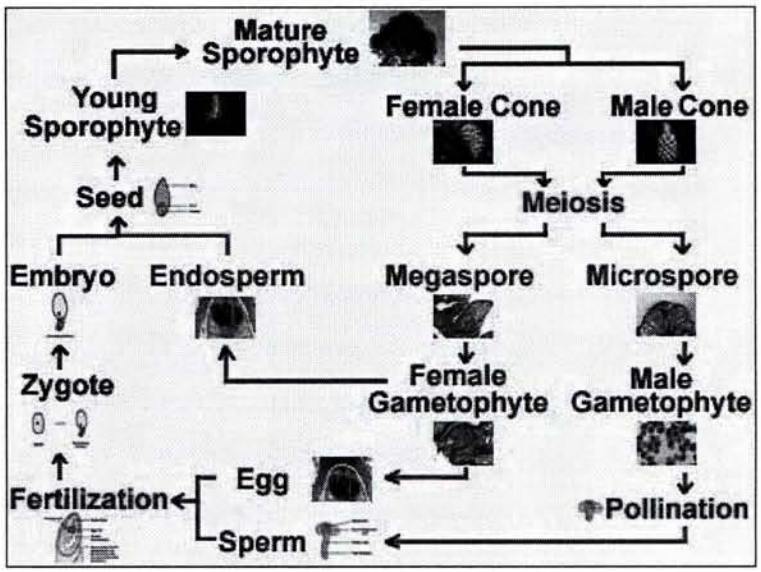


Figure A12 : Cycle de reproduction du pin maritime
(source : http://fybio.bio.usyd.edu.au/vle/L1/ResourceCentre/GraphicFiles/LCPine/LCPine.html)

A.2 L'espèce étudiée : le pin maritime

A.2.1 Présentation générale

Le pin maritime (*Pinus pinaster* Ait., figure A9) est une espèce autochtone du bassin méditerranéen occidental et de la façade atlantique du Maroc à la Bretagne (figure A10). Jeune, ce conifère a une ramure pyramidale qui devient plus ou moins étalée mais jamais aplatie. Haut de 40 m, il a un tronc droit ou un peu incurvé. L'écorce, épaisse même chez les individus jeunes, est d'un brun violacé et se détache en plaques. Les aiguilles, robustes, pouvant mesurer jusqu'à 20 cm de long, droites ou légèrement recourbées, sont groupées par deux dans une gaine longue de 2,5 cm et sont beaucoup plus denses à l'extrémité des rameaux. Les cônes, sont de forme allongée lorsqu'ils sont fermés, droits ou légèrement courbés et groupés par deux, ou plus, autour du rameau. Le pin maritime est une espèce monoïque, il possède donc des fleurs mâles et femelles distinctes sur le même arbre. Les graines ovales, aplaties, sont longues de 5 à 10 mm, noires d'un côté et gris taché de l'autre. Au sein de la graine, l'embryon est entouré d'un tissu nourricier appelé mégagamétophyte (figure A11). Ce tissu haploïde, formé lors de la formation de l'ovule avant la pollinisation, est d'origine maternelle. La figure A12 présente le cycle de reproduction du pin maritime.

Du point de vue écologique, le pin maritime présente une bonne adaptation vis-à-vis de conditions contrastées telles que des sols sableux acides et pauvres, soit gorgés d'eau en hiver (landes humides), soit très secs en été (landes sèches et dunes), et des hivers froids.

Le pin maritime est le conifère le plus utilisé en reboisement dans le sud-ouest de l'Europe où il représente 4 millions d'hectares, soit environ un millième de la forêt mondiale. Comparées au chiffre de la surface occupée, les données relatives à la récolte et aux sciages au niveau mondial (6‰ de la récolte et 1,2% des résineux) montrent que cette espèce est relativement bien utilisée et transformée. Ses utilisations principales concernent la production de bois d'œuvre (parquets, lambris, moulures, meubles, contreplaqué, charpentes et palettes) et de bois d'industrie (pâte et papier, panneaux de particules ou de fibres).

Le pin maritime est une composante majeure de la sylviculture française, où il représente 1,5 million d'hectares, soit 10% de la surface forestière nationale. Avec une productivité moyenne de 9 m^3/ha/an, il représente 30% de la production nationale de bois de résineux. En Aquitaine, il a non seulement un intérêt écologique (région française au taux de boisement le plus élevé : 45%) mais aussi un poids économique indiscutable. Historiquement, le pin maritime a été

planté au XIXème siècle dans les Landes pour fixer les dunes et assainir les marais. Cette initiative proposée par Bremontier à la fin du XVIIIème siècle, puis appuyée par la suite par une loi promulguée par Napoléon III, imposant aux communes de reboiser leurs "vacants", a conduit à augmenter la surface forestière des Landes de Gascogne pour dépasser un million d'hectares au début du XXème siècle.

Le pin maritime est également utilisé pour reboiser des surfaces importantes dans diverses régions du globe : par ordre d'importance des surfaces plantées on le retrouve en Australie, au Chili, en Afrique du Sud et en Nouvelle-Zélande.

A.2.2 Diversité génétique du pin maritime au sein de son aire de répartition

L'aire de distribution du pin maritime est discontinue et s'étend sur la façade atlantique du sud-ouest de l'Europe, de la Bretagne au Portugal, en passant par la côte Aquitaine et la Galice, et sur le pourtour méditerranéen, en Corse, en Italie, en Espagne, et au Maghreb (figure A10). Des tests de provenance réalisés par Illy (1966) ont permis de démontrer des caractéristiques différentes pour les différentes régions. Par exemple, on observe des contrastes marqués entre les populations portugaises et landaises en terme de résistance au froid. De même, les populations corses sont caractérisées par une meilleure rectitude du fût et une croissance plus lente que les populations landaises. On observe aussi que certaines populations sont très sensibles aux pathogènes, comme par exemple à l'insecte *Matsucoccus feytaudii* (Harfouche et al. 1995), alors que d'autres, comme par exemple les marocaines, sont résistantes.

Certains auteurs se sont employés à décrire la structure de la diversité génétique des populations de pin maritime à l'aide de plusieurs types de marqueurs, comme les terpènes (Baradat et Marpeau 1988), les protéines (Barhman et al. 1994), les isoenzymes (Petit et al. 1995, Salvador et al. 2000, Gonzalez-Martinez et al. 2001) et les marqueurs fondés sur l'ADN chloroplastique (Vendramin et al. 1998, Ribeiro et al. 2001, Burban et Petit 2003), mitochondrial (Burban et Petit 2003) et nucléaire (Mariette et al. 2001). La plupart des études concordent pour dire que l'aire de distribution du pin maritime est divisée en trois grands écotypes : un écotype atlantique, un écotype méditerranéen et un écotype maghrébin (figure A13). Ces trois écotypes proviendraient de trois grands refuges d'où a migré le pin maritime après la dernière période glacière. On peut noter également que ces trois écotypes, de par la forte fragmentation de l'aire de répartition du pin maritime, sont fortement différenciés (G_{st} variant de 0,15 à 0,20, Petit et al. 1995).

Figure A13 : Distribution des haplotypes mitochondriaux de pin maritime (gène *NAD*1, figure extraite de Burban et Petit 2003)

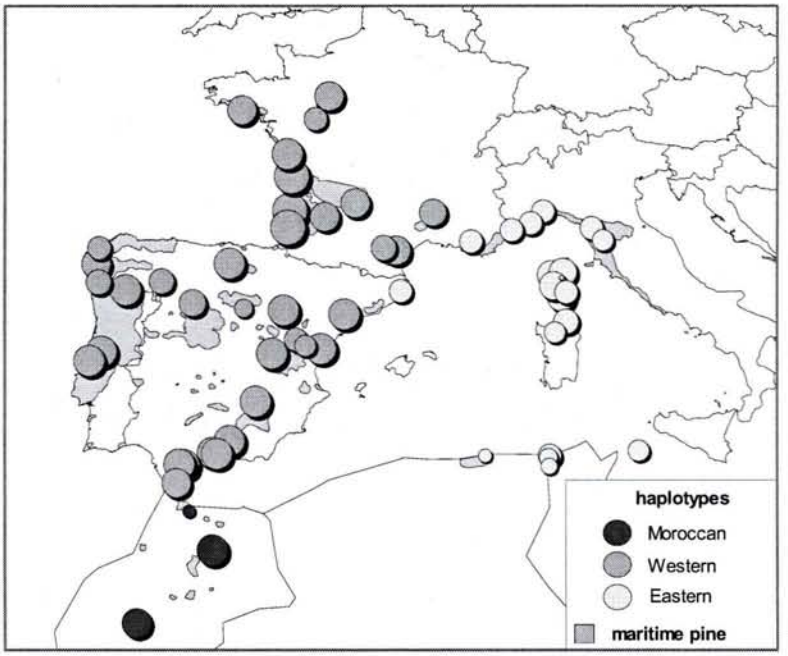
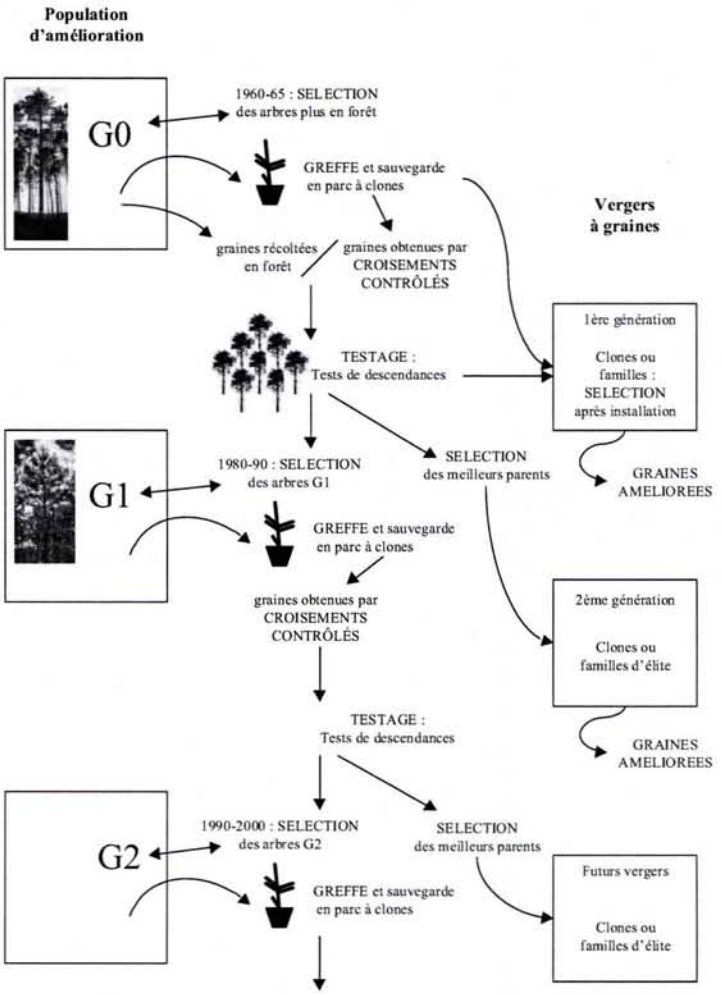


Figure A14 : Schéma présentant le programme d'amélioration du pin maritime.

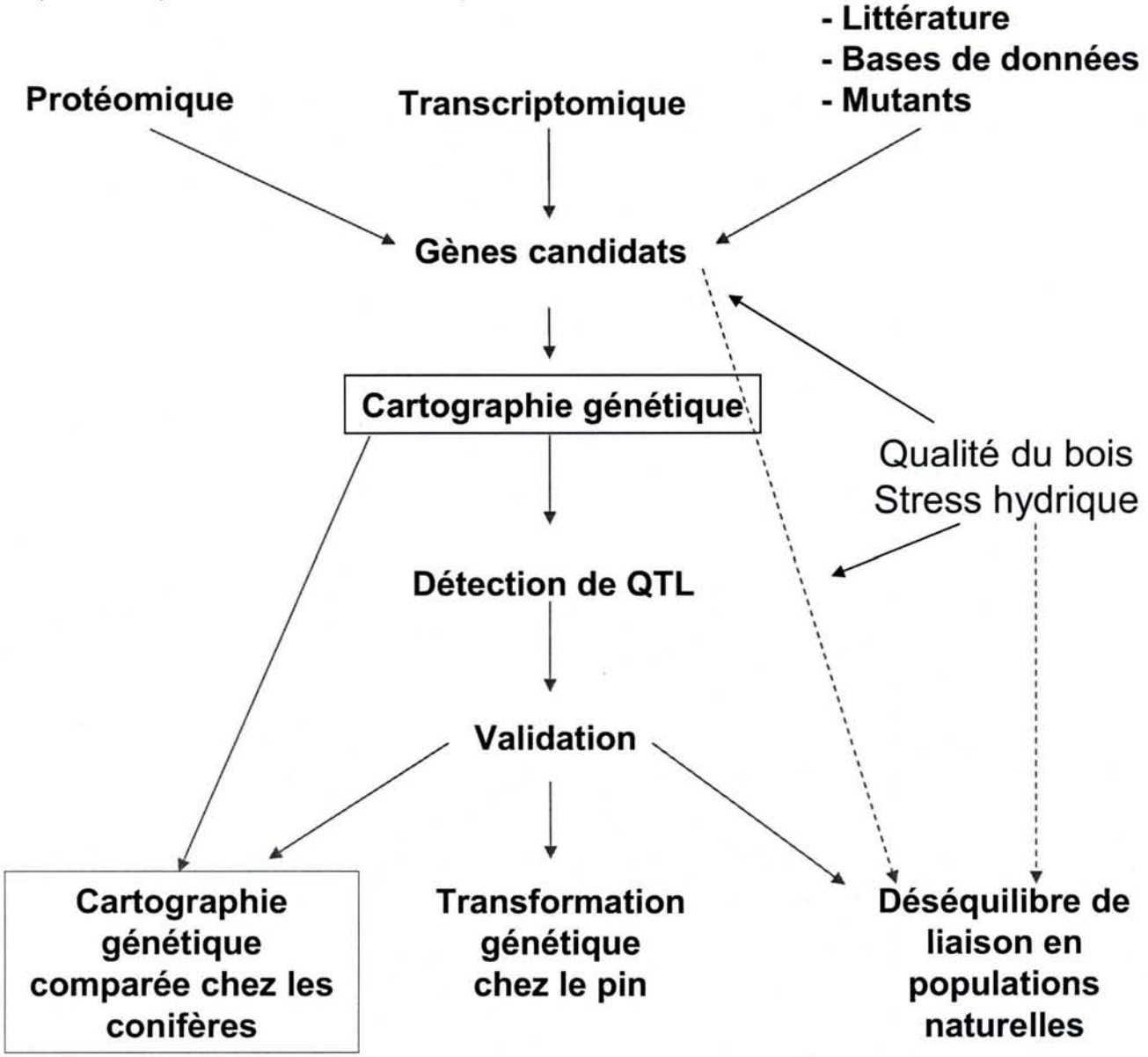


Un autre aspect important au niveau de la biodiversité du pin maritime est que les forêts de pin maritime sont particulièrement anthropisées. L'utilisation massive de cette espèce pour reboiser les Landes de Gascogne au cours des XIXème et XXème siècles en est un exemple marquant. Cette anthropisation des forêts de pin maritime doit être prise en compte afin de préserver la diversité de l'espèce. A l'heure actuelle, des changements de la structure génétique des populations de pin maritime peuvent être attribuables à l'action de l'homme, allant parfois jusqu'à des catastrophes naturelles, comme par exemple la perte de 100 000 ha de pin en hiver 1985 qui a principalement touché des peuplements constitués de pins portugais implantés après les grands incendies des années 40. Le choix des individus qui constitueront les peuplements de pin maritime de demain doit donc être fait de manière à concilier l'amélioration génétique de l'espèce et le maintien de sa diversité naturelle. Ainsi, le programme d'amélioration génétique du pin maritime doit non seulement tenir compte des critères de productivité et de qualité, mais aussi de la diversité afin de garder un potentiel suffisant d'adaptabilité de l'espèce en vue des changements climatiques annoncés.

A.2.3 Le programme d'amélioration génétique du pin maritime

L'INRA a démarré un programme d'amélioration génétique du pin maritime dans les années 60. Le schéma de ce programme de sélection est présenté à la figure A14. Une sélection massale d'"arbres plus" sur des critères de forme et de vigueur a été effectuée en forêt à travers le massif landais. Un total de 380 arbres élites a permis de constituer la première génération d'amélioration "G0". Deux cent nouveaux G0 ont été rajoutés dans les années 70 suite à une nouvelle campagne de sélection étendue à une zone périphérique de la zone landaise. Des critères de densité du bois ont alors été pris en compte en plus des critères précédents. D'autres campagnes ont été réalisées par la suite pour aboutir à un total de 635 G0. Ces arbres ont été greffés dans des parcs à clones. Ils ont été ensuite croisés entre eux de façon contrôlée, ce qui a abouti à une deuxième génération d'amélioration, composée de 1100 arbres "G1", également conservés en parcs à clones par greffage. Enfin, une génération "G2" a été obtenue en croisant des individus G1. Des gains génétiques considérables (+30%) ont été obtenus pour la vigueur et la forme des arbres (Pastuszka et al. 2002). Une optimisation des gains génétiques par unité de temps pourrait encore être obtenue en utilisant des prédicteurs précoces des critères cibles de la sélection. Pour cela, il faudra réduire le temps de sélection (actuellement de 12 ans) tout en augmentant l'héritabilité des prédicteurs. La mise en place d'un schéma de sélection assistée par marqueurs est envisagée pour des caractères liés à

Figure A14 : schéma du programme de génomique du pin maritime
(encadré : partie traitée dans la thèse)



la qualité du bois, ce qui a motivé l'initiation d'un programme de génomique chez cette espèce.

A.2.4 Le programme de génomique du pin maritime

Le programme de génomique du pin maritime est présenté à la figure A15. L'objectif de ce programme est d'identifier les gènes impliqués dans le contrôle de caractères important comme la qualité du bois ou l'adaptation de l'espèce aux contraintes de l'environnement (en particulier la réponse à un déficit d'alimentation en eau). Ce programme suit une approche partant du caractère complexe pour aboutir aux gènes contrôlant ce caractère.

Dans un premier temps des gènes et des protéines candidats sont identifiés par les méthodes d'études d'expression différentielle au niveau du transcriptome et du protéome. Ceci a constitué, par ordre chronologique, le travail de thèse de P. Costa (1999), C. Dubos (2001), G. Le Provost (2003), J. Paiva (en cours) et P. Chaumeil (en cours). Dans un souci de synthèse, et les méthodes mises en jeu étant les mêmes, je me limiterai à évoquer les méthodes en prenant l'exemple de la qualité du bois. Les mêmes méthodes ayant été appliquées pour l'étude de la réponse au stress hydrique. Les protéines exprimées différentiellement dans des types de bois distincts ont été identifiées grâce à l'utilisation de gels d'électrophorèse bidimentionnelle (Plomion et al. 2001). En parallèle, les gènes exprimés dans les différents types de bois ont été identifiés par la technique de ADNc-AFLP (Le Provost et al. 2003). La disponibilité de techniques puissantes comme les réseaux d'ADNc[4] à haute densité (*microarrays*[5]) et le séquençage systématique d'un grand nombre de gènes transcrits (*EST*[6]) permettent maintenant d'accéder à une vue plus exhaustive des gènes mis en jeu dans des conditions ou des phases de développement précis (Canton et al. 2004). Une fois les gènes candidats identifiés (on parle souvent de "candidats expressionnels"), il convient de vérifier s'ils sont impliqués dans les caractères d'intérêt par des approches fondées soit sur la colocalisation avec leur représentation discrète sur les chromosomes, les QTL (*Quantitative Trait Locus*), détectés sur des cartes génétiques, ou soit par des études d'association en populations naturelles. Ces études ont fait l'objet des thèses de P. Costa (1999), de D. Pot (2004) et de E. Eveno (en cours).

Ce travail de thèse qui concerne la mise au point de marqueurs moléculaires, l'établissement d'une carte génétique du pin maritime et la cartographie comparée chez les conifères, s'intègre

[4] ADNc : ADN complémentaire obtenu par rétrotranscription d'un fragment d'ARN.
[5] *Microarrays* : puces à ADNc servant par exemple à identifier des gènes différenciellement exprimés dans certaines conditions.
[6] EST : *Expressed Sequence Tags*. Fragment de séquence exprimée.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHEQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

à ce programme de génomique, à l'intersection de l'identification des gènes et des régions chromosomiques (QTL) impliquées dans le contrôle génétique de ces caractères.

A.3 Situation de la thèse, objectifs et problématiques de la thèse

Les principaux objectifs de ce travail de thèse étaient de développer des marqueurs moléculaires potentiellement orthologues pour réaliser une carte génétique du pin maritime, et la comparer avec celles d'autres espèces de conifères.

A.3.1 Construction d'une carte génétique du pin maritime

Avant de construire une carte génétique, il convient de choisir un bon pedigree ainsi qu'une méthode de cartographie adéquate. Le pedigree utilisé doit être assez informatif pour pouvoir cartographier un maximum de locus. Dans le cadre de cette thèse, deux pedigrees de pin maritime ont été utilisés : le premier (F2), à partir duquel une carte génétique avait déjà été construite (Plomion et al. 1995a), résulte de l'autofécondation d'un hybride issu du croisement entre un pin maritime de « race » corse et un pin maritime de « race » landaise. Néanmoins, la taille de ce pedigree et l'âge des descendants ne permettaient pas d'envisager la détection de QTL pour des caractères adultes comme la qualité du bois. C'est pour cela qu'une autre population de cartographie a été employée. Le second pedigree utilisé (G2) est une famille de plein-frères (adultes) résultant du croisement de quatre grands-parents de « race » landaise, choisis pour leurs différences selon des critères de vigueur et de forme.
Les stratégies de cartographie utilisée pour ces deux pedigrees sont différentes : pour le croisement F2, la méthode de cartographie utilise le tissu nourricier haploïde d'origine maternelle (mégagamétophyte). En ce qui concerne le croisement G2, la stratégie utilisée est celle du double-pseudo testcross (Grattapaglia et Sederoff 1994). Le principe de ces deux méthodes sera présenté au chapitre B.2.2.
Deux autres descendances de plein-frères de pin maritime ont été indirectement utilisées : 1/ le pedigree "UHD-MAP" (http://www.neiker.net/UHDfor/), développé par l'AFOCEL dans le cadre d'un projet européen, a permis de construire une autre carte génétique du pin maritime (Ritter et al. 2002). 2/ Un autre pedigree développé par l'INIA (Madrid, Espagne), dont la carte est en cours de construction (M. Cervera, communication personnelle). A court terme les quatre cartes seront alignées afin de réaliser une carte consensus du pin maritime.

A.3.2 Cartographie génétique comparée chez les conifères

Il est possible d'aligner les cartes génétiques d'espèces phylogénétiquement proches à l'aide de marqueurs communs (orthologues[7]). Ceci fait l'objet de la cartographie génétique comparée. Ainsi, un des objectifs de cette thèse est d'aligner la carte génétique du pin maritime avec celles d'autres espèces de la famille des Pinaceae. Etant donné que peu de connaissances ont été acquises sur la structure du génome des conifères, mais que des cartes génétiques sont disponibles pour certaines de ces espèces, la cartographie comparée parait prometteuse pour améliorer ce peu de connaissances. Si la composition et l'ordre des gènes sont conservés entre espèces, on pourra éventuellement déduire que le génome des conifères n'a pas évolué ou bien n'a pas connu de grands remaniements en terme de réarrangements chromosomiques depuis leur divergence. Le deuxième objectif s'inscrit dans un cadre plus appliqué de comparaison des cartes génétiques d'espèces de pin pouvant servir à "décloisonner" les études qui sont faites en parallèle chez ces différentes espèces. Aucune espèce modèle n'existant pour les conifères, la possibilité d'aligner leurs cartes génétiques permettra de vérifier ou de prédire la position d'un QTL ou d'un gène candidat.

A.3.3 Développement de marqueurs

Les deux objectifs précédents ont nécessité de développer des marqueurs moléculaires adaptés. Dans le cas de la cartographie génétique, les marqueurs doivent être rapides à mettre au point et en grand nombre, et de préférence répartis sur l'ensemble du génome. Un autre critère à prendre en compte est que les cartes génétiques ont comme application la détection de QTL. Ils doivent donc être le plus informatifs possible (codominance, multiallélisme).
Si possible les locus cartographiés doivent être transférables entre espèces ou tout simplement entre pedigrees différents. La transférabilité d'un type de marqueurs implique qu'il soit ou non utilisable en tant que marqueur orthologue. Les différents types de marqueurs utilisés dans le cadre de cette thèse sont décrits dans la partie B.1 : ce sont des marqueurs de type AFLP, des marqueurs basés sur des EST, des microsatellites ou des SNP.

[7] Orthologue : Deux locus sont qualifiés d'orthologues s'ils dérivent d'un ancêtre commun aux deux espèces auxquelles ils appartiennent et s'il n'y a pas eu de duplication de ce locus, auquel cas on peut avoir affaire à des locus paralogues.

A.3.4 Problématique

Au-delà des objectifs présentés ci-dessus, les questions auxquelles cette thèse tentera de répondre sont les suivantes.

Pour la cartographie génétique du pin maritime :
Quels types de marqueurs peuvent être utilisés pour construire une carte génétique saturée du pin maritime ?
Comment se répartissent-ils sur le génome ?

Pour la cartographie génétique comparée chez les conifères :
Peut-on aligner la carte génétique du pin maritime avec celles d'autres conifères ?
La structure du génome des conifères est-elle conservée entre espèces ?
Peut-on transférer des informations entre les cartes génétiques d'espèces de pin et si oui, quel type d'informations ?
Quels types de marqueurs peuvent servir comme marqueurs orthologues ?
Quels types de marqueurs peuvent être utilisés comme gènes candidats ?

Pour le développement de marqueurs moléculaires :
Quelle est la méthode qui peut être employée pour développer des marqueurs microsatellites chez le pin maritime ?
Quels sont les meilleurs marqueurs pour la carographie génétique ou la cartographie comparée ?

METHODES MISES EN JEU

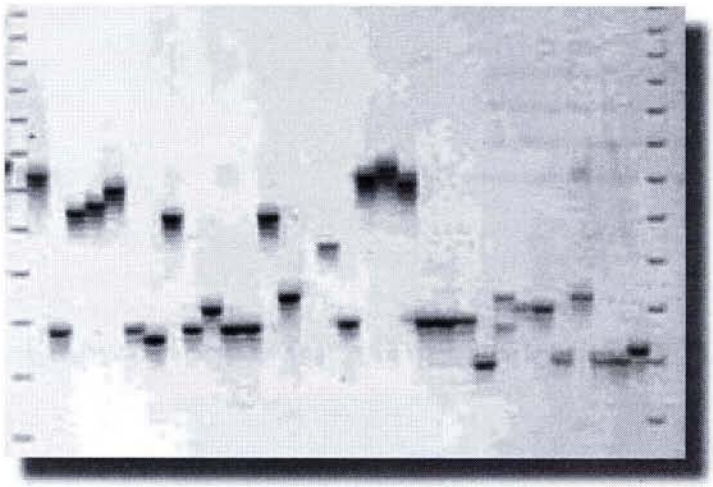
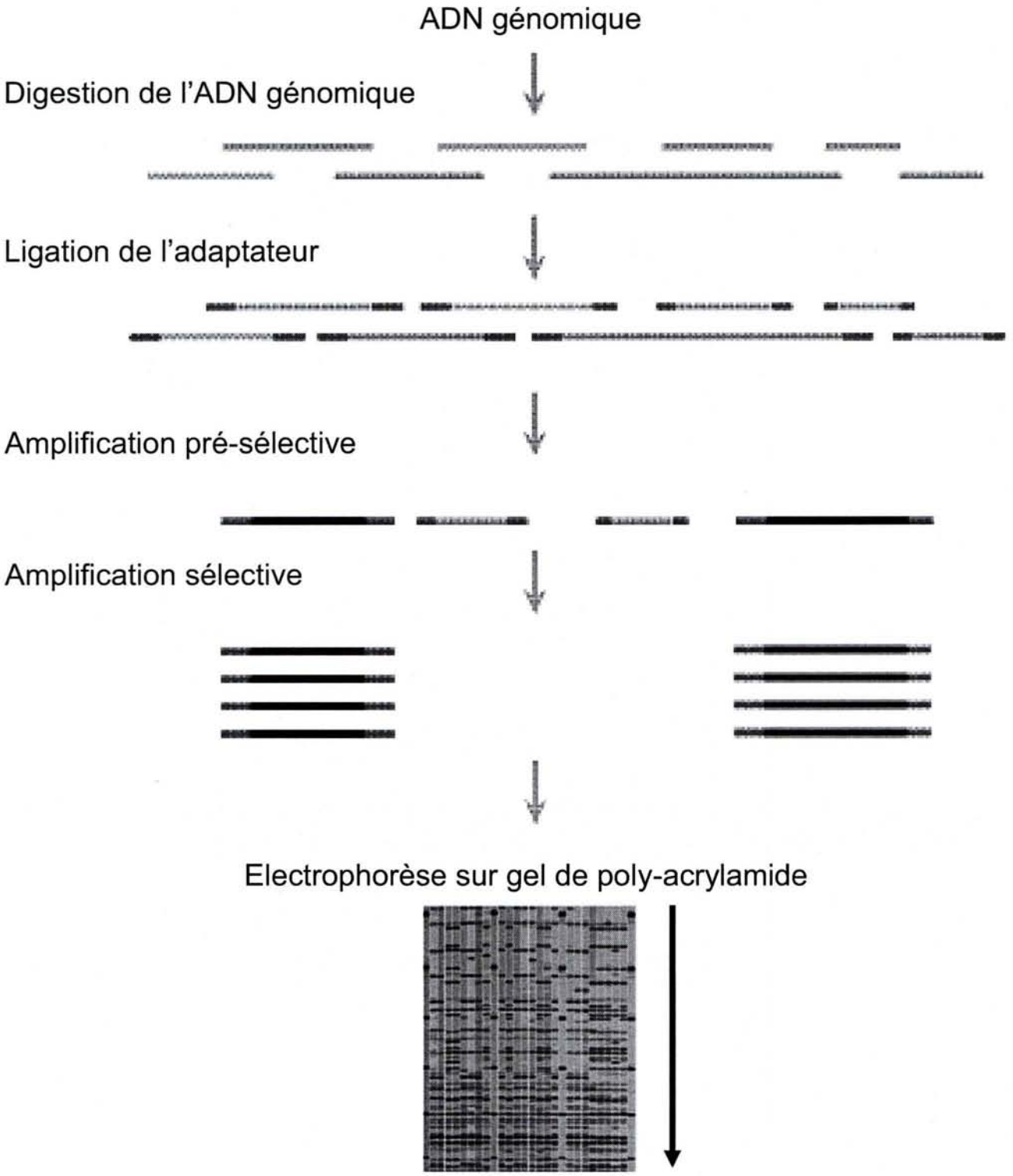


Figure B1 : Principales étapes de la technique AFLP



Partie B : Méthodes mises en jeu

B.1 Les marqueurs moléculaires

Les marqueurs moléculaires sont des marqueurs génétiques basés sur l'ADN, par comparaison avec les marqueurs morphologiques (ex: couleur des yeux), biochimiques (ex: composition en terpènes de l'aiguille de pin) ou protéiques (ex: isoenzymes). Ils sont présents en quantité presque infinie dans le génome. On peut les classer selon qu'ils sont anonymes ou de séquence connue, codants ou non-codants, révélés en masse ou simple-copie, basés sur des variations de longueur ou de séquence, dominants ou codominants, plus ou moins polymorphes ou utilisant l'hybridation d'ADN ou l'amplification par la technique PCR (de Vienne 1998). De ces critères découlent les approches utilisées pour leur développement et les méthodes moléculaires employées pour les révéler. Je me limiterai ici aux techniques employées dans le cadre de cette thèse.

B.1.1 Marqueurs anonymes révélés en masse : les marqueurs AFLP

Les marqueurs révélés en masse sont très utilisés car ils fournissent de nombreux locus par réaction. Outre les AFLP (*Amplified Fragment Length Polymorphism*), les plus utilisés sont les RAPD (*Random Amplified Polymorphic DNA*, Williams et al. 1990). La quantité de locus qu'ils fournissent est suffisante pour saturer l'ensemble du génome sans nécessiter au préalable de connaissance sur la séquence du génome. De plus, ils sont techniquement faciles et rapides à mettre en œuvre. Le plus gros inconvénient est qu'ils sont souvent difficilement reproductibles (surtout pour les RAPD) et non transférables entre espèces.

B.1.1.1 Description des marqueurs AFLP

La traduction littérale d'AFLP est "polymorphisme de longueur de fragments amplifiés". Elle est fondée sur l'amplification sélective de fragments de restriction d'ADN génomique. Cette méthode a été décrite pour la première fois par Vos et al. (1995). Brièvement, le principe, présenté à la figure B1, en est le suivant : l'ADN génomique est tout d'abord digéré en utilisant des enzymes de restriction (souvent *Eco*RI, *Pst*I ou *Mse*I), puis on ligue un adaptateur spécifiques des sites de restriction aux fragments générés. Une première

Tableau B1 : Liste des paires d'amorces AFLP utilisées pour construire les cartes génétiques des parents du pedigree G2. Voir annexe I, table I p630.

	Combinaison amorce/enzyme	Nombre de fragments amplifiés	Nombre de marqueurs 1:1 (a)	Nombre de marqueurs 3:1 (b)	Total (a+b)
1	E+ACA/M+CCAG	140	10	7	17
2	E+ACA/M+CCGA	130	9	7	16
3	E+ACG/M+CCGC	108	8	1	9
4	E+ACG/M+CCAG	60	9	5	14
5	E+ACG/M+CCGT	95	10	1	11
6	E+ACG/M+CCTA	56	4	4	8
7	E+ACG/M+CCCA	112	5	3	8
8	E+ACG/M+CCAA	118	12	1	13
9	E+ACG/M+CCTG	62	26	3	29
10	E+ACC/M+CCAG	140	13	5	18
11	E+ACC/M+CCTG	126	9	5	14
12	E+ACC/M+CCGT	70	7	1	8
13	E+ACC/M+CCTA	145	10	4	14
14	E+ACC/M+CCGA	110	8	2	10
15	E+ACT/M+CCGC	130	9	2	11
16	E+ACT/M+CCAG	110	11	1	12
17	E+ACT/M+CCTG	120	14	7	21
18	E+ACT/M+CCGT	130	7	5	12
19	E+ACT/M+CCCA	136	4	4	8
20	E+ACT/M+CCTA	105	9	6	15
21	E+ACAA/M+CCTA	95	6	9	15
22	E+ACAA/M+CCAC	100	6	5	11
23	E+ACAA/M+CCGC	140	9	3	12
24	E+ACAA/M+CCCA	140	12	5	17
25	E+ACAA/M+CCGA	110	4	5	9
26	E+ACAA/M+CCTT	136	17	4	21
27	E+ACAA/M+CCTG	70	9	4	13
28	E+ACAA/M+CCAG	75	10	4	14
29	E+ACAA/M+CCAT	110	12	4	16
30	E+ACAC/M+CCAA	100	9	2	11
31	E+ACAC/M+CCAT	130	16	10	26
32	E+ACAC/M+CCTA	100	7	2	9
33	E+ACAC/M+CCTT	100	12	8	20
34	E+ACAC/M+CCTC	90	9	3	12
35	E+ACAC/M+CCAG	123	9	3	12
36	E+ACAC/M+CCAC	100	13	4	17
37	E+ACAG/M+CCTG	114	15	4	19
38	E+ACAG/M+CCTA	107	9	3	12
39	E+ACAG/M+CCAT	104	18	9	27
40	E+ACAG/M+CCAA	99	14	3	17
41	E+ACAG/M+CCGA	120	6	6	12
42	E+ACAG/M+CCTC	110	5	7	12
43	E+ACAG/M+CCGT	130	3	6	9
44	E+ACAG/M+CCGC	145	9	5	14
45	E+ACAT/M+CCAG	115	10	7	17
46	E+ACAT/M+CCTA	110	15	9	24
47	E+ACAT/M+CCAT	132	7	6	13
48	E+ACAT/M+CCTC	143	20	9	29
49	E+ACAT/M+CCTG	105	2	6	8
50	E+ACAT/M+CCAC	138	13	9	22
51	E+ACAT/M+CCCA	145	6	8	14
52	E+ACAT/M+CCGA	115	7	7	14
TOTAL	-	5854	513	253	766

Figure B2 : Exemple de profil de gel AFLP. Les deux premières lignes contiennent les deux parents du croisement G2 et les lignes suivantes contiennent des descendants plein-frères.

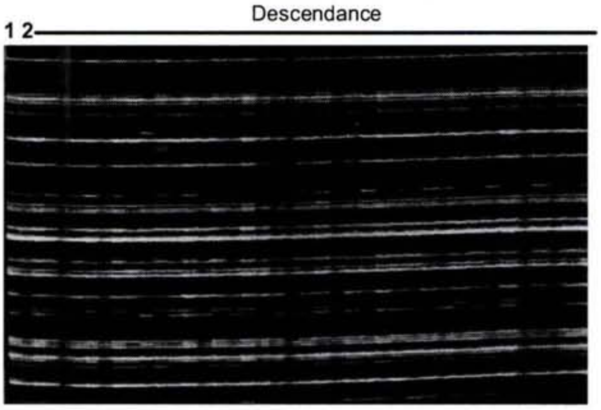


Figure B3 : Exemples de motifs microsatellites.

...AGCTGCGATTCAGTCGTACAGTCCAATGCTACAG**ATATATATATATATATATATATATATATATATATATATATATAT**GCTGACCTGGTGGTAACTG...

...AGCTGCGATTCAGTCGTACAGTCCAATGCTACAG**AAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAG**GCTGACCTGGTGGTAACTG...

...AGCTGCGATTCAGTCGTACAGTCCAATGCTACAG**AAACAAACAAACAAACAAACAAACAAACAAACAAAC**GCTGACCTGGTGGTAACTG...

...AGCTGCGATTCAGTCGTACAGTCCAATGCTACAG**CAGATCCAGATCCAGATCCAGATCCAGATCCAGATC**GCTGACCTGGTGGTAACTG...

réaction PCR est alors réalisée avec des amorces PCR spécifiques de l'adaptateur. Cette réaction est généralement qualifiée de "pré-sélective" car elle donne une quantité de fragments amplifiés difficilement interprétables. Une seconde réaction "spécifique" est donc réalisée en utilisant le produit PCR de la première réaction comme matrice et en employant des amorces PCR spécifiques des adaptateurs possédant des bases nucléotidiques supplémentaires, qui serviront à réduire le nombre de bandes amplifiées. Enfin, les produits de cette réaction sont déposés sur gel de polyacrylamide et révélés selon diverses méthodes (marquage radioactif, fluorescent ou au nitrate d'argent).

B.1.1.2 Protocole

Le protocole utilisé pour la mise au point des marqueurs AFLP chez le pin maritime a été développé et décrit en détails par Costa et al. (2000). Quelques modifications ont été apportées à ce protocole et sont décrites dans l'annexe I (*2.2 AFLP assay and gel electrophoresis*, p.629). Brièvement, les enzymes de restriction employées sont *Eco*RI et *Mse*I. Ces enzymes ont été choisies car elles ne sont pas sensibles à la méthylation, comme c'est le cas pour *Pst*I. De ce fait on peut faire l'hypothèse que l'on accède à toutes les régions du génome. Néanmoins, afin de s'affranchir au maximum de la grande complexité du génome du pin maritime, un total de 52 combinaisons enzyme/amorce (tableau B1) a été utilisé avec 2 bases nucléotidiques sélectives pour la première PCR et 3 à 4 bases sélectives pour la seconde. Les bandes ont été révélées en marquant radioactivement (^{33}P) les amorces spécifiques de *Eco*RI. La lecture des gels s'est effectuée de manière visuelle par deux personnes de façon indépendante (un exemple de gel se trouve figure B2).

B.1.2 Marqueurs anonymes fondés sur les séquences répétées : les microsatellites ou SSR

B.1.2.1 Description des microsatellites

Les microsatellites ou SSR (*Simple Sequence Repeat*) sont des séquences composées de courts motifs d'ADN (1 à 6 bases) répétés en tandem (figure B3). Ils sont très présents dans le génome des plantes et des animaux (Toth et al. 2000), chez qui ils sont fréquemment utilisés en tant que marqueurs moléculaires pour des études de cartographie génétique et de détection de QTL, ou en génétique des populations pour des études portant sur la diversité ou

Tableau B2 : Revue des publications décrivant le développement et le transfert de SSR chez les Pinaceae (tiré de l'annexe IV, table 1).

TGL: banque génomique d'ADN total; EGL: banque génomique enrichie; ELCL: banque enrichie en ADN peu répété; ECDL: banque d'ADNc enrichie; EUML: banque enrichie en ADN hypométhylé; EST: développement à partir de séquences d 'EST.

*: paires d'amorces dessinées à la fois à partir des banques EGL et TGL (non-différencié dans la publication);

**: paires d'amorces combinées avec celles de l'article de Elsik and Williams (2001) (non-différencié dans la publication);

Amplification chez d'autres espèces : H : hard pine sous-genre *Pinus*; S: soft pine sous-genre *Strobus*; SP: épicéa genre *Picea*; P : Pinaceae autre que *Pinus*; am: Americain; as: Asiatique; me: Pins Méditerranéens.

Espèce d'origine	Type de banque	Nombre de locus testés	Nombre de locus polymorphes simple copie	Référence	Nombre de locus testés dans une autre espèce de conifère	Nombre de locus transférés	Référence
Pinus radiata $^{H\text{-}am}$	EGL	2	2	Smith and Devey 1994	2	$2^{H}\ 0^{S}\ 0^{P}$	Echt et al. 1999
					2	$1^{H\text{-}me}$	Karhu et al. 2000
					2	$0^{H\text{-}me}$	Mariette et al. 2001
					2	$0^{H\text{-}am}$	Shepherd et al. 2002
Pinus radiata $^{H\text{-}am}$	TGL }	43*	2	Fisher et al. 1998			
	EGL }		11		7	$7^{H}\ 3^{S}\ 1^{P}$	Fisher et al. 1998
					4	$3^{H\text{-}me}\ 3^{H\text{-}am}\ 0^{S}\ 0^{P}$	Echt et al. 1999
					2	$0^{H\text{-}me}$	Mariette et al. 2001
					7	$4^{H\text{-}am}$	Shepherd et al. 2002
					20	$11^{H\text{-}am}$	Devey at al. 1999
Pinus radiata $^{H\text{-}am}$	EGL	50	10	Devey et al. 2002			
Pinus taeda $^{H\text{-}am}$	ELCL	18	16	Elsik et al. 2000b	7	$7^{H\text{-}am}\ 5^{H\text{-}me}$	Kutil and Williams 2001
					25**	$13^{H\text{-}am}$	Shepherd et al. 2002
					19**	$10^{H\text{-}me}$	Gonzalez-Martinez et al. 2003
Pinus taeda $^{H\text{-}am}$	ELCL	29	15	Elsik and Williams 2001			
	EGL	37	8				
Pinus taeda $^{H\text{-}am}$	ELCL	8	8	Kutil and Williams 2001	8	$8^{H\text{-}am}\ 2^{H\text{-}me}$	Kutil and Williams 2001
Pinus taeda $^{H\text{-}am}$	EUML	36	19	Zhou et al. 2002			
Pinus contorta $^{H\text{-}am}$	EGL	5	5	Hicks et al. 1998			
Pinus sylvestris $^{H\text{-}me}$	TGL	2	0	Kostia et al. 1995			
Pinus sylvestris $^{H\text{-}me}$	EGL	37	7	Soranzo et al. 1998	3	$3^{H\text{-}me}$	Gonzalez-Martinez et al. 2003
Pinus halepensis $^{H\text{-}me}$	EGL	25	8	Keys et al. 2000	8	$7^{H\text{-}me}$	Keys et al. 2000
					8	$1^{H\text{-}me}$	Mariette et al. 2001
Pinus pinaster $^{H\text{-}me}$	EGL	29	2	Mariette et al. 2001			
Pinus densiflora $^{H\text{-}as}$	EGL	14	6	Lian et al. 2000	6	$6^{H\text{-}as}\ 5^{H\text{-}am}\ 0^{S}$	Lian et al. 2000
Pinus strobus ^{S}	EGL	77	19	Echt et al. 1996	15	$12^{S}\ 0^{H}\ 0^{P}$	Echt et al. 1999
					28	3^{H}	Karhu et al. 2000
					4	0^{H}	Mariette et al. 2001
					5	0^{H}	Shepherd et a. 2002
Pinus strobus ^{S}	EGL	4	0	Echt et al. 1999	4	0^{S}	Echt et al. 1999
Picea sitchensis ^{SP}	EGL	7	4	van de Ven and Mac Nicol 1996			
Picea abies ^{SP}	EGL	36	7	Pfeiffer et al. 1997			
Picea abies ^{SP}	EGL	96	34	Paglia et al. 1998			
Picea abies ^{SP}	ECDL	6	6	Scotti et al. 2000			
Picea abies ^{SP}	EGL	55	16	Scotti et al. 2002a (tri)			
Picea abies ^{SP}	EGL	53	16	Scotti et al. 2002b			
Picea glauca ^{SP}	EGL	13	13	Hodgetts et al. 2001	13	12^{SP}	Hodgetts et al. 2001
Picea glauca ^{SP}	EGL	16	6	Rajora et al. 2001	6	6^{SP}	Rajora et al. 2001
Pseudotsuga menziesii ^{P}	EGL	102	48	Amarasinghe and Carlson 2002	50	31^{P}	Amarasinghe and Carlson 2002
Tsuga heterophylla ^{P}	EGL	16	11	Amarasinghe et al. 2003			
Cryptomeria japonica ^{P}	EST	3	2	Moriguchi et al. 2003			
	EGL }	67*	31				
	TGL }		1				
Total	-	776	333 (43%)		213	108 (50%)	

la gestion des ressources naturelles (Goldstein et Schlotterer 1999). Les variations que l'on peut observer au sein de ces séquences concernent des variations de nombre de répétitions des motifs de base. Le développement de ces marqueurs est techniquement difficile, car il doit passer par l'enrichissement d'une banque génomique en motifs répétés, puis par le clonage et le séquençage de nombreux fragments, et enfin par le développement d'amorces PCR et la mise au point des conditions d'amplification. D'autres méthodes ont été envisagées afin de contourner ces difficultés. Par exemple, des microsatellites peuvent être transférés d'une espèce à l'autre si celles-ci sont phylogénétiquement proches. Une autre possibilité consiste à chercher des motifs répétés dans les bases de données publiques. Ces trois approches ont été employées pour le pin maritime dans le cadre de cette thèse, et leur efficacité pour développer des microsatellites sera discutée par la suite.

B.1.2.2 Les microsatellites chez les Pinaceae

Comme pour de nombreux organismes, des microsatellites ont été développés chez les conifères, et en particulier chez les Pinaceae. La liste des études portant sur les microsatellites chez une ou plusieurs espèces de Pinaceae est présentée au tableau B2. La première remarque que l'on peut tirer de cette synthèse est que le nombre de locus mis au point reste faible (333) si l'on considère le nombre d'études (26) et d'espèces (14) considérées, et que ce nombre, même en cumulant les résultats obtenus chez toutes les espèces, ne permettrait pas de saturer une carte génétique ! Si on regarde plus en détail les études où des microsatellites ont été développés à partir d'une banque enrichie, on se rend compte que cette quantité est très faible par rapport aux efforts fournis en terme de travail de laboratoire. De plus, la plupart de ces études montrent une quantité importante d'amplifications PCR non-spécifiques présentant des profils électrophorétiques multi-bandes. Ceci peut être expliqué par la complexité du génome des conifères, riche en séquences répétées. On peut donc imaginer que les amorces oligonucléotidiques peuvent s'hybrider, lors de la PCR, à plusieurs endroits du génome. Certains auteurs ont cherché à optimiser le protocole d'enrichissement d'une banque génomique afin d'éliminer les régions hautement répétées. Elsik et Williams (2001) et Kutil et Williams (2001) ont séparé les fractions très répétées des fractions simple-copie du génome en utilisant des techniques basées sur la cinétique de réassociation de l'ADN (courbes CoT). De même, Scotti et al. (2002) ont utilisé la technique du Dot Blot[8], et Zhou et al. (2002) ont

[8] Dot Blot : Technique basée sur l'hybridation d'ADN cloné non-digéré.

sélectionné les fractions hypo-méthylées du génome. Ces méthodes ont permis d'augmenter considérablement l'efficacité du développement de microsatellites chez les conifères, en réduisant le nombre d'amplifications non-spécifiques. Néanmoins, elles représentent toujours des étapes supplémentaires au protocole de développement et sont techniquement longues et coûteuses à mettre en oeuvre.

Un autre résultat étonnant présenté au tableau B2 est que les microsatellites de pin sont difficilement transférables entre espèces. Ceci peut être expliqué par l'origine très ancienne des pins. Tout d'abord on peut s'attendre à ce que les séquences flanquant les SSR et correspondant généralement aux sites d'hybridation des amorces PCR ont subi des mutations et divergent fortement entre espèces. Ensuite, sachant que les taux d'évolution des microsatellites sont très forts par rapport à d'autres régions du génome, on peut s'attendre à ce que des locus microsatellites soient apparus ou aient disparu depuis la divergence des espèces de pin. De plus, étant donné le temps d'évolution du génome des conifères, on peut aussi s'attendre à ce que des phénomènes de duplication se soient produits, provoquant l'apparition de locus paralogues. Tout ceci peut expliquer le fait qu'il est difficile de transférer des microsatellites entre espèces du genre *Pinus* ou même à l'intérieur d'un sous-genre (ex: 36% d'amplification des microsatellites de *P. taeda* chez *P. pinaster*, espèces du sous-genre *Pinus*, Gonzalez-Martinez et al. 2003) et presque impossible de les transférer entre sous genre ou entre genre (Echt et al. 1999).

B.1.2.3 Les méthodes de développement des microsatellites utilisées chez le pin maritime

B.1.2.3.1 Transfert entre espèces proches

L'approche qui peut paraître la plus intéressante pour un chercheur qui souhaite développer des microsatellites chez l'espèce qu'il étudie est de transférer les locus mis au point chez une espèce phylogénétiquement proche. L'aspect le plus séduisant est qu'il n'y a pas besoin de construire de banque génomique ou de disposer de données de séquences. Chronologiquement, cette stratégie a été employée en premier pour développer des microsatellites chez le pin maritime. Une recherche bibliographique des locus développés chez d'autres espèces de pin a été réalisée et ceux-ci ont été testés chez le pin maritime. Ces locus étaient originaires de *Pinus radiata* (107 locus, C. Echt et T. Richardson, Forest Research Institute, NZ, communication personnelle; 7 locus, G. Moran, CSIRO, Australie; 2

Tableau B3 : Liste des références utilisant des banques d'EST pour développer des microsatellites chez les plantes.

Espèce	Référence
Vigne	Scott et al. 2000
Arabidopsis	Cardle et al. 2000
Riz	Temnykh et al. 2000
Riz	Cho et al. 2000
Riz	Temnykh et al. 2001
Canne à sucre	Cordeiro et al. 2001
Arabidopsis	Morgante et al. 2002
Céréales	Kantety et al. 2002
Blé	Eujayl et al. 2002
Céréales	Varshney et al. 2002
Céréales	Holton et al. 2002
Céréales	Gao et al. 2003
Orge	Thiel et al. 2003
Luzerne	Eujayl et al. 2003
Blé	Gupta et al. 2003

locus, Smith et Devey 1994; 2 locus, Fisher et al. 1998), *Pinus halepensis* (25 locus, Keys et al. 2000), *Pinus sylvestris* (7 locus, Soranzo et al. 1998), *Pinus strobus* (11 locus, Echt et al. 1996) et *Pinus taeda* (3 locus, Elsik et al. 2000). Les protocoles utilisés pour l'amplification et la révélation de ces microsatellites sont présentés à l'annexe IV pour les locus obtenus au Forest Research Institute de *P. radiata* et à l'annexe II pour les autres espèces.

B.1.2.3.2 Recherche de microsatellites dans les banques d'EST

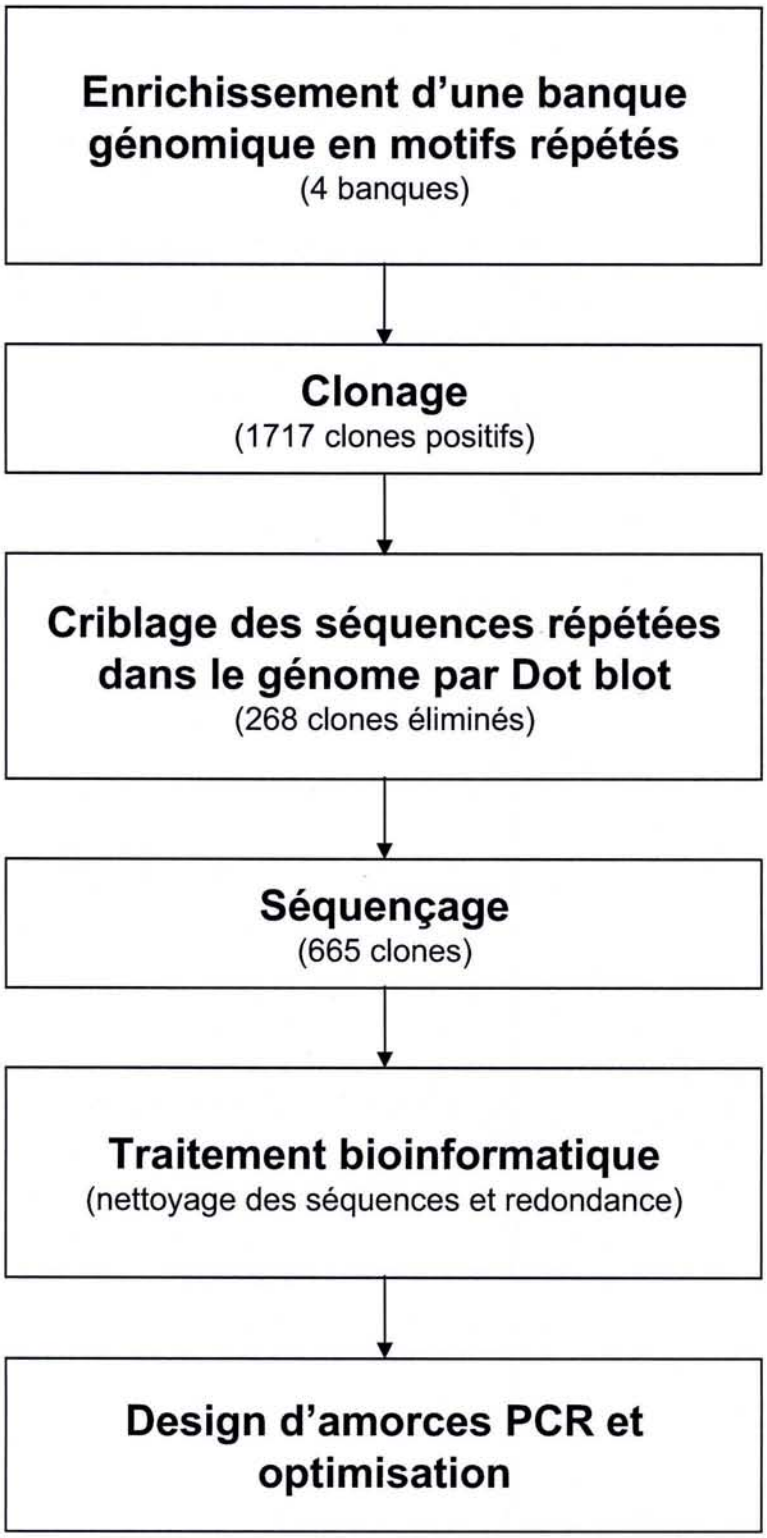
Une approche alternative au développement d'une banque enrichie consiste à rechercher des motifs répétés dans les bases de données de séquences. Etant donné que le génome de peu d'espèces de plantes a été séquencé (pour l'instant), mais que, par contre, de grandes quantités d'EST sont disponibles, de nombreuses études se sont employées à rechercher des motifs microsatellites dans ces banques d'EST. Une liste de ces études menées chez les plantes est présentée au tableau B3. Comme des EST étaient disponibles pour deux espèces de pin (*P. taeda* et *P. pinaster*) j'ai également initié cette approche.

Pour cela un logiciel développé chez le riz (Temnykh et al. 2001) a été utilisé : *SSRIT* (téléchargeable à : ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl). Ce logiciel, écrit en *Perl*, permet de trouver la position et le type de motif de microsatellites dans des EST au format *FASTA*. Afin que les résultats issus de cette analyse représentent un lot de séquences non-redondantes (c'est-à-dire que l'on ne puisse pas retrouver plusieurs fois la même séquence), les séquences des unigènes[9] (l'assemblage en contigs des EST) de *Pinus taeda* et de *Pinus pinaster* ont été utilisées plutôt que les EST eux-mêmes. Outre le fait que l'on travaille avec un lot de séquences uniques, l'utilisation de ces unigènes permet d'accéder à des séquences plus longues et de meilleure qualité. De plus, ces séquences correspondant à des gènes, on peut déduire la position des microsatellites à l'intérieur de ceux-ci. Pour cela le logiciel *SSRIT* a été combiné avec le logiciel *FrameD* (http://www.toulouse.inra.fr/FrameD.html, Schiex et al. 2003) qui sert à prédire la position des régions traduites et non traduites d'un gène.

Le lot de séquences utilisées comportait 20 377 séquences de *P. taeda* (8 070 contigs + 12 377 singletons, http://web.ahc.umn.edu/biodata/nsfpine/contig_dir16/) et 7 894 séquences de *P. pinaster* (2 893 contigs + 5 001 singletons). Ces deux unigènes étaient issus de 75 047 EST

[9] Unigène : Représentation non-redondante d'un jeu de séquences codantes.

Figure B4 : Protocole de développement d'une banque enrichie en motifs microsatellites chez le pin maritime.



Encadré B1 : Extraction d'ADN génomique de pin au Chlorure de Césium

Solutions :

Tampon d'homogénéisation :
(0,3 M saccharose, 50 mM Tris-HCl, pH 8.8, 5 mM MgCl2)
3,03 g Trisma Base (pH8.8)
0,51 g $MgCl_2$
51,35 g saccharose
5 g PVP-40
qsp 500 ml (H_2O)
250 µl EtBr (bromure d'ethidium, 10 mg/ml)

Tampon de lyse nucléaire :
(50 mM Tris-HCl, pH 8, 20 mM EDTA)

Tampon d'équilibrage :
75 g CsCl (chlorure de césium)
75 ml H_2O
2,82 ml EtBr
Ajuster l'indice de réfraction à 1,3895

Protocole :

1/ Broyer les aiguilles dans de l'azote liquide (mortier) et transférer dans un grand tube (250 ml)
2/ Ajouter 50 ml de tampon d 'homogénéisation et mélanger (spatule)
3/ Broyer à l'aide d'un polytron (2-3 min) afin d'obtenir une mixture fluide et épaisse
4/ Filtrer à travers de la gaze et centrifuger 10 min à 8000 rpm (le culot contient le noyau et les organelles)
5/ Resuspendre le culot dans du tampon de lyse nucléaire et rajouter du sarkosyl (conc. finale 1%)
6/ Ajouter du Chlorure de Césium (CsCl) à 1g/ml et centrifuger 10 min à 10 000 rpm
7/ Récupérer le surnageant et le transférer dans une éprouvette
8/ Rajouter du EtBr pour obtenir une solution à 300 µg/ml
9/ Ajuster la solution à 1,3895 (indice de réfraction) en rajoutant de l'eau ou du CsCl
10/ Remplir des tubes de centrifugation (verticaux) et équilibrer avec de la solution d'équilibrage
11/ Centrifuger 72 heures à 40 000 rpm
12/ Détecter la bande fluorescente dans le tube sous lumière UV
13/ Collecter doucement l'ADN avec une pipette Pasteur
14/ Extraire le EtBr en ajoutant du NaCl saturé en isopropanol : mixer (2-3 min par inversion) et récupérer le surnageant (isopropanol)
15/ Répéter l'extraction au moins 3 fois (la solution ne doit plus être rose)
16/ Diluer avec un volume d'eau et ajouter deux volumes d'éthanol (95%)
17/ Enrouler l'ADN sur une tige de verre (ex: pipette Pasteur brûlée au bout)
18/ Laisser l 'éthanol s 'évaporer et resuspendre dans 5 ml de tampon 10 mM Tris, pH 8, 0,1 mM EDTA
19/ Ajouter 1/10 volume de 10X protéinase K et incuber 1-2 heures à 60°C
20/ Refroidir et faire deux extractions au phénol : chloroforme
21/ Extraire à l'éther au moins 3 fois ou jusqu'à ce que le surnageant soit clair et évaporer l'éther résiduel
22/ Ajouter 1/2 volume d'acétate d 'ammonium (7,5 M) et deux volumes d'éthanol (95%) froid
23/ Enrouler l'ADN à nouveau (voir étape 16)
24/ Laver avec de l'éthanol (95%) et laisser évaporer
25/ Resuspendre dans du tampon 10 mM Tris, pH 8, 0,1 mM EDTA

Responsable	Institut	Pays
Almeida M-H.	ISA Lisbonne	Portugal
Byrne M.	CALM Kensington	Australie
Cervera M-T.	CIFOR-INIA Madrid	Espagne
Favre J-M.	UHP Nancy	France
Harvengt L.	AFOCEL Nangis	France
Rocheta M.	IBET Lisbonne	Portugal
Vendramin G.G.	CNR Florence	Italie
Plomion C.	INRA Bordeaux	France

Tableau B4 : Membres du consortium international pour le développement de marqueurs microsatellites chez le pin maritime.

de *P. taeda* (banques d'ADNc développées à partir de xylème) et de 18 498 EST de *P. pinaster* (banques d'ADNc développées à partir de xylème et de racines).
Une fois la détection de microsatellite réalisée, des amorces PCR ont été dessinées à l'aide du logiciel *Primer 3.0* (http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi). Les séquences des amorces, les conditions PCR et de révélation des fragments sont détaillées dans l'annexe IV et sont disponibles à l'adresse : http://www.pierroton.inra.fr/genetics/SSR/index.php.

B.1.2.3.3 Développement de banques génomiques enrichies en microsatellites

La solution la plus efficace en terme de nombre de locus informatifs, mais qui a chronologiquement été utilisée à l'issue des deux approches présentées ci-dessus (à cause de sa lourdeur technique) consiste à développer une banque enrichie en motifs répétés. Les différentes étapes de cette approche sont présentées à la figure B4. Afin de réduire le coût de développement et de partager le travail lors de certaines étapes, huit équipes intéressées par l'obtention de microsatellites chez le pin maritime se sont regroupées dans le cadre d'un consortium international (tableau B4). Néanmoins, une grande partie du travail de développement et la coordination du projet ont été réalisées dans le cadre de mon travail de thèse.
La première étape, que j'ai réalisée à Pierroton, a consisté en l'extraction d'ADN génomique de pin maritime. Afin de disposer d'une grande quantité d'ADN génomique de très bonne qualité, la méthode employée a été celle de l'extraction au chlorure de césium. En effet, cette première étape est très importante car, de la qualité de l'ADN génomique de départ, découlent toutes les étapes ultérieures (enrichissement, clonage, etc...). D'où l'utilisation de cette méthode d'extraction longue mais très efficace. Le protocole d'extraction au chlorure de césium sur aiguille a été fourni par C. Echt et est présenté dans l'encadré B1. Le matériel végétal de base pour cette extraction a consisté en un mélange d'aiguilles fraîches issues de plusieurs provenances de pin maritime (Landes, Italie, Espagne, Portugal, Maroc, Corse), ceci afin de détecter des variations de longueur des microsatellites *in silico* lors de l'étape d'assemblage des séquences obtenues.
La seconde étape a consisté en l'enrichissement d'une banque génomique en motifs répétés. Cette étape n'a pas été réalisée à Pierroton, mais par une entreprise privée (Genetic Identification Service, Chartsworth, USA) spécialisée dans ce travail, chez qui l'ADN génomique total de pin maritime a été envoyé. Le protocole détaillé utilisé n'a pas été

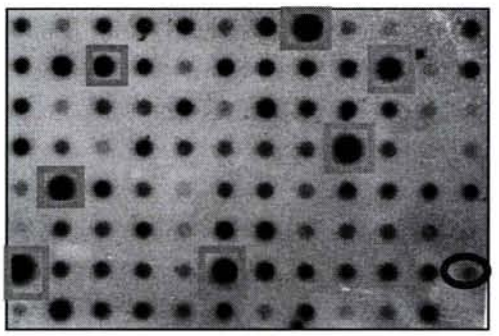


Figure B5 : Exemple d'hybridation des inserts (PCR sur colonies) avec l'ADN génomique de pin maritime. Les spots encadrés on été écartés de l'étape de séquençage. Le spot entouré correspond au contrôle : gène faiblement répété (Chalcone synthase).

Encadré B2 : Protocole d'hybridation des membranes pour le Dot Blot

Solutions :

SDS 20% :
20 g SDS
qsp 1 l H_2O

SSC 20X (pH 8) :
175,3 g NaCl
44,1 g Citrate de sodium
qsp 1 l H_2O

Denhardt's 50X :
2,5 g Ficoll
2,5 g PVPsoluble
qsp 250 ml H_2O

Tampon de pré- et d'hybridation :
(SSC 5X, Denhardt's 5X, SDS 0,5%)
80 ml SSC 20X
32 ml Denhardt's 50X
8 ml SDS 20%
qsp 320 ml H_2O

Protocole :

1/ Mettre le four à hybridation à préchauffer à 65°C
2/ Préhybridation (5-6 heures, sous agitation) : mettre 30 ml de tampon par tube + 300 µl d'ADN de sperme de saumon dénaturé
3/ Hybridation (toute la nuit, sous agitation) : mettre 20 ml de tampon + 200 µl d 'ADN de sperme de saumon dénaturé + 100 µl de sonde dénaturée (marquée au ^{33}P)
4/ Rincer les membranes avec 20 ml de solution SSC 2X, SDS 0,5% (une fois, température ambiante)
5/ Laver les membranes avec 25 ml de solution SSC 2X, SDS 0,5% (deux fois, 65°C, 15 min)
6/ Emballer les membranes dans du film cellophane et les mettre dans la cassette avec un film autoradiographique (au moins 48 h)

communiqué par ce laboratoire, mais le principe général en est le suivant : l'ADN génomique a été digéré par une enzyme de restriction, puis dénaturé. Ces fragments simples brins ont ensuite été hybridés à des motifs répétés (généralement liés à une membrane ou à une colonne). Un lavage permet d'éliminer tous les fragments ne contenant pas de motif répété et ne s'hybridant donc pas. Il ne reste plus qu'à les insérer dans un vecteur de clonage. Les motifs répétés choisis étaient AC, AG, AAG et AAAT. Pour chacune de ces banques, environ 8 000 vecteurs (pUC-19) contenant des inserts de taille voisine de 600 pb et supposés contenir des microsatellites ont été envoyés dans notre laboratoire.

L'étape suivante a été réalisée à Pierroton : les vecteurs ont été transformés dans des bactéries *E. coli* (Kit TOPO cloning, Invitrogene, Carlsbad, USA) et 672 (Banque A, AT), 576 (Banque B, AC), 384 (Banque C, AAG) et 288 (Banque D, AAAT) colonies ont été repiquées dans du milieu LB + glycérol 15% et conservées à -80°C. La totalité des clones des 4 banques a été dupliquée et envoyée à tous les membres du consortium. Les fragments insérés dans les vecteurs ont été amplifiés par PCR en utilisant les amorces M13 universelles *forward* et *reverse* afin de vérifier la présence de l'insert pour chaque colonie. Les conditions PCR sont celle préconisées par le fournisseur du kit de clonage (Kit TOPO cloning, Invitrogene). 601, 487, 367 et 262 clones se sont révélés positifs pour les banques A, B, C et D, respectivement.

Les produits PCR obtenus ont alors été dénaturés et fixés sur membrane pour cribler les microsatellites les moins répétés dans le génome (figure B5). Chaque membrane de 96 *Dot Blot* a été répétée deux fois. L'ADN génomique total de pin maritime a été marqué au ^{33}P, dénaturé, puis hybridé sur les membranes. Le principe de la technique du Dot Blot est le même que celui utilisé pour les études d'expression différentielle de transcrits. Il a été utilisé de manière efficace par Scotti et al. (2002b) chez l'épicéa. Le protocole utilisé (encadré B2) est le même que celui décrit par Le Provost (2003) pour les macro-arrays. Un témoin a été placé sur la membrane : il correspond à un clone d'EST de pin maritime dont le gène est simple copie (Chalcone synthase). Les 268 clones (15%) présentant des signaux forts sont ceux qui se retrouvent en grand nombre de copies dans le génome. Ils ont donc été éliminés et ne sont pas pris en compte pour le design d'amorces PCR étant donné que leur amplification peut provoquer des profils multi-bandes.

L'étape suivante de séquençage a été partagée entre les différents laboratoires du consortium. Au total 859 séquences issues de 665 clones ont été obtenues. Une fois les séquences obtenues, elles ont été rassemblées à Pierroton où les étapes bioinformatiques de vérification de la redondance des séquences, et leur alignement, ont été réalisés. Au préalable, toutes les séquences ont été "nettoyées" : les séquences de vecteur et les régions de qualité insuffisante

Figure B6 : Structure de la base MySQL contenant les données relatives aux microsatellites développés à partir d'EST

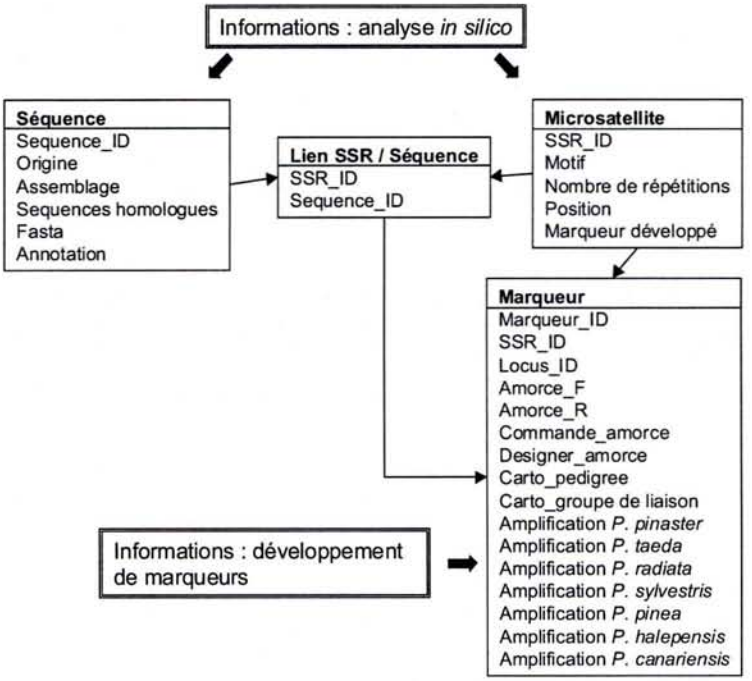
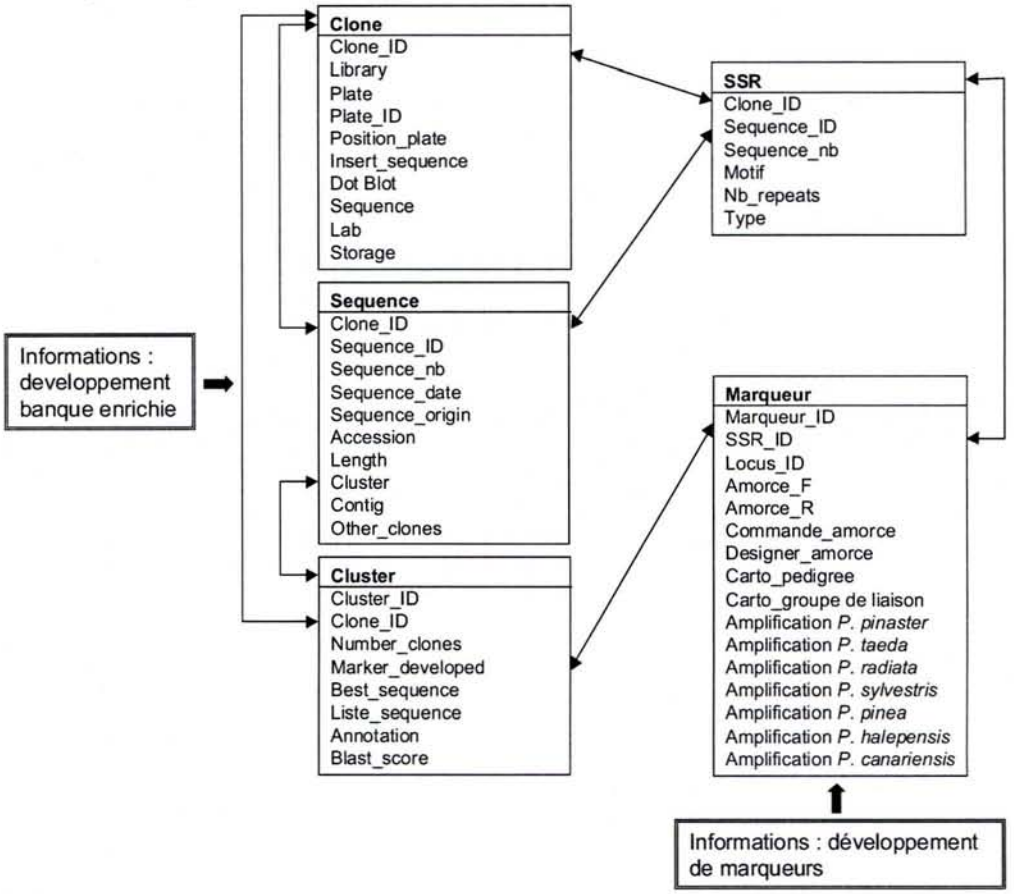


Figure B7 : Structure de la base MySQL contenant les données relatives à la banque enrichie en microsatellites pour le pin maritime



S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

(score Phred < 20) ont été éliminées. Le logiciel utilisé pour réaliser le traitement bioinformatique des séquences est le logiciel *StackPack* (Christoffels et al. 2001). A l'origine, ce logiciel sert à assembler les séquences issues de projets de séquençage d'EST. Le principe est le même sauf que dans le cas présent les séquences contiennent des motifs microsatellites. *StackPack* est intéressant car il "masque" les séquences de faible complexité, comme par exemple les microsatellites. Ainsi, l'assemblage se fait à partir des séquences flanquant les microsatellites. En effet dans le cas où les microsatellites ne seraient pas masqués toutes les séquences pourraient s'aligner entre elles et résulter en un seul contig[10].

Une fois les séquences assemblées en contigs, j'ai dessiné des amorces PCR oligonucléotidiques. Pour ce faire, le logiciel *Primer 3.0* (http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi) a été utilisé. Les paramètres de *Primer 3.0* utilisés sont décrits dans l'annexe IV.

B.1.2.4 Développement de bases de données pour les microsatellites de pin maritime

Une base de données contenant les informations sur les microsatellites contenus dans les séquences codantes de pin maritime et de pin taeda, et ceux issus de la banque enrichie de pin maritime a été créée et mise en accès libre sur internet (http://www.pierroton.inra.fr/genetics/SSR/index.php).

Une base de données MySQL a été créée pour chacun de ces types de microsatellites. Les figures B6 et B7 représentent les schémas de structure des deux bases de données. L'interface internet permettant d'interroger ces bases de données a été écrite en PHP.

B.1.3 Marqueurs fondés sur les séquences codantes : les EST

B.1.3.1 Qu'est-ce qu'un EST ?

La méthode la plus efficace pour avoir accès à des séquences codantes est le séquençage systématique de fragments de gènes exprimés. Ces fragments, de longueur excédant rarement 1000 pb, sont généralement appelés des étiquettes de séquences exprimées ou EST (*Expressed Sequence Tags*). A l'heure actuelle, chez les conifères, les EST

[10] Contig : Jeu de séquences homologues chevauchantes.

Figure B8 : Exemple de profil obtenu sur gel de poly-acrylamide (8%) : 2 parents + 22 descendants (plein-frères).

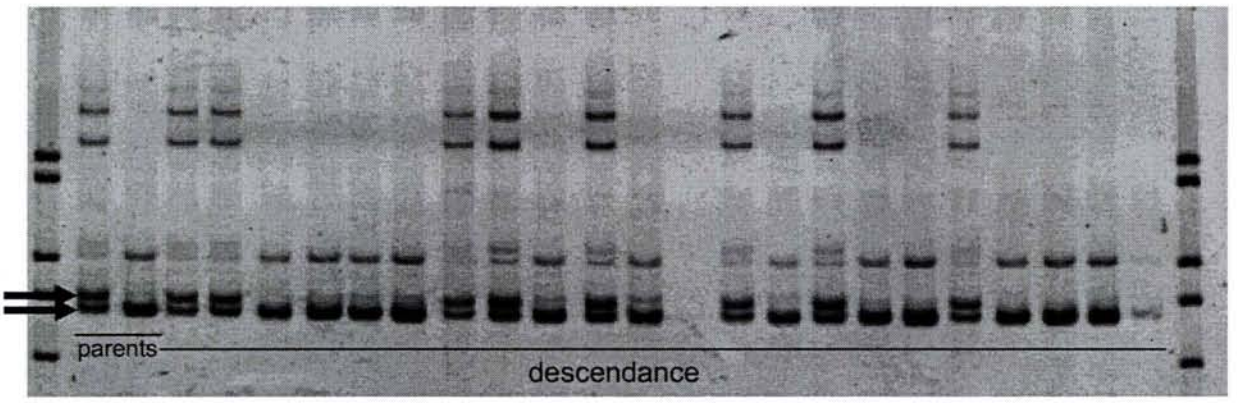
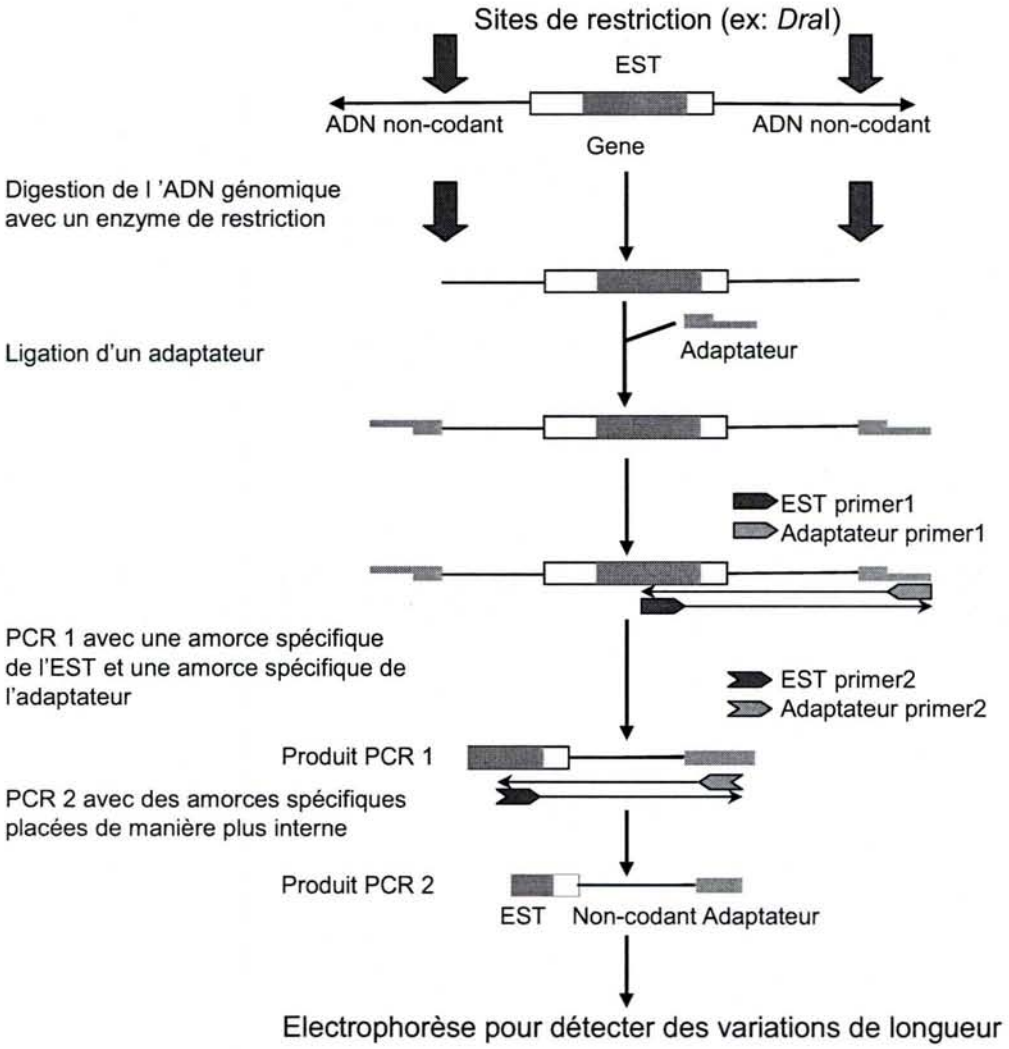


Figure B9 : Description de la méthode décrite par Cato et al. 2001 et utilisée pour la cartographie d'EST chez le pin maritime.



disponibles dans les bases de données publiques (Genbank, EMBL) sont principalement issus de *Pinus taeda* et de *Pinus pinaster*. La mise à jour de ces bases de données datant de janvier 2004 (dbEST, mise à jour du 9 janvier 2004, http://www.ncbi.nlm.nih.gov/dbEST/index.html) fait état de 110 622 EST de *P. taeda* et de 15 719 EST de *P. pinaster*. Ces quantités sont à comparer, entre autres, avec celles plus importantes des EST publiques disponibles chez l'homme (5 438 178), la souris (3 995 783), ou le blé (549 985). Néanmoins, ces EST de pin constituent une source remarquable de données pour l'identification de gènes et le développement de marqueurs moléculaires. Pour cela, il a fallu mettre en place différentes méthodes moléculaires et bioinformatiques permettant de détecter du polymorphisme dans ces séquences. Une des méthodes possibles a été décrite plus haut (identification de microsatellites dans les EST). Je ne reviendrai donc pas sur cette approche.

B.1.3.2 Méthodes de détection de polymorphisme sur les EST

B.1.3.2.1 Méthodes basées sur la longueur des fragments d'ADN

La méthode la plus simple est d'observer des variations de taille de fragments amplifiés par PCR avec des amorces spécifiques du gène. Ces variations sont généralement mises en évidence par électrophorèse sur des gels d'agarose (2%) ou de polyacrylamide (4-8%). Cette méthode a été utilisée pour cartographier des EST, dont la liste se trouve à l'annexe III. Elle a été employée comme première étape préalable à l'utilisation des autres techniques (DGGE et SSCP). Un exemple de profil de gel présentant des variations de longueur est présenté à la figure B8.

Les variations de longueur de fragment peuvent être dues à des insertions-délétions (indels) au sein de la séquence. Une méthode de révélation de ces polymorphismes de longueur a été utilisée dans le cadre de cette thèse. Néanmoins, étant donné que les EST sont généralement de faible taille (autour de 600 pb), les chances de rencontrer des indels sont réduites. De plus, le caractère codant des EST les rend potentiellement moins variables car ils sont soumis à la sélection. L'idée est donc d'explorer les séquences bordant ces EST, non-codantes et supposées être plus variables. Une méthode comparable au principe de la marche chromosomique[11] a ainsi été mise au point chez *Pinus radiata* par Cato et al. (2001). Cette méthode est très proche de la méthode d'AFLP. Cependant, elle utilise des amorces

[11] Marche chromosomique (ou chromosome walking) : Méthode consistant à produire des clones chevauchants afin d'étudier une séquence qui serait trop grande pour être clonée individuellement.

spécifiques de la séquence d'un gène (ou d'un EST). La figure B9 explique cette méthode. Dans un premier temps l'ADN génomique est digéré par des enzymes de restriction reconnaissant des sites riches en AT (*Alu*I, *Dra*I, *Eco*RV, *Ssp*I). Les fragments de restriction sont ensuite ligués à un adaptateur. Une première PCR est réalisée avec une amorce spécifique d'un EST et une amorce spécifique de l'adaptateur. Cette PCR étant peu spécifique et pouvant entraîner des profils multibandes, une deuxième PCR emboîtée est réalisée avec deux autres amorces placées de manière plus interne sur l'EST et l'adaptateur.
Afin de mettre au point cette méthode sur le pin maritime, j'ai utilisé le même protocole et les mêmes amorces (adaptateur et EST) que ceux décrits par Cato et al. (2001), en supposant que les séquences codantes de *Pinus radiata* utilisées par ces auteurs différaient peu de celles de *Pinus pinaster*.

B.1.3.2.2 Méthodes basées sur la séquence d'ADN

La méthode la plus directe pour détecter des variations dans la séquence d'ADN consiste à séquencer directement des fragments d'ADN. Par exemple, on peut séquencer le même EST sur plusieurs individus différents, puis les aligner. On observe alors des variations de types insertion-délétion ou des variations de séquence qui, lorsqu'elles concernent une seule base, sont appelées SNP (*Single Nucleotide Polymorphism*). Néanmoins cette méthode est assez lourde à réaliser car le séquençage d'un fragment peut passer par une étape de clonage (en particulier si on veut avoir accès à un haplotype et pouvoir différencier les deux allèles d'un hétérozygote). D'autres méthodes sont donc généralement employées. Je m'appuierai plus particulièrement sur des méthodes utilisées dans le cadre de cette thèse et fondée sur la conformation de fragments d'ADN lors de leur migration en gel de polyacrylamide. Deux méthodes ont été utilisées : la méthode SSCP (*Single Strand Conformation Polymorphism*) et la méthode DGGE (*Denaturing Gradient Gel Electrophoresis*). Néanmoins, l'intérêt récent de l'utilisation de SNP pour des études de cartographie génétique ou des études d'association avec des caractères quantitatifs a amené au développement d'un grand nombre de techniques moléculaires plus puissantes que la SSCP et la DGGE permettant de détecter ou de révéler des SNP. Des revues complètes de ces techniques ont été publiées par Kwok (2001) et Jenkins et Gibson (2002). Parmi elles, l'extension d'amorces (Kuppuswamy et al. 1991, Syvanen 1999) permet de génotyper un grand nombre de SNP de position connue. Elle consiste à marquer la base où se trouve le polymorphisme avec un fluorochrome (chaque allèle aura une "couleur" différente), placée en

Figure B10 : Description de la méthode DGGE.
Les fragments d'ADN double-brin sont soumis à une migration dans des conditions de plus en plus dénaturantes. Au fur et à mesure de la migration, les fragments double-brin peuvent s'ouvrir, prendre une conformation en « Y » et être freinés. Une différence sera alors décelée entre les deux allèles (■ et ●). Une bande sera observée à deux différents niveaux pour les génotypes homozygotes, et de deux à quatre bandes pourront être observées pour les génotypes hétérozygotes, ceci étant dû à la formation d'hétéroduplexes (appariement des brins d'ADN hétérologues). Un exemple de gel DGGE correspondant à la conformation décrite est présenté.

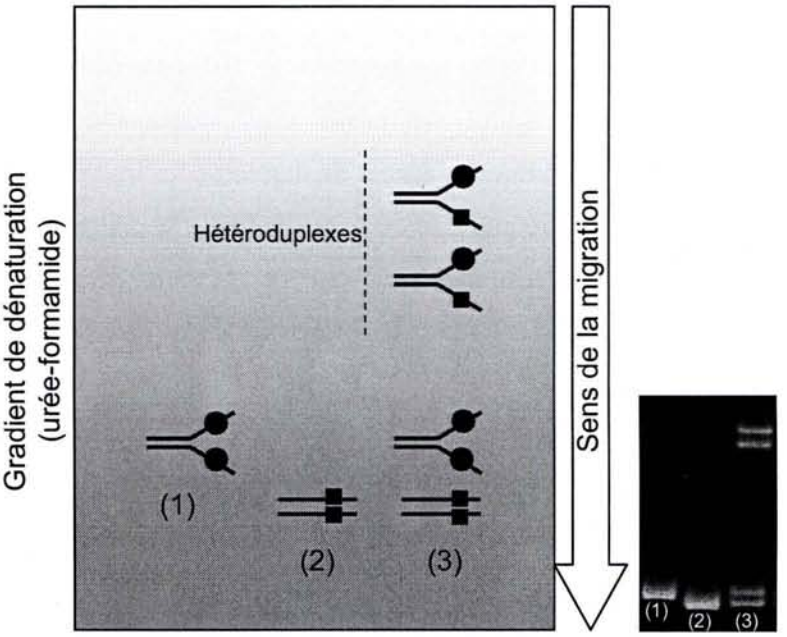


Figure B11 : Appareil DCode utilisé pour la migration des fragments d'ADN dans le cadre de la technique DGGE (BioRad™)

extrémité d'un oligonucléotide spécifique de la séquence bordante. Les fragments amplifiés sont alors analysés selon différentes méthodes à haut débit comme la chromatographie DHPLC (Giordano et al. 2001), la révélation sur séquenceur automatique (SNapShot™, Applied Biosystems), la polarisation différente des fluorochromes (chimie Acyclo-Prime™, Victor 3.0, Perkin Elmer), ou l'hybridation sur des puces à ADN (Hazen et Kay 2003).

B.1.3.2.2.1 Electrophorèse en gradient de dénaturation : la technique DGGE (*Denaturing Gradient Gel Electrophoresis*)

B.1.3.2.2.1.1 Principe

La DGGE (Myers *et al.* 1987) est basée sur le principe de l'électrophorèse sur gel de polyacrylamide dénaturant. Elle permet de détecter des changements d'une base nucléotidique à l'intérieur d'un fragment d'ADN. L'ADN double-brin est soumis à un gradient de dénaturation parallèle au sens de migration, ce qui provoque la séparation de ces fragments en domaines. La température de fusion (Tm) de ces domaines dépend directement de leur séquence (figure B10). Quand les conditions de dénaturation les plus basses sont atteintes, l'ADN partiellement dénaturé se retrouve sous forme de molécule possédant des "branches" freinant sa mobilité dans le gel de polyacrylamide. Les conditions dénaturantes sont créées par la température du tampon de migration (58°C) et le gradient résulte d'une concentration croissante en urée et formamide.

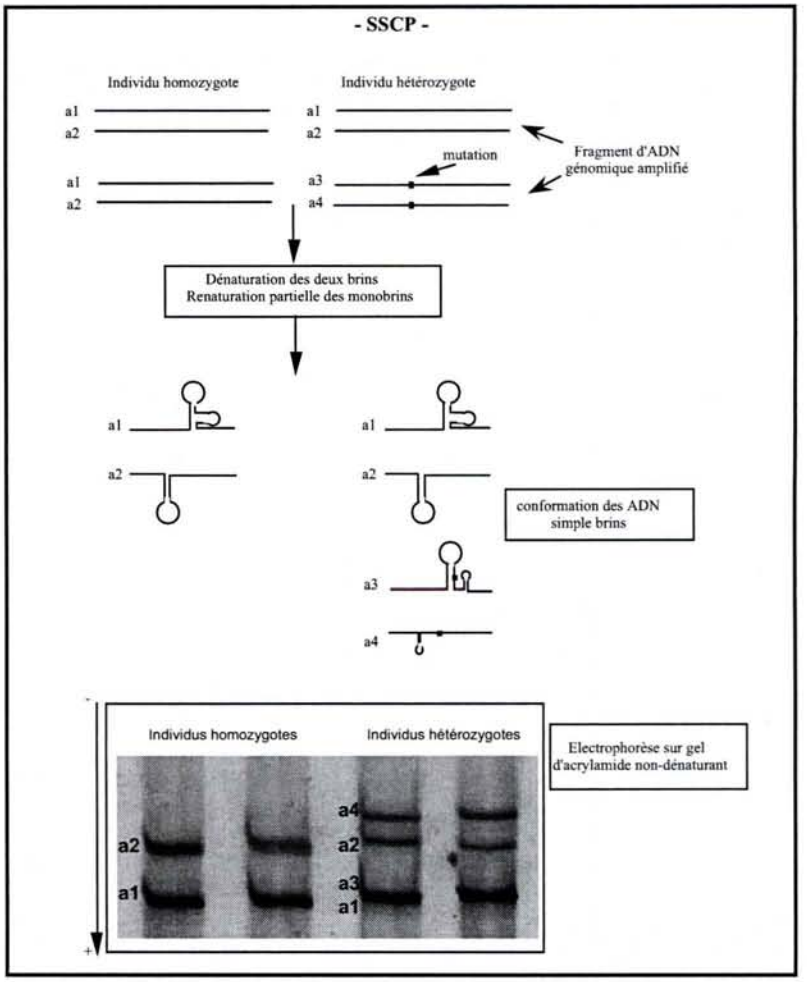
Cette technique peut révéler la présence d'hétéroduplexes[12]. Ils sont le résultat d'un mésappariemment de l'ADN double-brin. On observe ce phénomène dans le cas d'individus hétérozygotes (figure B10) : les segments différents des chromosomes homologues s'associent entre eux, mais leur conformation est telle qu'ils sont plus retardés que les homoduplexes lors de la migration. Les hétéroduplexes augmentent la résolution de la DGGE. En effet, leur analyse peut aider à résoudre le génotype des homoduplexes (Temesguen *et al.* 2000).

B.1.3.2.2.1.2 Composition du gel :

Un appareil commercialisé par Bio-Rad ™ (figure B11) permet de créer le gradient de dénaturation. Une solution de polyacrylamide dénaturante (6% ou 10% selon la taille des

[12] Hétéroduplexes : ADN double brin formé par la réassociation de deux fragments d'ADN simple brin différents.

Figure B12 : Description de la méthode SSCP.
Les fragments d'ADN sont d'abord dénaturés et maintenus sur de la glace jusqu'à leur dépôt sur le gel. Les conditions de migration sont non-dénaturantes, ce qui provoque un repliement des molécules d'ADN sur elles-mêmes. Les différentes conformations obtenues vont être plus ou moins freinées lors de la migration.



fragments amplifiés) est mélangée à une solution dénaturante à 80% contenant de l'urée et du formamide. Le gradient le plus souvent utilisé est un gradient dont la teneur en urée-formamide est de 15% en haut et 45% en bas du gel. Ces valeurs peuvent varier en fonction du locus, mais la plupart des locus utilisés ont été choisis pour présenter un profil simple sous ces conditions. Les conditions de DGGE des différents locus utilisés sont présentées dans l'annexe III.

B.1.3.2.2.2 Polymorphisme de conformation d'ADN simple brin : la technique SSCP (Single Strand Conformation Polymorphism)

A l'inverse de la DGGE, la technique SSCP (Orita *et al.* 1989) consiste à faire migrer des fragments d'ADN dénaturés dans un gel de polyacrylamide non-dénaturant. Avant le dépôt, l'ADN double brin est dénaturé par chauffage à 94°C pendant 6 minutes, puis rapidement refroidi (sur de la glace pendant 2 minutes), les molécules simple brin n'ont pas le temps de se réassocier entre elles, et forment une structure secondaire stable par des réassociations au niveau de zones de séquences complémentaires. Ces molécules peuvent prendre des conformations différentes décelables par une différence de migration lors de l'électrophorèse (figure B12). Les conditions de migration (voltage, température, composition du gel) utilisées pour la cartographie d'EST grâce à la SSCP sont décrites par Plomion et al. (1999) et à l'annexe III.

B.1.3.2.3 Méthode bioinformatique de détection de polymorphisme nucléotidique

Les bases de données d'EST sont redondantes car elles peuvent contenir les mêmes séquences homologues en plusieurs copies. En les alignant, on se rend compte que des variations existent entre les différentes séquences. On peut, par exemple, détecter des indels, des microsatellites ayant des nombres différents de répétition, ou bien du polymorphisme affectant une seule base (SNP). Les SNP sont d'ailleurs la source la plus commune de polymorphisme à l'intérieur d'un génome. Ce constat permet d'envisager la détection *in silico* de ces polymorphismes à l'intérieur de gènes, en se servant des bases d'EST. Cette stratégie a été employée chez le pin maritime pour lequel 18 498 EST sont disponibles. Pour cela un système automatique de recherche de SNP utilisant une série de logiciels a été mis au point.

Figure B13 : Description de la méthode automatique utilisée pour la détection de SNP dans les bases de données d'EST de pin maritime

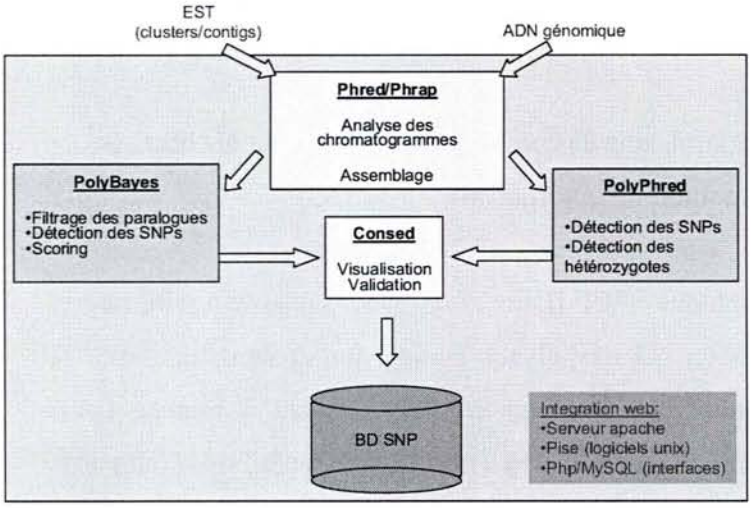


Figure B14 : Localisation dans l'aire de répartition du pin maritime des 30 mégagamétophytes utilisés pour la détection des SNP de référence.

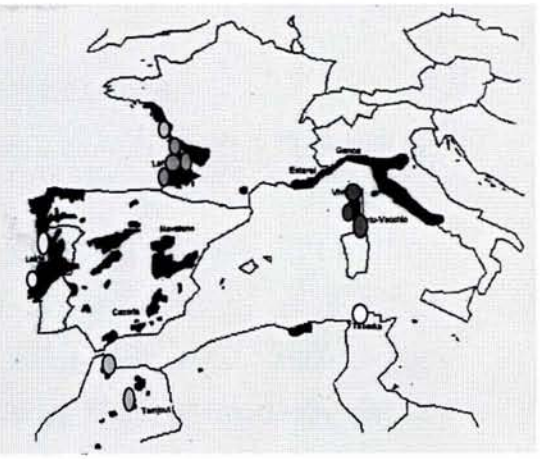


Tableau B5 : Description des 65 « vrais » SNP utilisés pour mettre au point la méthode automatique. Identifiant du gène, taille du fragment séquencé, comparaison entre les deux méthodes, efficacité de la méthode automatique et nombres de faux positifs (SNP détectés uniquement par la méthode automatique).
C4H: Trans-cinnamate-4-monooxygenase; CAD: Cinnamyl alcohol dehydrogenase (deux fragments F1 et F2); PAL: Phenylalanine ammonia lyase; Korrigan: endo-1,4-b-D-glucanase; GRP: Glycine-rich protein; MYB-like TF: MYB-like transcription factor; ACC oxidase: amino-cyclopropane-carboxylic acid oxidase; CesAn: Cellulose synthase; CCoAOMT: Caffeoyl CoA O-methyltransferase; AGP: Arabinogalactan protein.

Nom du gène	Taille du fragment (pb)	Nombre de vrais SNP détectés par inspection visuelle	Accession dans dbSNP	Nombre de total SNP détectés par la méthode automatique	Nombre de vrais SNP détectés par la méthode automatique	Nombre de vrais SNP non-détectés	Nombre de faux positifs
C4H	539	10	ss16208982;8985-8987;8991;8994-8998;9000	10	10	0	0
CAD-F1	537	4	ss16209001;9005;9008;9010	3	3	1	0
CAD-F2	523	4	ss16209011;9012;9015;9016	4	3	1	1
PAL	547	9	ss16209020-9025;9028;9029;9032	7	7	2	0
Korrigan	937	5	ss12709589-ss12709593	7	5	0	2
GRP	479	9	ss12709575-ss12709585	8	8	1	0
MYB-like TF	494	2	ss12709586-ss12709587	3	2	0	1
ACC oxydase	270	1	ss12709588	1	1	0	0
CesA7	553	2	ss12709573-ss12709574	2	2	0	0
CesA4	489	1	ss12709572	0	0	1	0
CesA3	490	7	ss12709565-ss12709571	6	6	1	0
CCoAOMT	492	7	ss16209076-ss16209082	3	3	4	0
AGP	321	4	ss16209070;9071;9073;9074	4	4	0	0
Total	6671	65		58	54	11	4

Les détails concernant la mise au point de cette chaîne d'analyse sont présentés à la figure B13 ainsi qu'à l'annexe V.

Brièvement, pour chaque contig, les chromatogrammes (fichiers bruts issus du séquenceur automatique) ont été ré-analysés (relecture des bases et assignation d'un score de qualité), ré-assemblés et alignés grâce au logiciel *Phred-phrap* (Ewing et al. 1998, Ewing and Green 1998, http://www.phrap.org). Ces alignements ont ensuite été traités par le logiciel *Polybayes* (Marth et al. 1999) qui a la particularité de détecter les variations de séquences de type SNP. De plus, ce dernier associe à chaque site polymorphe une valeur de probabilité indiquant les chances que ce SNP soit réel. Cette valeur, notée P_{SNP}, dépend de la qualité de lecture de la base nucléotidique considérée et du nombre de séquences présentant un polymorphisme.

La mise au point de cette méthode automatique de détection de SNP passe par l'évaluation de son efficacité. Pour cela, un lot de 65 vrais SNP détectés visuellement à l'aide d'un logiciel d'alignement (*BioEdit*, téléchargeable à : http://www.mbio.ncsu.edu/BioEdit/bioedit.html) a été utilisé comme jeu de référence. Ces SNP ont été détectés sur 13 fragments de gènes (tableau B5) séquencés sur 30 mégagamétophytes provenant de la totalité de l'aire naturelle du pin maritime (figure B14). L'ensemble de ces fragments couvrait 6 671 pb, la fréquence des SNP était de 1 SNP / 102,6 pb. Différents paramètres de *Phred-Phrap* (ex: critères d'assemblage et d'alignement plus ou moins stricts) et de *Polybayes* (valeur seuil de P_{SNP} pour tenir compte d'un SNP) ont été réglés de manière à retrouver un maximum de SNP de référence tout en évitant la détection de faux-positifs.

Outre l'efficacité de détection de cette méthode automatique, il était important d'évaluer les probabilités de détecter des SNP dans la base d'EST. Cette probabilité dépend de trois paramètres. Tout d'abord elle dépend de la fréquence allélique (P_A) des différents variants des SNP. Cette valeur ne peut pas être connue dans une base d'EST, étant donné que ceux-ci sont mélangés et peuvent être issus de plusieurs individus, et du fait que cette fréquence est dépendante de la population utilisée. Le second paramètre important est le nombre d'EST par contig (e). Enfin le troisième concerne le nombre de génomes haploïdes (g) utilisés pour construire la banque d'ADNc. Par exemple, si on utilise un seul individu diploïde, $g = 2$. Le détail des calculs ayant servi à déterminer la probabilité d'observer un SNP est présenté à l'annexe V.

B.2 Cartographie génétique

B.2.1 Principe de la cartographie génétique

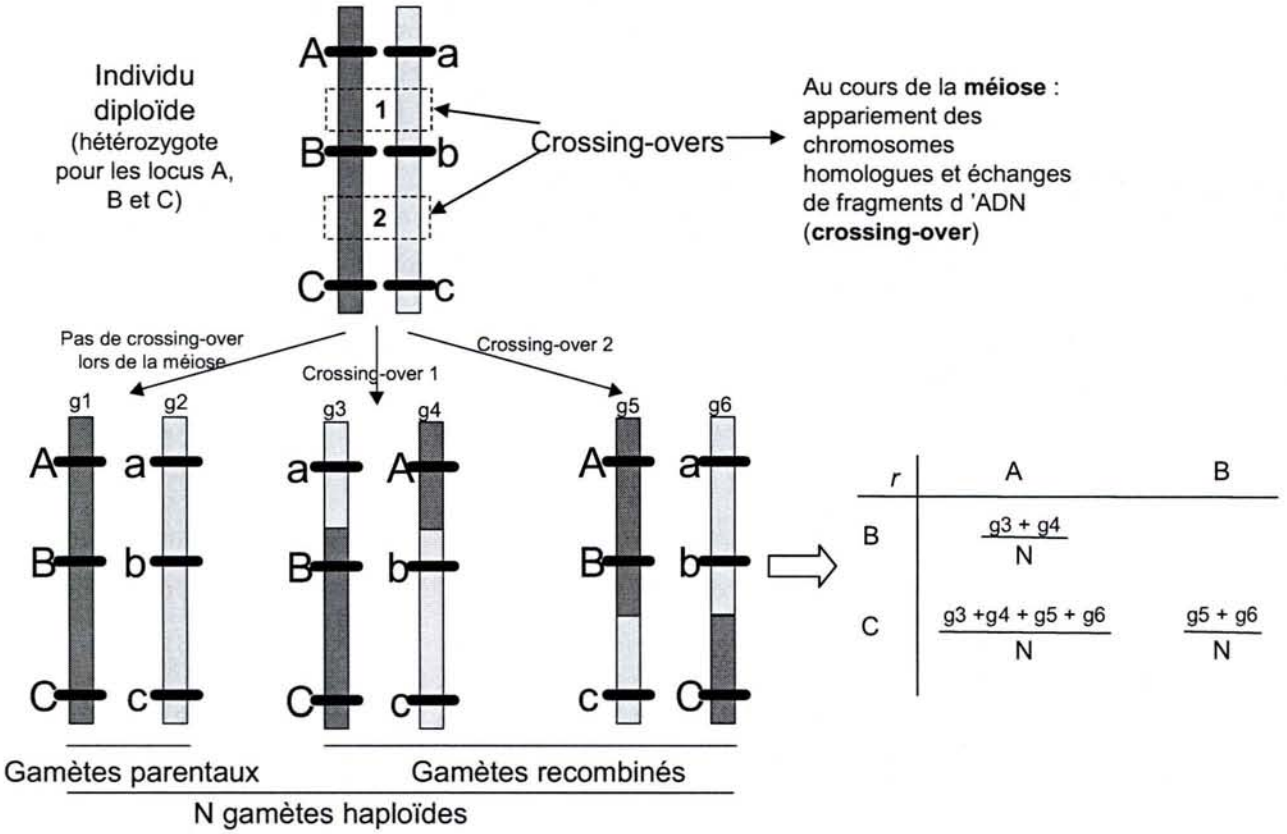
Le principe de la cartographie génétique est décrit en détails par D. de Vienne (1998). Brièvement, une carte génétique est une représentation simplifiée du génome où des marqueurs sont placés les uns par rapport aux autres le long de groupes de liaison, représentant les chromosomes. La distance génétique entre deux marqueurs est proportionnelle à la probabilité qu'un événement de recombinaison (crossing-over) se produise entre ces marqueurs lors de la méiose. Elle est mesurée en centiMorgan (cM), du nom de Morgan qui fut le premier à décrire la liaison entre locus chez la drosophile (Morgan 1911).

La construction d'une carte génétique passe tout d'abord par l'obtention d'une population en ségrégation pour différents locus. Le cas idéal est d'avoir un nombre important d'individus dans cette population afin d'augmenter le nombre de gamètes recombinés et donc de mieux estimer les distances entre locus. Ceci est possible chez les plantes, où des croisements contrôlés peuvent être réalisés pour obtenir des descendances ayant de grands effectifs. Une méthode de choix chez les plantes annuelles consiste à obtenir des lignées isogéniques recombinantes (de Vienne 1998) obtenues par autofécondations successives. Néanmoins, ce genre de descendance est techniquement très difficile à obtenir chez les gymnospermes si l'on considère leur floraison tardive (environ 8 ans pour le pin maritime) et leur caractère fortement allogame. D'autres stratégies ont dû être employées pour ces espèces, comme par exemple l'utilisation de familles de plein-frères ou de demi-frères.

Une fois la population de cartographie disponible, il est nécessaire de caractériser tous les individus de cette population à l'aide de marqueurs génétiques. Ces marqueurs doivent être dans un nombre suffisant pour pouvoir saturer le génome, c'est-à-dire faire en sorte que tous les points du génome soient liés à au moins un autre marqueur de la carte. Dans ce cas, on se retrouve avec autant de groupes de liaison que de chromosomes. Depuis une vingtaine d'années, l'essor de la biologie moléculaire a permis de fournir un nombre presque illimité de marqueurs moléculaires.

Une fois ces données obtenues, des outils statistiques et mathématiques permettent de tester la liaison entre marqueurs, de calculer les distances les séparant et de les ordonner dans les groupes de liaison (Liu 1998).

Figure B15 : Principe de la cartographie génétique : calcul du taux de recombinaison entre locus. Le taux de recombinaison (*r*) entre deux locus est la proportion de gamètes recombinés par rapport au nombre total de gamètes formés (N). Il y a indépendance entre les locus lorsque le nombre de gamètes de types parental est égal au nombre de gamètes de type recombiné (*r*=0,5).



Encadré B3 : Test d'indépendance des locus : le LOD score.

Valeur LOD (lod score) = $Z(r)$

↓

Logarithme de :

$$\frac{\text{Probabilité que 2 marqueurs soient liés [à une valeur de r, } L(r)]}{\text{Probabilité qu 'ils ne soient pas liés [} r = 0{,}5,\ L(1/2)]}$$

↓

$$Z(r) = \log_{10} [\, L(r) / L(1/2) \,]$$

Lorsque $Z(r) \geq 3$: l'absence de liaison est rejétée (**liaison**)

Lorsque $Z(r) < 3$: l 'absence de liaison est acceptée

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Le premier test effectué est un test du χ^2 qui vise à tester la ségrégation de chaque marqueur. Lorsque la ségrégation d'un marqueur est significativement différente des proportions mendéliennes attendues, on parle de distorsion de ségrégation. Ceci peut être expliqué, d'un point de vue biologique, par la liaison physique du marqueur avec un locus létal ou semi-létal, soit plus simplement, d'un point de vue statistique, par un échantillonnage trop faible.

Le second test consiste à tester l'hypothèse d'indépendance pour chaque paire de locus (hypothèse H0: "$r = 0,5$" ou "autant de gamètes parentaux que de gamètes recombinés") afin de tester leur liaison (figure B15). La méthode la plus employée est celle du LOD score (*Logarithm of the Odds ratio*, Morton 1955, encadré B3). Ce paramètre s'obtient en calculant le rapport de vraisemblance testant la probabilité que les deux locus soient liés contre la probabilité qu'ils ne le soient pas.

L'étape suivante consiste à calculer la distance génétique entre chaque paire de locus. Celle-ci dépend directement du taux de recombinaison (r), qui est la proportion de gamètes recombinants par rapport au nombre total de méioses. Toutefois, le taux de recombinaison est une représentation biaisée de la distance génétique étant donné que plusieurs crossing-over peuvent avoir lieu entre deux locus. Ainsi plusieurs fonctions de cartographie ont été définies pour tenir compte des recombinaisons multiples. Haldane (1919) fut le premier à proposer une fonction. Sa distance s'écrit de la manière suivante :

$$m_H = -50 \ln(1-2r),$$

où m_H est la distance génétique de Haldane et r le taux de recombinaison.

Cette formulation a été améliorée par Kosambi (1944), dont la fonction tient compte de l'interférence (phénomène dû à l'encombrement spatial d'un chiasma empêchant d'avoir deux crossing-over très proches). Sa fonction s'écrit :

$$m_K = 25 \ln[(1+2r)/(1-2r)],$$

où m_K est la distance génétique de Kosambi.

Une fois les distances calculées entre chaque couple de marqueurs, il faut ordonner les marqueurs à l'intérieur de chaque groupe de liaison à l'aide d'un test multipoint. Etant donné que la construction d'une carte génétique nécessite un grand nombre de tests statistiques sur un grand nombre de locus, des logiciels spécifiques ont été développés. Une liste de logiciels de cartographie génétique est présentée au tableau B6, mais les plus couramment utilisés sont JOINMAP (Stam 1993) et MAPMAKER (Lander 1987). Nous reviendrons plus en détails sur les spécificités de ces logiciels dans la partie concernant les résultats de cartographie chez le pin maritime.

Tableau B6 : Liste des logiciels de cartographie génétique.

logiciel	Fonctions	Type de populations	Plate-forme	Accessibilité	Auteur-Réf.
MAPMAKER	Analyse de liaisons construction de carte	F2,backcross,RIL,DH	Unix,MAC	domaine public	E.Lander,P.Green, J. Abrahamson, A.Barlow,M.Daly,S.Lincoln,L.Newburg //ftp-genome.wi.mit.edu/ distribution/ software/ mapmaker3
JOINMAP	Analyse de liaisons construction de carte Comparaison de cartes	F2,backcross,RIL,DH, F1 outbreds	Unix,MAC,PC	payant	P.Stam (Wageningen) http://www.cpro.dlo.nl/cbw/
LINKAGE (FASTLINK)	Analyse de liaisons	Pedigrees variés	Unix, PC	domaine public	M.Lathrop ; J. Ott, J. Lalouel, C. Julier http://linkage.rockefeller.edu/software/linkage
MAPL	Analyse de liaisons construction de carte	F2,backcross,RIL,DH	PC	domaine public	Y. Ukaï (Japon) http://peach.ab.au-tokyo.ac.jp/~ukai ftp://peach.ab.au-tokyo.ac.jp
GMENDEL	Analyse de liaisons construction de carte	F2,backcross,RIL,DH, F1 outbreds	Unix	domaine public	S.Knapp sknapp@helix.css.orst.edu
CARTHAGENE	Création de cartes consensus	F2,backcross,RIL, F2 intercross, F1 outbreds (phase connue)	Unix	domaine public	P. Chabrier, C. Caspin, T. Schiex http://www.inra/fr/bia/T/Carthagene
MAPCOMP	Comparaison de cartes	F2,backcross,RIL,DH, F1 outbreds	PC	domaine public	R. van Berloo (Wageningen) http://www.dpw.wau.nl/pv/PUB/MapComp/
PGRI	Analyse de liaisons construction de carte, Carte consensus	F2, backcross, RIL, DH, F1 outbreds, open pollinated	Unix	?	B. Liu benliu@unity.ncsu.edu
TETRAMAP	Génotypes parentaux Calcul du taux de double-réduction Analyse de liaison	F1 allogames autotétraploïdes	PC	BioSS, SCRI, Dundee, UK	C. Hackett hacke@scri.sari.ac.uk
OUTMAP	Construction de carte	F1 outbreds	PC	Payant 330 AU$	D.Whitaker, E.R. Williams http://www.ffp.csiro.au/software/outmap

Figure B16 : Alignement de la carte physique (gauche) et de la carte génétique (droite) pour le chromosome 7 de l'Homme (NIH/CEPH Collaborative Mapping Group 1992).

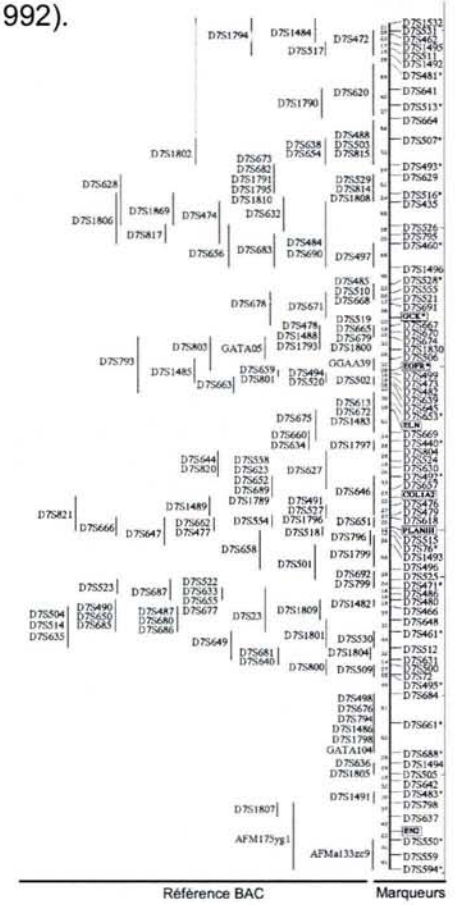
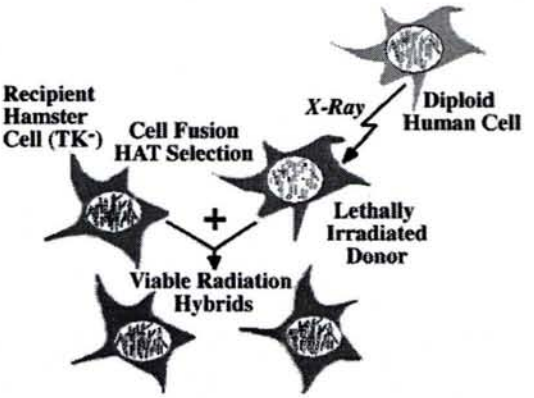
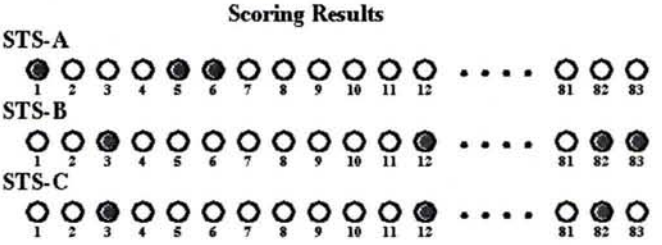


Figure B17 : Principe de la cartographie du génome de l'Homme par la technique des hybrides d'irradiation.

1/ Construction des hybrides par cassure de l'ADN aux rayons X.



2/ « Génotypage » des panels d'hybrides : présence / absence des locus.



La cartographie génétique est souvent mise en parallèle avec la cartographie physique. La métrique utilisée par cette dernière est le nombre de bases nucléotidiques de longs fragments d'ADN, généralement clonés dans des BAC ou des YAC[13], l'objectif ultime étant d'obtenir la séquence complète du génome considéré. Les cartes obtenues par ces deux approches complémentaires peuvent être alignées en cartographiant des séquences connues sur la carte génétique ou en localisant des marqueurs génétiques sur la carte physique (figure B16). On peut aussi assigner les groupes de liaison de la carte génétique aux chromosomes en hybridant des sondes nucléotidiques *in situ* (technique FISH[14]).

Une autre méthode puissante de cartographie du génome peut être utilisée : la cartographie d'hybrides d'irradiation (Cox et al. 1990). Cette méthode est fondée sur la présence de marqueurs dans des fragments d'ADN coupés artificiellement par une exposition à des rayons X (on peut comparer cela à une méiose, figure B17). Elle est très employée chez les animaux (Hawken et al. 1999) du fait qu'elle ne nécessite pas, d'une part, de population de cartographie possédant de nombreux descendants et d'autre part de polymorphisme (présence/absence des marqueurs dans les clones). Néanmoins, cette méthode ne sera pas discutée ultérieurement car elle n'est que peu utilisée chez les plantes, pour des raisons techniques (difficulté d'obtenir des hybrides somatiques).

Le séquençage complet du génome d'un arbre forestier, le peuplier, a été réalisé dans le cadre d'un projet financé par le DOE (*Department of Energy*, USA). La séquence complète des 19 chromosomes du peuplier, disponible depuis fin 2003 (http://genome.jgi-psf.org/poplar0/poplar0.home.html), sert de référence pour les études réalisées chez les arbres forestiers (Bradshaw et al. 2000, Bhalerao et al. 2003, Brunner et al. 2004). Néanmoins, aucun projet de séquençage complet du génome d'un conifère n'est envisagé étant donné sa taille et la quantité de séquences répétées. La cartographie génétique semble donc être la seule alternative efficace qui s'offre à l'étude du génome des conifères.

---

[13] BAC/YAC : *Bacterial / Yeast Artificial Chromosome*. Vecteurs bactériaux ou de levure permettant de cloner de très grands fragments d'ADN.

[14] FISH : *Fluorescent In Situ Hybridization*. Technique basée sur l'hybridation sur les chromosomes de fragments d'ADN marqués de manière fluorescente.

B.2.2 Les stratégies de cartographie génétique utilisées chez les arbres forestiers

B.2.2.1 Stratégies utilisant le mégagamétophyte des conifères

Chez les gymnospermes, le tissu nourricier haploïde entourant les jeunes plantules est d'origine maternelle. Ce tissu, appelé mégagamétophyte ou endosperme (figure A11), peut permettre de construire la carte génétique du parent femelle, aussi bien dans le cas d'un croisement "*open*" (plusieurs pères inconnus, descendance de demi-frères) que d'un croisement dont le père est connu (descendance de plein-frères), ou d'une autofécondation (descendance F2). En effet, on peut observer une ségrégation de type 1:1 dans la descendance pour les locus hétérozygotes chez la mère (figure B18). Le caractère haploïde du mégagamétophyte implique que ce type de ségrégation est valable aussi bien pour les marqueurs codominants que pour les marqueurs dominants. Cette méthode a été très utilisée pour obtenir les premières cartes génétiques de conifères à l'aide de marqueurs de type RAPD ou AFLP (Tuilseram et al. 1992, Nelson et al. 1993, Kaya et Neale 1995, Yazdani et al. 1995, Paglia et al. 1998, Travis et al. 1998, Krutovskii et al. 1998, Remington et al. 1999, Costa et al. 2000, Hayashi et al. 2001, Hayashi et al. 2002, Gosselin et al. 2002).
Les quantités d'ADN extraites d'un mégagamétophyte sont suffisantes pour la mise en oeuvre de techniques de marquage fournissant de nombreux locus par réaction (marqueurs multiplex), comme c'est le cas des RAPD et des AFLP. Néanmoins ces quantités d'ADN semblent trop limitées pour le génotypage[15] de marqueurs "simple-locus", comme par exemple les microsatellites ou les marqueurs géniques. Une autre méthode a donc été développée pour cartographier des arbres adultes.

B.2.2.2 La stratégie du double-pseudo testcross

Pour les mêmes raisons qu'il est difficile d'obtenir des RILs chez les arbres forestiers, il est difficilement envisageable de réaliser de "vrais" backcross. Néanmoins, on peut rechercher des configurations de type testcross au niveau des marqueurs dans des descendances F1. En effet, si un locus est hétérozygote chez un parent P1 et homozygote chez l'autre (P2) on se retrouve virtuellement dans le cas d'un testcross au niveau de ce locus. Dans ce cas, la descendance F1 ségrègera dans les proportions 1:1 et ce locus sera informatif pour P1. D'autres types de ségrégation peuvent aussi être observés en fonction du caractère

[15] Génotypage : détermination des variants génétiques à un site dans un échantillon d'ADN.

Figure B18 : Stratégies utilisées pour la cartographie génétique chez les gymnospermes :

A/ Stratégie du double-pseudo backcross
B/ Stratégie utilisant le mégagamétophyte haploïde (cas particulier d'une autofécondation)

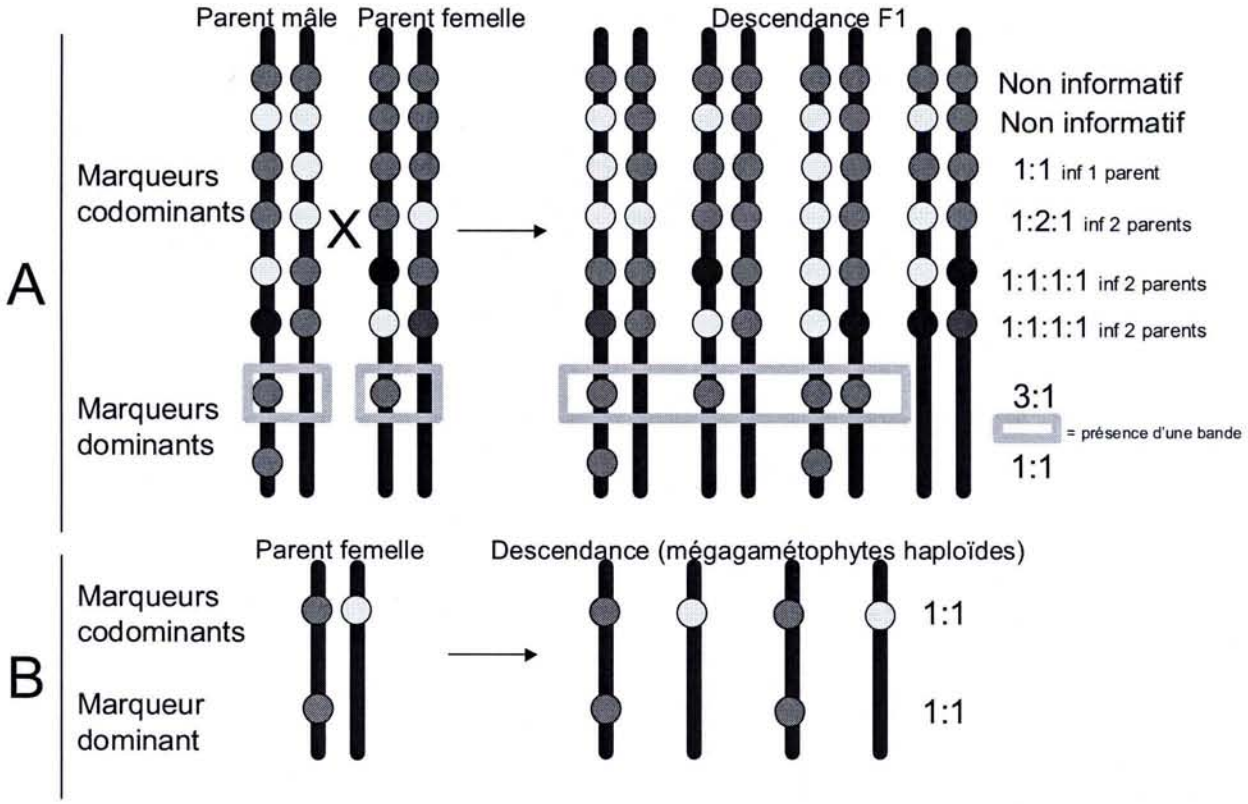
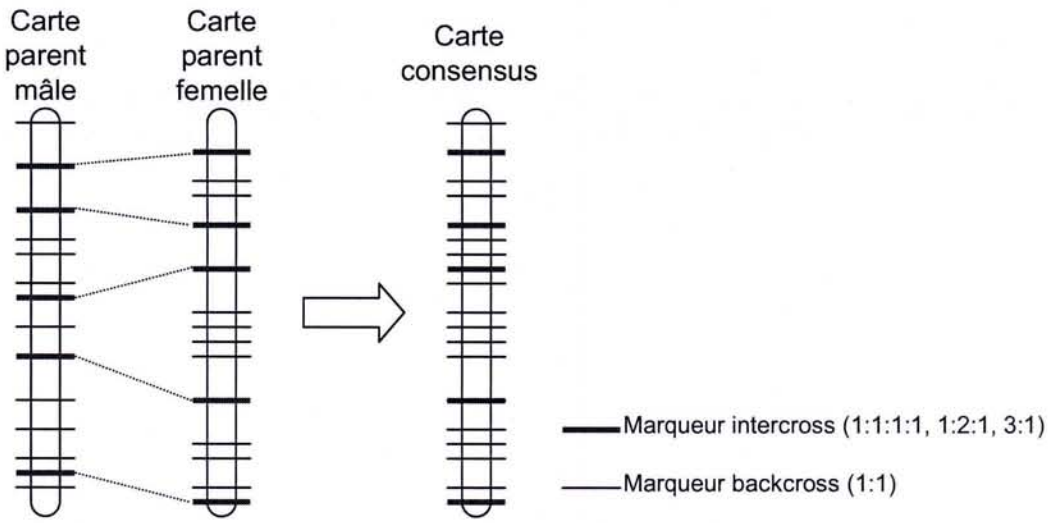
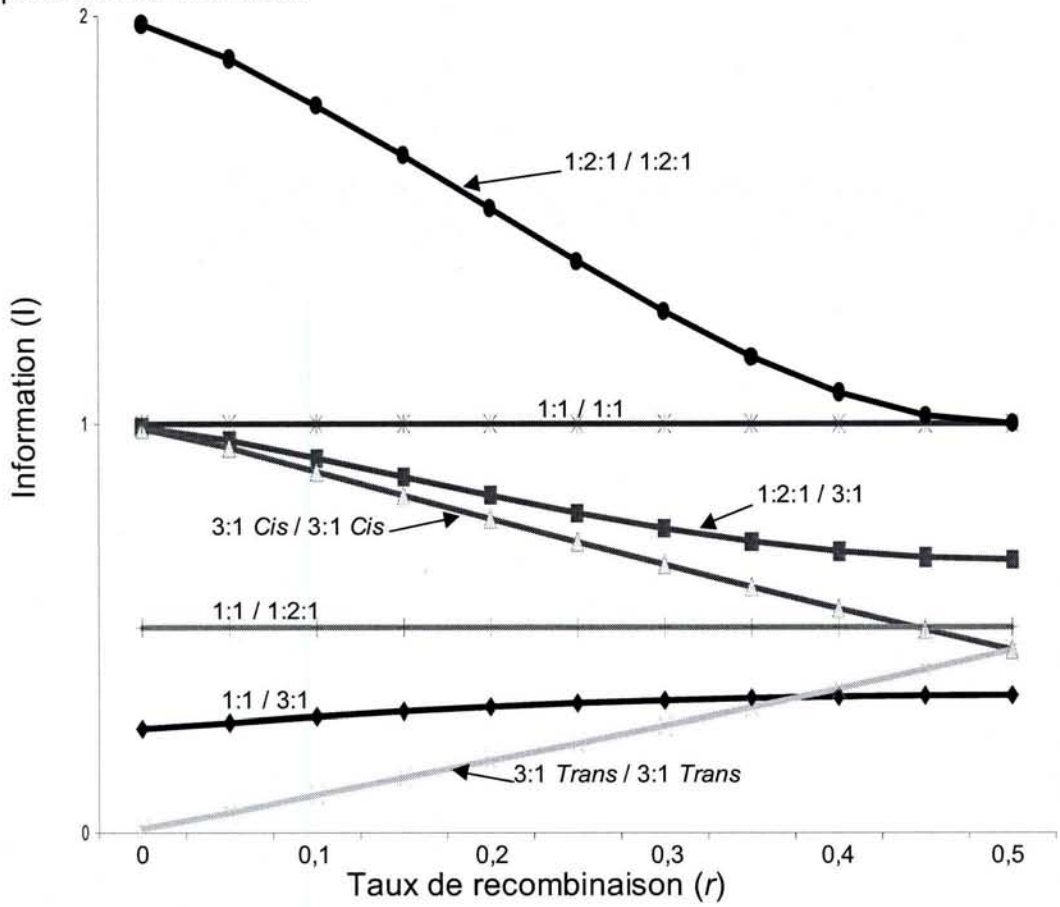


Figure B19 : Stratégie du double-pseudo backcross : construction des deux cartes parentales et d'une carte consensus.



Encadré B4 : Information du taux de recombinaison en fonction du type de ségrégation. Dans la figure ci-dessous, l'information a été calculée pour les différents couples de marqueurs possibles. L'information obtenue avec des marqueurs de type backcross (1:1 / 1:1, en rouge) a été prise comme référence.



Equation des fonctions d'information I(*r*) correspondant à différentes configuration de couples de marqueurs dans différentes types de croisement :

Croisement de type backcross :
$I(r)_{1:1/1:2:1} = 1 / r(1-r)$ (Allard 1956)

Croisement de type F2 :
Cas de 2 marqueurs codominants :
$I(r)_{1:2:1/1:2:1} = 2(1-3r+3r^2) / r(1-r)(1-2r+2r^2)$ (Allard 1956)
Cas de 2 marqueurs dominants en position *Cis* :
$I(r)_{3:1/3:1} = 2(3-4r+2r^2) / r(2-r)(3-2r+r^2)$ (Ritter et al. 1990)
Cas de 2 marqueurs dominants en position *Trans* :
$I(r)_{3:1/3:1} = 2(1+2r^2) / (2+r^2)(1-r^2)$ (Allard 1956)
Cas d'un marqueur dominant et d 'une marqueur codominant :
$I(r)_{3:1/1:2:1} = 2 + (1-r)^2 / r(2-r) + (1-2r)^2 / 2(1-r+r^2) + r^2 / (1-r)^2 + (1-2r)^2 / 2r(1-r)$ (Allard 1956)

Croisement entre deux parents hétérozygotes (double-pseudo testcross) :
Cas d'un marqueur backross et d'un marqueur codominant :
$I(r)_{1:2:1/1:1} = 1 / 2r(1-r)$ (Plomion et al. 1997)
Cas d'un marqueur backross et d'un marqueur dominant :
$I(r)_{3:1/1:1} = (1+2r-2r^2) / 2r(1-r^2)(2-r)$ (Ritter et al. 1990)

dominant ou non du marqueur considéré, ou bien du nombre d'allèles présents chez les parents. Ces différentes configurations sont présentées à la figure B18. L'obtention d'un nombre suffisant de marqueurs hétérozygotes chez chacun des parents permet de construire leur carte. Il est à noter que dans le cas de locus ayant des ségrégations de type 1:2:1, 1:1:1:1 ou 3:1, les deux parents étant informatifs, on peut aligner les groupes homologues mâle et femelle et ainsi réaliser une carte consensus du croisement (figure B19). Cependant, il convient de noter que ces différents types de ségrégation n'ont pas la même information quant à la précision du taux de recombinaison (encadré B4).

La stratégie du double-pseudo testcross chez les arbres forestiers présente l'avantage que ces derniers étant fortement hétérozygotes, on peut avoir accès à un nombre important de locus informatifs. De plus, on peut mesurer les caractères adultes pour la détection de QTL, ce qui aurait nécessité de stocker les mégagamétophytes dans le cas de la stratégie présentée plus tôt. Enfin, les quantités d'ADN ne sont plus limitantes. Cette méthode a été mise au point pour la première fois chez les arbres forestiers par Grattapaglia et Sederoff (1994) et a été appliquée aussi bien chez les gymnospermes (Devey et al. 1994, Groover et al. 1994, Kubisiak et al. 1995, Devey et al. 1996, Gocmen et al. 1996, Jermstad et al. 1998, Sewell et al. 1999, Arcade et al. 2000, Lerceteau et al. 2000, Iwata et al. 2001, Wilcox et al. 2001, Komulainen et al. 2003, Yin et al. 2003) que chez les angiospermes (Verhaegen et Plomion 1996, Barreneche et al. 1998, Marques et al. 1998, Cervera et al. 2001, Casasoli et al. 2001), en utilisant de nombreux types de marqueurs différents. Cette méthodologie a également été grandement utilisée chez les arbres fruitiers (Liu MengJun 1998) comme par exemple pour le pommier (Hemmat et al. 1994, Liebhard et al. 2003) ou le pêcher (Foolad et al. 1995, Foulongne et al. 2003).

B.2.3 Description des populations utilisées pour la construction des cartes génétiques de pin maritime et d'autres espèces de conifères

B.2.3.1 Population "F2"

Le croisement d'un pin maritime de provenance corse (C10) et d'un pin maritime de provenance landaise (L146) a abouti à une descendance F1 d'où a été sélectionné un individu hybride (H12; figure B20). Etant donné que les provenances auxquelles appartiennent les deux grands-parents présentent des caractéristiques phénotypiques très contrastées (forte croissance en hauteur, tronc sinueux, résistance au ravageur *Matsucoccus feytaudii* pour la

Figure B20 : Description du croisement F2.

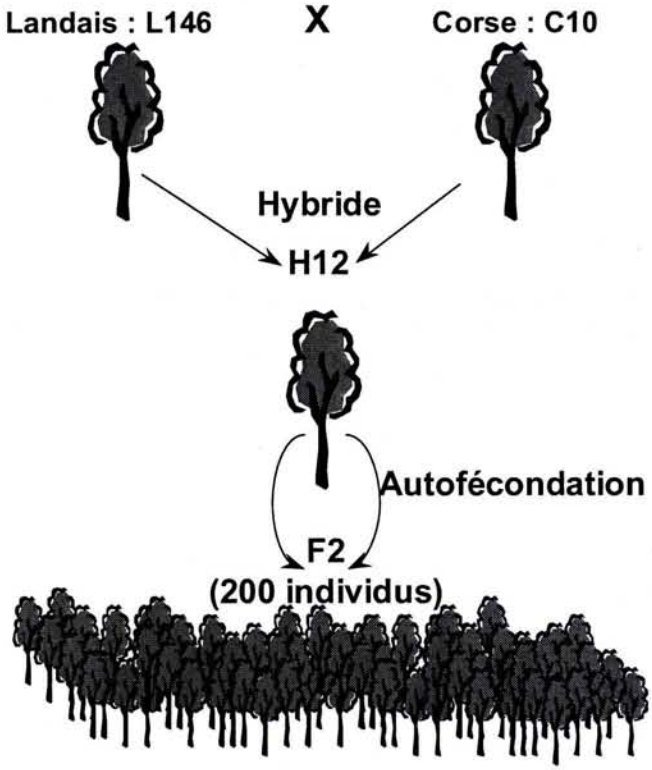
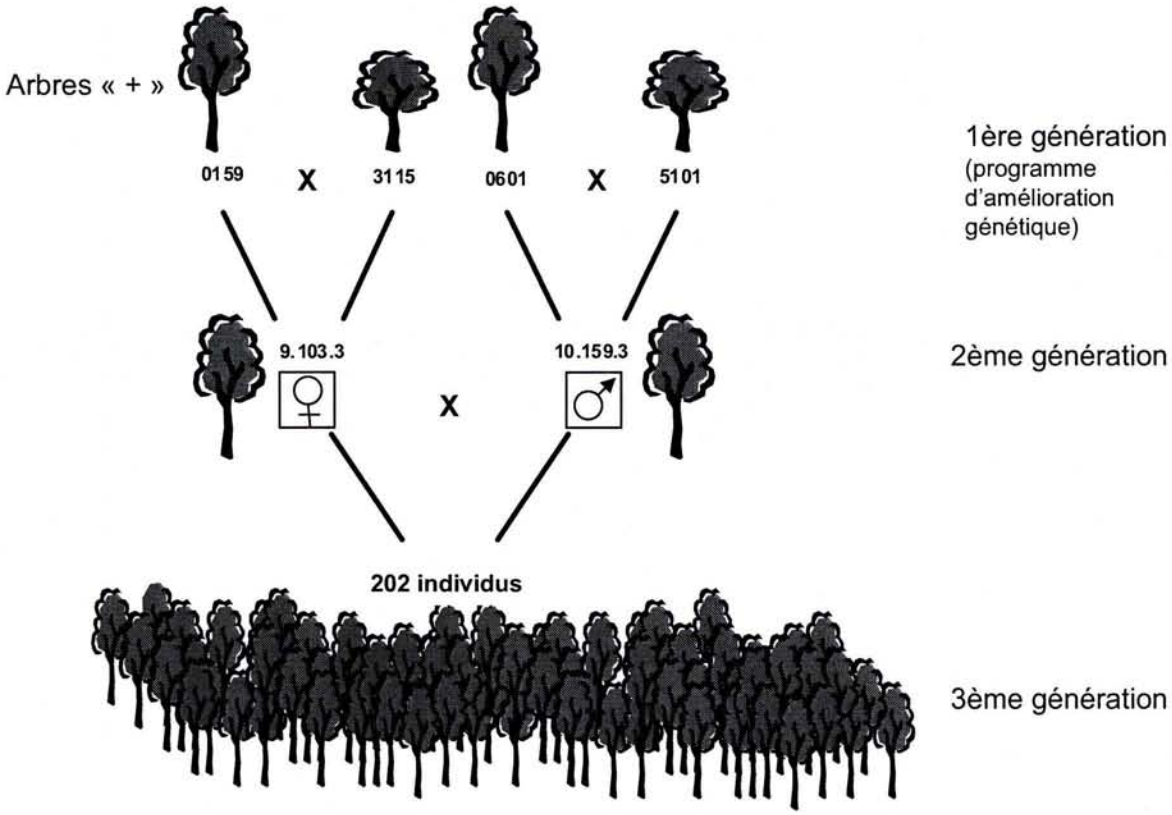


Figure B21 : Description du croisement G2.



race landaise *vs.* croissance modérée, tronc rectiligne et sensibilité à *M. feytaudii* pour la race corse), ainsi qu'un taux de différenciation élevé (G_{ST} = 0,092 en utilisant des marqueurs SSR, Mariette et al. 2002), et que, de plus, le pin maritime présente une forte diversité intra-population (H_T = 0,832 en utilisant des marqueurs SSR, Mariette et al. 2002), on peut s'attendre à ce que H12 soit hautement hétérozygote. De ce fait on se rapproche de la conformation obtenue par le croisement de deux individus issus de RILs. Une première carte génétique saturée de H12 a été obtenue à l'aide de RAPD et de protéines par Plomion et al. (1995a) à partir de mégagamétophytes issus d'une pollinisation libre et d'une autofécondation sur H12. Une seconde carte fondée sur uniquement l'autofécondation de H12 a été décrite par Costa et al. (2000), cette fois-ci sur 200 individus et en utilisant des marqueurs AFLP, RAPD et protéiques. Les deux cartes comportant des marqueurs communs ont pu être alignées (disponibles à l'URL : http://www.pierroton.inra.fr/genetics/pinus/map1.html).
Les embryons descendant de l'autofécondation de H12 ont été plantés et du matériel diploïde (greffé) est maintenant disponible pour ce croisement. Les marqueurs jalons RAPD de la carte décrite par Plomion et al. (1995a) ont été génotypés sur ce matériel diploïde (Plomion et O'Malley 1996) afin de pouvoir rajouter des marqueurs sur cette carte sans être limité en terme de quantité d'ADN.

B.2.3.2 Population "G2"

La population de cartographie "G2" (figure B21) correspond à une famille de 202 plein-frères plantés en automne 1982 à Malente (Gironde, France) et appartenant à la 3ème génération du programme d'amélioration génétique du pin maritime (figure A14). Les arbres ont été abattus en 1997 et greffés afin de pouvoir toujours accéder à du matériel végétal. Une description détaillée de ce croisement se trouve à l'annexe I (Paragraphe *2.1. Mapping population*). La stratégie de cartographie pour cette population est celle du double-pseudo testcross. Il est important de noter que contrairement à la population "F2" qui provient du croisement de grand-parents d'origines très éloignées, la population "G2" a été obtenue à partir du croisement de 4 grands-parents landais (appartenant donc à une même provenance). Néanmoins, étant donné la forte diversité génique observée à l'intérieur de la population landaise (H_T = 0,786 pour des marqueurs SSR, Mariette et al. 2002), on peut tout de même s'attendre à ce que la population "G2" soit polymorphe pour un grand nombre de marqueurs. De plus les 4 grands-parents ont été sélectionnés pour des caractéristiques phénotypiques (notamment la croissance) les différenciant.

Tableau B7 : Cartes génétiques utilisées dans le cadre du projet de cartographie génétique comparée chez les conifères (CCGP)

Espèces	Type de population	Nombre de groupes de liaison	Référence
Pinus taeda (*reference*)	F1	12	Sewell et al. 1999
Pinus pinaster	F1	12	Chagné et al. 2002, annexe I
Pinus sylvestris	F1	12	Komulainen et al. 2003
Pinus elliottii	F1	15	Brown et al. 2001
Pinus radiata	F1	12	Wilcox et al. 2001 / Devey et al. 1999
Pseudotsuga menziesii	F1	17	Jermstad et al. 1998
Picea abies	F1	13	Scotti et al. en préparation, annexe IV

B.2.3.3 Les autres cartes utilisées dans le cadre de la cartographie comparée chez les conifères

L'étape préalable à la cartographie génétique comparée est la construction de cartes génétiques saturées pour différentes espèces apparentées. Une description des cartes génétiques de sept espèces de Pinaceae utilisées dans le cadre d'un projet de cartographie génétique comparée chez les conifères (Conifer Comparative Genome Project, 1999-2003, http://dendrome.ucdavis.edu/ccgp/), est présentée au tableau B7. Cinq espèces de pins ont été considérées (*P. elliottii*, *P. pinaster*, *P. radiata*, *P. sylvestris*, *P. taeda*) ainsi que le sapin de Douglas (*Pseudotsuga menziesii*) et l'épicéa commun (*Picea abies*). La carte génétique de *Pinus taeda* (Sewell et al. 1999) est celle qui a été utilisée comme carte de référence, étant donné que la plupart des marqueurs orthologues ont été développés chez cette espèce. Il est à noter que cette carte est une carte consensus issue des données de deux cartes distinctes (pedigree *base* et *qtl*, Devey et al. 1994 et Groover et al. 1994).

La comparaison des cartes génétiques du pin maritime et de l'épicéa commun a aussi fait l'objet d'un projet européen (projet ANACONGEN, 1997-2000). Ce projet a permis la construction des cartes génétiques du pin maritime (annexe I) et de l'épicéa commun (annexe VI).

Enfin, un autre projet européen, indépendant de ce travail de thèse (UHD-MAP, 2000-2004, www.neiker.net/UHDfor/), a aussi consisté à comparer les cartes génétiques d'espèces de Pinaceae européennes (pin maritime, pin sylvestre et épicéa commun).

Dans le cadre de cette thèse, une page web a été réalisée afin de dresser une liste, la plus exhaustive possible, des cartes génétiques publiées chez les conifères : http://www.pierroton.inra.fr/genetics/labo/mapreview.html.

B.2.4 Méthodes et logiciels utilisés pour la construction des cartes génétiques de pin maritime

B.2.4.1 Protocole utilisé pour la construction de la carte de H12 à partir de tissus diploïde et haploïde

La carte de l'hybride H12 obtenue à partir des données issues des mégagamétophytes a été construire à l'aide du logiciel *MAPMAKER* v2.0 pour *MacIntosh* (Lander et al. 1987),

selon le protocole décrit par Plomion et al. (1995a). Brièvement, des groupes ont été obtenus avec un LOD score minimum de 6 et un taux de recombinaison maximum de 0,30 (fonction *group*). Les marqueurs qui se sont retrouvés non-liés en utilisant ces critères ont été éventuellement rattachés à des groupes de liaisons en utilisant la fonction near. Les marqueurs ont alors été ordonnés à l'intérieur de chaque groupe grâce à une analyse multi-point (fonction *first order*) en tenant compte de la vraisemblance de l'ordre obtenu (*Log-likelihood*). Il est important de noter que tous les marqueurs ont été recodés dans les deux phases de liaison, celle-ci étant inconnue. *MAPMAKER* a été utilisé pour faire les cartes des deux phases à partir des données diploïdes. Il a donc par contre utilisé le logiciel *JOINMAP* (Stam et al. 1993), pour calculer les taux de recombinaison entre marqueurs ségrégeant 1:1 et 3:1 pour réaliser une carte regroupant l'ensemble de l'information diploïde et haploïde.

B.2.4.2 Protocole utilisé pour la construction de la carte génétique de la G2

Le protocole employé pour la construction de la carte génétique de la population "G2" est détaillé dans l'annexe I (Paragraphe *2.3 Mapping procedure*). Brièvement, les marqueurs 1:1 informatifs ont été regroupés chez chaque parent, et une carte génétique a été obtenue pour chacun d'entre eux en utilisant le logiciel *MAPMAKER* (même protocole que celui décrit plus haut). Une carte consensus a ensuite été obtenue en rajoutant les marqueurs intercross 3:1 grâce au logiciel *JOINMAP v3.0* (van Ooijen et Voorips 2001). Etant donné le faible niveau d'information entre marqueurs 1:1 et 3:1 (encadré B4), il a fallu trier les marqueurs intercross les plus informatifs, à savoir que seuls les marqueurs 3:1 liés sur les deux cartes ont été retenus. De plus une attention particulière a été portée sur le fait que ces marqueurs 3:1 ne modifient pas l'ordre obtenu à l'aide des marqueurs 1:1.

B.2.4.3 Protocole utilisé pour l'ajout de nouveaux marqueurs et pour tester l'ordre de ces marqueurs

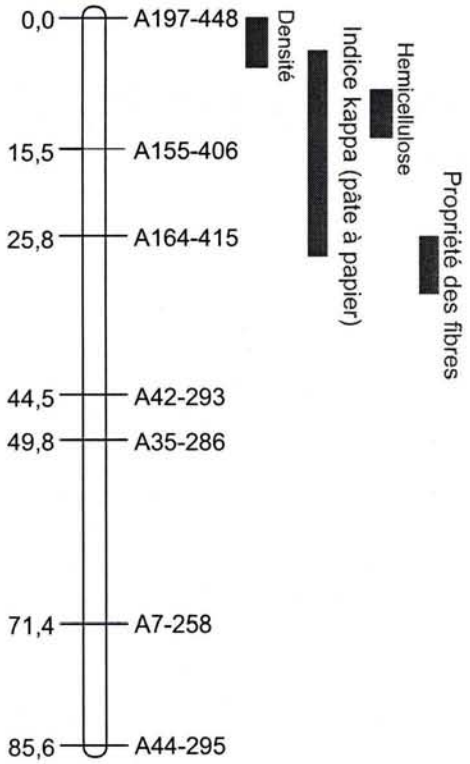
Le protocole employé pour rajouter des marqueurs un à un dans une des cartes génétiques F2 ou G2 est présenté dans l'annexe III pour le cas particulier des marqueurs ESTP. Brièvement, chaque marqueur additionnel est ajouté au fichier de données, puis localisé sur la carte avec le logiciel *JOINMAP v3.0*. Pour chaque groupe de liaison, le logiciel *CARTHAGENE* (Schiex et al. 1997, http://www.inra.fr/bia/T/CarthaGene/) a permis de vérifier l'ordre des marqueurs. En effet, ce logiciel permet d'optimiser l'ordre des marqueurs à

l'intérieur de chaque groupe à l'aide d'algorithmes d'optimisation basés sur différents modèles statistiques. Par exemple, l'ordre des marqueurs sur la carte a été calculé par analogie avec un des modèles les plus étudiés en informatique, le problème du *Wandering Salesman* (un voyageur doit visiter toutes les villes d'un pays tout en parcourant une distance minimale). Les algorithmes utilisés pour résoudre ce problème consistent par exemple à insérer les marqueurs un par un à la carte en considérant leur LOD ou leur distance 2-points, en commençant par la distance ou le LOD le plus fort vers le plus faible (méthodes *nicemapl* et *nicemapd*). D'autres méthodes consistent à permuter les marqueurs entre eux (*flip* ou *polish*), ou à perturber volontairement l'ordre des marqueurs (*annealing* ou *taboo search*) jusqu'à obtenir une meilleure vraisemblance de l'ordre obtenu (*log-likelihood*).
Les cartes génétiques obtenues ont été mises en forme grâce au logiciel *MapChart* (Voorips 2001).

B.2.5 Détection de QTL chez le pin maritime

Une des principales applications des cartes génétiques concerne l'identification des régions chromosomiques contrôlant des caractères quantitatifs. Ces régions sont appelées QTL (*Quantitative Trait Loci*). De la même façon que pour les marqueurs moléculaires à hérédité mendélienne, des données quantitatives ont été mesurées sur les individus ayant servi à construire les cartes "F2" et "G2". Plusieurs types de caractères ont été pris en compte dans les deux pedigrees de pin maritime.
Des QTL impliqués dans la réponse à un stress hydrique ont été localisés sur la carte "F2" dans le cadre de la thèse de P. Costa (1999). Les caractères ont été mesurés sur des jeunes plants (2 ans) de pin maritime soumis à un stress hydrique (pas d'arrosage pendant 6 semaines) et sont principalement des paramètres physiologiques comme la consommation en eau, les potentiels hydrique et osmotique, la teneur relative en eau de la plante, les échanges gazeux, l'efficacité d'utilisation de l'eau, les teneurs en minéraux, la croissance et la biomasse. Trois régions chromosomiques contenant plusieurs QTL à effets forts à modérés ont été identifiées sur les groupes 6, 8 et 12. De plus, sur ces mêmes groupes, des colocalisations entre les QTL détectés et des protéines différentiellement exprimées dans les mêmes conditions hydriques (présentées sous forme de PQL, *Protein Quantitative Loci*) ont été mises en évidence, suggérant l'importance de ces protéines dans le contrôle génétique de la réponse à un déficit hydrique. Une autre étude comparable a été réalisée sur le croisement "G2" par Brendel et al. (2002). Des QTL impliqués dans l'efficacité d'utilisation de l'eau ont été

Figure B22 : Exemple de cluster de QTL observé pour des caractères liés à la qualité du bois (groupe de liaison 12 mâle, extrait de la thèse de D. Pot 2004)



identifiés sur plusieurs groupes de liaison des cartes mâle et femelle. Lors de cette étude aucune co-localisation des QTL d'efficacité d'utilisation de l'eau (mesuré grâce à la discrimination isotopique du ^{13}C) et de QTL de croissance radiale de l'arbre (mesurée grâce à la largeur des cernes de croissance) n'a été observée.

En complément de ces études portant sur la compréhension de la réponse génétique du pin maritime à un stress abiotique, des études ont été menées sur la compréhension des mécanismes moléculaires et génétiques contrôlant la qualité du bois. Ces études ont fait l'objet de deux projets européens (projet GENIALITY, 1998-2001 et projet GEMINI, 2000-2004) dont les résultats concernant l'architecture génétique de ces caractères (détection de QTL, identification de gènes candidats, études d'association) ont été présentés dans la thèse de D. Pot (2004). Des caractères impliqués dans les caractéristiques anatomiques et physiques (ex: densité du bois, morphologie des fibres) et chimiques (taux de lignine, de cellulose et d'hemicellulose) du bois ont été mesurés sur les arbres adultes (15 ans) de la famille G2. Les mesures de densité ont été réalisées au niveau de plusieurs cernes de croissance afin d'apprécier la stabilité des QTL détectés au cours de différentes années, ainsi qu'en distinguant le bois de printemps (ou bois initial) du bois d'été (ou bois final) à l'intérieur d'un même cerne afin d'apprécier la stabilité des QTL au cours de la saison de végétation. Des co-localisations entre QTL suggèrent la présence de "hot spots" génomiques pour le contrôle des propriétés du bois chez le pin maritime. En effet, des clusters de QTL ont été mis en évidence sur les cartes mâle et femelle sur les groupes de liaison 8f, 4m, 12f, 12m, 2m, 10f, ou 5m (figure B22).

Des QTL impliqués dans la qualité du bois ont aussi été détectés dans un autre pedigree de pin maritime (Markussen et al. 2003). L'absence de marqueurs communs entre les deux cartes n'a pas permis de comparer la position des QTL obtenus dans ces deux études.

Figure B23 : Cartographie génétique comparée : macrosynténie entre les génomes de l'Homme et de la souris. Exemple de reconstruction du génome de la souris à l'aide de blocs issus du génome humain (Hudson et al. 2001).

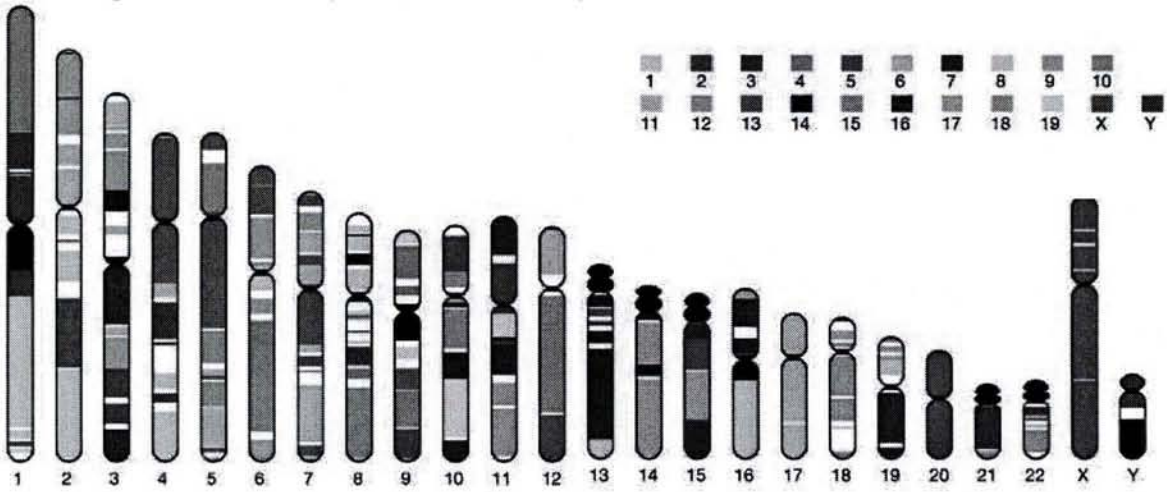


Figure B24 : Cartographie génétique comparée : microsynténie. Exemple de comparaison entre deux fragments d'ADN entre *Ricketsia conorii* (haut) et *R. prowazekii* (bas), d 'après Ogata et al. (2001). Il est à noter que des duplications de gènes ou des inversions peuvent intervenir à ce niveau (sens des flêches).

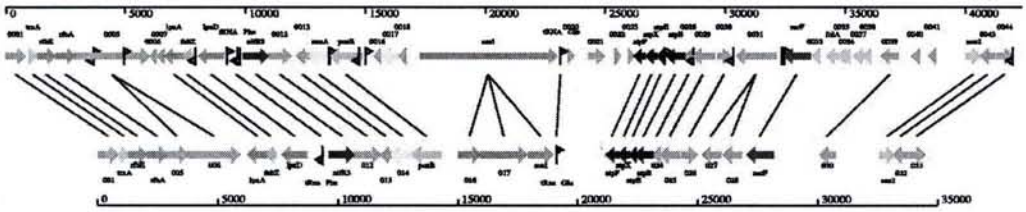
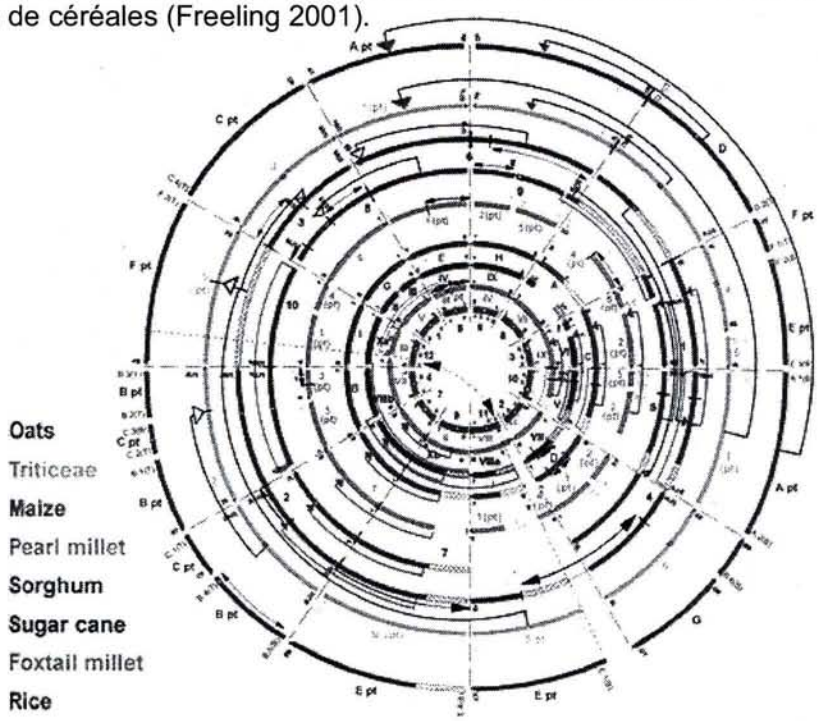


Figure B25 : Cartographie génétique comparée : comparaison entre les cartes génétiques d'espèces de céréales (Freeling 2001).



B.3 Cartographie génétique comparée

B.3.1 Définition et exemple chez les plantes

Depuis les années 80, grâce à l'essor des techniques de marquage moléculaire, des cartes génétiques ont été construites chez de nombreuses espèces, aussi bien animales que végétales (de Vienne, 1998). La disponibilité de ces cartes a très vite été utilisée comme outil de comparaison des génomes. Les deux objectifs principaux de la cartographie génétique comparée sont, d'un point de vue évolutif, de reconstruire la structure du génome ancestral de genres, voire de familles phylogénétiquement proches, et d'un point de vue pratique, de transférer les informations relatives aux cartes génétiques à des espèces apparentées, comme par exemple la prédiction de gènes ou de régions (QTL) impliquées dans le contrôle génétique de caractères d'intérêts agronomiques. Cette discipline est fondée sur des études de conservation de la synténie[16], elle-même subdivisée en macro-synténie s'intéressant au contenu et à l'ordre de marqueurs ou de gènes sur des groupes de liaison homologues (figure B23), et en micro-synténie s'intéressant à la comparaison des séquences nucléotidiques des gènes orthologues (figure B24). Dans le cadre de cette thèse, c'est au niveau de la macrosynténie que je me suis placé.

Les génomes de différentes familles de plantes ont été comparés. Chez les céréales, plusieurs auteurs (Moore et al. 1995, Gale et Devos 1998, Keller et Feuillet 2000, Devos et Gale 2000, Freeling 2001), ont montré une bonne conservation de la synténie au niveau de grands fragments chromosomiques. Une représentation en cercle des cartes génétiques à haute densité de marqueurs construites pour le riz, le millet, la canne à sucre, le maïs, l'avoine, le blé et l'orge est présentée à la figure B25. Une représentation de l'alignement de ces cartes peut également être obtenue *via* la base de données Gramene (Ware et al. 2002, http://www.gramene.org). La cartographie génétique comparée entre les espèces de la famille des Poacées a permis de mieux comprendre l'évolution de leur génome et a pu révéler des réarrangements de la structure (inversion, duplication, translocation, délétion) et du nombre (polyploïdisation[17]) de leurs chromosomes. Des résultats similaires ont été obtenus chez les plantes dicotylédones, comme chez les Crucifères (*Arabidopsis*, Brassicacées : Kowalski et al. 1994, Lagercrantz et al. 1996, Cavell et al. 1998, Lan et al. 2000, Babula et al. 2003), les

[16] Conservation de la synténie : au moins deux paires de gènes orthologues se trouvant sur le même chromosome définissent un segment où la synténie est conservée.

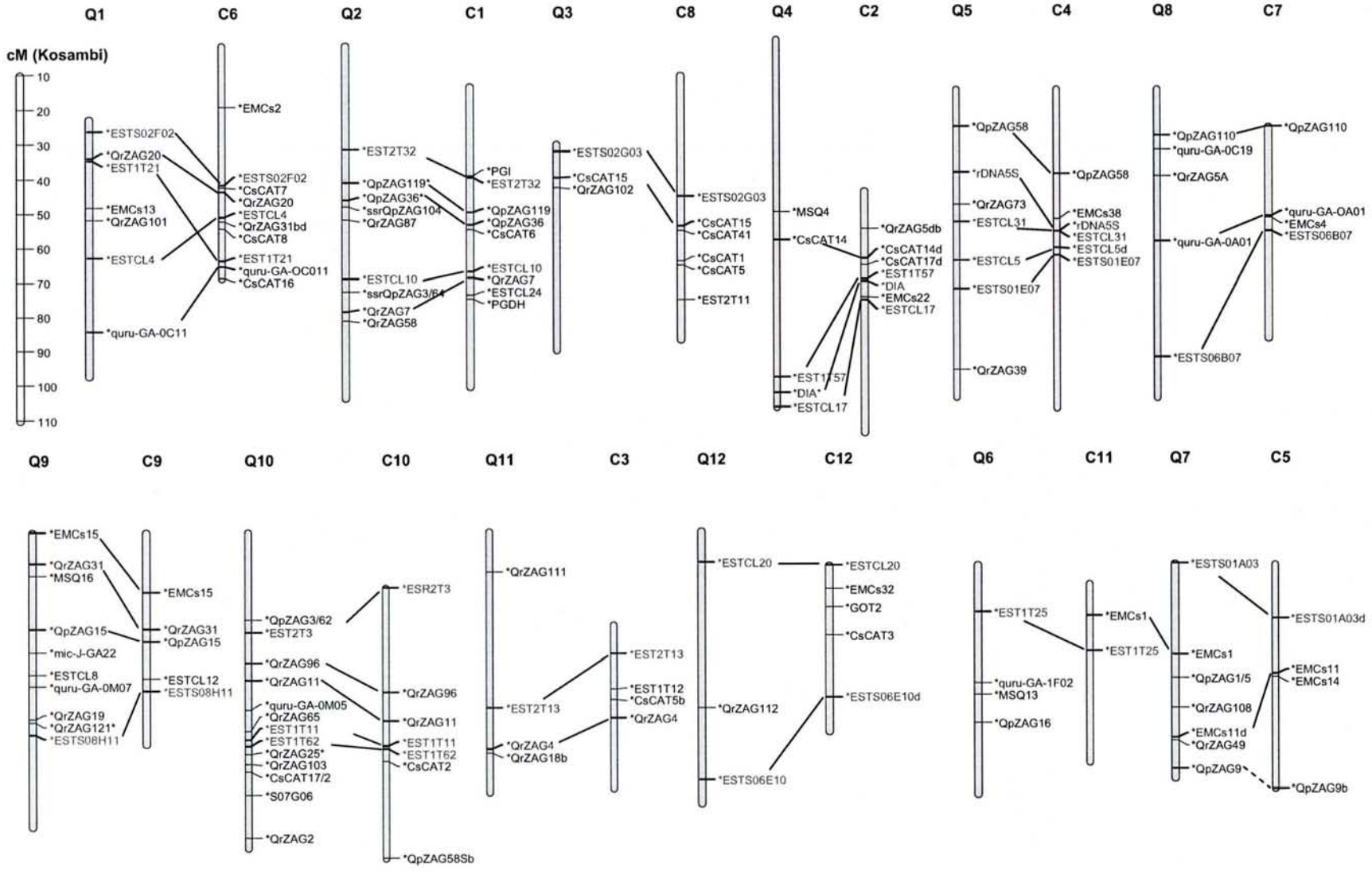
[17] Polyploïdisation : cas où un cellule ou un organisme a trois ou plus de trois jeu de chromosomes.

Solanacées (pomme de terre, tomate, piment : Tanksley et al. 1988, 1992, Livingstone et al. 1999, Doganlar et al. 2002a), les légumineuses (haricot, pois, lentille : Weeden et al. 1992, Menancio-Hautea et al. 1993, Humphry et al. 2002) ou les espèces pérennes fruitières (Sosinski 2000, Foulongne et al. 2003, Ruiz et Asins 2003). Ces résultats, combinés avec les efforts de séquençage systématique réalisés chez des espèces modèles comme *Arabidopsis* pour les dicotylédones (The Arabidopsis Genome Initiative 2000) ou le riz pour les monocotylédones (Yu et al. 2001) devraient permettre à terme d'utiliser l'information (par exemple le clonage positionnel[18] d'un gène d'intérêt) obtenue chez une espèce particulièrement bien étudiée chez d'autres espèces proches ne bénéficiant pas d'efforts de recherches aussi soutenus.

Du point de vue des applications de la cartographie génétique comparée, de nombreuses études ont montré l'intérêt de cette discipline. Je me limiterai à deux de ces exemples. Doganlar et al. (2002b) ont observé une conservation de la position de QTL contrôlant la couleur, le poids et la forme du fruit chez l'aubergine, la tomate, la pomme de terre et le poivron, ce qui démontrerait que les mêmes gènes seraient impliqués dans le contrôle de ces caractères chez les Solanaceae. D'un point de vue évolutif, ceci a permis aux auteurs de faire l'hypothèse que les mêmes gènes pourraient être soumis aux pressions de sélection provoquées par la domestication de ces espèces. Un autre exemple peut être pris chez les céréales où Sutton et al. (2003) ont utilisé les connaissances sur la synténie entre les génomes du blé et du riz pour identifier des gènes candidats impliqués dans le contrôle de la méiose chez le blé. Dans cette étude, les gènes de la zone chromosomique du riz qui était homologue à la région d'intérêt du blé (délétion *ph2a*), ont servi à retrouver des gènes candidats, par homologie de séquence avec les EST de blé. Cet exemple est très intéressant car le blé est une espèce à gros génome, pour lequel aucun projet de séquençage complet n'est envisagé pour l'instant, mais qui peut bénéficier des informations de séquences obtenues chez une espèce modèle à "petit" génome comme le riz, grâce à la cartographie génétique comparée.

[18] Clonage positionnel : méthode qui utilise les techniques de cartographie physique et de séquençage, pour identifier un gène dont la mutation est responsable de la variation d'un caractère, à partir de la connaissance de sa localisation dans le génome.

Figure B26 : Comparaison des cartes génétiques du chataignier (C) et du chêne pédonculé (Q) à l'aide de marqueurs microsatellites (bleu) et EST (rouge). Données fournies gracieusement par M. Casasoli.



B.3.2 Cartographie génétique comparée chez les arbres forestiers

A la différence des espèces annuelles de grande culture, la cartographie comparée des espèces forestières ne fait que débuter.

Des études de cartographie comparée ont été réalisées chez les espèces d'arbres forestiers angiospermes. L'utilisation d'hybrides interspécifiques de chênes (Saintagne 2003), de peuplier (Cervera et al. 2001) ou d'eucalyptus (Marques et al. 2001, Myburg et al. 2003) peut permettre d'avoir un aperçu de la conservation des génomes des espèces ayant servi à faire ces croisements. Dans ce cas, la construction des cartes parentales (stratégie du double-pseudo testcross, voir Partie B.2.2) et leur alignement grâce à des marqueurs intercross peut permettre de donner une idée sur la conservation de l'ordre des marqueurs entre ces espèces. Au-delà de ces études préliminaires, les cartes génétiques d'espèces de la famille des Fagaceae (chêne, châtaignier, hêtre) ont été alignées en utilisant des marqueurs orthologues de type microsatellites (Barreneche et al. 2004) ou basés sur des EST (M. Casasoli, communication personnelle, figure B26). Malgré la faible densité de marqueurs orthologues cartographiés, les génomes de *Quercus robur* et de *Castanea sativa* semblent être colinéaires[19].

Le cas des conifères et plus particulièrement celui des Pinaceae semble plutôt favorable à la cartographie génétique comparée. Ces espèces sont toutes diploïdes et la structure de leur génome est très conservée en terme de caryotype : à part *Pseudotsuga menziesii* pour lequel n=13 et *Pseudolarix amabilis* et *Pseudolarix kaempferi* pour lesquels n=22, tous les genres de la famille des Pinaceae, incluant la centaine d'espèces du genre *Pinus*, ont un caryotype présentant n=12 chromosomes métacentriques ou submétacentriques (voir Partie A.1.4). Une première étude comparative menée entre *Pinus taeda* et *Pinus radiata* a aussi montré une bonne colinéarité entre ces deux génomes (Devey *et al.* 1999). Une autre étude est aussi en cours de réalisation à l'Université de Laval (Canada) afin de comparer les cartes génétiques d'espèces d'épicéa canadiennes (*Picea glauca* et *Picea mariana*, Pelgas et al. 2004). Les premiers résultats montrent une bonne conservation de la synténie chez ces espèces.

[19] Colinéarité : il y a colinéarité quand au moins deux paires de locus orthologues sont dans le même ordre.

RESULTATS ET DISCUSSION

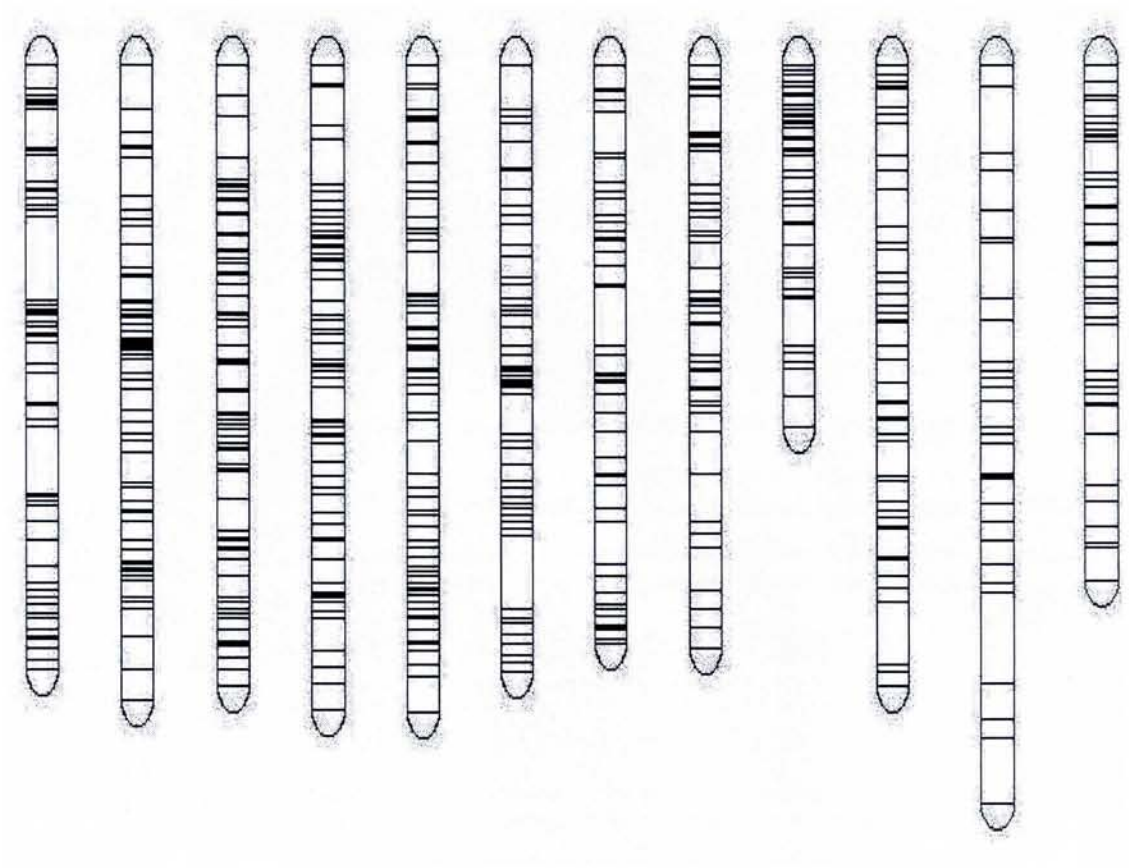


Tableau C1 : Evaluation des différents types de marqueurs pour la construction d'une carte génétique du pin maritime, pour la cartographie génétique comparée chez les conifères, en tant que gènes candidats et pour la détection de QTL.

	Développement	Cartographie génétique	Cartographie comparée	Détection QTL	Gène candidat
AFLP	?	?	?	?	?
SSR	?	?	?	?	?
SSR-ADNc	?	?	?	?	?
EST	?	?	?	?	?
SNP	?	?	?	?	?

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Partie C : Résultats et discussion

Je présenterai dans cette troisième partie les résultats obtenus concernant le développement de marqueurs moléculaires ainsi que la construction d'une carte génétique pour le pin maritime et sa comparaison avec les cartes d'autres espèces de conifères. La plupart des résultats sont présentés en détails et discutés dans les six articles référencés dans le texte et placés en annexe (I-VI).

C.1 Mise au point de marqueurs moléculaires chez le pin maritime et évaluation de leur utilité pour la cartographie génétique, la cartographie comparée et la détection de QTL

Les objectifs de la thèse ont été présentés en introduction. Le chapitre C.1 a pour but de discuter les avantages et inconvénients des différents marqueurs pour la cartographie génétique du pin maritime et la cartographie comparée des conifères. Le tableau C1 constituera le fil conducteur de cette discussion et sera rempli au fur et à mesure.

C.1.1 Marqueurs AFLP

Les résultats concernant l'obtention de marqueurs AFLP sont présentés à l'annexe I.

C.1.1.1 Obtention de marqueurs AFLP et évaluation de leur utilité pour la cartographie génétique du pin maritime

Un total de 52 combinaisons amorce/enzyme (tableau B1) a été utilisé pour générer 766 locus non-distordus ($p<0,01$) et polymorphes sur 90 arbres de la descendance G2. Un tiers (253) de ces marqueurs montrait une ségrégation de type 3:1. Le reste des marqueurs (ségrégation 1:1) était informatif pour le parent mâle (251) ou pour le parent femelle (262). Le nombre de marqueurs obtenus a été suffisant pour obtenir les cartes saturées des deux parents du croisement G2 (annexe I et partie C.2). De par la quantité de locus qu'ils fournissent, à faible coût et sans mise au point technique difficile, les marqueurs AFLP constituent des marqueurs de choix pour établir rapidement une carte génétique saturée (tableau C1).

Tableau C1 : Evaluation des marqueurs AFLP.

	Développement	Cartographie génétique	Cartographie comparée	Détection QTL	GC positionnel
AFLP	Développement techniquement simple, rapide et de faible coût Pas de connaissance au préalable du génome nécessaire	Nombre de marqueurs suffisant pour saturer le génome Répartition uniforme sur le génome	Non-transférable entre espèces	Marqueur dominant : moins informatif Possible en configuration backcross pour chaque parent	En équilibre de liaison avec les polymorphismes fonctionnels en population naturelle Non-codant
SSR	?	?	?	?	?
SSR-ADNc	?	?	?	?	?
EST	?	?	?	?	?
SNP	?	?	?	?	?

Le choix des combinaisons amorce/enzyme a été réalisé de telle sorte que les locus obtenus aient des chances d'être répartis uniformément sur le génome. En effet, la combinaison (*Eco*RI-*Mse*I) est moins sensible à la méthylation de l'ADN que d'autres combinaisons, comme par exemple celles utilisant l'enzyme de restriction *Pst*I. Ces dernières peuvent en effet cibler des régions faiblement répétées réparties en clusters au sein du génome, comme cela a été démontré chez le maïs (Vuylsteke et al. 1999). En revanche, les combinaisons *Eco*RI-*Mse*I utilisées présentent l'inconvénient de montrer des profils complexes et donc plus difficiles à interpréter, du fait de la nature complexe du génome des conifères. De ce fait, les amorces utilisées possédaient deux bases nucléotidiques sélectives pour la PCR pré-sélective et trois ou quatre pour la seconde PCR. Ainsi, le nombre de fragments variait de 60 à 145 (dont 8 à 29 polymorphes selon la combinaison).

C.1.1.2 Evaluation des AFLP pour la détection de QTL

Une des applications principales des cartes génétiques est l'identification de zones chromosomiques contrôlant des caractères quantitatifs : les QTL (*Quantitative Trait Locus*). La puissance de détection des QTL est fortement dépendante du type de marqueur (Liu 1998). Il est nécessaire pour pouvoir détecter des QTL de manière efficace de posséder des marqueurs polymorphes, bien répartis sur l'ensemble du génome et si possible codominants. Un choix judicieux de plusieurs combinaisons a permis de génotyper à moindre coût les 110 génotypes supplémentaires dont on disposait pour la détection de QTL dans le pedigree G2 (Pot 2004). Nous disposions donc de marqueurs bien répartis sur le génome. Néanmoins, les marqueurs AFLP sont dominants et biallélíques. Dans un pedigree de plein-frères, la seule manière de construire une carte consensus est d'utiliser les marqueurs ségrégeant dans les proportions 3:1. Or ce type de ségrégation ne permet pas de décrire exactement les effets génétiques des allèles de ces marqueurs au niveau d'un QTL car on ne peut pas distinguer les hétérozygotes des homozygotes pour l'allèle dominant. C'est pour cette raison que des marqueurs codominants ont été utilisés plutôt que des AFLP dans certaines études menées chez les arbres forestiers (Sewell et al. 1999). Néanmoins, dans le cadre du programme de génomique du pin maritime, la détection de QTL contrôlant les caractères liés à la qualité du bois a été réalisée indépendamment sur les deux cartes parentales du croisement G2 (Pot 2004), à partir des marqueurs AFLP ségrégeant dans les proportions 1:1.

Tableau C2 : Bilan du transfert de marqueurs microsatellites entre espèces de pins.

Espèce d'origine	Nombre de paires d'amorces testées	Nombre de paires d'amorces ayant donné une amplification simple bande	Nombre de paires d'amorces ayant donné une amplification multibande	Nombre de paires d'amorces polymorphes	Nombre de locus polymorphes	Annexe	Référence
Pinus radiata	11	1	7	0	0	II II II	G. Moran, communication personnelle Smith et Devey 1994 Fisher et al. 1998
Pinus radiata	107	58	8	10	12	IV	C. Echt, communication personnelle
Pinus strobus	11	8	3	0	0	II	Echt et al. 1996
Pinus halepensis	25	5	12	1	1	II	Keys et al. 2000
Pinus sylvestris	7	3	0	0	0	-	Soranzo et al. 1998
Pinus taeda	3	3	0	0	0	-	Elsik et Williams 2001
Total	164	78	30	11	13		

C.1.1.3 Evaluation des AFLP la cartographie comparée chez les conifères

Pour ce qui est de l'utilisation des marqueurs AFLP comme marqueurs orthologues pour la cartographie génétique comparée, le caractère non-codant et multilocus des AFLP pose un problème majeur pour le transfert de ces marqueurs entre espèces. Aucune comparaison de deux cartes génétiques d'espèces différentes à l'aide d'AFLP n'a été publiée jusqu'ici. Myburg et al. (2003) ont utilisé des marqueurs AFLP pour aligner les cartes parentales d'un hybride F1 issu du croisement entre deux espèces (*Eucalyptus grandis* et *E. globulus*), mais on ne peut parler de marqueurs orthologues *sensu stricto* car ces marqueurs sont spécifiques au croisement analysé et ne peuvent pas être utilisés pour d'autres espèces du genre *Eucalyptus*. Néanmoins, les marqueurs AFLP peuvent être utilisés pour comparer des cartes génétiques au sein d'une même espèce (Vuylsteke et al. 1999), mais la nature biallélique de ces marqueurs reste un handicap majeur car ils ne sont pas forcément toujours informatifs d'un pedigree à l'autre. Une alternative pour développer des marqueurs orthologues issus de la technique AFLP pourrait être l'utilisation de la méthode d'ADNc-AFLP (Bachem et al. 1996) pour cartographier des gènes exprimés. Cette approche a par exemple été utilisée chez le manioc par Suarez et al. (2000).

C.1.2 Marqueurs microsatellites

C.1.2.1 Transfert de marqueurs microsatellites génomiques entre espèces de pin

Les résultats sur le transfert de marqueurs microsatellites issus d'espèces proches de *P. pinaster* ont été publiés et sont présentés aux annexes II et IV.

C.1.2.1.1 Quelques marqueurs polymorphes

Au total, 164 paires d'amorces, développées chez d'autres espèces de pins, ont été testées chez le pin maritime (tableau C2). 78 marqueurs microsatellites sur 164 testés (47,5%) ont donné un produit d'amplification PCR de type simple bande chez le pin maritime. Le reste des paires d'amorces PCR n'a pas amplifié (34,5%) ou a donné des profils multibande (18%) sur gel d'agarose. Il est important de noter que les conditions d'amplification PCR de ces locus ont nécessité une mise au point fine et laborieuse de différents paramètres de la PCR dont la

température d'hybridation des amorces, le nombre de cycles de la PCR, la concentration en $MgCl_2$ (cofacteur de la *Taq* polymerase), la quantité d'ADN par réaction ou bien l'utilisation d'un *touchdown*[20] (Chou et al. 1992) au début de la réaction.

Les 78 paires d'amorces qui ont amplifié des profils « simple bande[21] » chez le pin maritime ont été testées sur les parents de la famille G2 et quatre de leurs descendants, ainsi que sur les grands-parents C10 et L146, l'hybride H12 et quatre de ses descendants F2, afin d'évaluer leur polymorphisme. Seules une paire d'amorces issue de *P. halepensis* et dix paires d'amorces issues de *P. radiata* se sont révélées polymorphes. Ces résultats montrant un faible nombre de locus polymorphes transférables (11 / 164, 7%) sont en accord avec d'autres résultats obtenus chez les pins par d'autres auteurs (tableau B2).

Ceci illustre les difficultés qui sont rencontrées pour transférer des microsatellites chez les conifères. Le transfert de locus microsatellites chez les conifères est caractérisé par plusieurs contraintes dues principalement à l'ancienneté et à la taille du génome de ces espèces (voir parties A.1.2 et A.1.4). Tout d'abord des mutations ont pu se produire au niveau des régions flanquant les motifs microsatellites (où s'hybrident les amorces PCR). Les pins ayant divergé des autres Pinaceae il y environ 140 Ma, on peut s'attendre à ce que ces zones ne soient pas conservées. A titre d'exemple, Kostia et al. (1995) ont observé de grandes différences dans les séquences flanquant les microsatellites chez le pin sylvestre.

Une autre explication des problèmes rencontrés lorsqu'on transfère des microsatellites entre espèces de pins porte sur la structure du génome des conifères. Morgante et al. (2002) ont montré que les microsatellites étaient préférentiellement associés aux séquences peu répétées chez les plantes. L'hypothèse peut s'appliquer également aux conifères : la proportion de microsatellites dans les séquences peu répétées serait plus importante que dans les régions hautement répétées. Néanmoins, étant donné que le génome des conifères contient une fraction hautement répétée plus importante par rapport aux autres familles de plantes, le nombre de microsatellites associés à cette fraction serait aussi beaucoup plus important. La probabilité de tomber sur un microsatellite situé dans les séquences répétées serait donc plus élevée que chez d'autres plantes. Ceci est cohérent avec le fait que l'on observe plus de profils multibandes chez les conifères et que le transfert de locus entre espèces est plus difficile.

Une fois que les locus ont pu être amplifiés avec succès chez une espèce différente, il faut encore que ces locus soient polymorphes afin d'être considérés comme des marqueurs

[20] *Touchdown* : méthode qui consiste à diminuer progressivement la température d'hybridation des amorces lors des premiers cycles de la PCR (ex : diminution de 1°C pour les 10 premiers cycles).

[21] Simple bande : cette expression sera utilisée dans ce manuscrit pour désigner des profils électrophorétiques peu complexes où un nombre réduit de fragments d'ADN est visualisé suite à la PCR. Ce terme est à opposer au terme « multibande » également employé dans ce manuscrit.

moléculaires. Là encore, les résultats obtenus montrent un faible nombre de locus polymorphes dans les pedigrees de cartographie utilisés. Ceci peut s'expliquer par le fait que des mutations ont pu se produire à l'intérieur des motifs répétés, créant des motifs imparfaits connus pour être moins polymorphes (Goldstein et Schlotterer 1999). Le cas extrême peut être que ces phénomènes aboutissent à la disparition du motif répété lui-même. Ce phénomène a été confirmé chez les pins par Kutil et Williams (2001) qui ont montré que le transfert de microsatellites pouvait être meilleur en utilisant des motifs trinucléotides parfaits. Une autre explication peut être que le motif répété s'est étendu et est devenu polymorphe uniquement dans l'espèce d'origine.

Nos résultats de transfert interspécifique de microsatellites peuvent être comparés à ceux obtenus chez d'autres espèces d'un même genre. Scott et al. (2003) ont montré que des microsatellites pouvaient être conservés chez les gymnospermes en transférant 6 locus microsatellites sur 8 testés entre des espèces du genre *Araucaria* ayant divergé il y a plus de 200 Ma. Chez les dicotylédones, des microsatellites développés chez le soja ont été transférés chez d'autres espèces du genre *Glycine* (Peakall et al. 1998). Chez ces espèces, 65% des locus microsatellites testés ont pu être transférés dans des espèces proches et 85% de ces locus transférés présentaient un motif microsatellite conservé et polymorphe. Chez les arbres fruitiers, 12 sur 41 (29,3%) des microsatellites mis au point chez le pêcher ont amplifié sur la totalité de 8 espèces de Rosaceae testées (genres *Prunus*, *Malus* et *Fragaria* ; Dirlewanger et al. 2002). Chez les monocotylédones, Chen et al. (2002) ont observé que des microsatellites issus du riz étaient facilement transférables chez d'autres espèces du genre *Oryza*, mais néanmoins peu transférables vers d'autres genres (maïs, blé). Pour les Triticeae, des microsatellites issus du blé hexaploïde ont été transférés avec succès chez 14 espèces du groupe *Triticum/Aegilops* (Sharma et al. 2002). Le nombre de locus microsatellite transférables entre genres différents est néanmoins modéré chez les plantes, par rapport aux résultats obtenus chez les animaux. L'exemple le plus frappant est le transfert de 17 locus microsatellites entre des espèces de poissons ayant divergé il y a 470 Ma (Rico et al. 1996) !

Tous ces résultats montrent que le transfert de microsatellites chez les plantes reste difficile en particulier chez les conifères. C'est pourquoi d'autres approches ont été employées pour développer des marqueurs orthologues transférables chez les conifères, comme par exemple l'utilisation des séquences codantes, plus conservées (cf C.1.2.2).

Figure C1 : Locus ssrNZPR1702 (26 descendants plein-frères) : profil sur gel de polyacrylamide montrant deux bandes ségrégeant à deux niveaux différents.

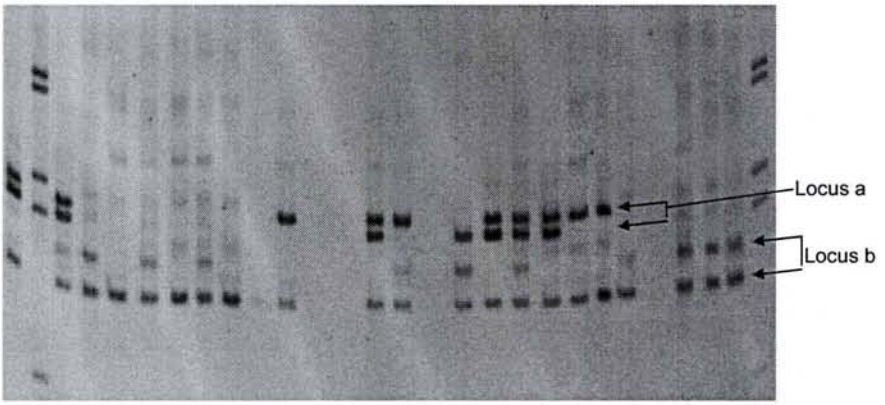


Tableau C3 : Homologie de séquence entre *P. radiata* et *P. pinaster* pour les locus microsatellites issus de *P. radiata*.

Locus	Motif répété	Amorce F	Amorce R	Temperature d'hybridation des amorces (°C)	Taille attendue (pb)	Homologie de séquence
NZPR1078	AC_{10}	tggtgatcaagccttttcc	gttgatgagtgatggcatgg	53	342	91,5%
NZPR114	CA_{15} .. CA_{13} TA_{22}	aagatgacccacatgaagtttgg	ggagctttataacatatctcgatgc	56	193	88,2%
NZPR1702_b	AC_{15} CA_{13}... AT_5	tatgattggaccattggggt	ccaaaccctcctccacatatc	53	187	Pas d'homologie
NZPR413	TG_{23} GT_6	tgaacctcgatggaatagcc	cccgccttgcatcaatta	53	253	89,1%
NZPR472	AC13	gagaaaattcaaccaccgga	ggttgtagggcagtgaatcc	53	309	89,4%
NZPR544	CA_5 AC_{12} TA_5	gcgatgtgcaacccttgata	tgctattccgtcaaaaaccc	56	286	86,1%
NZPR823	AC_{57}	tatcgggagcaagttatgcc	tgcactctttttcgtctcca	53	296	92,5%

Tableau C1 : Evaluation du transfert de microsatellites entre espèces de pins.

	Développement	**Cartographie génétique**	**Cartographie comparée**	**Détection QTL**	**GC positionnel**
AFLP	++	++	--	+/-	--
SSR (transfert)	Pas besoin de connaissance sur les séquences contenant des microsatellites	Marqueurs informatifs mais moins polymorphes que dans l'espèce d'origine Peu de locus polymorphes	Marqueurs peu transférables entre espèces (en terme de nombre de locus transférés avec succès) Locus orthologue	Bons marqueurs pour la détection de QTL (codominants, multialléliques)	Même problème que pour les AFLP
SSR (banque enrichie)	?	?	?	?	?
SSR-ADNc	?	?	?	?	?
EST	?	?	?	?	?
SNP	?	?	?	?	?

C.1.2.1.2 Transférabilité et orthologie des marqueurs microsatellites : application en cartographie comparée chez les conifères

Treize locus microsatellites génomiques ont pu être cartographiés sur au moins une des deux cartes génétiques du pin maritime. Le nombre de locus polymorphes est supérieur au nombre de paires d'amorces testées car, pour deux d'entre elles (NZPR823 et NZPR1702), deux locus polymorphes ont été révélés (figure C1). Ces locus ont tout de même été pris en compte car ils ont pu être clairement génotypés dans les descendances utilisées.

Les locus qui ont été cartographiés à la fois sur la carte de l'hybride H12 et sur la carte consensus de la G2 ont pu servir à aligner ces deux cartes (voir partie C.2.2). De la même manière, 9 locus étant déjà cartographiés sur la carte de *Pinus radiata*, ils ont pu servir comme marqueurs orthologues pour identifier les groupes de liaison homologues entre *P. radiata* et *P. Pinaster*.

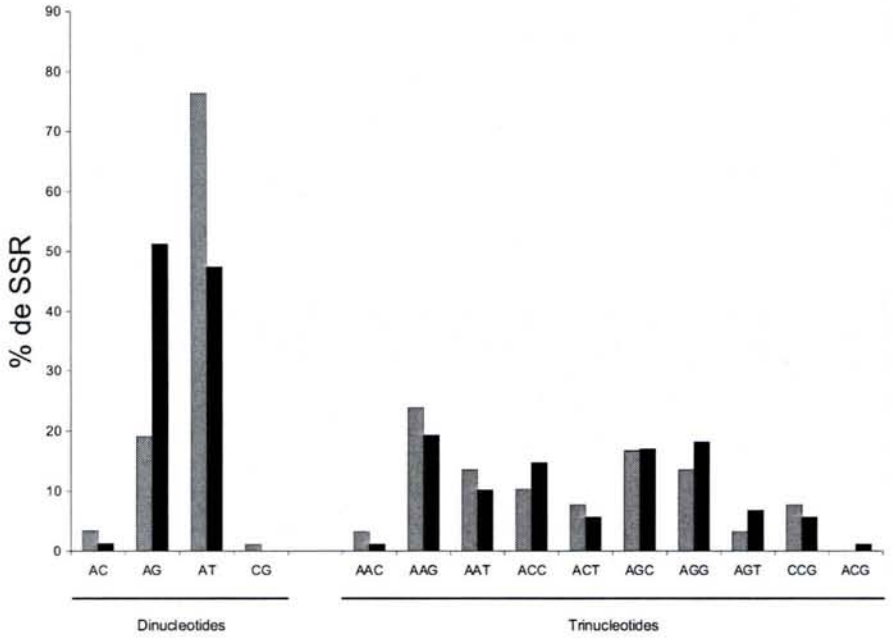
Sept locus ont été séquencés chez le pin maritime et leurs séquences ont été comparées à celles de l'espèce d'origine afin de vérifier l'orthologie de ces marqueurs. Le tableau C3, tiré de l'annexe IV, montre de bonnes similarités de séquence pour 6 marqueurs sur 7. En effet, le locus NZPR1702b ne présente pas d'homologie de séquence entre *P. pinaster* et *P. radiata* et ne contient pas de motif répété chez le pin maritime. De plus ce locus présente deux bandes distinctes sur gel d'agarose, ce qui laisse supposer qu'il est présent en deux copies dans le génome du pin maritime.

Le tableau C1 récapitule les conclusions qui peuvent être tirées sur le transfert de microsatellites chez les conifères. Cette approche reste longue et fastidieuse au regard du nombre de locus fournis *in fine*. Etant donné que la plupart des paires d'amorces provenant d'espèces de pins ont été épuisées, d'autres approches ont dû être envisagées pour développer des marqueurs microsatellites pour le pin maritime.

Tableau C4 : Résultats de la détection de motifs répétés dans les unigènes de *P. taeda* et *P. pinaster*.

	Pinus taeda			*Pinus pinaster*		
	Contigs	Singletons	Total	Contigs	Singletons	Total
Nombre de séquences analysées	8 070	12 307	20 377	2 893	5 001	7 894
Nombre de motifs dinucléotide (n>7)	41	48	89	41	35	76
Nombre de motifs trinucléotide (n>5)	103	56	159	43	44	87
Nombre de motifs tétranucleotide (n>5)	1	2	3	3	2	5
Nombre total	145	106	251	87	81	168
Enrichissement (%)			1,2%			2,1%

Figure C2 : Types de motifs répétés dans les unigènes de *P. taeda* (en gris) et *P. pinaster* (en noir).



C.1.2.2 Détection *in silico* de marqueurs microsatellites

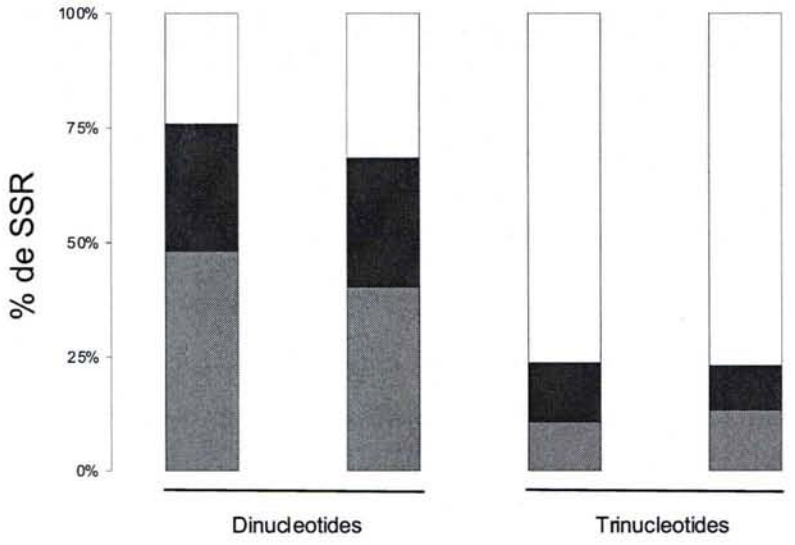
Les méthodes et les résultats obtenus par cette approche ont été publiés et sont décrits à l'annexe IV.

C.1.2.2.1 Quels types de motifs répétés se trouvent dans les régions codantes des conifères ?

La disponibilité d'un grand nombre d'EST pour des espèces de pin (principalement *P. taeda* et *P. pinaster*) permet d'envisager de détecter des motifs répétés dans ces séquences afin de développer des marqueurs microsatellites plus conservés et donc plus facilement transférables entre espèces. Le logiciel *SSRIT* a permis de faire cette détection à partir des deux lots de séquences codantes non-redondantes (unigènes) issus de l'assemblage des EST de *P. taeda* et de *P. pinaster*.

Au total, 251 et 168 motifs répétés ont été trouvés dans les unigènes de *P. taeda* et de *P. pinaster*, correspondant à des taux d'enrichissement des banques d'EST de 1,2% et 2,1%, respectivement (tableau C4). Etant donné que les longueurs cumulées des unigènes de *P. taeda* et de *P. pinaster* étaient de 8 900 kb et 4 400 kb, la probabilité de retrouver un motif répété dans les séquences codantes de pin peut être estimée à un motif microsatellite toutes les 26,2 à 35,5 kb. Ces résultats peuvent être comparés aux résultats obtenus à partir de banques génomiques chez d'autres espèces de pins : Kostia et al. (1995) ont obtenu un microsatellite tous les 100 à 500 kb chez *Pinus sylvestris*, ces valeurs variant en fonction du type de motif considéré (CT, GT, AT). De la même façon, Fisher et al. (1998) ont observé des valeurs encore plus faibles de l'ordre d'un microsatellite tous les 1,3 à 4,8 Mb chez *Pinus radiata*. Même si ces études ne sont basées que sur quelques types de motifs microsatellites, ces observations montrent que les séquences codantes possèdent plus de motifs répétés que le reste du génome. Considérant que le génome des conifères est composé d'une très grande quantité de séquences hautement répétées et que la partie codante représente environ 0,1% de la totalité du génome, il semble que les microsatellites soient préférentiellement associés aux régions faiblement répétées. Ceci a été observé chez les plantes (Morgante et al. 2002) où des proportions de motifs répétés comparables ont été observées aussi bien dans les régions répétées, simple copie que codantes chez Arabidopsis et le maïs. Néanmoins, ces espèces ayant une proportion d'ADN répété moins importante par rapport aux conifères, le développement de microsatellites pour ces espèces est plus facile et produit moins de locus

Figure C3 : Position des motifs répétés dans les séquences traduites (en blanc), 3'UTR (gris clair), 5'UTR (gris foncé), chez *P. taeda* (à gauche) et *P. pinaster* (à droite). Les séquences pour lesquelles aucun ORF n'a été détecté ne sont pas présentées.



multibandes. Une étude ayant mis en jeu une banque d'ADNc enrichie en motif dinucléotides a été publiée chez *Picea abies* (Scotti et al. 2000) et a mis en évidence peu de motifs répétés dans les régions codantes (environ un motif répété tous les 1250 gènes). Néanmoins cette étude a utilisé des conditions beaucoup plus stringeantes, ceci étant dû aux conditions d'hybridation et de lavage, au cours de l'étape d'enrichissemement, qui ont pu sélectionner uniquement les motifs les plus longs.

En ce qui concerne les types de motifs détectés, les trinucléotides (63% chez *P. taeda* et 51% chez *P. pinaster*) sont plus nombreux que les dinucléotides (36% chez *P. taeda* et 45% chez *P. pinaster*). Les autres motifs ne sont que très peu représentés (1% et 3% de tétranucléotides). Les trinucléotides ont été également décrits comme étant les motifs les plus abondants dans les séquences codantes d'autres espèces de plantes et d'animaux. En ce qui concerne les types de motif, les plus abondants sont AAG, AGC et AGG (figure C2), ce qui est également en accord avec les résultats trouvés chez les dicotylédones. Néanmoins, d'autres types de motifs ont été trouvés en abondance chez les céréales (par exemple CCG, Kantety et al. 2002).

Le logiciel *FrameD* (Schiex et al. 2003) a permis de prédire la position des cadres de lecture traduits en protéine (*Open Reading Frame*, ORF) au sein des séquences des deux unigènes de pin et ainsi, combiné avec *SSRIT*, de prédire la position des microsatellites dans les régions codantes. Auparavant un jeu d'entraînement a dû être développé pour que le logiciel *FrameD*, qui utilise des chaînes de Markov (Salzberg et al. 1998) pour prédire la position des ORF, puisse différencier les séquences traduites des non-traduites. Pour cela, 65 séquences correspondant à des gènes pleine longueur d'espèces de pin (*P. taeda*, *P. radiata*, *P. pinaster*) et auxquelles j'ai pu accéder par le logiciel SRS (http://www.infobiogen.fr/srs) ont été utilisées.

Les résultats obtenus sont présentés à la figure C3 et montrent que les trinucléotides sont plutôt localisés dans les ORF, alors que les dinucléotides se situent préférentiellement dans les régions non-traduites en 5' et 3' (*Untranslated regions*, UTR). Ces résultats ne sont pas surprenants car les trinucléotides présents dans les ORF peuvent correspondre à des codons et être traduits en acides aminés, qui sont eux-mêmes répétés au sein de la protéine. De la même manière, Morgante et al. (2002) ont montré une sélection positive pour les trinucléotides dans les séquences codantes d'Arabidopsis, et Metzgar et al. (2000) ont expliqué cette forte présence de trinucléotides par l'effet de fortes pressions de sélection pour ces motifs. En effet, une modification du nombre de répétitions pour des motifs mono-, di-, tétra ou pentanucléotidiques peut décaler le cadre de lecture et provoquer un changement dans la

Tableau C5 : Information sur les 53 locus SSR-ADNc amplifiant dans l'espèce d'origine : motifs, position dans le gène, conditions PCR et amplification chez sept espèces de pins.

Identification : * : numérotation des contigs de *Pinus taeda*; ** : accession GenBank; *** : numérotation des contigs de *Pinus pinaster*.

Position : UTR : région non traduite; ORF : cadre ouvert de lecture (région traduite); NP : pas de protéine.

Température d'annealing (PCR) : température d'hybridation ou *touchdown* utilisé (ex: 60-50).

Amplification : *Pp* : *Pinus pinaster*; *Pt* : *Pinus taeda*; *Pr* : *Pinus radiata*; *Ps* : *Pinus sylvestris*; *Ph* : *Pinus halepensis*; *Ppi* : *Pinus pinea*; *Pc* : *Pinus canariensis*; + : amplification; - : pas d'amplification; na :pas de données.

Information sur les locus									Amplification						
Locus	Identification	Motif répété	Nombre de répétitions	Position	Amorce F	Amorce R	Température d'annealing (°C)	Taille attendue (pb)	*Pp*	*Pt*	*Pr*	*Ps*	*Ph*	*Ppi*	*Pc*
RPtest1	Contig4518*	AAT	7	5' UTR	gatcgttattcctccctgcca	ttcgatatcctccctgcttg	50	125	+	+	+	+	+	+	+
RPtest5	Contig6309*	AAC	6	ORF	acaacaataataacgggggc	acgctttagatcctcctgca	55	197	+	+	+	+	+	+	+
RPtest6	Contig3845*	TGC	5	ORF	aggattccaacagcatcacc	ctgaacatgaagcgcagtgt	55	147	+	+	+	+	+	+	+
RPtest8	Contig8048*	CCG	6	ORF	ggtgcgagattgaaattcgt	tttgcagtctgttgcctttg	60-50	196	-	+	+	na	na	na	na
RPtest9	Contig1667*	AGC	10	ORF	ccagacaacccaaatgaagg	gcctgctatcgaatccagaa	51	289	+	+	+	+	+	+	+
RPtest11	Contig3631*	ATC	7	3' UTR	aggatgcctatgatatgcgc	aaccataacaaaagcggtcg	56	213	+	+	+	+	-	+	+
RPtest13	AA739656**	CTG	5	ORF	gatttttcaggaagaccccc	tgtaaggcacaagccctctt	51	277	+	+	+	+	-	-	+
RPtest15	Contig8064*	ACC	6	ORF	gaacgtggttatggcggtag	ccagggacagttaccagcat	56	246	+	+	+	+	+	+	+
RPtest16	AA739818**	AGT	5	ORF	cagaaatggcgtccaaattc	accccacttatatccccagc	56	132	+	+	+	+	-	+	+
RPtest20	Contig6393*	AGC	5	ORF	gttcccactcaagggttgaa	acatcatttgttgccgcata	56	259	+	+	+	-	-	-	-
RPtgbLP5	AF013805**	AAT	6	5' UTR	agaggttccaaacgagaggt	tcgacttctgatttctttacatga	60-50	176	-	+	-	na	na	na	na
ssrPp_cn524	Contig524***	AG	14	5' UTR	cgattgtttttgccttttaagc	aaatatggcggggtgtgc	50	156	+	+	+	+	+	+	-
ssrPt_AA739797	AA739797**	AT	11	3' UTR	actttgcggtgaatcagacc	aaagtaaggctgcttgcatga	51	281	+	+	+	-	-	-	-
ssrPt_AW010960	AW010960**	AT	9	ORF	atcgactaggcatcaggtgg	tcctcgtagcccagcttttta	49	225	+	+	+	+	+	+	+
ssrPt_AW225917	AW225917**	AT	9	3' UTR	tgcattgaaaaatacagcgg	attatgtacgaggccccaca	49	198	+	+	+	+	+	+	+
ssrPt_AW981642	AW981642**	AAG	7	ORF	gtggcacagggttttctgat	caaaccttcggtagcctcac	60-50	245	-	+	+	na	na	na	na
ssrPt_AW981772	AW981772**	CCT	4	ORF	gatcctgttcctcctcctcc	cctggacagaaacagcaaca	49	266	+	+	+	+	+	+	+
ssrPt_BF049767	BF049767**	AG	22	ORF	ttttgggtcgtaggaacctg	taaaacgggtgtctcttcgg	51	227	+	+	+	+	+	+	-
ssrPt_BF778306	BF778306**	AG	7	NP	gaagatggagacgaagcagg	tttgcagtctgttgcctttg	60-50	172	-	+	+	na	na	na	na
ssrPt_ctg1376	Contig1376*	AT	20	NP	cgatattatggattttgcttgtga	aaatgcatgccaaacttaaatac	60-50	145	+	+	+	+	-	-	-
ssrPt_ctg1525	Contig1525*	AGG	7	ORF	ttgaaaccatataagcaatgcc	aggacctgggtaaggaggc	60-50	173	+	+	+	+	+	+	+
ssrPt_ctg16480	Contig16480*	AAAT	13	NP	ctaaaacatcggtcggaagc	atttagtccaggccatgtcg	60-50	151	+	+	na	na	na	na	na
ssrPt_ctg16811	Contig16811*	AT	11	5' UTR	tccatgatgttgcagattgg	gtgttccccaatggtctgtc	56	199	+	+	+	-	-	-	+
ssrPt_ctg17607	Contig17607*	AAG	9	ORF	cgccattaatatgcctaccg	atctctgcgctgcttgaagt	54	225	+	+	+	+	+	+	+
ssrPt_ctg18103	Contig18103*	AT	10	NP	cctggattcatttgtggctaa	catgccaacttcttgcattg	60	184	+	+	-	+	+	+	+
ssrPt_ctg2300	Contig2300*	CCG	6	ORF	cactttgcgagagactgcac	acgctgaaggaaatcgagaa	49	173	+	+	+	+	+	+	+
ssrPt_ctg275	Contig275*	AT	16	3' UTR	acggagatatattgctggcg	aaagaataacgtgaaacaaaccc	60-50	137	+	+	+	+	-	-	-
ssrPt_ctg3021	Contig3021*	AGC	14	ORF	ctcagattcctccaaatgcg	catgcaacatatgcaaaccg	60-50	234	+	+	+	+	+	+	+
ssrPt_ctg3089	Contig3089*	AT	17	NP	ctttcttcacgttggacttctt	ttagccatggagagtgcaga	45	482	-	+	+	+	+	+	+
ssrPt_ctg3754	Contig3754*	AGC	6	5' UTR	tctttgggtttctggagtgg	gctgttgctgttgttcttgg	60-50	421	+	+	+	+	+	+	+
ssrPt_ctg4363	Contig4363*	AT	10	3' UTR	taataattcaagccaccccg	agcaggctaataacaacacgc	60-50	100	+	+	+	+	+	+	+
ssrPt_ctg4487a	Contig4487*	CCG	5	ORF	tctgctgtgtggacaaacct	ttcttggctcaaaatctcgg	60-50	155	+	+	+	+	+	-	-
ssrPt_ctg4487b	Contig4487*	CCG	10	3' UTR	atgacgcattatcaggggaa	ttgcacagaaagcaggtttg	45	254	+	+	+	+	+	-	-
ssrPt_ctg4698	Contig4698*	ATC	10	ORF	cgaaaaggtggttctgatgg	ttttccgctggatttaccac	49	246	+	+	+	+	+	+	+
ssrPt_ctg5167	Contig5167*	AAC	7	ORF	tgcagagagattcgatggg	attttggtttgtttgctggc	60-50	293	+	+	+	+	+	+	+
ssrPt_ctg5333	Contig5333*	AGC	7	ORF	gaaggagtcggcgataacag	gggaattcgacctgtgaaga	49	163	+	+	+	+	-	-	-
ssrPt_ctg6390	Contig6390*	AAG	8	5' UTR	atccacgacttgtcgacgc	atcaaccaacttaggcagcg	45	440	-	+	+	-	+	+	+
ssrPt_ctg64	Contig64*	CCG	7	ORF	ggaagctgttacaagtgcgg	atcgagaagagaggaagggc	60-50	284	+	+	+	+	+	+	+
ssrPt_ctg7024	Contig7024*	AAG	7	ORF	gggaattctgaaagacaaggg	aacttaccatcgagagcccc	60-50	277	+	+	+	+	+	+	-
ssrPt_ctg7081	Contig7081*	AAG	7	ORF	gtcatccacgttcattggc	tcacaactgaccaaactgcc	60-50	442	+	+	+	+	+	+	-
ssrPt_ctg7141	Contig7141*	CGG	8	ORF	gaatgacgcattatcagggg	tcacctttctcacctctgcc	45	381	-	+	+	+	+	+	+
ssrPt_ctg7170	Contig7170*	AGC	5	ORF	ggtttttcgatttctgaggc	aacaggtgtgcaaatagccc	60-50	385	+	+	+	+	+	+	-
ssrPt_ctg7425	Contig7425*	AAG	6	ORF	aataagaccccagaggagcc	gacgtctttcaccaaatcgc	60-50	384	+	+	+	+	+	-	-
ssrPt_ctg7444	Contig7444*	AT	10	5' UTR	tcttcaccatcggtttctcc	tggatctgtcacctcctcatc	58	285	+	+	+	+	+	+	+
ssrPt_ctg7731	Contig7731*	AT	12	5' UTR	agtggtgaagggtccatctg	gcataacacaaaagccagca	51	217	+	+	+	+	+	+	+
ssrPt_ctg7824	Contig7824*	AT	12	3' UTR	tgacctgtcttgtgagacgc	ttttgaaacagattgcagcc	60-50	501	+	+	+	+	+	-	-
ssrPt_ctg7867	Contig7867*	CCG	6	5' UTR	ggtcgtggaggaggtaggg	actgataacagctgccccc	45	154	+	+	+	+	-	+	+
ssrPt_ctg8064	Contig8064*	ACC	6	ORF	gaacgtggttatggcggtag	tcgtggcaactatctcctcc	50	147	+	+	+	+	+	+	+
ssrPt_ctg865	Contig865*	AT	15	3' UTR	tttcagaagctcccgatttg	cttgtggacatggttaatgaag	45	232	+	+	+	+	+	+	+
ssrPt_ctg8767	Contig8767*	AGC	8	ORF	tggggaaaaatggcatacat	ggagcagacacccatggact	55	180	+	+	+	-	-	-	-
ssrPt_ctg9249	Contig9249*	AAG	7	5' UTR	ctgctccctcagctcttcc	agacgtcactgccattaccc	55	156	+	+	-	+	+	+	+
ssrPt_ctg946	Contig946*	AGG	9	3' UTR	tatcaggtataggcctccgc	aaataggagcccttctggga	53	287	+	+	+	-	-	-	-
ssrPt_ctg988	Contig988*	AT	7	3' UTR	taataattcaagccaccccg	aacattttgcacgatagccc	51	319	+	+	+	-	-	-	-
							Amplification (%)		86,8	100	94,2	85,4	72,9	70,8	64,6

fonction de la protéine, ce qui n'est pas le cas pour les trinucléotides (ainsi que les hexa-, même s'ils ne sont que très peu présents).

L'alignement des EST des contigs de *P. taeda* et de *P. pinaster* ont parfois révélé du polymorphisme au niveau des motifs répétés. Par exemple pour *P. taeda*, l'alignement des 32 contigs ayant au minimum 10 EST et contenant des motifs répétés a révélé 8 contigs possédant du polymorphisme en terme de nombre de répétitions. Cette recherche de polymorphisme *in silico* peut permettre de valider le fait que ces motifs répétés sont bel et bien des microsatellites, avant de passer à l'étape de développement de marqueur.

C.1.2.2.2 Développement d'amorces PCR, transférabilité, polymorphisme et orthologie des microsatellites de régions codantes (SSR-ADNc)

Cinquante-deux paires d'amorces sur 69 développées à partir de séquences de *P. taeda* et une paire d'amorces sur trois développées à partir de *P. pinaster* ont donné un produit d'amplification PCR simple bande chez l'espèce d'origine. Les conditions PCR pour ces locus sont présentées au tableau 5 de l'annexe IV. Il est à noter que les conditions PCR ont dû être mises au point, mais de manière moins systématique que pour les microsatellites génomiques. La plupart du temps, une condition PCR consistant en un *touchdown* allant de 60°C à 50°C, suivi de 30 cycles à 50°C, a suffi à amplifier un grand nombre de locus, ce qui laisse espérer que des locus pourront être multiplexés afin d'augmenter le débit de génotypage. Ces dernières optimisations vont bientôt débuter au sein du laboratoire de Pierroton et restent à réaliser en vue de l'utilisation des SSR-ADNc en génétique des populations ou pour faire des tests variétaux chez le pin maritime.

Globalement, nos résultats montrent que la mise au point de microsatellites à partir de séquences codantes est plus efficace que dans le cas des microsatellites transférés d'autres espèces, comme je l'ai décrit dans le chapitre précédent. Ceci peut s'expliquer tout d'abord par le fait que les séquences issues de l'assemblage et l'alignement des EST pour construire les unigènes utilisés pour détecter les motifs répétés permettent de dessiner des amorces de meilleure qualité. Ensuite, la même explication que pour le chapitre précédent peut être appliquée : alors que dans le cas des microsatellites génomiques non-codants les chances de se retrouver dans des régions répétées étaient élevées, dans le cas présent des séquences codantes le problème ne se pose pas. Il n'est donc pas étonnant d'observer que les conditions PCR étaient plus simples à mettre au point et que moins d'amplifications multibandes ont été observées (seulement 5 paires d'amorces sur les 19 non retenues ont amplifié des profils

Tableau C6 : Description des locus SSR-ADNc polymorphes et comparaison avec des microsatellites issus de *P. radiata*, *P. halepensis* et *P. pinaster*. Localisation sur les cartes G2 et F2, évaluation sur deux autres pedigrees et diversité génétique.

M : monomorphe; P : polymorphe; UL : non lié; H : hétérozygotie; A : nombre d'allèles; * : ce locus n'a pas été pris en compte pour la comparaison des paramètres de diversité entre SSR-ADNc et SSR génomiques.

		Pedigree de cartographie				Diversité génétique	
Type de marqueur	Locus ID	*INRA-G2*	*INRA-F2*	*AFOCEL-F1*	*INIA-F1*	*H*	*A*
SSR-ADNc	ssrPp_cn524	6	1	P	M	0.81	5
	ssrPt_ctg275	P/UL	P/UL	P	P	0.74	8
	ssrPt_ctg4363	M	12	P	M	0.68	4
	ssrPt_ctg7824	10	M	M	M	0.35	2
	ssrPt_ctg988	11	M	P	M	0.55	3
	RPtEST11	5	2	P	M	0.74	4
	RPtEST13	10	M	M	M	0.66	3
	ssrPt_ctg1525	M	11	M	M	0.16	2
	ssrPt_ctg64	3	3	M	P	0.68	4
SSR *P. radiata*	NZPR1078	2	7	P	M	0.68	4
	NZPR114	M	5	M	P	0.68	5
	NZPR1702_b	11	6	P	M	0.38*	2*
	NZPR413	4	8	P	P	0.58	4
	NZPR472	1	M	P	P	0.67	4
	NZPR544	M	3	M	P	0.41	4
	NZPR823_a	5	M	P	P	0.67	3
SSR *P. pinaster* et *P. halepensis*	FRPp91	1	9	P	P	0.85	9
	FRPp94	10	5	P	P	0.80	8
	ITPh4516	3	3	P	P	0.84	8

Tableau C1 : Evaluation des marqueurs microsatellites issus de la détection de motifs répétés dans les bases de données d'EST.

	Développement	**Cartographie génétique**	**Cartographie comparée**	**Détection QTL**	**GC positionnel**
AFLP	++	++	--	+/-	--
SSR (transfert)	+	+/-	+/-	++	--
SSR-ADNc	Pas de développement nécessaire (détection de motifs répétés *in silico)*	Marqueurs informatifs mais moins polymorphes que les microsatellites génomiques	Marqueurs transférables entre espèces de pin	Bons marqueurs pour la détection de QTL (codominants, multialléliques)	Marqueurs codants pouvant correspondre à des gènes candidats
SSR (banque enrichie)	?	?	?	?	?
EST	?	?	?	?	?
SNP	?	?	?	?	?

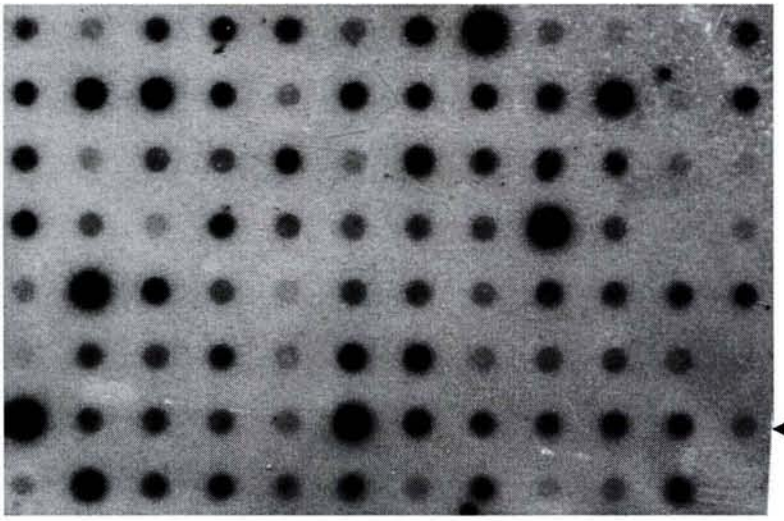
multibandes). Néanmoins, les familles multigéniques présentes en grand nombre chez les conifères (Kinlaw et Neale 1997) peuvent produire des profils multibandes. De plus la présence d'introns au milieu du produit d'amplification attendu peut provoquer l'échec de la réaction PCR, ce qui est surement le cas pour les 14 paires d'amorces n'ayant pas donné d'amplification.

De la même manière que pour les microsatellites génomiques issus d'autres espèces de pin, l'amplification des microsatellites détectés dans les EST a été testée chez plusieurs espèces de pins. Au total sept espèces du sous-genre *Pinus* ont été utilisées : *P. pinaster*, *P.taeda*, *P. radiata*, *P. sylvestris*, *P. halepensis*, *P. canariensis* et *P. pinea*. Les taux d'amplification de ces locus varient de 94,2% pour *P. radiata* à 64,6% pour *P. canariensis* (tableau C5). Ils sont supérieurs à ceux obtenus dans le cas du transfert de microsatellites génomiques (voir chapitre précédent). De plus, ces taux d'amplification sont comparables aux taux d'amplification interspécifique obtenus avec des marqueurs basés sur des EST utilisés dans le cadre du projet de cartographie génétique comparée chez les conifères (Brown et al. 2001) sur les mêmes espèces. Néanmoins, il faudrait séquencer les fragments obtenus dans le cas présent afin de confirmer que les locus amplifiés sont bien orthologues et que les motifs microsatellites sont toujours présents au sein des séquences.

Quarante-six locus SSR-ADNc ont montré une amplification simple bande chez le pin maritime. Sur ces 46 locus, 9 (19,5%) ont pu être cartographiés sur au moins un des deux pedigrees considérés (ssrPp_cn524, ssrPt_ctg275, ssrPt_ctg1525, ssrPt_ctg4363, ssrPt_ctg7824, ssrPt_ctg988, ssrPt_ctg64, RPtEST11, RPtEST13 ; tableau C6). Ce nombre de locus polymorphes est faible par rapport à ce que l'on pourrait attendre de la part de locus hyper-variables comme les microsatellites. Néanmoins, il semble que la position des SSR dans les régions codantes puisse jouer sur le taux de polymorphisme. En effet, pour les motifs dinucléotides se trouvant dans les régions UTR le nombre de locus polymorphes passe à 55%. Ces résultats montrent donc l'utilité de l'annotation des motifs microsatellites détectés dans les EST à l'aide du logiciel *FrameD*.

Afin de comparer leur niveau de diversité, les 9 SSR-ADNc cartographiés, ainsi que les 9 SSRs génomiques mis au point chez le pin maritime par transfert ont été génotypés sur 26 individus appartenant à la première génération du programme d'amélioration du pin maritime. La diversité génétique (hétérozygotie de Nei, Nei 1973) et la richesse allélique (A) ont été calculées grâce au logiciel Arlequin v2.000 (Schneider et al. 2000). Le tableau C6 présente les valeurs obtenues pour ces locus. Une différence significative a été obtenue pour *H* et *A* entre les SSR-ADNc et les SSR génomiques. Ces résultats montrent que des forces de sélection

Figure C4 : Exemple de Dot Blot sur la banque C. Les spots de forte intensité ont été éliminés des étapes suivantes de séquençage.



Témoin : *Chalcone synthase*

Tableau C7 : Bilan des étapes de séquençage et d'analyse bioinformatique pour les banques enrichies en motifs répétés.

	Dinucléotides	Trinucléotides	Tétranucléotides
Nombre de clones séquencés	515	126	24
Nombre de séquences	700	132	27
Nombre de clusters	54	6	-
Nombre de contigs	151	25	-
Nombre de contigs ayant un SSR	84	21	-
Nombre de singletons	518	79	27
Nombre de singletons ayant un SSR	295	22	0
Nombre de paires d'amorces	58	21	-
Nombre de locus polymorphes	15	0	-

peuvent agir sur ces locus. Des tests de neutralité de ces marqueurs pour des caractères adaptatifs pourraient permettre de valider cette hypothèse. Aucune différence significative n'a été observée entre les marqueurs SSR-ADNc dont les motifs se trouvaient dans les ORF ou les UTR, ou entre les motifs di- et trinucléotides. Pour les marqueurs microsatellites issus d'ADN génomique, des différences significatives ont été obtenues au niveau de la diversité entre les locus issus de *P. radiata* et ceux issus de *P. halepensis* (un locus) et *P. pinaster* (deux locus). Ce résultat peut être expliqué par les distances phylogénétiques qui séparent les espèces considérées. Même si ces trois espèces appartiennent au même sous genre *Pinus*, *P. pinaster* et *P. halepensis* sont deux espèces proches provenant du bassin méditerranéen alors que *P. radiata* provient de la côte pacifique des Etats-Unis. Il n'est pas étonnant d'observer une diminution du polymorphisme d'un marqueur lorsqu'on s'éloigne de l'espèce d'origine pour laquelle ce marqueur a été développé. De tels résultats ont été démontrés chez les pins par Echt et al. (1999).

En guise de conclusion sur ce type de marqueur (tableau C1), les SSR-ADNc fournissent une alternative intéressante aux microsatellites génomiques (qu'ils soient transférés ou développés à partir d'une banque génomique enrichie). Avec très peu d'efforts de développement, ils donnent accès à des marqueurs génétiques polymorphes et potentiellement orthologues. La disponibilité de nouveaux EST chez les pins (100 000 EST de racine pour *P. taeda*) qui viennent d'être obtenus par une équipe américaine (J. Dean, communication personnelle) permet d'envisager de mettre au point de nouveaux marqueurs de ce type.

C.1.2.3 Développement de banques génomiques enrichies en motifs microsatellites

Malgré les caractéristiques très intéressantes des SSR-ADNc (très bonne transférabilité et taux de polymorphisme correct), le nombre de locus développés reste insuffisant si l'on souhaite saturer une carte génétique avec des marqueurs microsatellites. Il en va de même pour les locus microsatellites transférés depuis d'autres espèces. C'est pour ces raisons que des banques enrichies en motifs microsatellites ont récemment été développées chez le pin maritime. Comme je l'ai décrit dans la partie B.1.2.3.3, un consortium international a été créé pour répondre à cet objectif.

Au total, 1920 transformations de cellules compétentes ont été réalisées correspondant à 20 plaques ELISA de 96 puits (7 pour la banque AG, 6 pour la banque AC, 4 pour la banque AAG et 3 pour la banque AAAT). Sur ces 1920 clonages, 1717 se sont révélés correspondre à

Tableau C8 : Description des amorces PCR pour les microsatellites dinucléotidiques des banques A et B de *P. pinaster* : amplification et cartographie dans les pedigree F2 et G2.

Locus	Amorce F	Amorce R	Taille attendue (pb)	Polymorphisme G2	Polymorphisme F2
A1C03	TGGTAAAGGAAGGAATACAAAGTG	AAGTGGTGCCCTACGATGAG	172	**10**	M
A3A03	GAGTTTCCTGCATTGTGTGC	CTCTTCCTTTGCCTCTCAGG	199	M	M
A3B03	GGAGAGACGACATCAAACTGG	AGATACGAACATGCCACTGC	227	M	M
A3B05a	GCGACGACTGTGAAGTTAGG	GGGAGCCTGACAGAAATAACC	195	M	M
A3B05b	AAGTGGCATGCGTCACTATG	AAGCTTGACCATTAGGAGTAGTCT	131	**7**	M
A3D04	CCTTAACCGTCCCATCTCAA	GCTTGACTATCAATGCCATGAA	150	**5**	M
A3D08	TTTTGGGTCTGTGACATTGC	CAGGCATGCAAGCTTATCG	268	M	M
A3F03	GCAGGCATGCAAGCTTATC	GCTCTTGCTCTCCAACTTCTACTC	251	M	M
A3G05	CATAACCCACCGTAGTAGATGC	GGAGATTAAACACGTTAACTGATGC	288	M	M
A4A07	ACCTATGGTTTTCGGCATGG	ACACAAAGGTTTTCCCTTGG	274	M	M
A4A09	GCCTAACTGTTTATGCAGAATGG	GGACCAGTTTCCCAATCTAAGG	286	M	M
A4B01	GCTAGGTATGGGCTTTGTCC	TCTTTATATGCAGGGTTGAACG	133	P?	P?
A4C03	TTCCATCAGGATACCCAAGC	CATGGGAAATTTGCTTTTGG	271	M	M
A4C06	GGATATTGCATCGCTTACAGG	TCCAACTATCGGCACTTTCC	199	M	M
A4F06	TTCCATGACCCAACTATAGGC	GGTTGAATGAACAACCTCAGC	274	**4**	**Non lié**
A4H08	ATACCAGAGGCACACAGTGG	GCGTACAAAGCACAATAAAACC	285	M	M
A5A11	ATCACAATTGGCTTGGTTCC	TGCCACAAAGTAAGGACAGG	299	**1**	**9**
A5A12	TAAGATCACACGCCAGATGC	CCCAACCCAACCTATCTAAGG	239	M	M
A5B01	ACAACGCACCACAATAGTCC	GTGTCGAACAGTCGCATAGC	222	**12** (2 locus)	M
A5B04	TGTTGGAAACACTGGTGTGG	TAACAAACGACCCATTGTGG	274	M	M
A5B07	CGTGCAGATCACAGAATTTGG	CAATTACAACCATCAGAGTGACC	247	P?	M
A5B10	TCTTGGCTCTCAGCTCTTGG	GAACTTGGGTAGAGCAACTCG	261	**2**	**Non lié**
A5C07	GAGGGTGAGAGATTGACAGAGG	GCAGGCATGCAAGCTTAAC	258	M	M
A5E11	TGTATGTTTTGGTGGTTGTCG	ATCGGCAGGAGAAACAATCC	197	P?	P?
A5F03	GAGCAGATGGAAACGAAAGG	TGTGTGCTCCTCCTCTTTGG	123	M	M
A5G07	CACCGCGATTAGACTTGTAGG	CATCATCGCATGTAAAAGATCC	205	M	M
A5G09	AAGCTTCCCATGAGATCAAGTC	CGTATAACTGCCCACTGCAC	144	M	M
A5G11	CTCAGGCCGTATGAGATTCG	TCATCATCACATGAAGAGATCC	172	M	M
A6A02	TGCCCTCATTCGTCTCTAGC	GGGTGTTTGTGTTGTTTTGC	203	M	M
A6A03	GTGCTACAGCGAGCATAACC	TGTATCGAACAGTCGCATAGC	253	No amplification	
A6A04	CGGATGTATCAGGGTGAACG	CAATGGGTCGTTTGTTATGC	224	M	M
A6A09	GCTGAACCCTTTGAACAACC	ATATGGGTCAGGGCGATAGG	270	No amplification	
A6D04	ACCATATAGGTCTTAGTGGATG	TCTTGCTTGCATCAATGTCC	157	P?	P?
A6D12	CCATCAGAAACCCTCAGATAATG	TGAGGCATGTGAGCCATTG	295	M	M
A6E05	GCAACAAACCCTACCGAACC	TTACGCTGACATACCCATCG	278	**12**	**12**
A6E08	TTGTTTTAGACCAGGTGACTAGG	CCAGGTGGATTGTTTCTAGGG	159	**1**	**9**
A6F03	CCTGAAAATCGACGGATCG	ATGGTATTTTGCGGGTTGC	256	**10**	**5**
A6F10	CCATTGGATAGATCGTCAGAATAG	GTGACTTCGCTGTCTTGCTG	257	M	**8**
A6H10	CCATTGGATAGATCGTCAGAATAG	GTGACTTCGCTGTCTTGCTG	257	M	**8**
A7A04	ATATCGCAATTGATGGATGG	TATCGCTAATCGGCAATCG	185	M	M
A7E03	TAATGTACCCAACCCCTTCG	TTGCACCTCATTAGTCAACG	117	**4**	**8**
A7E07	CCTTGCCCGTTAGGAATAGC	CCGGGTTAGTGTAGATTAGATGC	213	M	M
B3A01	CATGTCCACGAAGAAACAGC	TTTTGCCCTAGAATGCAACC	218	M	M
B3D02	TTGTTAAGTCGGCTTGTGG	CAAATTAAAGGGAGGGAAAAGG	262	M	M
B3E01	TGTGAGATCTACCCCCATGC	ATGGGGACTTTCGAGAGAGG	195	M	M
B4C01	TTCTTTCCCTCCCTCTCTCC	ATATTGACCGGGATTTGAGG	138	M	M
B4C06	GCAAGAGGACACAAGATGTGG	GCAACATGATGACGAGATGC	295	M	M
B4D01	GGTGATGCCTCGAGTAATGG	CATGCAAGCTTAACAGTATGC	173	No amplification	
B4D04	TCGACCTCAAGTGTCACAGG	GAGGAAGGAAAAGGGAGAGG	179	M	M
B4D05	TTTCGGCATCACAACAGC	GTTTGGAAGCTGGAAGTTGG	212	M	M
B4D05	CAACTTCCAGCTTCCAAACC	ATGCGGAATTGAAGGAAAGG	134	M	M
B4E05	GGAGGTCTTCTCCTCATTTGG	GCCCTAAATCTGCGTCTTCC	159	M	M
B4E09	TTTTGCTAGTCTCTCCCTGTCC	AGAGATAGGGGATGGGAAGC	243	M	M
B4F04	CGAACAAGCATGGAGATCG	CCTCTGCTTGGCTTTACTGC	174	M	M
B4F08	GCACTTTGATTGTTGTCATCG	GTGGCTGATGTCCAAATGC	189	**3**	**3**
B5A02	TCTCAGTGGGTATGGGTATGC	CTTCCACAAGGCACTAAAAGG	183	M	M
B5A03	CACTTTCCAGTATGAGAAGGATATG	CCCACATTGTTTTCTTTTTAAATC	264	M	M
B5B02	AAGGAGAGGAAGATGGATGGA	TGAGTGCAAAACTTGAGGACA	286	M	M

des clones positifs contenant un insert. Ces inserts ont été amplifiés par PCR à l'aide des amorces universelles M13 et les produits d'amplification ont servi au criblage du nombre de copies par Dot Blot. Les fragments produits ont été dénaturés et fixés sur une membrane. De l'ADN génomique total de pin maritime marqué radioactivement a été hybridé sur les 1717 inserts et les clones montrant un signal d'hybridation intense ont été éliminés pour les étapes suivantes. Afin de quantifier ces signaux, un témoin négatif avait été utilisé : le gène de la *Chalcone synthase*, supposé être présent en une seule copie dans le génome du pin maritime. L'analyse des autoradiogrammes a été réalisée à l'œil sans emploi d'un logiciel pour quantifier les signaux d'hybridation car la plupart de ceux-ci étaient faibles (figure C4). Au total, 268 clones (15%) ont été écartés, ce qui porte le nombre de clones intéressants pour le séquençage à 1449. Cette méthode du Dot Blot a été employée avec succès chez l'épicéa par Scotti et al. (2002) qui ont montré qu'elle améliorait considérablement les chances d'obtenir des locus simple bande par amplification PCR.

L'étape suivante a consisté à séquencer une partie des clones contenant un insert et apparemment peu répétés dans le génome. 515 clones dinucléotides, 126 trinucléotides et 24 tétranucléotides ont été séquencés.

La description des séquences réalisées est présentée au tableau C7. La totalité des 859 séquences (le nombre de séquences est supérieur à celui des clones séquencés car les inserts ont été séquencés en 5' et 3' pour certains clones) a été rassemblée par type de motif (di-, tri- et tétranucléotide) afin de réaliser leur traitement bioinformatique. Tout d'abord chaque séquence a été débarrassée des fragments de vecteurs et des zones de faible qualité (score *Phred* < 20). Les séquences ont ensuite été introduites dans le « pipeline » d'assemblage de séquences nucléotidiques basé sur *StackPack* (Christoffels et al. 2001). Des clusters ont été obtenus ainsi que des contigs. Ces deux types de groupes de séquences représentent des niveaux hiérarchiques de l'assemblage, correspondant à des niveaux plus ou moins strigeants de comparaison des séquences par les logiciels composant la chaîne de programme *StackPack* : un cluster peut contenir plusieurs contigs qui eux-mêmes sont le résultats de l'alignement de plusieurs séquences. Une séquence consensus a été réalisée pour chaque contig. Les séquences non incluses dans les clusters sont des singletons. Pour les clones supposés contenir des tétranucléotides aucun assemblage n'a été réalisé car la plupart des séquences étaient de très mauvaise qualité ou ne contenaient pas de motif répété. Au total 518 singletons, 54 clusters et 151 contigs ont été obtenus pour les séquences des banques A et B. En ce qui concerne la banque C, 79 singletons ont été obtenus ainsi que 6 clusters et 25 contigs. Pour les dinucléotides, 84 contigs sur 151 (55%) contenaient un motif répété et pour

Tableau C9 : Description des locus développés à partir de la banque enrichie en motifs trinucléotides et résultats des analyses blast.

Locus	Cluster	Contig	Amorce F	Amorce R	Taille attendue (pb)	Blastx	Blastn nr	Blastn est	Amplification
C2D10	cl2	2	TGGTCGTCCTTTCATAATCG	GGCTGCTAGTCCATCAATCC	307	no hit	no hit	NXRV058_C07_F 9e-43 (ctg11529)	Oui
C2H04	cl3	3	CTGTTTGGTTGTTCCCATCC	GAGTAGCGGCACTTTTAGCC	247	no hit	no hit	NXRV_022_A12_F 6e-07	Oui
C2G09	cl5	5	TTGCCTGGTGAAAGTCTCG	GGCAATCACGAAGAAGATGG	291	no hit	no hit	NXRV_022_A12_F 6e-13 (ctg13951)	Oui
C2A12	cl5	6	GGCAATCACGAAGAAAATGG	TTCGACGCCTACTTCTACGC	245	no hit	no hit	NXRV_022_A12_F 5e-13	Oui
C2A06	cl5	7	GACAATCGCGAAGAAGATGG	GAAAGTCTCGCTTCGAGTCC	202	no hit	no hit	NXRV_022_A12_F 7e-09	Oui
C2B09	cl5	8	GACAATCGCGAAGAAAATGG	TGCCTTCCTCTTCCACTTCG	225	no hit	no hit	no hit	Oui
C2C09	cl5	9	GTGAAAGTCCCGCAAGTACC	TGTTCCTCCTTGCATAATCG	178	no hit	no hit	no hit	Oui
C2E10	cl5	10	TTGCCTGGTGAAAGTCTCG	TCTACCCTTCCCAACTGAGG	305	no hit	no hit	NXRV_022_A12_F 8e-06	Oui
C2F03	cl5	11	TGGTCTTCCTTGCATAAGTGG	GAAAGTCTCGCTTCGAGTACC	185	no hit	no hit	no hit	Non
C2H01	cl5	12	TTGTTCTCGACCTCGGTAGC	CTCAACGTCAGGAGAACTTGG	266	no hit	no hit	no hit	Faible
C3B08	cl5	13	AACCTCATGTTGATTCCAAGC	GACAATCGCGAAGAAAATGG	281	no hit	no hit	NXRV_022_A12_F 0,0012	Oui
C3D07	cl5	14	TAAGGTGGCACAACTTCAGG	TCGTCATGTTGATTCCAAGG	207	no hit	no hit	no hit	Oui
C3E08	cl5	15	TTGTTCTCGACCTCGGTAGC	TTGGGGTAGGTAGCAAGAGG	247	no hit	no hit	NXRV_022_A12_F 0,001	Oui
C3H02	cl5	16	TGAAAGTTCCGCAAGTACCC	GAAGAAAATGGTCCTCCTTGC	191	no hit	no hit	NXRV_022_A12_F 2e-06	Oui
C2G04	cl5	17	TGGGAATACCAGGACAATCG	TGAACCTCTCCCCTTCTTCG	249	no hit	no hit	NXRV_022_A12_F 9e-07	Oui
C1D03	cl5	19	AAGCCTACTTCTACGGCTTCC	TCGTTGCATAATCGACAAGG	213	no hit	no hit	no hit	Oui
C1F07	cl5	20	AATCGCGAAGAAAATGATCC	ATTCCAAGGTGGACCTCTCC	241	no hit	no hit	NXRV_022_A12_F 3e-05	Oui
C3A02	cl5	21	TGTGATCTCTCCCCTTCGAC	TCCTCCTTGCATAATCGACA	262	no hit	no hit	no hit	Oui
C1A02	cl5	23	GCGAAAGTCTTACTTCGAGTACC	ATCTAGGACCCTGGGAATGC	227	no hit	no hit	NXRV_022_A12_F 3e-08	Oui
C3A05	cl6	24	GGCAAAGGCAAAGGTAAAGG	AAGCCCACTGTCCACATACC	313	no hit	no hit	*BX250187 Pinus pinaster* differenciating xylem adult 5e-41	Oui
C3C06	cl6	25	GAAAGGCAAAGGGAAGAAGC	CAAAACGACCTGCAAAGTCC	231	*putative polyprotein Oryza sativa* 2e-09	*Arabidopsis thaliana* glyoxal oxidase 3e-04	*BX250187 Pinus pinaster* differenciating xylem adult 4e-60	Oui

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

les trinucléotides, 21 contigs sur 25 (84%). Pour ce qui est des singletons, 22 séquences sur 79 (33%) contenaient un microsatellite pour la banque C, pour les banques A et B, 295 sur 518 (57%) contenaient un microsatellite. Les résultats obtenus sur les taux d'enrichissement des banques di- et trinucléotides sont plutôt surprenants étant donné que la banque a été réalisée par un entreprise spécialisée dans la construction de banque enrichie en microsatellites. Néanmoins, d'autres erreurs et des contaminations ont pu se produire durant les autres étapes de la réalisation de cette banque (clonage, séquençage). Une autre explication pour le faible taux d'enrichissement des singletons peut être que les séquences réalisées étaient trop courtes pour atteindre le motif microsatellite. Enfin, et ceci a été noté lors de l'étape de séquençage, il semble que la qualité des séquences chute considérablement en présence d'un microsatellite.

Les 422 contigs et singletons contenant des microsatellites ont été utilisés pour dessiner des amorces PCR à l'aide du logiciel *Primer 3.0*. Pour certaines séquences, aucune paire d'amorce n'a pu être réalisée du fait que les microsatellites se trouvaient à l'extrémité de la séquence. Au total, 58 paires d'amorces ont pu être dessinées pour les dinucléotides et 21 paires d'amorces pour les trinucléotides. La liste et la description de ces amorces sont présentées aux tableaux C8 et C9.

Pour les trinucléotides, 19 paires d'amorces ont amplifié un profil simple bande de la taille attendue chez le pin maritime. Seul le locus ssrPp_C2F03 n'a pas amplifié et le locus ssrPp_C2H01 a donné un produit d'amplification de très faible intensité. Sur les 19 locus amplifiés, aucun n'a révélé de polymorphisme, ni dans les quatre pedigrees de cartographie utilisés, ni sur les 26 individus appartenant à la première génération du programme d'amélioration du pin maritime. Ces résultats sont surprenants car d'autres auteurs ont développé des marqueurs microsatellites trinucléotides chez des espèces de pins (Kutil et Williams 2001) et ont obtenu du polymorphisme. Une recherche de séquences homologues dans les bases de données de séquences génomiques (*Blastn vs.* « nr[22] »), d'EST (*Blastn vs.* « EST_other[23] ») et de protéines (*Blastx vs.* « Swissprot[24] ») a été réalisée sur les séquences des clones ayant servi à développer ces amorces. L'hypothèse faite est que si ces microsatellites trinucléotidiques appartiennent à des séquences codantes, ils peuvent être moins polymorphes. Aucune homologie de séquence n'a cependant été détectée à partir des analyses contre les bases de données « Swissprot » et « nr » (séquences génomiques). Cependant, des scores significatifs ont été obtenus avec les bases d'EST (tableau C9). Ceci peut expliquer l'absence de polymorphisme pour ces paires d'amorces. Cependant,

[22] nr : base de données de séquences nucléotidiques sans EST ni STS (nr = non redundant, ce qui n'est plus le cas actuellement).
[23] EST_other : bases de données de séquences d'EST de tous les organismes sauf souris et Homme.
[24] Swissprot : base de données de séquences protéiques.

Tableau C1 : Evaluation des marqueurs microsatellites développés à partir d'une banque enrichie.

	Développement	Cartographie génétique	Cartographie comparée	Détection QTL	GC positionnel
AFLP	++	++	--	+/-	--
SSR (transfert)	+	+/-	+/-	++	--
SSR-ADNc	++	+/-	+	+	+
SSR (banque enrichie)	Développement long et coûteux	Marqueurs informatifs et hautement polymorphes	Marqueurs peu transférables entre espèces (en terme de nombre de locus transférés avec succès) Locus orthologue	Bons marqueurs pour la détection de QTL (codominants, multialléliques)	Idem SSR transfert
EST	?	?	?	?	?
SNP	?	?	?	?	?

considérant que des motifs trinucléotides répétés développés à partir de séquences d'EST ont présenté du polymorphisme (voir chapitre précédent), cette hypothèse peut être en partie rejetée. Une autre explication peut être donnée en relation avec l'assemblage des séquences de la banque enrichie en motifs trinucléotidiques. La plupart des séquences appartiennent au Cluster 5, ce qui indique que le même locus monomorphe a pu servir plusieurs fois pour dessiner des amorces. Ainsi, ce ne sont pas 19 locus qui sont monomorphes, mais effectivement quatre locus correspondant aux 4 clusters utilisés pour le design des paires d'amorces. Les chances de ne pas observer de polymorphisme pour un lot réduit de quatre locus ne sont pas nulles. Dix autres paires d'amorces ont été dessinées à partir des séquences des singletons de la banque enrichie en trinucléotides et devraient être testées prochainement. Enfin, il serait intéressant de vérifier la présence de motifs répétés dans les produits d'amplification.

En ce qui concerne les paires d'amorces développées à partir de la banque enrichie en dinucléotides, 55 paires d'amorces sur 58 (95%) ont donné un produit d'amplification simple bande de taille attendue chez le pin maritime. Sur ces 55 locus, 15 (27%) ont révélé du polymorphisme dans au moins un des deux pedigrees de cartographie. Les détails sur ces locus sont présentés au tableau C8. Ces locus sont en cours de génotypage en population naturelle afin d'évaluer leur diversité. Six locus sur 15 ont pu être cartographiés à la fois sur les cartes F2 et G2 et ont servi à identifier les groupes de liaison homologues.

Les informations relatives à chaque clone sont présentées dans la base de donnée *MySQL* accessible à l'adresse : http://www.pierroton.inra.fr/genetics/SSR/index.php, à savoir les informations concernant les étapes de clonage, de Dot Blot, de séquençage, d'assemblage et de développement d'amorces PCR.

Le tableau C1 dresse un bilan du développement de microsatellites chez le pin maritime à partir de banques génomiques enrichies en motifs répétés. La longueur du développement de ces banques, due aux nombreuses étapes nécessaires, rend cette approche difficile à envisager pour une espèce pour laquelle peu de moyens financiers sont disponibles.

Tableau C10 : Lots d'amorces utilisés pour la cartographie d'EST.

Lot de marqueurs	Nombre de marqueurs testés	Taux d'amplification	Taux de polymorphisme (G2)	Méthode de détection du polymorphisme	Références
Lot #1	94	86% (81/94)	30% (25/81)	PAGE - DGGE	Temesguen et al. 2001
				PAGE - DGGE	Brown et al. 2001
Lot #2	65	81% (53/65)	22% (12/53)	PAGE - DGGE	Brown et al. 2003
Lot #3	90	52% (47/90)	31% (15/47)	PAGE - SSCP	http://www.pierroton.inra.fr/genetics/pinus/primers.html
Lot #4	60	63% (38/60)	5% (2/38)	PAGE	Cato et al. 2001
Total	309	70% (217/309)	24% (54/217)	-	-

C.1.3 Développement de marqueurs codants

C.1.3.1 Des EST polymorphes révélés par DGGE et SSCP et EST-AFLP

Les résultats obtenus pour les marqueurs basés sur les EST sont présentés à l'annexe III.

Au total 309 paires d'amorces ont été développées. Elles peuvent être divisées en plusieurs lots (tableau C10). Le premier lot est issu des EST développés dans le cadre du projet CCGP de cartographie génétique comparée chez les conifères (Temesgen et al. 2001, Brown et al. 2001). Le second lot est issu du développement d'une autre série de paires d'amorces PCR servant à la fois de marqueurs orthologues pour la cartographie génétique comparée chez les conifères et pour la cartographie de gènes candidats impliqués dans les caractères liés à la qualité du bois (Brown et al. 2003). Le troisième lot est issu de paires d'amorces développées à partir d'EST de différentes espèces de conifères dont le pin maritime. Certaines de ces paires d'amorces ont été décrites par Plomion et al. (1999). La totalité des amorces est accessible à l'adresse : http://www.pierroton.inra.fr/genetics/pinus/primers.html. Le dernier lot est issu d'EST développés chez le pin radiata et décrits par Cato et al. (2001). Pour la détection du polymorphisme, les deux premiers lots utilisaient la technique DGGE, le troisième lot la technique SSCP et le quatrième une technique inspirée du *chromosome walking*.

Deux-cent-dix-sept paires d'amorces sur 309 (70%) ont amplifié chez le pin maritime. La plupart des marqueurs ont permis d'amplifier une seule bande sur gel d'agarose, ce qui indique que des locus orthologues ont pu être obtenus en grande majorité. Cependant, l'amplification à partir de la paire d'amorces PtIFG_9036 a permis d'amplifier deux bandes distinctes sur gel d'agarose (240 et 280 pb). L'orthologie de ces locus amplifiés a été vérifiée en séquençant les produits PCR amplifiés. De fortes homologies de séquences ont été obtenues, à l'exception des locus PtIFG_8436_b et PtIFG_9036_b. Les raisons de ces identités de séquences plus faibles seront discutées plus tard dans ce manuscrit (chapitre C.3.1).

Pour le premier lot d'amorces, des résultats similaires ont été obtenus entre *P. taeda* et *P. elliottii* où de fortes homologies de séquences et des amplifications simple bande ont également été obtenues (Brown et al. 2001). Les amorces PCR ont été choisies suivant plusieurs critères : tout d'abord elles ont été développées à partir de clones d'ADNc présentant des profils simples lors des études utilisant des marqueurs RFLP chez *P. taeda*. De

Tableau C11 : Description des EST cartographié dans le pedigree G2.

DGGE : Denaturing Gradient Gel Electrophoresis; PAGE : Polyacrylamide Gel Electrophoresis; SSCP : Single Strand Conformation Polymorphism; ND : non dénaturant; UF : Urée-Formamide; RT : température de migration; * p > 0,01; ** p > 0,05.

Locus	Fonction	Méthode de détection	Acrylamide	Conditions de migration	Type de ségrégation	Distorsion	Groupe de liaison
PtIFG_1623	NS1-associated protein	PAGE	4%	ND	1:1	-	2
PtIFG_1643	ABI1 gene product (phosphatase)	DGGE	6%	15-45% UF	1:1:1:1	*	10
PtIFG_1764	-	DGGE	6%	15-45% UF	1:1	-	6
PtIFG_2274	Adenylyl cyclase	DGGE	6%	15-45% UF	1:1	-	4
PtIFG_2358	Phenylalanine tRNA synthetase	DGGE	10%	15-45% UF	1:1	-	6
PtIFG_2781	Glucose-induced repressor	PAGE	10%	ND	1:1:1:1	*	8
PtIFG_464	Aquaporin	PAGE	4%	ND	1:1	-	2
PtIFG_606	entE gene	DGGE	6%	15-45% UF	1:1	-	6
PtIFG_8415	-	DGGE	10%	15-45% UF	1:1	-	9
PtIFG_8436	Ribosomal protein 40S S3A	DGGE	10%	15-45% UF	1:1	-	7
PtIFG_8702	Thioredoxin	DGGE	10%	15-45% UF	1:1	-	6
PtIFG_8898	Testis mitotic checkpoint	DGGE	6%	15-45% UF	1:1	-	4
PtIFG_8907	Peroxidase cationic	DGGE	10%	15-45% UF	1:1:1:1	-	8
PtIFG_893	Nonspecific lipid transfer protein	PAGE	6%	ND	1:2:1	-	5
PtIFG_8939	Ribosomal protein 40S S16	DGGE	10%	10-30% UF	1:1:1:1	-	2
PtIFG_8972	Plasma membrane protein	DGGE	10%	15-45% UF	1:1	-	6
PtIFG_9036	Ribosomal protein L37	DGGE	10%	15-45% UF	1:1	-	7
PtIFG_9044	Ribosomal protein 40S S27	DGGE	10%	15-45% UF	1:1	*	6
PtIFG_9092	Nonspecific lipid transfer protein	DGGE	10%	15-45% UF	1:1	-	5
PtIFG_9151	Cucumber basic protein	DGGE	10%	15-45% UF	1:1	**	7
PtIFG_8429	3-DMOcytidylyltransferase	DGGE	10%	15-45% UF	1:1	-	4
PtIFG_8656	GF6P aminotransferase	DGGE	10%	15-45% UF	1:1	-	U
PtIFG_8779	Histone H3	DGGE	10%	15-45% UF	1:2:1	-	5
PtIFG_9136	Ribosomal protein S11	DGGE	10%	15-45% UF	1:1:1:1	-	3
PtIFG_1584	Deoxychalcone synthase	PAGE	8%	ND	1:1	-	4
PtIFG_8580	Embryogenesis abundant protein	PAGE	8%	ND	1:2:1	-	10
PtNCS_17G4	Endoglucanase	DGGE	10%	15-45% UF	1:1	-	U
PtNCS_1CAB7E	ACC oxidase	PAGE	10%	ND	1:1	-	U
PtNCS_22B8	Glycine hydroxymethyltransferase	DGGE	10%	15-45% UF	1:1	-	3
PtNCS_2C11E	Cinnamyl alcohol dehydrogenase	DGGE	10%	15-45% UF	1:1	-	9
PtNCS_C4H-1	trans-cinnamate 4-hydroxylase	DGGE	10%	15-45% UF	1:1:1:1	-	3
PtNCS_6C5A	Isoflavone reductase homolog	DGGE	10%	15-45% UF	1:1	-	4
PtNCS_7C4C	30S ribosomal protein	PAGE	10%	ND	1:1	-	U
PtTX_p14A9	Arabinogalactane-like protein	DGGE	10%	15-45% UF	1:1	-	3
PtNCS_PtaAGP6	Arabinogalactane-like protein	DGGE	6%	25-55% UF	1:1:1:1	-	5
PtNCS_ctg3	Caffeoyl CoA O-methyltransferase	DGGE	10%	15-45% UF	1:1	-	6
PpINR_CHS	Chalcone synthase	DGGE	10%	15-45% UF	1:1	-	2
PsUF1_NIR	Nitrite reductase	DGGE	10%	15-45% UF	1:1	-	7
PpINR_AN01E4	Glycin Decarboxylase	SSCP	8%	4°C RT	1:1	**	7
PpINR_AS01C7	Heat Shock Protein (HSP)	PAGE	8%	ND	1:2:1	-	10
PpINR_AS01F3	50S ribosomal protein	SSCP	8%	4°C RT	1:2:1	-	4
PpINR_AS01G01	Initiation factor	SSCP	8%	4°C RT	3:1	-	1
PpINR_AS01H04	RuBP Carboxylase	PAGE	8%	ND	1:1	-	10
PtMTU_PtCW1	Ferritin	SSCP	8%	4°C RT	1:1	*	5
PtMTU_PtCW2	Peroxidase	Agarose	2%	ND	1:1	-	8
AbWS2_AG3.18	Pinene Synthase	Agarose	2%	ND	1:1	-	8
PpINR_Pp.ap12	-	SSCP	8%	15°C RT	3:1	-	8
PpINR_Pp.ap23	-	Agarose	2%	ND	1:1	-	10
PpINR_Pp.ap9	Hypothetical zinc finger protein	SSCP	8%	4°C RT	1:1	-	6
PpINR_RS01D05	MYB-related protein	PAGE	8%	ND	1:1	-	7
PpINR_RS01G05	Unknown protein	Agarose	2%	ND	1:1	-	1
PpINR_SODChl	Superoxide dismutase	SSCP	8%	4°C RT	1:2:1	-	10
PrFRI_PrMC2	Male cone protein 1 precursor	PAGE	8%	ND	1:1	-	6
PrFRI_PthCAB	Chlorophyll a/b binding protein	PAGE	8%	ND	1:1	-	2

plus, afin d'éviter l'amplification de paralogues, les amorces ont été placées en 3'UTR afin d'être le plus spécifique possible d'un membre d'une famille multigénique.

Cinquante-quatre paires d'amorces sur 217 ont révélé du polymorphisme dans le croisement G2. Le tableau C11 présente la liste des amorces polymorphes ainsi que la technique employée pour révéler ce polymorphisme et les conditions PCR utilisées. Les données de génotypage de ces marqueurs polymorphes ont été intégrées dans les données de la carte G2 et les locus ont été cartographiés avec un LOD minimum de 6. Quatre locus n'ont pas pu être liés, même en diminuant les LOD. Il est intéressant de voir que deux de ces locus sont également non-liés chez *P. taeda*.

Les locus amplifiant chez le pin maritime également ont été testés sur le croisement F2. Dix-sept locus étaient polymorphes et 11 ont pu être cartographiés (tableau C12). Le fait que les six locus restants ne sont pas liés sur la carte peut avoir une explication statistique : la carte de l'hybride H12 avait été réalisée à l'aide de marqueurs RAPD (marqueurs dominants bialléliques) ségrégeant dans des proportions 3:1 dans la famille F2 (diploïde). Les EST polymorphes (codominants ou dominants) ségrégeaient quant à eux dans des proportions 1:2:1 ou 3:1 dans cette descendance. Nous avons indiqué dans l'encadré B4 l'information sur le taux de recombinaison « deux points » (inverse de la variance sur l'estimation du taux de recombinaison) pour de nombreuses configurations. On peut remarquer la faible information obtenue pour des paires de marqueurs ségrégeant en 3:1 / 1:2:1 (courbe rose située en dessous de celle obtenue pour des paires de marqueurs ségrégeant dans des proportions 1:1 / 1:1). L'information est encore plus faible pour des paires de marqueurs ségrégeant en 3:1 et 3:1 (courbe rouge 3:1 *Cis* / 3:1 *Cis*). Il n'est donc pas surprenant d'obtenir une forte variance sur le taux de recombinaison et donc une absence de liaison significative pour de nombreux locus. On peut par ailleurs invoquer la taille de l'échantillon. Alors que 192 F2 avait été génotypés avec les RAPD, seuls 49 à 80 individus l'ont été avec les EST.

La plupart des locus polymorphes ont été révélés par les techniques DGGE ou SSCP. Ceci montre que pour des fragments très courts (200-600 pb) ces techniques sont très efficaces. Guldberg et Guttler (1993) ont montré que la DGGE est très puissante et peut détecter des variations de l'ordre d'une seule base dans des fragments amplifiés. La présence d'hétéroduplexes a également permis d'interpréter certains profils complexes sans avoir à remettre au point de nouvelles conditions de gradient ou de migration. Néanmoins, à l'époque des séquenceurs automatiques, *microarrays* ou autres techniques à haut débit, les deux méthodes (SSCP et DGGE) ne fournissent que peu de génotypes par jour (pour une cuve DGGE, compter environ 48 génotypages par jour). En effet, le protocole complet

Tableau C12 : Description des EST cartographié dans le pedigree F2.
DGGE : *Denaturing Gradient Gel Electrophoresis*; SSCP : *Single Strand Conformation Polymorphism*; ** p > 0,05; † : problème de cartographie entre G2 et F2 décelé grâce à la carte de *P. taeda*.

	Méthode de détection	Acrylamide	Condition PCR	Ségrégation	Distorsion	Groupe de liaison F2	G2	*P. taeda*
PpINR_AN01B11	SSCP	8%	54°C	3:1	-	7	M	M
PpINR_AS01C7	SSCP	8%	54°C	1:2:1	-	5	M	M
PpINR_AS01D10	SSCP	8%	64°C	1:2:1	-	8	M	M
PpINR_AS01F3	SSCP	8%	54°C	1:2:1	-	Non lié	4	M
PpINR_AS01G01	SSCP	8%	48°C	1:2:1	-	4†	1	1
PpINR_RN01F06	SSCP	8%	52°C	3:1	-	11	M	M
PmLU_SB18	DGGE	6%	50°C	1:2:1	**	Non lié	M	M
PmLU_SB31	DGGE	6%	50°C	1:2:1	-	Non lié	M	M
PtNCS_20G2	DGGE	10%	52°C	1:2:1	**	2	M	-
PtNCS_AGP6	DGGE	6%	48°C	1:2:1	**	Non lié	4	4
PtNCS_CCoAOMT	DGGE	6%	55°C	1:2:1	-	Non lié	6	6
PtIFG_8580	DGGE	10%	54°C	1:2:1	-	5	10	10
PtIFG_8596	DGGE	10%	50°C	3:1	-	9†	M	5
PtIFG_8647	DGGE	10%	52°C	1:2:1	-	1	M	6
PtIFG_9113	DGGE	10%	52°C	1:2:1	-	Non lié	M	8
PtNCS_23E4	DGGE	10%	53°C	1:2:1	-	1	M	-
PtIFG_9092	DGGE	10%	52°C	1:2:1	-	2	M	5

Tableau C1 : Evaluation des marqueurs basés sur les EST.

	Développement	**Cartographie génétique**	**Cartographie comparée**	**Détection QTL**	**GC positionnel**
AFLP	++	++	--	+/-	--
SSR (transfert)	+	+/-	+/-	++	--
SSR-ADNc	++	+/-	+	+	+
SSR (banque enrichie)	--	+	+	++	--
EST (SSCP, DGGE)	Utilisation des EST de conifères disponibles (bases de données)	Polymorphisme détecté par des méthodes à "moyen débit" (SSCP, DGGE, gel d'acrylamide)	Marqueurs orthologues hautement transférables Problème de locus paralogues (familles multigéniques)	Bons marqueurs pour la détection de QTL (codominants, multialléliques)	Marqueurs codants pouvant correspondre à des gènes candidats
SNP	?	?	?	?	?

(amplification PCR, préparation des gels, migration, coloration) prend toute une journée de travail. De plus ces techniques ne permettent pas d'envisager de multiplexer des locus.
En ce qui concerne la technique décrite par Cato et al. (2001), les résultats sont très décevants : beaucoup d'efforts ont été placés dans l'optimisation de cette technique chez le pin maritime mais en vain. Seulement deux locus sur 60 testés ont pu être cartographiés et un grand nombre de locus présentait des profils de gel ségrégeant dans des proportions non mendéliennes. De plus, la mise en œuvre de cette technique nécessite beaucoup d'étapes (une digestion, une ligation et deux PCR emboîtées), ce qui la rend difficilement applicable à très grande échelle. De ce fait une autre approche a été envisagée pour détecter du polymorphisme dans les gènes d'intérêt : cette approche est basée sur l'utilisation des polymorphismes ponctuels, aussi appelés SNP.
Pour conclure sur les marqueurs moléculaires basés sur les EST (tableau C1), ces marqueurs sont beaucoup plus intéressants que les marqueurs microsatellites et AFLP pour l'obtention de locus orthologues transférables entre espèces de pins. En revanche, la mise en évidence du polymorphisme est de faible efficacité en terme de polymorphisme et de débit de génotypage. De ce fait, une autre approche présentée dans le chapitre suivant a été mise en œuvre afin de détecter des polymorphismes dans ces séquences codantes.

C.1.3.2 Les polymorphismes ponctuels (*Single Nucleotide Polymorphism*, SNP)

Les premiers résultats présentés n'ont pas été obtenus dans le cadre de cette thèse mais dans le cadre des travaux de thèse de D. Pot (2004) et de J. Tibbits (Université de Melbourne, thèse en cours). Ils sont synthétisés ici car la technique utilisée (extension d'amorce, Kuppuswamy et al. 1991) a permis de cartographier des gènes. La seconde partie de ce chapitre (C.1.3.2.2) est présentée en détails à l'annexe V.

C.1.3.2.1 Cartographie de gènes candidats liés à la formation du bois par détection de SNP

Des gènes candidats pour la qualité du bois ont été identifiés selon plusieurs critères : tout d'abord des gènes candidats « expressionnels » qui sont différentiellement exprimés dans des types de bois différents (bois initial *vs.* bois final, bois opposé *vs.* bois compressé, bois juvénile *vs.* bois mature, Le Provost 2003 ; Sheree Cato, FRI, communication personnelle),

Tableau C13 : Description des SNP cartographiés dans le pedigree G2. Les détails concernant l'expression et l'implication de ces gènes dans les voies de biosynthèse liées à la formation du bois sont décrits par Plomion et al. 2001.

Locus	Fonction	Expression différentielle	Voie de biosynthèse	SNP	Groupe de liaison
Fructokinase	Fructokinase	non	Cellulose	Indel	2
AGP_1	Arabinogalactane protéine	oui	Cellulose	G/C	4
PAL	Phenylalanine ammonia lyase	oui	Lignines	G/C	6
CcoAOMT	Caffeoyl coenzyme A O-méthyltransférase	oui	Lignines	A/G	6
CesA01	Cellulose synthase	non	Cellulose	A/G	6
Sh2	Putative déhydrine	oui	-	A/G	8
CAD	Cinnamyl alcool déshydrogénase	oui	Lignines	A/G	9
SuSy	Sucrose synthase	non	Cellulose	A/T	10
Korrigan	Endo-1,4-ß-Glucanase	non	Cellulose	A/G	12
CesA07	Cellulose synthase	non	Cellulose	G/C	Non lié

Tableau C14 : Résultats de la détection de SNP dans 13 gènes avec le pipeline de détection de SNP et par inspection visuelle : gènes utilisés, taille des fragments séquencés, accession dbSNP, comparaison des deux méthodes, efficacité de la méthode automatique et nombre de faux positifs. Nomenclature des gènes : C4H: *Trans-cinnamate-4-monooxygenase*; CAD: *Cinnamyl alcohol dehydrogenase* (deux fragments F1 et F2); PAL: *Phenylalanine ammonia lyase*; Korrigan: *endo-1,4-β-D-glucanase*; GRP: *Glycine-rich protein*; MYB-like TF: *MYB-like transcription factor*; ACC oxidase: *amino-cyclopropane-carboxylic acid oxidase*; CesAn: *Cellulose synthase*; CCoAOMT: *Caffeoyl CoA O-methyltransferase*; AGP: *Arabinogalactan protein*.

Nom du gène	Taille du fragment (pb)	Nombre de vrais SNP détectés par inspection visuelle	Accession dans dbSNP	Nombre de total SNP détectés par la méthode automatique	Nombre de vrais SNP détectés par la méthode automatique	Nombre de vrais SNP non-détectés	Nombre de faux positifs
C4H	539	10	ss16208982;8985-8987;8991;8994-8998;9000	10	10	0	0
CAD-F1	537	4	ss16209001;9005;9008;9010	3	3	1	0
CAD-F2	523	4	ss16209011;9012;9015;9016	4	3	1	1
PAL	547	9	ss16209020-9025;9028;9029;9032	7	7	2	0
Korrigan	937	5	ss12709589-ss12709593	7	5	0	2
GRP	479	9	ss12709575-ss12709585	8	8	1	0
MYB-like TF	494	2	ss12709586-ss12709587	3	2	0	1
ACC oxydase	270	1	ss12709588	1	1	0	0
CesA7	553	2	ss12709573-ss12709574	2	2	0	0
CesA4	489	1	ss12709572	0	0	1	0
CesA3	490	7	ss12709565-ss12709571	6	6	1	0
CCoAOMT	492	7	ss16209076-ss16209082	3	3	4	0
AGP	321	4	ss16209070;9071;9073;9074	4	4	0	0
Total	6671	65		58	54	11	4

ensuite des gènes appartenant à des voies de biosynthèses importantes pour la formation du bois (biosynthèse des lignines et des celluloses, Plomion et al. 2001), et enfin un gène dont un mutant déficient a été identifié chez Arabidopsis (*Korrigan*, Nicol et al. 1998). Ces gènes sont présentés au tableau C13.

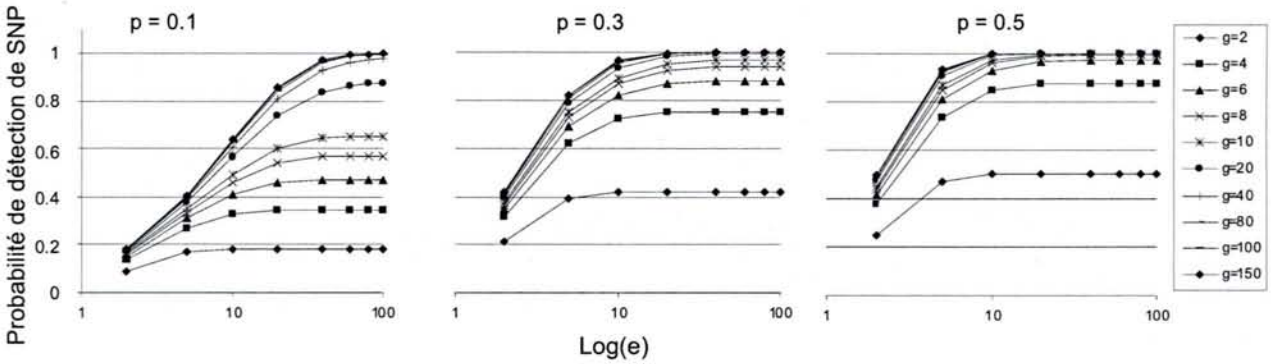
La technique utilisée pour révéler le polymorphisme est la technique d'extension d'amorces. Dix locus se sont révélés polymorphes dans le croisement G2. La position sur la carte génétique de neuf de ces locus est présentée à la figure C18 (préfixe snp_), le locus CesA07 étant resté non lié aux marqueurs de la carte jalon.

L'utilisation de marqueurs basés sur les SNP dans des séquences codantes a permis de cartographier des gènes candidats sur la carte du croisement G2. Ces marqueurs peuvent de plus être utilisés pour des études d'association en population naturelle (E. Eveno, thèse en cours) en utilisant des techniques de génotypage à haut débit. Cependant, ces gènes ont dû être séquencés sur de nombreux fragments afin de détecter des SNP. Une méthode automatique de détection de SNP dans les bases d'EST disponibles pour le pin maritime a donc été mise au point dans le but d'éviter cette étape préliminaire de séquençage.

C.1.3.2.2 Détection automatique de SNP dans les bases d'EST de pin maritime

La partie la plus longue dans le développement de SNP est la détection de ces SNP. Généralement il est nécessaire de séquencer les mêmes fragments sur des individus différents, puis d'aligner les séquences obtenues afin de détecter des SNP. La disponibilité de nombreuses séquences d'EST dans les bases de données publiques peut fournir des SNP en grande quantité. Pour cela un système automatique de détection de SNP a été mis au point à partir des logiciels *Phred-Phrap* et *Polybayes*. Un lot de 65 vrais SNP a été utilisé pour paramétrer ce système automatique (tableau C14). Les données utilisées proviennent de 13 fragments de gènes séquencés sur 30 mégagamétophytes issus de l'aire naturelle du pin maritime. Cinquante huit SNP ont été détectés avec une valeur $P_{SNP} > 0.99$ (valeur donnée par *Polybayes* et dépendant de la qualité des séquences et du nombre de variants au niveau du site polymorphe). Sur ces 58 SNP, 54 correspondaient à de vrais SNP du lot de référence, ce qui montre que le système automatique a une efficacité de 83% de détection de vrais SNP. Les 11 SNP non détectés correspondaient tous à des SNP ayant des allèles rares ($p<0,1$) au sein des 30 mégagamétophytes. Ainsi, en considérant les SNP ayant une fréquence supérieure à 10%, le pipeline détecte 98% des vrais SNP. Quatre SNP ayant une P_{SNP} significative se sont

Figure C5 : Probabilité de détection de SNP dans une base de données d'EST en fonction (i) du nombre de génome haploïdes (2<g<150), (ii) du nombre de séquences dans un contig (2<e<100) et (iii) de la fréquence allélique des SNP (0,1< P_A <0,5).



Encadré C1 : Calcul de la probabilité de détection de SNP dans la base de données d'EST.

Soit X_A la variable aléatoire correspondant au nombre d'allèles A dans "e" ESTs, Y_A la variable aléatoire correspondant au nombre d'allèles A dans "g" génomes, et P_A la fréquence allélique d'un allèle du SNP dans la population où ont été échantillonés les individus ayant servis à faire la banque de'ADNc (en considérant tous les SNP bialléliques). La probabilité de détecter des SNP est égale à :

$1-[\ P(X_A=0, or, X_A=e)\ ] = 1-P(X_A=0)-P(X_A=e)$, avec $P(X_A = e)$ la probabilité que tous les EST aient l'allèle A, et $P(XA = 0)$ la probabilité que tous les EST aient l'allèle alternatif. Pour calculer ces probabilités nous avons utilisé le théorème de Bayes :

$$P(X_A=0)=\sum_{y_A=1}^{g} P(Y_A=y_A)\times P(X_A=0/Y_A=y_A)$$

et $$P(X_A=e)=\sum_{y_A=1}^{g} P(Y_A=y_A)\times P(X_A=e/Y_A=y_A)$$

avec $$P(Y_A=y_A)=\beta(y_A,g,P_A)=\frac{g!P_A^{\,y_A}(1-P_A)^{g-y_A}}{(g-y_A)!y_A!}$$

et $$P(X_A=x_A/Y_A=y_A)=\beta(x_A,e,\frac{y_A}{g})$$

ce qui donne $$P(X_A=0/Y_A=y_A)=\left[\frac{1-y_A}{g}\right]^e$$

et $$P(X_A=e/Y_A=y_A)=\left[\frac{y_A}{g}\right]^e$$

La probabilité de distribution β correspond à la loi binomiale.

révélés être de faux positifs. Ces faux positifs se trouvaient dans des régions où les séquences étaient de mauvaise qualité.

Disposant d'un système automatique capable de détecter des SNP de façon efficace, nous les avons recherchés dans la base de données d'EST de pin maritime. La probabilité de détecter des SNP dans une base d'EST a d'abord été calculée (encadré C1) : en fonction du nombre d'EST dans les contigs considérés pour la détection (e), du nombre de génomes haploïdes qui ont servi à construire la banque (g) et de la fréquence allélique du SNP (P_A). La figure C5 présente la probabilité de détecter des SNP en fonction du nombre d'EST par contigs et pour plusieurs valeurs de nombre de génome haploïde et diverses fréquences alléliques. Pour des SNP que l'on qualifiera de « fréquents » ($P_A \geq 0,3$), le nombre d'EST par contig et le nombre de génomes haploïdes ne sont pas limitants car les courbes obtenues atteignent rapidement un plateau proche de 100% de chances de retrouver le SNP. En revanche pour les allèles « plus rares » ($P_A \leq 0,1$), le nombre d'EST par contig et de génomes haploïdes devient limitant.

En ce qui concerne la banque d'EST de pin maritime, une partie a été construite avec quatre arbres (banque xylème de 10 000 EST, $g = 8$), tandis qu'une autre partie (environ 8 000 EST) a été construite à partir d'une centaine d'individus (banque racines, $g > 200$). De ce fait, la valeur de g ne doit pas être limitante dans le cas des EST de pin maritime. Etant donné que la fréquence allélique de chaque SNP n'est pas accessible, le seul paramètre qui peut être ajusté est le nombre d'EST par contig. Pour un nombre minimal de $e = 4$, la probabilité de détecter un SNP varie entre 0,4 (pour $P_A = 0,1$) et 0,95 (pour $P_A = 0,5$). De ce fait, nous avons choisi cette valeur pour détecter des SNP dans la base de données d'EST de pin maritime, conscient que certains SNP ayant des allèles rares pouvaient rester indétectables. En ce qui concerne le nombre maximal d'EST par contig, Batley et al. (2003) dans une étude similaire réalisée chez le maïs ont restreint leur recherche à un maximum de 20 EST par contig. En effet, des erreurs peuvent se produire au cours des étapes de construction de la banque d'ADNc, en particulier au moment de la rétrotranscription de l'ARN. Bebenek et al. (1993) ont montré que les *reverse transcriptase* commettent souvent ce genre d'erreurs. Au niveau des séquences d'EST, ces erreurs ne sont pas discernables car les bases erronées possèdent des scores de qualité de séquence optimaux étant donné que les erreurs ont eu lieu avant l'étape de séquençage. Une manière de prévenir ce genre d'artéfacts serait de valider les SNP détectés en fonction du nombre de fois qu'ils apparaissent dans l'alignement. Ces erreurs de transcription sont censées se produire de manière aléatoire n'importe où dans les fragments d'ADN, la probabilité de retrouver ces erreurs au même endroit chez plusieurs individus est donc très faible. De plus, les SNP pourraient également être validés par l'étude des

Figure C6 : Comparaison entre la quantité de SNP détectés par la méthode automatique avec la quantité espérée considérant la fréquence des SNP dans le lot de SNP de référence (1 SNP / 102 pb) et la longueur totale des séquences utilisées.

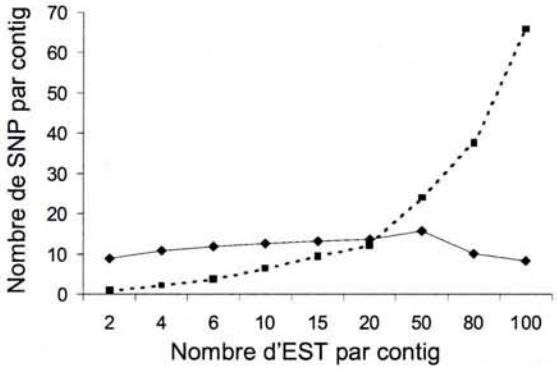


Tableau C15 : Résultats de la détection de SNP dans la base de données de séquences d'EST de pin maritime.

Nombre de contigs avec 4<e<20		940
Taille cumulée des séquences		924 216 pb
Nombre de contigs sans SNP		568
Fréquence		1 SNP / 660 pb
Nombre de SNP détectés (P_{SNP}>0.99)		1400
Plus de 2 allèles		0
Transition	ag	374
	ct	444
	total	818
Transversion	ac	148
	at	167
	cg	122
	gt	145
	total	582

Figure C1 : Evaluation des marqueurs SNP.

	Développement	Cartographie génétique	Cartographie comparée	Détection QTL	GC positionnel
AFLP	++	++	--	+/-	--
SSR (transfert)	+	+/-	+/-	++	--
SSR-ADNc	++	+/-	+	+	+
SSR (banque enrichie)	--	+	+/-	++	--
EST	+	+	++	+	++
SNP	Nécessité de séquencer les mêmes fragments pour plusieurs individus Détection *in silico* à partir des séquences d'EST	Taux de polymorphisme important (1 SNP / 100 pb) Techniques de génotypage à "haut débit"	Site polymorphes non conservés entre espèce Possibilité de trouver un autre SNP dans un gène orthologue chez une espèce proche	Marqueurs le plus souvent biallélliques	Idem EST

coségrégations au sein des séquences avec des SNP proches. Cette méthode basée sur les haplotypes a été proposée par Batley et al. (2003) sur le maïs et par Kota et al. (2003) sur l'orge.

Au niveau de nos données, la fréquence d'apparition des SNP au sein des séquences a été comparée pour plusieurs valeurs d'EST par contig avec la fréquence attendue, correspondant à la fréquence observée dans le lot de vrais SNP (1 SNP / 102 pb). La figure C6 nous montre que le nombre de SNP détectés augmente exponentiellement lorsque le nombre d'EST par contig augmente, ce qui suggère la présence de faux positifs dus aux erreurs de transcription.

Au total, 940 contigs possédaient entre 4 et 20 contigs et ont permis la détection de 1400 SNP ayant une valeur de P_{SNP} supérieure à 0,99 (*Polybayes*). 568 contigs ne contenaient pas de SNP. Les résultats de cette détection sont résumés au tableau C15 et les SNP détectés ont été publiés dans dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). La totalité des SNP détectés sont bialléliques. La fréquence de SNP calculée sur la longueur cumulée des 372 contigs est de un SNP tous les 660 pb, ce qui est supérieur à ce qui avait été observé sur le lot de SNP de références ou bien chez d'autres espèces de plantes dans le cadre d'études similaires (Batley et al. 2003). A l'inverse des séquences génomiques alignées au niveau d'un même fragment, les EST peuvent être alignés entre eux par chevauchement d'une partie de leur séquence. Ainsi, il est fréquent de n'observer que deux séquences alignées dans certains contigs, ce qui réduit considérablement les chances de détecter des SNP.

En ce qui concerne les types de SNP détectés, les transitions représentent 58,4% des résultats, ce qui est en accord avec les résultats obtenus chez d'autres espèces (Batley et al. 2003).

Pour conclure sur l'approche basée sur la détection *in silico* de SNP dans les bases d'EST (tableau C1), cette stratégie a fourni un nombre important de SNP situés dans des séquences de fonction connue. Ceci permettra d'avoir accès à une ressource importante de marqueurs pour des études de cartographie génétique et des études d'association.

Figure C7 : Exemple d'alignement des cartes femelle et mâle pour le groupe de liaison 12 du croisement G2. Les trois cartes sont reliées entre elles par des marqueurs intercross 3:1.

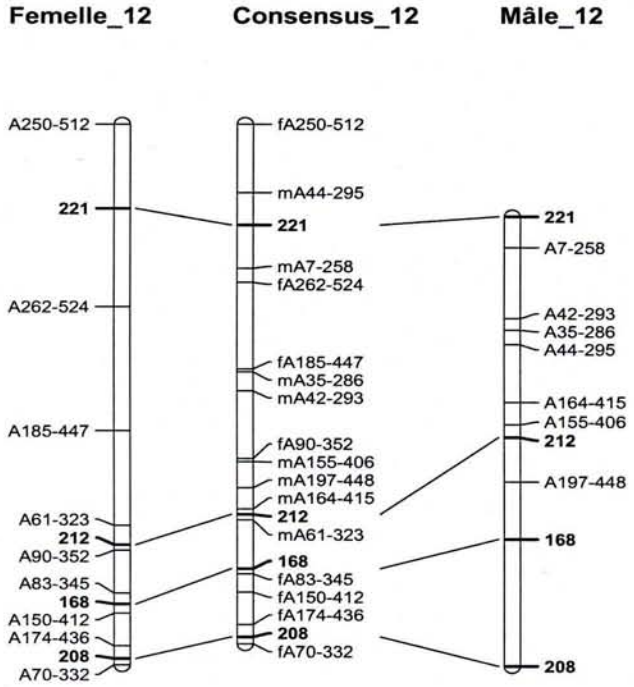


Tableau C16 : Tailles génétiques totales et nombre de groupe de liaison obtenus pour les cartes du parent femelle, mâle et de la carte consensus, à l'aide de deux logiciels.

	JOINMAP	*MAPMAKER*
Femelle	1218 cM (12 LG)	1807 cM (12 LG)
Mâle	1297 cM (15 LG)	1541 cM (16 LG)
Consensus	1407 cM (12 LG)	-

Figure C9 : Alignement des cartes G2 et F2 : groupes homologues. L'intensité des lignes reliant les deux cartes est proportionnelle au nombre de locus communs. La numérotation des groupes de liaison de la carte G2 suit celle des cartes d'autres espèces de conifères.

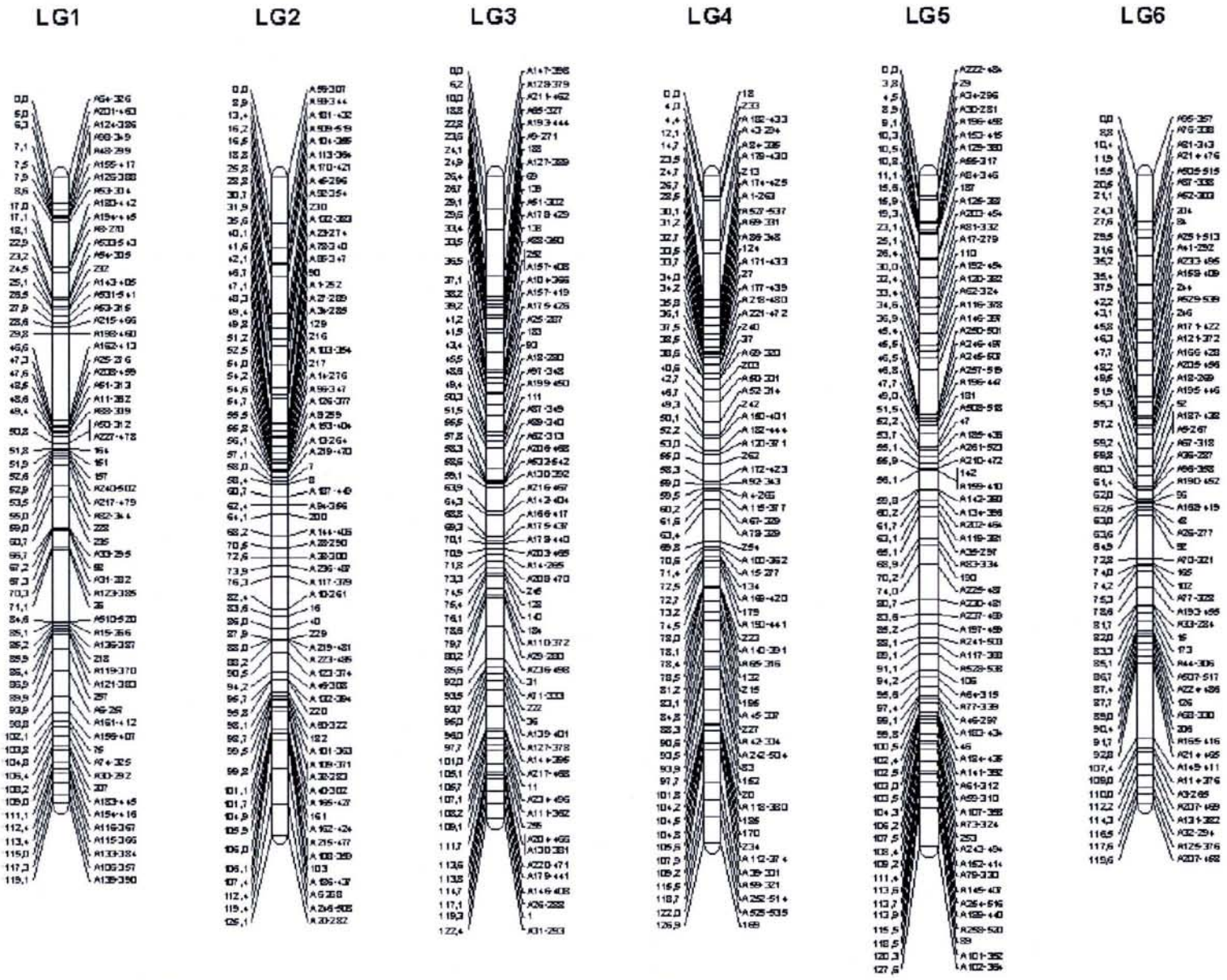
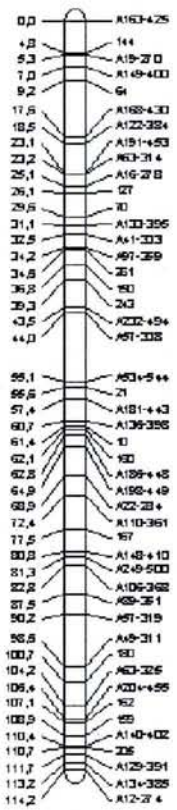
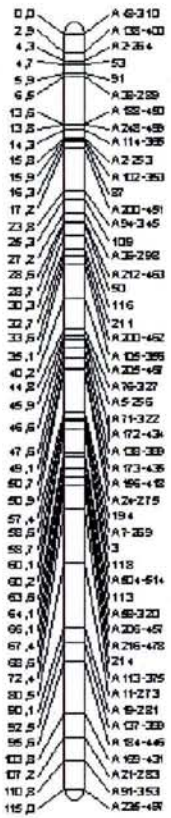


Figure C8 : Carte génétique consensus du croisement G2.

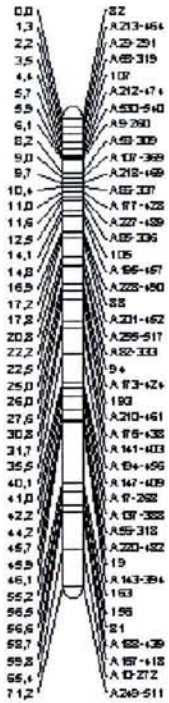
LG7



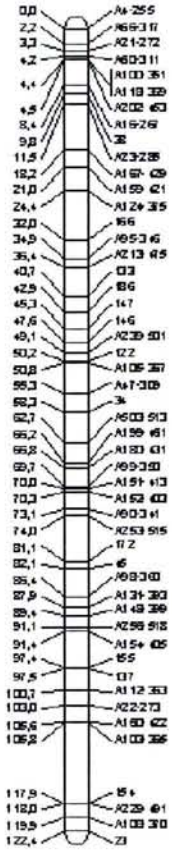
LG8



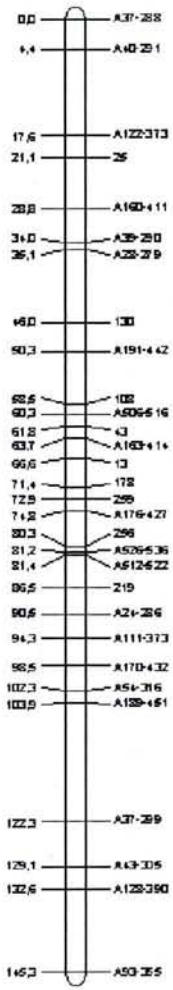
LG9



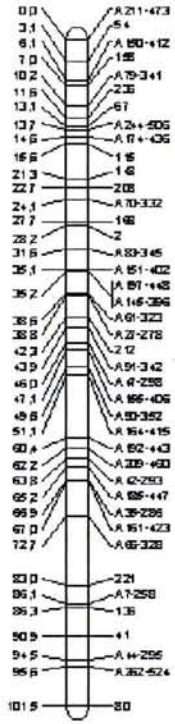
LG10



LG11



LG12



C.2 Cartographie génétique chez le pin maritime

C.2.1 Etablissement d'une carte génétique à partir de la descendance G2

Les résultats concernant la construction d'une carte génétique saturée du croisement G2 sont présentés à l'annexe I.

Les cartes saturées des parents du croisement G2 ont été construites en utilisant des marqueurs AFLP (partie C.1.1). Un total de 251 marqueurs ségrégeant dans les proportions 1:1 a permis de construire la carte du parent mâle et 262 marqueurs ségrégeant dans les proportions 1:1 ont permis de construire la carte du parent femelle. Ces deux cartes ont respectivement des longueurs totales de 1297 cM et 1218 cM (Kosambi, *JOINMAP*). La carte du parent mâle comporte 15 groupes de liaison alors que celle de la femelle est saturée (12 groupes de liaison). Il est intéressant de noter que les résultats obtenus à l'aide du logiciel *MAPMAKER* varient légèrement en terme de taille génétique par rapport à ceux obtenus grâce à *JOINMAP* (tableau C16). Par exemple, la carte du parent mâle comportait 16 groupes (1541 cM) avec *MAPMAKER* et 15 groupes (1297 cM) avec *JOINMAP*. Des résultats similaires ont été observés chez l'orge (Qi et al. 1996) et le pin taeda (Sewell et al. 1999). Ce phénomène est dû au fait que les deux logiciels utilisent des méthodes différentes pour calculer les interférences.

L'utilisation de marqueurs ségrégeant dans les proportions 3:1 cartographiés à la fois sur les deux cartes grâce à *JOINMAP*, ont permis d'aligner les cartes parentales (figure C7) et de réaliser une carte consensus qui a une longueur totale de 1407 cM (figure C8).

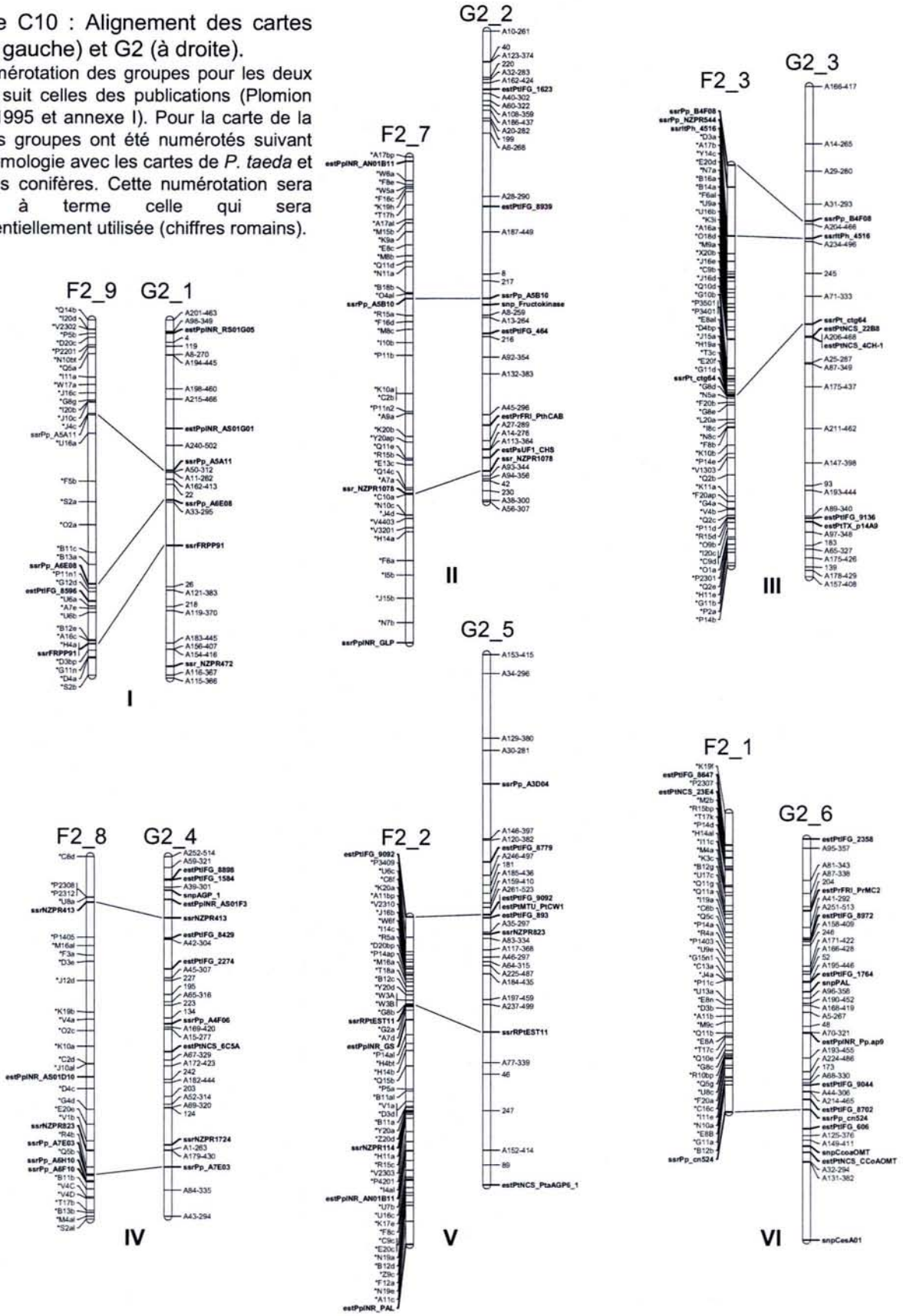
C.2.2 Vers une carte génétique consensus du pin maritime

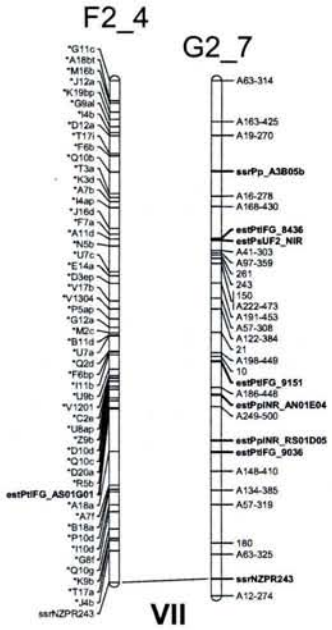
C.2.2.1 Comparaison des cartes F2 et G2

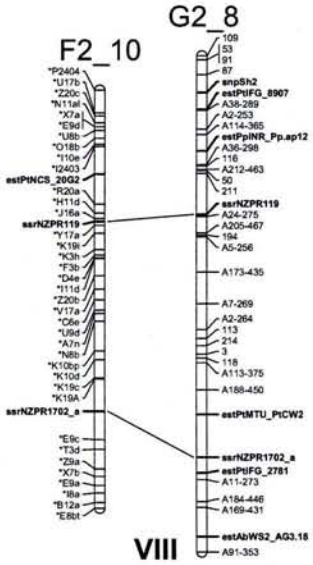
Vingt-trois locus comprenant 19 microsatellites et 4 EST ont pu être cartographiés à la fois sur la carte du croisement G2 et sur la carte F2. Onze groupes de liaison sur 12 ont pu être alignés grâce à ces marqueurs. La figure C9 présente les groupes homologues entre les deux cartes en utilisant leur numérotation (Plomion et al. 1995 pour la F2, annexe I pour la G2). Idéalement la numérotation de la carte F2 devrait disparaître afin d'éviter toute ambiguïté. En effet, la numérotation de la carte G2 est homologue à celle de la carte de référence du projet

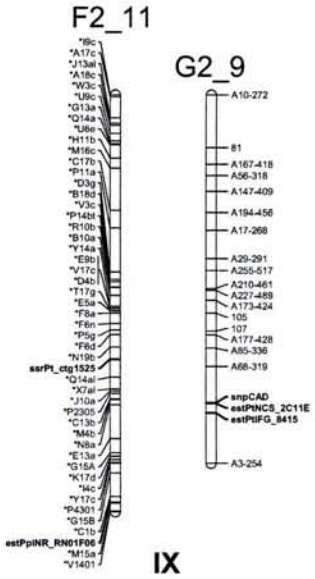
Figure C10 : Alignement des cartes F2 (à gauche) et G2 (à droite).

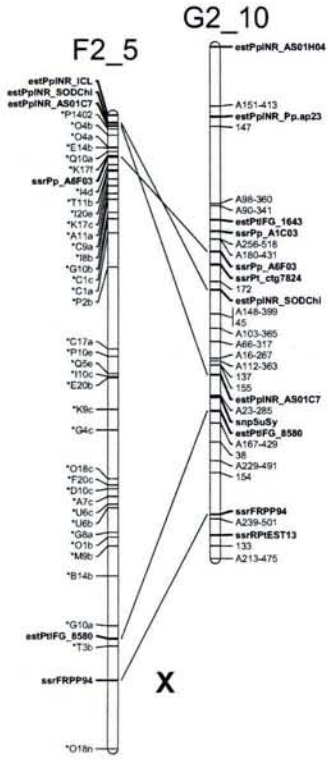
La numérotation des groupes pour les deux cartes suit celles des publications (Plomion et al. 1995 et annexe I). Pour la carte de la G2, les groupes ont été numérotés suivant leur homologie avec les cartes de *P. taeda* et d'autres conifères. Cette numérotation sera donc à terme celle qui sera préférentiellement utilisée (chiffres romains).

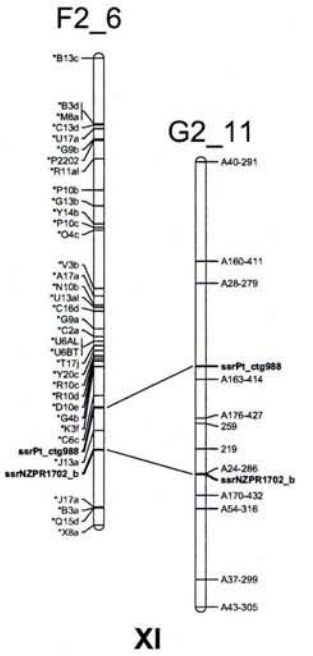


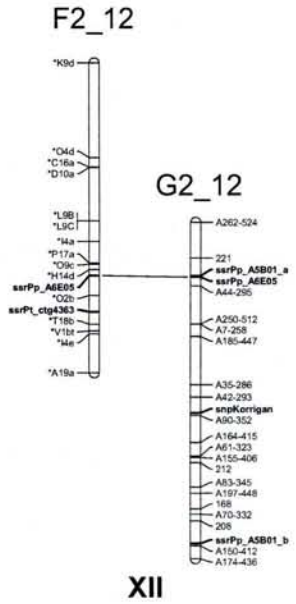

F2_4
G2_7
VII


F2_10
G2_8
VIII


F2_11
G2_9
IX


F2_5
G2_10
X


F2_6
G2_11
XI


F2_12
G2_12
XII

de cartographie génétique comparée chez les conifères (*P. taeda*, Brown et al. 2001), il est donc plus judicieux d'utiliser cette nomenclature.

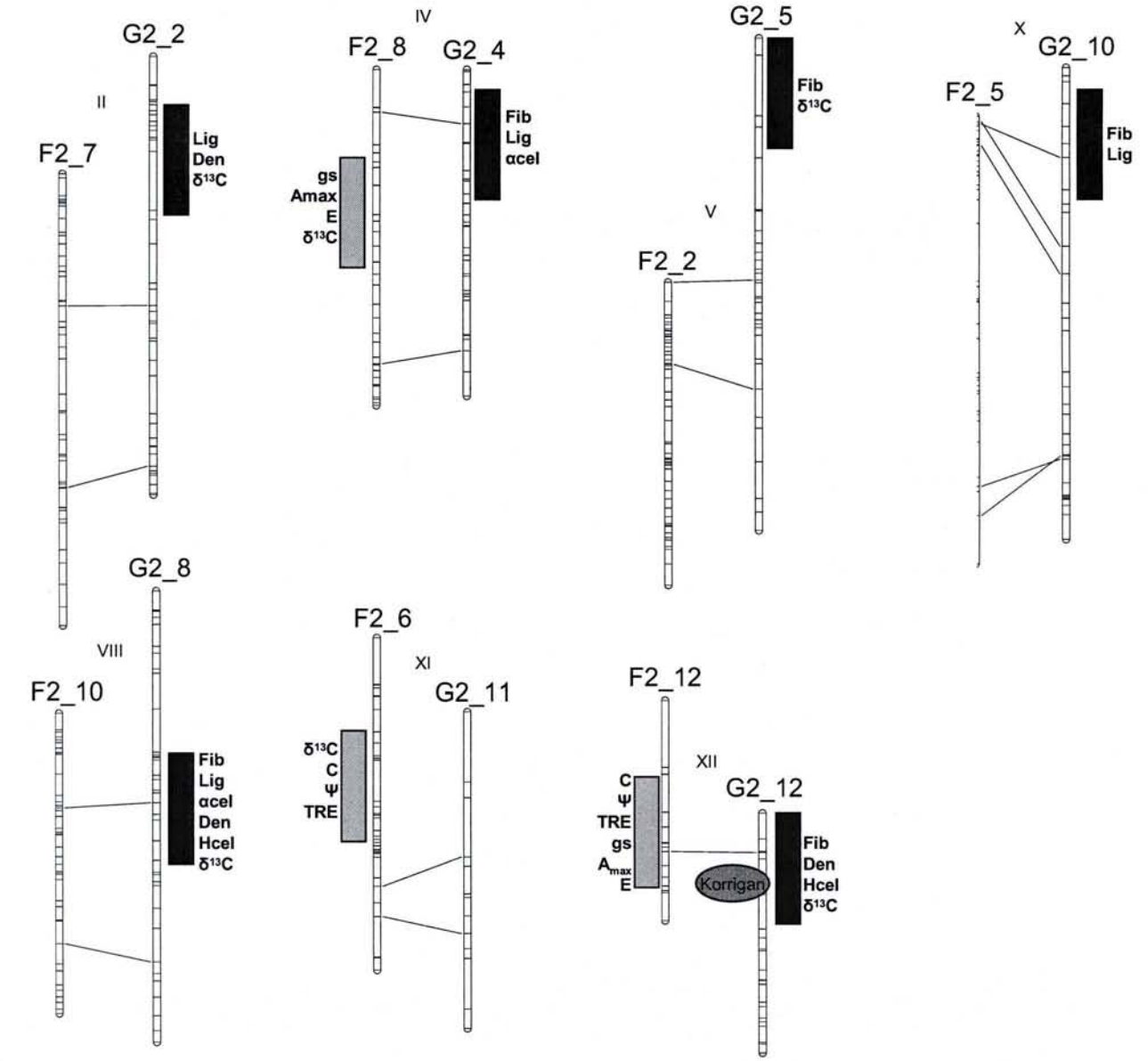
La figure C10 présente l'alignement des deux cartes F2 et G2. Il est à noter que certains groupes de liaison ont un seul marqueur commun. Cela ne suffit pas à assigner une homologie de manière sûre étant donné que les chances de cartographier des locus paralogues ne sont pas nulles. De la sorte, certains locus semblent ne pas être localisés sur le bon groupe de liaison. Par exemple, les locus ssrPp_A5B11, ssrPp_A6E08 et ssrFRPP91 se situent sur le groupe 1 du croisement G2, ce qui nous permet de penser que le groupe G2_1 est homologue au groupe F2_9. En revanche le locus estPpINR_AS01G01 se trouve sur les groupes G2_1 et F2_4, ce qui indique qu'il est sûrement le résultat d'une amplification PCR non-spécifique. Ce résultat doit néanmoins être confirmé par séquençage des produits PCR obtenus sur les deux croisements.

Dans certains cas, la cartographie génétique comparée avec d'autres espèces de conifères a permis de confirmer l'homologie des groupes de liaison (figure C18). Par exemple, les locus développés et cartographiés *P. taeda* (EST) peuvent apporter ce genre d'information. En prenant l'exemple du groupe F2_1, on se rend compte que le locus estPtIFG_8647 n'est pas polymorphe dans la G2. Ce locus est cartographié sur le groupe 6 de *P. taeda*, ce qui confirme que le groupe F2_1 est bien homologue avec le groupe G2_6. De la même manière, le locus estPtIFG_9092 cartographié sur la F2 confirme l'homologie entre les groupes 5 de la G2 et de *P. taeda*. En revanche d'autres locus posent problème. C'est le cas du locus estPpINR_PAL cartographié dans la F2 (groupe 2 homologue au groupe 5 de la G2), qui devrait être retrouvé sur le groupe 6 où il est localisé chez *P. taeda*. Ceci peut être expliqué par l'amplification d'un locus paralogue car la *Phenyalanine Ammonia-Lyase* appartient à une famille multigénique chez les pins, comme cela a été démontré chez *Pinus banksiana* (Butland et al. 1998). De la même manière, le locus estPtIFG_8596 est localisé sur la carte de la F2 sur le groupe 9 qui est homologue au groupe 1 de la G2, alors que ce locus est attendu sur le groupe 5 de *P. taeda*. Dans ce cas, on remarque que ce marqueur est seulement lié à la carte avec un LOD maximum de 2,2.

C.2.2.2 Y a-t-il des QTL communs entre les cartes F2 et G2 ?

Une des manières de valider la position d'un QTL est de vérifier sa localisation sur des cartes génétiques différentes. Une telle approche a été décrite par exemple chez *P. taeda* par Brown et al. (2003) et au niveau interspécifique entre *P. taeda* et *P. pinaster* (annexe III et

Figure C11 : Comparaison de la position de clusters de QTL contrôlant des caractères liés à la résistance à un stress hydrique et à la qualité du bois entre les cartes G2 et F2.



Légende

Cluster de QTL contrôlant la réponse à un stress hydrique

- **C** : Consommation en eau*
- **Ψ** : Potentiel hydrique*
- **TRE** : Teneur relative en eau*
- **gs** : Conductance stomatique*
- **A_{max}** : Taux de photosynthèse maximum*
- **E** : Transpiration*
- **δ13C** : Efficience d'utilisation de l'eau*,**

Cluster de QTL liés à la qualité du bois

- **Lig** : Taux de lignine***
- **αcel** : Taux de cellulose***
- **Hcel** : Taux d'hémicellulose***
- **Den** : Densité du bois***
- **Fib** : Propriétés des fibres***

* : Costa 1999
** : Brendel et al. 2002
*** : Pot 2004

partie C.3.3). De la même manière, l'alignement des cartes G2 et F2 pourrait permettre de vérifier la position de QTL afin de les valider. Néanmoins, ces deux cartes ont été utilisées avec des objectifs différents. Comme je l'ai présenté en introduction, des QTL de réponse à un stress osmotique ont été cartographiés sur la F2 (Costa 1999) alors que des QTL contrôlant des caractères liés à la qualité du bois ont été localisés sur la G2 (Pot 2004). Ces caractères semblent donc difficiles à comparer. Cependant, la physiologie particulière des arbres forestiers peut éventuellement nous permettre de les comparer. En effet, le bois d'été a des propriétés différentes du bois de printemps. Ces différences peuvent être en particulier liées à la disponibilité en eau qui est très contrastée selon les saisons pour le pin maritime (en particulier dans les podzols très drainants du plateau des Landes de Gascogne). Les mêmes mécanismes moléculaires et physiologiques peuvent donc être mis en jeu lors de ces deux processus *a priori* indépendants.

Les thèses réalisées par Costa (1999) et Pot (2004) sur le pin maritime ont révélé des « hot spots » de QTL sur les deux cartes génétiques du pin maritime (voir chapitre B.2.5). La figure C11 présente un alignement simplifié des deux cartes avec la localisation de ces régions. Deux groupes présentent des co-localisations pour les QTL de qualité du bois et liés au stress hydrique. Au niveau du groupe 12, des QTL codant pour les propriétés des fibres, la densité du bois, les taux de lignine et d'hémicellulose, la transpiration foliaire, la consommation en eau, le potentiel hydrique, la teneur relative en eau, la photosynthèse et la conductance stomatique co-localisent entre les deux cartes ainsi qu'avec le gène *Korrigan*. Ce gène a été identifié à partir d'un mutant d'*Arabidopsis thaliana* (Nicol et al. 1998) et plusieurs études ont montré qu'il est impliqué dans la voie de biosynthèse des polysaccharides. De plus, Pot (2004) a montré que ce gène est soumis à une sélection purificatrice au niveau moléculaire et qu'il est associé avec des caractères comme les taux de cellulose, d'hémicellulose et de lignine, la densité du bois et le $\delta^{13}C$ (corrélé à l'efficacité d'utilisation de l'eau). Ceci suggère que ce gène, ou du moins cette région chromosomique, joue un rôle important dans l'adaptation du pin maritime à son environnement.

C.2.3 Qu'est-ce que la cartographie génétique peut nous apprendre sur la structure du génome des conifères ?

Outre la conservation de la structure des chromosomes entre espèces révélée grâce à la cartographie génétique comparée (se reporter à la partie C.3), la disponibilité d'une carte génétique peut fournir des informations sur la structure du génome des conifères. Deux

Figure C12 : description d'un appareillage de cytométrie de flux.

Les cellules sont dispersées dans du sérum physiologique et passent à travers un faisceau de lumière laser.

La diminution de l'intensité du faisceau direct permet de mesurer la masse cellulaire ou la masse de DNA.

Les cellules qui émettent de la fluorescence sont reconnues par une cellule photoélectrique particulière.

Enfin, on peut éventuellement séparer les cellules selon leur charge électrique : séparateur de cellules.

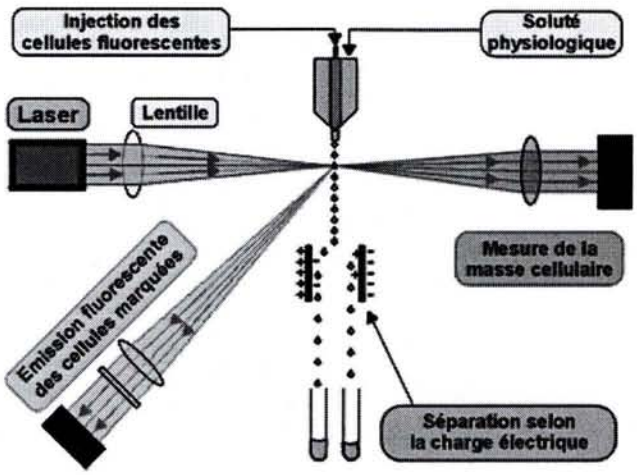


Figure C13 : Relation entre le nombre de crossing-over et la taille physique moyenne par chromosome pour 15 espèces de plantes. Les correspondance des numéros d'espèces se trouvent dans le tableau C17.

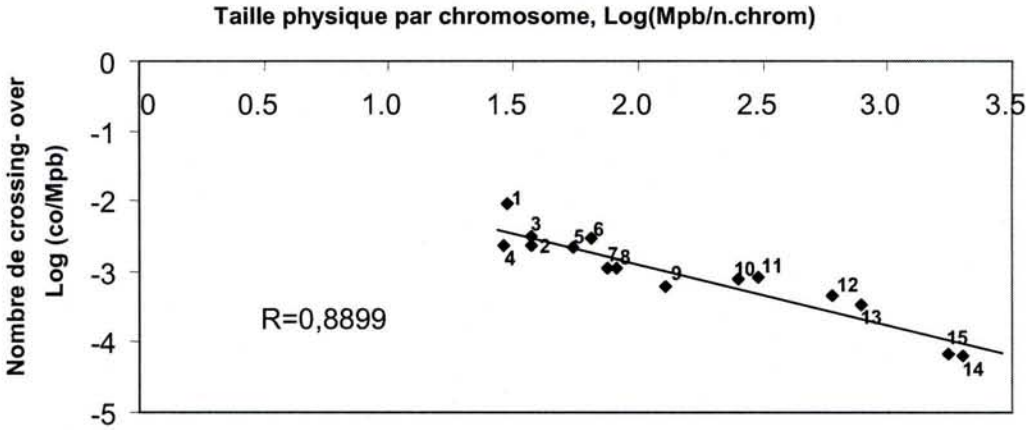


Tableau C17 : Relation entre la taille physique et la taille génétique chez 15 espèces de plantes.
Pour retrouver les références citées, se référer à l'annexe I.

	Espèce	Taille physique du génome haploïde (Mb)	Taille génétique (cM) (MAPMAKER)	Nombre de chromosomes (n)	Taille moyenne des chromosomes (cM)	Rapport taille physique / génétique (Mb/cM)
1	*Arabidopsis thaliana*	150[4]	675[50]	5	135	0,22
2	*Prunus persica*	300[4]	712[20]	8	90	0,42
3	*Oryza sativa*	450[4]	1490[2]	12	125	0,3
4	*Populus deltoides*	550[4]	2300[16]	19	121	0,23
5	*Eucalyptus grandis*	600[4]	1370[64]	11	125	0,43
6	*Brassica rapa*	650[4]	1850[56]	10	185	0,35
7	*Quercus robur*	900[4]	1200[6]	12	100	0,75
8	*Lycopersicon esculentum*	980[4]	1280[57]	12	107	0,76
9	*Solanum tuberosum*	1540[4]	1120[28]	12	93	1,37
10	*Zea mays*	2500[4]	1860[13]	10	186	1,34
11	*Lactuca sativa*	2730[4]	1950[27]	9	217	1,4
12	*Triticum tauschii*	4200[5]	1330[38]	7	190	3,15
13	*Hordeum vulgare*	5500[5]	1250[29]	7	178	4,4
14	*Pinus taeda*	21000[4]	1700[19]	12	141	12,35
15	*Pinus pinaster*	25500	1850[18]	12	154	13,78

thèmes ont été abordés : tout d'abord la comparaison des tailles génétiques et physiques du génome du pin maritime avec celles d'autres espèces de plantes, puis l'étude de la répartition de différents types de marqueurs sur le génome d'une autre espèce de conifère l'épicéa commun (*Picea abies* Karst.).

C.2.3.1 La relation entre tailles génétique et physique

La taille physique du génome d'une espèce peut être mesurée par la cytométrie de flux (figure C12). Brièvement, le principe de cette méthode est le suivant : une suspension monocellulaire est traitée par un colorant fluorescent qui reflète le contenu en ADN, puis les cellules sont passées au travers d'un rayon laser qui excite le colorant. La fluorescence émise est ensuite analysée par une cellule photoélectrique, puis comparée avec un standard afin d'estimer la quantité d'ADN.

La taille physique du génome du pin maritime a été calculée suivant cette méthode dont le protocole et les résultats obtenus sont présentés à l'annexe I. Cette taille a été estimée à 51,49 pg d'ADN par cellule diploïde, ce qui correspond à un génome haploïde d'une longueur approximative de 25 700 Mb (1 pg = $9{,}65.10^8$ pb, Arumuganathan et Earle 1991). De nombreux auteurs ont également montré que le génome des conifères était très important comparé à d'autres espèces de plantes ou d'animaux (voir partie A.1.4). Wakayima et al. (1993) ont en particulier observé une corrélation entre la taille du génome et la réponse à des conditions environnementales (stress hydrique) pour certaines espèces de conifères et ont émis l'hypothèse que l'augmentation de la quantité d'ADN par cellule augmentait la taille des trachéides, ce qui permettait d'augmenter la conductivité de celles-ci.

Il est intéressant de comparer la taille physique du génome du pin maritime avec celle d'autres espèces de plantes. Le tableau C17 donne les tailles physiques et génétiques du génome de 15 espèces de plantes, en allant du petit génome d'*Arabidopsis thaliana* (150 Mb) aux mégagénomes de *Pinus taeda* (21 000 Mb) et de *Pinus pinaster* (25 700 Mb). Le rapport taille physique / taille génétique a été calculé pour ces espèces. Pour l'ensemble des espèces considérées, le nombre de crossing-over par chromosome est à peu près constant. La figure C13 montre une corrélation forte et négative entre le nombre de crossing-over (calculé grâce à la taille moyenne d'un groupe de liaison en cM pour chaque espèce) et la taille physique par centiMorgan. Ceci montre que les espèces qui ont des petits chromosomes recombinent plus par unité de taille physique. Ceci a des conséquences sur les études d'association en populations naturelles et suggèrent que les conifères, pour un même nombre de générations,

pourraient présenter des déséquilibres de liaison sur de plus grandes distances que les autres espèces. Néanmoins ces résultats sont à considérer avec prudence car ils supposent que toutes les régions du génome recombinent de la même manière, ce qui n'est pas le cas (Heslop-Harrison 2000, Fu et al. 2002). De plus, il faut tenir compte du régime de reproduction des espèces pour pouvoir conclure sur la relation entre le nombre de recombinaisons et le déséquilibre de liaison. En effet, les gymnospermes sont des organismes exclusivement allogames et les observations faites sur quelques gènes (voie de biosynthèse des lignines) montrent que les déséquilibres de liaison ne sont conservés qu'à de très courtes distances chez le pin maritime (Garnier-Géré, communication personnelle). Des résultats similaires montrent des fenêtres réduites de déséquilibre de liaison chez le maïs et chez l'Homme par rapport aux fenêtres plus larges observées chez Arabidopsis (Rafalski et Morgante 2004).

Le calcul du rapport taille physique / taille génétique permet aussi d'accéder au nombre moyen de bases nucléotidiques par unité de distance génétique. Ainsi 1 cM correspond en moyenne à 0,22 Mb pour *A. thaliana* alors que pour *P. pinaster* 1 cM correspond à 13,78 Mb ! Ces résultats ont des conséquences importantes pour les stratégies qui seront mises en place dans le cadre du programme de génomique chez les pins. Une telle distance physique ne permet pas d'envisager d'utiliser le clonage positionnel comme méthode d'identification des gènes sous-jacents aux QTL (ex : riz, Song et al. 1995 ; tomate, Frary et al. 2000). Cet inconvénient impose la mise en place d'une approche adaptée pour trouver des polymorphismes qui pourront servir de diagnostic pour faire de la sélection assistée par marqueurs. L'approche la mieux adaptée est l'approche gène candidat (Pfliegler et al. 2001) qui consiste à développer des marqueurs diagnostics basés directement sur le gène (voire le polymorphisme) responsable de la variation observée au niveau du caractère quantitatif. C'est cette approche qui est utilisée chez le pin maritime, comme je l'ai décrit dans la partie A.

C.2.3.2 Répartition de différents marqueurs moléculaires sur le génome de l'épicéa commun (*Picea abies* Karst.)

C.2.3.2.1 Avant-propos

Les résultats présentés dans cette partie s'éloignent du sujet de cette thèse étant donné qu'ils concernent la construction d'une carte génétique de l'épicéa commun. De plus ils ne sont pas directement présentés dans les objectifs ou dans les méthodes de ma thèse. Néanmoins cette partie est tout de même présentée dans ce manuscrit car une partie de ce travail a été réalisé

lors de ma thèse, comme par exemple la cartographie d'EST ou l'analyse en ségrégation de l'ensemble des marqueurs et la construction de la carte. Par ailleurs, les résultats obtenus apportent plus que la simple construction d'une carte génétique car non seulement cette carte a pu être utilisée pour comparer les cartes d'espèces de pin avec celle d'une espèce d'un autre genre de la famille des Pinaceae (ce qui est en relation directe avec le sujet de thèse), mais en plus, les types de marqueurs ayant servi à construire cette carte étant très divers, des informations originales ont pu être tirées de l'étude de leur répartition sur la carte génétique. En faisant l'hypothèse de la conservation de la synténie entre les génomes des pins et des épicéas (Troggio et al. 2004), ces informations peuvent en effet servir à mieux comprendre la structure du génome du pin maritime.

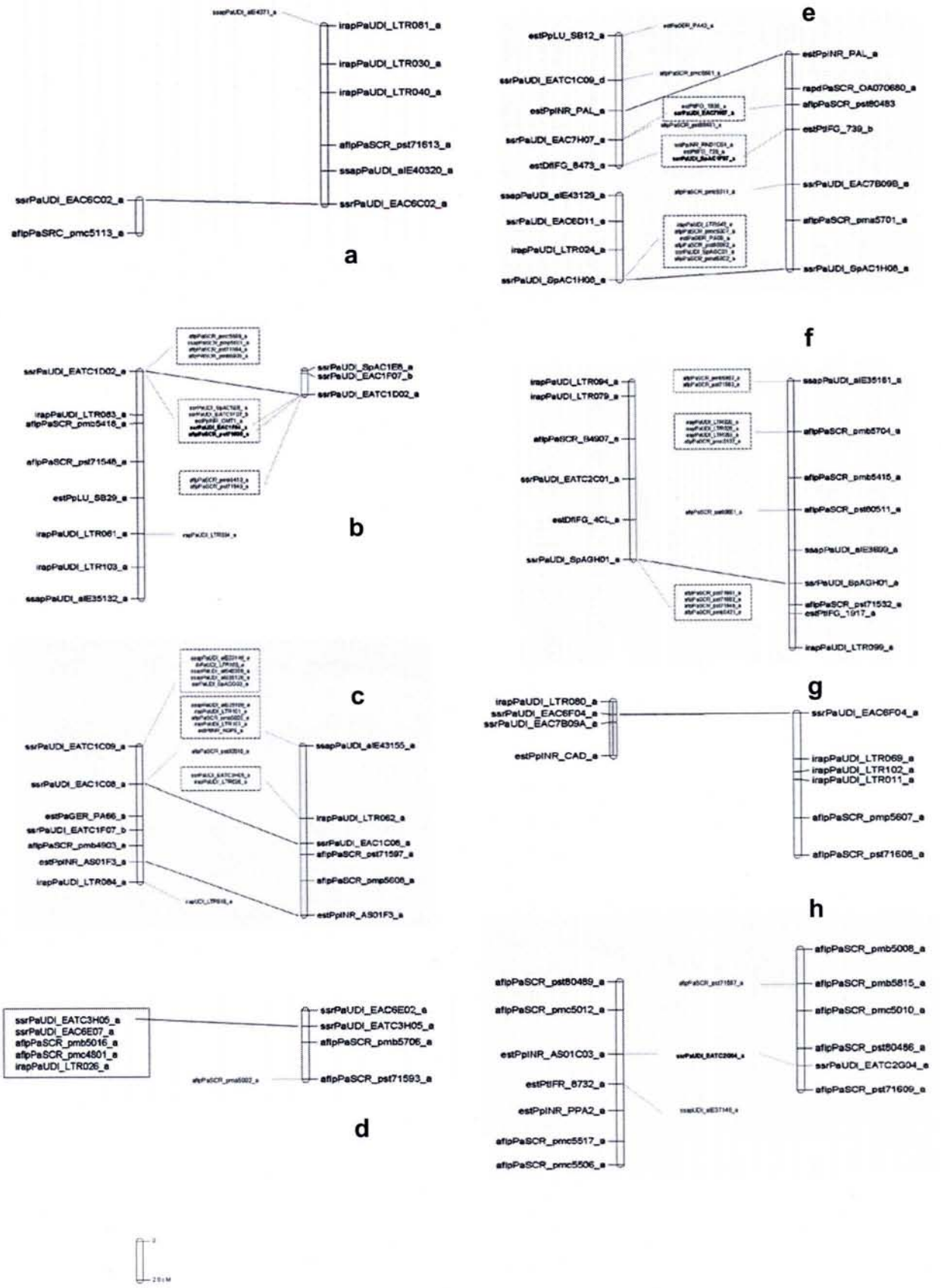
Les détails sur les protocoles utilisés pour le développement des différents types de marqueurs et sur la construction de la carte se trouvent à l'annexe VI.

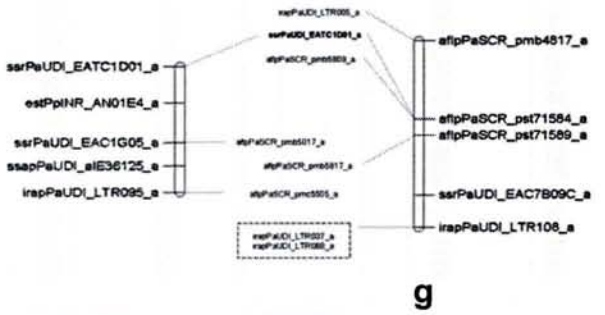
C.2.3.2.2 Divers types de marqueurs moléculaires

Etant donné qu'aucune carte saturée de *Picea abies* n'avait pu être obtenue jusqu'à présent[25] (Binelli et Bucci 1994, Bucci et al. 1997, Paglia et al. 1998, Troggio et al. 2001), les marqueurs moléculaires utilisés pour construire cette carte ont été choisis de manière à couvrir la plus grande diversité possible de locus en terme de nombre de copies, de fonction et de distribution dans le génome. Par exemple des marqueurs S-SAP (*Sequence-Specific Amplified Polymorphism*, Waugh et al. 1997) ont été employés. Le principe de cette méthode est assez similaire à celui de la méthode AFLP, à ceci près que les S-SAP utilisent des amorces flanquant des rétrotransposons, ce qui fait qu'ils sont associés aux régions hautement répétées du génome. Dans ce cas précis les régions LTR (*Long Terminal Repeat*) de rétroéléments des familles *Alisei* d'épicéa et *Copia* de pin ont été utilisées. En ce qui concerne les marqueurs AFLP, l'enzyme *Pst*I, qui est sensible à la méthylation de l'ADN et qui a donc tendance à être associée aux régions faiblement répétées, a été utilisée. Des marqueurs IRAP (*Inter-Retrotransposon Amplified Polymorphism*, Kalendar et al. 1999) ont également été utilisés. Cette technique consiste à amplifier les régions se trouvant entre des rétroéléments en utilisant des paires d'amorces situées dans les régions LTR (encadré C2). Enfin, des microsatellites et des EST ont été utilisés, eux-mêmes associés aux régions peu répétées du génome.

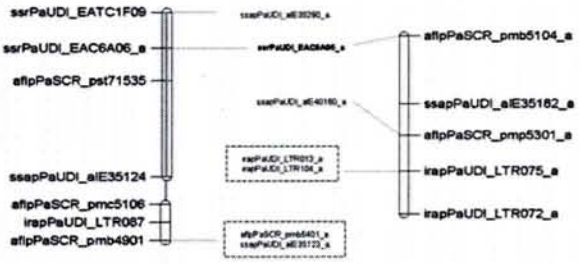
[25] Une carte génétique saturée de l'épicéa vient d'être publiée par l'équipe de génétique de J-M. Favre à l'Université de Nancy (Achéré et al. 2004).

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

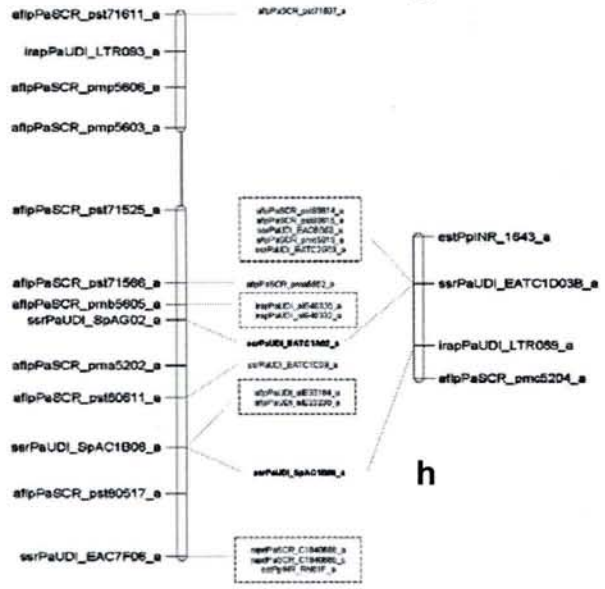




g

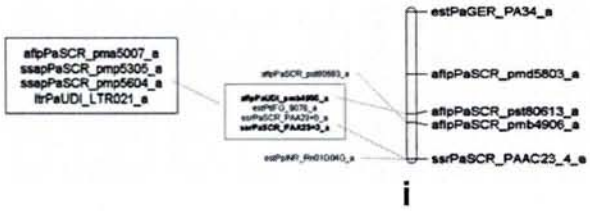


k

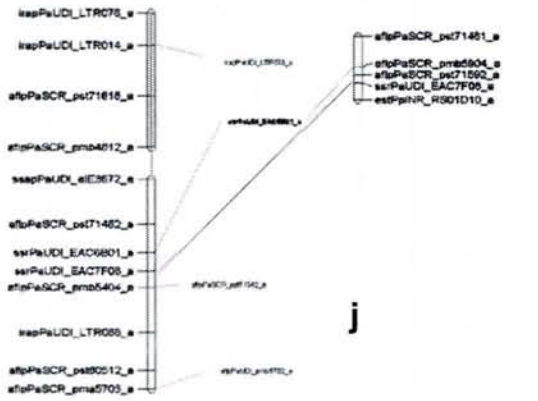


h

Figure C14 : carte génétique de l'épicéa commun (annexe VI).



i



j

C.2.3.2.3 Une répartition non-aléatoire de différents types de marqueurs moléculaires

Une carte génétique (mâle, femelle) de l'épicéa commun comprenant 13 groupes de liaison a été obtenue (figure C14). La description de cette carte est détaillée à l'annexe VI.

Les différents types de marqueurs ont été divisés en cinq classes : EST, SSR, AFLP, IRAP, S-SAP. La première observation qui a été faite est que toutes les classes de marqueurs ne contribuent pas de la même manière à la carte génétique : la proportion observée de marqueurs liés à la carte est plus importante que celle espérée pour les EST, alors que le cas inverse se produit pour les S-SAP et les IRAP (test G et test χ^2 significatifs). Ce résultat va dans le sens de l'hypothèse que les recombinaisons ne se produisent pas de manière homogène le long du génome et qu'elles ont tendance à se produire dans les régions riches en gènes, comme cela a été démontré chez le maïs (Fu et al. 2002). Ces résultats ont une implication sur le déséquilibre de liaison : on peut s'attendre à ce que ce dernier ne soit pas conservé sur de grandes distances dans les régions riches en gènes.

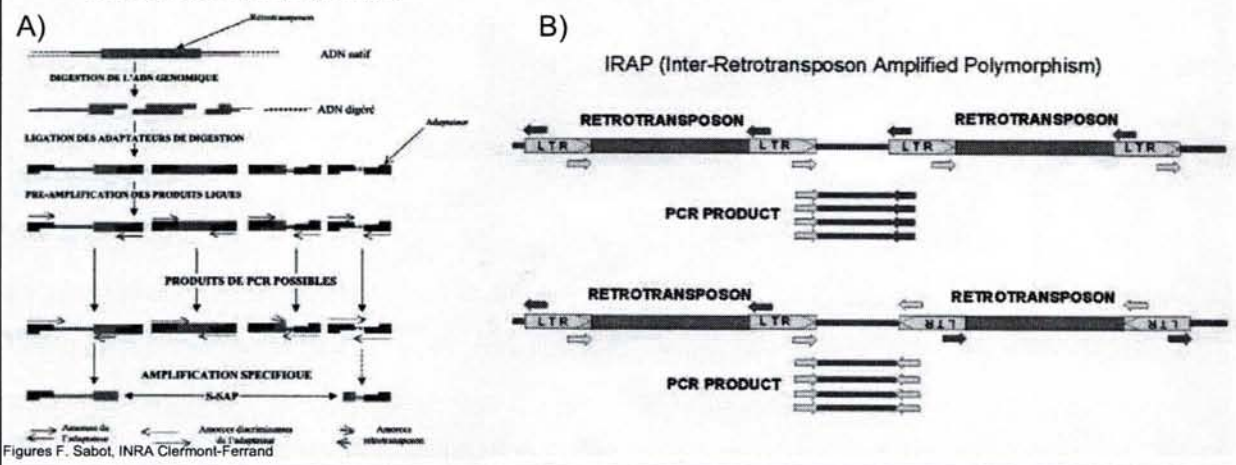
L'hypothèse de répartition non-aléatoire des types de marqueurs sur les groupes de liaison a été vérifiée par autocorrélation spatiale. Cette méthode statistique, développée à l'origine pour des études de géographie (Cliff et Ord 1973), permet de répondre à la question suivante : la similitude entre deux objets (locus) est-elle une fonction décroissante de leur distance respective (distance génétique) ? En d'autres termes : est-ce que les différents types de marqueurs cartographiés ont tendance à se regrouper, et si oui lesquels ? Il y a autocorrélation positive si des objets sont préférentiellement proches les uns des autres par rapport à l'hypothèse de distribution aléatoire de ces objets dans l'espace. Une autocorrélation spatiale positive sera observée dans le cas où des marqueurs ont tendance à se regrouper. A l'inverse, une autocorrélation spatiale négative traduira que ces objets ont tendance à « se repousser ».

L'ordonnancement des marqueurs obtenu par la fonction *first order* de *MAPMAKER* a servi de données sources pour l'analyse des autocorrélations spatiales. L'ordre des marqueurs dans ce cas n'est pas optimal pour les marqueurs fortement liés, mais en l'occurrence l'ordre importe peu. Ce qui est important c'est la distance entre les marqueurs. Au total, 15 (5 pour les 5 classes de marqueurs et 4x5/2 pour les couples de marqueurs différents) autocorrélogrammes permettant de comparer les cinq classes de marqueurs ont été réalisés. Neuf de ces 15 graphiques montraient des valeurs significativement différentes à une distribution aléatoire des marqueurs (figure C15).

Encadré C2 : Principe des techniques S-SAP et IRAP.

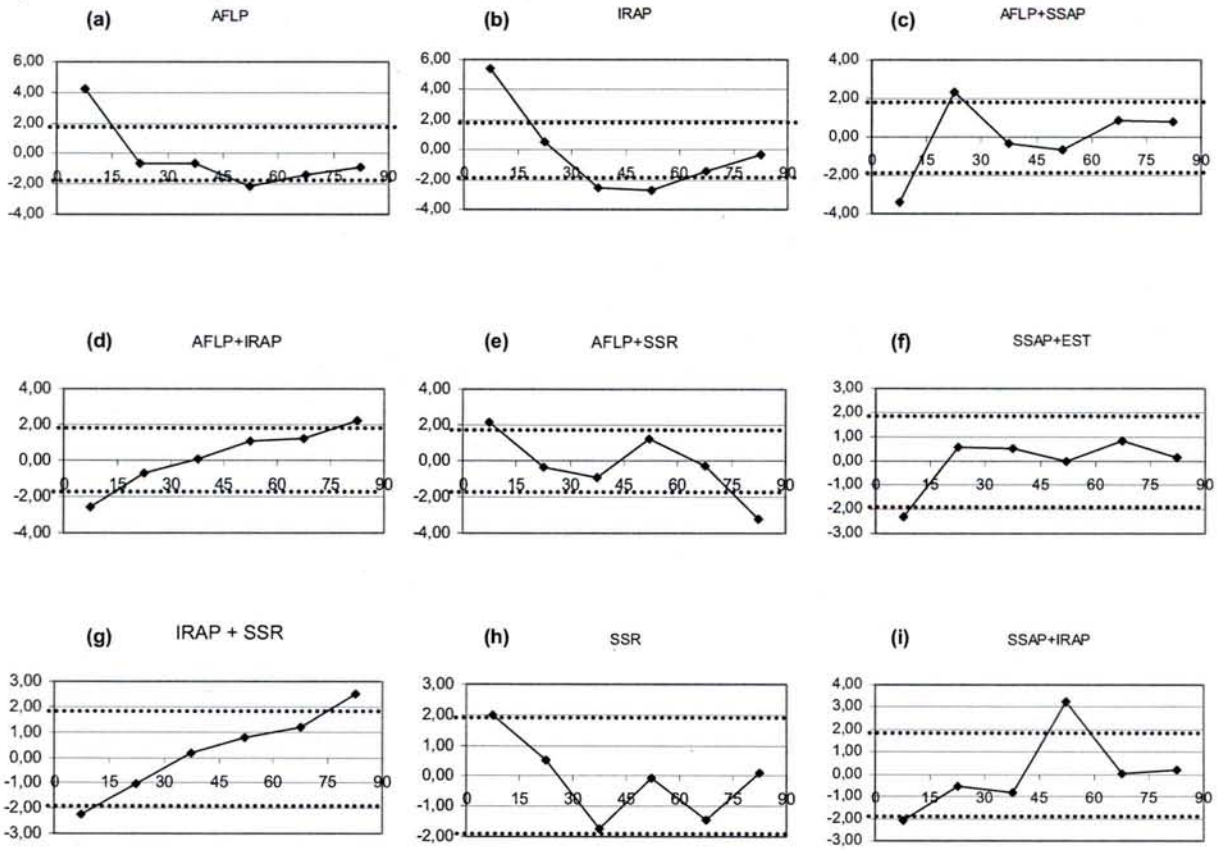
A) S-SAP : *Single-Sequence Amplified Polymorphism*. Cette technique est dérivée de la technique AFLP, à part qu'elle utilise des amorces spécifiques d'un adaptateur et de rétrotransposons.

B) IRAP : *Inter-Retrotransposon Amplified Polumorphism*. Cette technique est basée sur une amplification entre deux rétrotransposons.



Figures F. Sabot, INRA Clermont-Ferrand

Figure C15 : Graphiques montrant des autocorrélations significatives pour six couples de type de marqueurs et deux types de marqueurs seuls. Le graphique obtenu pour les SSR seuls est aussi présenté. Pour chaque graphique les ordonnées représentent l'écart type des autocorrélations et les abscisses les classes de distance en cM. Les limites des valeurs significatives sont présentées en pointillé.



Les différents graphiques présentés fournissent les informations suivantes : tout d'abord, il semble que les AFLP (a), les IRAP (b) et les SSR (h) ont tendance à former des clusters (autocorrélation positive à faible distance) et que les AFLP ont tendance à ne pas se trouver loin les uns des autres (autocorrélation négative à grande distance). Les AFLP semblent ne pas former de cluster avec les rétro-transposons (c,d), alors qu'ils peuvent se trouver proches de microsatellites (e). Les IRAP (g) semblent se séparer des microsatellites et les S-SAP (f) semblent se séparer des EST. Tous ces résultats montrent que les régions répétées du génome, représentées par les IRAP et les S-SAP qui sont liés à des rétrotransposons, sont séparées des régions peu répétées où se trouve les microsatellites et les gènes. Cette approche montre donc que les différents types de marqueurs ne sont pas répartis de manière aléatoire le long du génome. L'hypothèse peut donc être faite que des « continents » de séquences peu répétées et riches en microsatellites et/ou en gènes seraient isolés au milieu « d'océans » de séquences hautement répétées. De la même manière, Feuillet et Keller (2002) ont montré que le génome des céréales contenait des régions riches en gènes et des régions plus pauvres en gènes, où celles-ci sont séparées par de grandes distances pouvant contenir des régions répétées.

Ces observations ont des conséquences importantes sur les programmes de génomique chez les gymnospermes comme l'épicéa et le pin maritime. Les progrès techniques considérables réalisés au cours des dernières années en matière de biologie moléculaire ont permis d'envisager le séquençage complet de nombreuses espèces, dont un des exemples est celui du génome du peuplier, quatre fois plus grand que celui d'Arabidopsis, mais néanmoins séquencé en moins de deux ans (Brunner et al. 2004). D'un point de vue technique, le séquençage du génome d'un conifère ne semblerait pas devoir poser plus de problèmes que pour le peuplier et ce projet ne serait juste qu'une affaire de temps. Cependant, les résultats obtenus sur la répartition des différentes classes de séquences sur le génome de l'épicéa dévoilent les difficultés qui pourraient être rencontrées dans le cas d'un projet de séquençage du génome d'un conifère. En effet, l'isolement des séquences peu complexes par de grandes régions très répétées compliquerait non seulement les étapes de clonage mais surtout les étapes d'assemblage des séquences obtenues sur la carte physique. Les mêmes difficultés ont été rencontrées dans le cas du maïs, une espèce à relativement gros génome (2 500 Mb), où le séquençage et l'assemblage des clones BAC contenant des éléments répétés se sont révélés impossibles (San Miguel et Bennetzen 1998). Ces inconvénients peuvent être éventuellement contournés par l'utilisation de méthodes permettant d'accéder aux fractions peu répétées du génome. L'approche qui a été exclusivement employée chez les conifères jusqu'ici est le développement de banques d'ADNc qui donnent accès non seulement à la partie peu répétée

du génome mais permet aussi de découvrir les gènes impliqués dans des processus biologiques ciblés (par exemple la formation du bois ou la réponse aux contraintes biotiques ou abiotiques). De nombreux EST d'espèces de conifères ont ainsi été générés (par exemple chez le pin taeda, Kirst et al. 2003). Néanmoins, les banques d'ADNc ont l'inconvénient de ne présenter qu'une image partielle et restreinte des gènes d'une espèce car les banques développées dépendent du type de tissu utilisé et du stade de développement. D'autres méthodes basées sur la méthylation de l'ADN peuvent être utilisées : l'ADN répété est souvent plus méthylé que l'ADN non-répété. Certains auteurs ont donc employé des enzymes sensibles à la méthylation pour construire des banques génomiques enrichies en séquences simple copie. Néanmoins, cette approche n'est pas exhaustive non plus car la méthylation peut être assez hétérogène selon l'organisme, le stade de développement ou les régions du génome considérés (revue par Heslop-Harrison 2000). C'est pour cela que d'autres auteurs ont décidé d'utiliser la technique des courbes de *Cot*, basée sur la cinétique de réassociation de l'ADN, pour séparer les fractions plus ou moins répétées du génome. Cette méthode a permis de réaliser la construction d'une banque BAC de manière très efficace pour le sorgho (Petersen et al. 2002), où les fractions hautement répétées, faiblement répétées et simple copie ont pu être séparées avec succès. L'application de cette technique chez les conifères pourrait permettre de caractériser aussi bien les fractions répétées que les régions simple copie, voire même de séquencer la totalité de ces dernières.

Figure C17 : Profils de gels d'acrylamide et DGGE présentant les deux locus paralogues amplifiés par la paire d'amorces PtIFG_9036.
a) PtIFG_9036 – Gel de polyacrylamide non-dénaturant – 4 individus *P. pinaster* + 4 individus *P. taeda.*
b) PtIFG_9036 - DGGE - 16 descendants plein-frères.

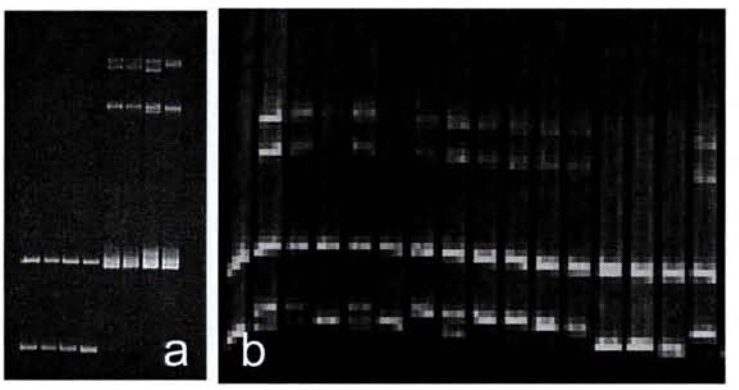


Tableau C18 : Identités de séquences entre *P. taeda* et *P. pinaster* au niveau de 7 locus EST.

Locus	Identité de séquence
PtIFG_8436	235 / 256 (91,7%)
PtIFG_9036_a	274 / 280 (97,8%)
PtIFG_9036_b	227 / 280 (81,1%)
PtIFG_9151	272 / 291 (93,4%)
PtNCS_4CH1	481 / 489 (98,3%)
PtTX_p14A9	507 / 518 (97,8%)
PtNCS_22B8	225 / 228 (98,6%)

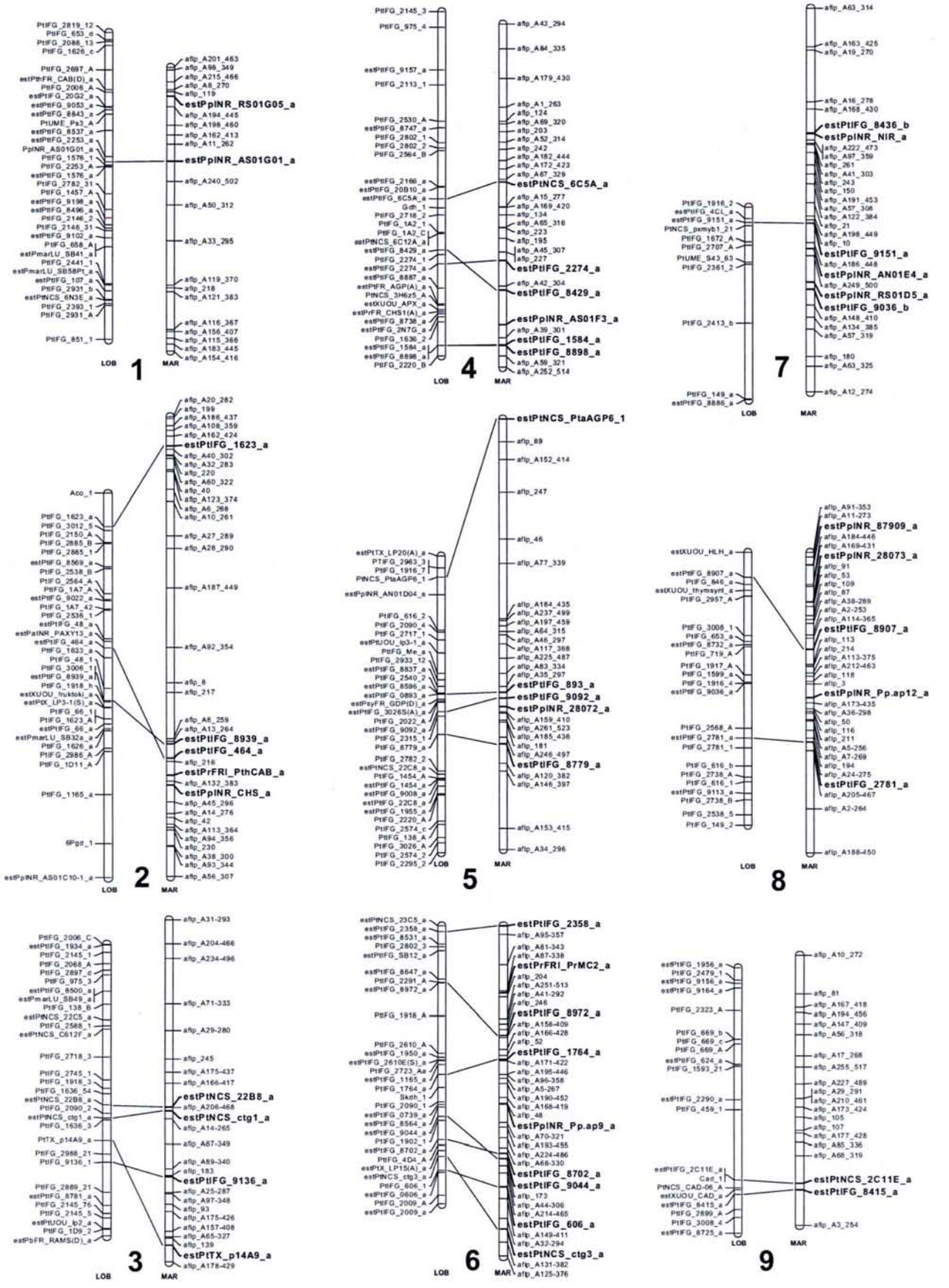
C.3 Cartographie comparée chez les Pinaceae

La comparaison des cartes génétiques de *P. taeda* et *P. pinaster*, ainsi que de l'identification de QTL et de gènes candidats communs entre les deux espèces sont présentées à l'annexe III.

C.3.1 Comparaison des cartes génétiques de Pinus taeda et de Pinus pinaster

Trente-deux locus EST ont pu être cartographiés à la fois sur la carte génétique du croisement G2 de *P. pinaster* et sur la carte de référence de *P. taeda*. Un alignement de ces deux cartes est présenté à la figure C16. Dix groupes de liaisons sur 12 ont pu être alignés. Seuls les groupes 11 et 12 n'ont pu être alignés. La composition et l'ordre des marqueurs semblent être conservés entre les deux espèces.

Deux locus sont localisés sur des groupes non homologues (PtIFG_8436_b et PtIFG_9036_b). Une première hypothèse serait l'existence d'une probable translocation des fragments chromosomiques contenant ces locus. La seconde hypothèse, qui parait la plus probable, est que des locus paralogues ont été amplifiés par PCR et cartographiés. Une seule bande a été obtenue pour la paire d'amorces PtIFG_8436, alors que deux bandes distinctes ont été obtenues sur gels d'agarose et de polyacrylamide pour la paire d'amorces PtIFG_9036. Les produits d'amplification des ces deux locus ont été clonés et séquencés, ainsi que la bande additionnelle monomorphe de PtIFG_9036. Les séquences obtenues chez *P. pinaster* ont été comparées avec celles de *P. taeda* sur lesquelles les amorces PCR ont ensuite été dessinées. Les séquences de 4 autres fragments amplifiés dont les positions ne posent pas de problème ont été également réalisées. Le tableau C18 montre les identités de séquences de ces locus. On remarque tout d'abord que pour le locus PtIFG_8436_b, l'homologie de séquence est de 91,8%, valeur inférieure à celle qui a été observée pour d'autres gènes orthologues de positions identiques sur les deux cartes génétiques (PtIFG_9151, PtNCS_C4H-1, PtTX_p14A9, PtNCS_22B8). Ceci indique que la paire d'amorces PtIFG_8436 a pu amplifier un locus paralogue chez le pin maritime et il n'est donc pas étonnant de retrouver ce locus sur un autre groupe de liaison. Pour la paire d'amorces PtIFG_9036 (figure C17), l'identité de séquence du locus PtIFG_9036_a (non informatif chez le pin maritime) est plus élevée (97,8%) que celle du locus PtIFG_9036_b (81,1%) cartographié sur un groupe non homologue. Ces données montrent donc également qu'un locus paralogue a été cartographié.


1
PtIFG_2819_12
PtIFG_653_d
PtIFG_2086_13
PtIFG_1626_c
PtIFG_2697_A
estPtrFR_CAB(D)_a
PtIFG_2006_A
estPtIFG_2032_a
estPtIFG_9053_a
estPtIFG_8843_a
PtUME_Pa3_A
estPtIFG_8537_a
estPtIFG_2253_a
PpINR_AS01G01_a
PtIFG_1576_1
PtIFG_2253_A
estPtIFG_1576_a
PtIFG_2782_31
PtIFG_1457_A
estPtIFG_9198_a
estPtIFG_8496_a
PtIFG_2146_2
PtIFG_2146_31
estPtIFG_9102_a
PtIFG_658_A
estPmarLU_SB41_a
PtIFG_2441_1
estPmarLU_SB58Pt_a
estPtIFG_107_a
PtIFG_2931_b
estPtNCS_6N3E_a
PtIFG_2393_1
PtIFG_2931_A
PtIFG_851_1
LOB
aflp_A201_463
aflp_A98_349
aflp_A215_466
aflp_A8_270
aflp_119
estPpINR_RS01G05_a
aflp_A194_445
aflp_A198_460
aflp_A162_413
aflp_A11_262
estPpINR_AS01G01_a
aflp_A240_502
aflp_A50_312
aflp_A33_295
aflp_A119_370
aflp_218
aflp_A121_383
aflp_A116_367
aflp_A156_407
aflp_A115_366
aflp_A183_445
aflp_A154_416
MAR
2
Aco_1
PtIFG_1623_a
PtIFG_3012_5
PtIFG_2150_A
PtIFG_2885_B
PtIFG_2885_1
estPtIFG_8569_a
PtIFG_2538_B
PtIFG_2564_A
PtIFG_1A7_A
estPtIFG_9022_a
PtIFG_1A7_42
PtIFG_2536_1
estPtIFG_48_a
estPtINR_PAXY13_a
estPtIFG_464_a
PtIFG_1633_a
PtIFG_48_1
PtIFG_3006_1
estPtIFG_8939_a
PtIFG_1918_h
estXUOU_fruktoki_a
estPtX_LP3-1(S)_a
PtIFG_66_1
PtIFG_1623_A
estPtIFG_66_a
estPmarLU_SB32a_a
PtIFG_1626_a
PtIFG_2985_A
PtIFG_1D11_A
PtIFG_1165_a
6Pgd_1
estPpINR_AS01C10-1_a
LOB
aflp_A20_282
aflp_199
aflp_A186_437
aflp_A108_359
aflp_A162_424
estPtIFG_1623_a
aflp_A40_302
aflp_A32_283
aflp_220
aflp_A60_322
aflp_40
aflp_A123_374
aflp_A6_268
aflp_A10_261
aflp_A27_289
aflp_A28_290
aflp_A187_449
aflp_A92_354
aflp_8
aflp_217
aflp_A8_259
aflp_A13_264
estPtIFG_8939_a
estPtIFG_464_a
aflp_216
estPrFRI_PthCAB_a
aflp_A132_383
estPpINR_CHS_a
aflp_A45_296
aflp_A14_276
aflp_42
aflp_A113_364
aflp_A94_356
aflp_230
aflp_A38_300
aflp_A93_344
aflp_A56_307
MAR
3
PtIFG_2006_C
estPtIFG_1934_a
PtIFG_2145_1
PtIFG_2068_A
PtIFG_2897_c
PtIFG_975_3
estPtIFG_8500_a
estPmarLU_SB49_a
PtIFG_138_B
estPtNCS_22C5_a
PtIFG_2588_1
estPtNCS_C612F_a
PtIFG_2718_3
PtIFG_2745_1
PtIFG_1918_3
PtIFG_1636_54
estPtNCS_22B8_a
PtIFG_2090_2
estPtNCS_ctg1_a
PtIFG_1636_3
PtTX_p14A9_a
PtIFG_2966_21
PtIFG_9136_1
PtIFG_2889_21
estPtIFG_8781_a
PtIFG_2145_76
PtIFG_2145_5
estPtUOU_lp2_a
PtIFG_1D9_2
estPbFR_RAMS(D)_a
LOB
aflp_A31-293
aflp_A204-466
aflp_A234-496
aflp_A71-333
aflp_A29-280
aflp_245
aflp_A175-437
aflp_A166-417
estPtNCS_22B8_a
aflp_A206-468
estPtNCS_ctg1_a
aflp_A14-265
aflp_A87-349
aflp_A89-340
aflp_183
estPtIFG_9136_a
aflp_A25-287
aflp_A97-348
aflp_93
aflp_A175-426
aflp_A157-408
aflp_A65-327
aflp_139
estPtTX_p14A9_a
aflp_A178-429
MAR
4
PtIFG_2145_3
PtIFG_975_4
estPtIFG_9157_a
PtIFG_2113_1
PtIFG_2530_A
estPtIFG_8747_a
PtIFG_2802_1
PtIFG_2802_2
PtIFG_2564_B
estPtIFG_2166_a
estPtIFG_20810_a
estPtIFG_6C5A_a
Gdh_1
PtIFG_2718_2
PtIFG_1A2_1
PtIFG_1A2_C
estPtNCS_6C12A_a
estPtIFG_8429_a
PtIFG_2274_1
estPtIFG_2274_a
estPtIFG_8887_a
estPtFR_AGP(A)_a
PtNCS_3H6z5_A
estXUOU_APX_a
estPrFR_CHS1(A)_a
estPtIFG_8738_a
estPtIFG_2N7G_a
PtIFG_1636_2
estPtIFG_1584_a
estPtIFG_8898_a
PtIFG_2220_B
LOB
aflp_A43_294
aflp_A84_335
aflp_A179_430
aflp_A1_263
aflp_124
aflp_A69_320
aflp_203
aflp_A52_314
aflp_242
aflp_A182_444
aflp_A172_423
aflp_A67_329
estPtNCS_6C5A_a
aflp_A15_277
aflp_A169_420
aflp_134
aflp_A65_316
aflp_223
aflp_195
aflp_A45_307
aflp_227
estPtIFG_2274_a
aflp_A42_304
estPtIFG_8429_a
estPpINR_AS01F3_a
aflp_A39_301
estPtIFG_1584_a
estPtIFG_8898_a
aflp_A59_321
aflp_A252_514
MAR
5
estPtX_LP20(A)_a
PtIFG_2963_3
PtIFG_1916_7
PtNCS_PtaAGP6_1
estPpINR_AN01D04_a
PtIFG_616_2
PtIFG_2090_4
PtIFG_2717_1
estPtUOU_lp3-1_a
PtIFG_Me_a
PtIFG_2933_12
estPtIFG_8837_a
PtIFG_2540_2
estPtIFG_8586_a
estPtIFG_0893_a
estPsyFR_GDP(D)_a
estPtIFG_3026S(A)_a
PtIFG_2022_A
estPtIFG_9092_a
PtIFG_2315_1
PtIFG_8779_a
PtIFG_2782_2
estPtNCS_22C8_a
PtIFG_1454_A
estPtIFG_1454_a
estPtIFG_9008_a
estPtIFG_22C8_a
estPtIFG_1955_a
PtIFG_2220_A
PtIFG_2574_c
PtIFG_138_A
PtIFG_3026_A
PtIFG_2574_2
PtIFG_2295_2
LOB
estPtNCS_PtaAGP6_1
aflp_89
aflp_A152_414
aflp_247
aflp_46
aflp_A77_339
aflp_A184_435
aflp_A237_499
aflp_A197_459
aflp_A64_315
aflp_A46_297
aflp_A117_368
aflp_A225_487
aflp_A83_334
aflp_A35_297
estPtIFG_893_a
estPtIFG_9092_a
estPpINR_28072_a
aflp_A159_410
aflp_A261_523
aflp_A185_436
aflp_181
aflp_A246_497
estPtIFG_8779_a
aflp_A120_382
aflp_A146_397
aflp_A153_415
aflp_A34_296
MAR
6
estPtNCS_23C5_a
estPtIFG_2358_a
estPtIFG_8531_a
PtIFG_2802_3
estPtIFG_SB12_a
estPtIFG_8647_a
PtIFG_2291_A
estPtIFG_8972_a
PtIFG_1918_A
PtIFG_2610_A
estPtIFG_1950_a
estPtIFG_2610E(S)_a
PtIFG_2723_Aa
estPtIFG_1165_a
PtIFG_1764_a
Skdh_1
PtIFG_2090_1
estPtIFG_0739_a
estPtIFG_8564_a
estPtIFG_9044_a
PtIFG_1902_1
estPtIFG_8702_a
PtIFG_4D4_A
estPtX_LP15(A)_a
estPtNCS_ctg3_a
PtIFG_606_1
estPtIFG_0606_a
PtIFG_2009_A
estPtIFG_2009_a
LOB
estPtIFG_2358_a
aflp_A95-357
aflp_A81-343
aflp_A87-338
estPrFRI_PrMC2_a
aflp_204
aflp_A251-513
aflp_A41-292
aflp_246
estPtIFG_8972_a
aflp_A158-409
aflp_A166-428
aflp_52
estPtIFG_1764_a
aflp_A171-422
aflp_A195-446
aflp_A96-358
aflp_A5-267
aflp_A190-452
aflp_A168-419
aflp_48
estPpINR_Pp.ap9_a
aflp_A70-321
aflp_A193-455
aflp_A224-486
aflp_A68-330
estPtIFG_8702_a
estPtIFG_9044_a
aflp_173
aflp_A44-306
aflp_A214-465
estPtIFG_606_a
aflp_A149-411
aflp_A32-294
estPtNCS_ctg3_a
aflp_A131-382
aflp_A125-376
MAR
7
PtIFG_1916_2
estPtIFG_4CL_a
estPtIFG_9151_a
PtNCS_pxmyb1_21
PtIFG_1672_A
PtIFG_2707_A
PtUME_S43_63
PtIFG_2361_2
PtIFG_2413_b
PtIFG_149_a
estPtIFG_8886_a
LOB
aflp_A63_314
aflp_A163_425
aflp_A19_270
aflp_A16_278
aflp_A168_430
estPtIFG_8436_b
estPpINR_NIR_a
aflp_A222_473
aflp_A97_359
aflp_261
aflp_A41_303
aflp_243
aflp_150
aflp_A191_453
aflp_A57_306
aflp_A122_384
aflp_21
aflp_A198_449
aflp_10
estPtIFG_9151_a
aflp_A186_448
estPpINR_AN01E4_a
aflp_A249_500
estPpINR_RS01D5_a
estPtIFG_9036_b
aflp_A148_410
aflp_A134_385
aflp_A57_319
aflp_180
aflp_A63_325
aflp_A12_274
MAR
8
estXUOU_HLH_a
estPtIFG_8907_a
PtIFG_846_a
estXUOU_thymsynt_a
PtIFG_2957_A
PtIFG_3008_1
PtIFG_653_a
estPtIFG_8732_a
PtIFG_719_A
PtIFG_1917_A
PtIFG_1599_a
PtIFG_1916_4
estPtIFG_9036_a
PtIFG_2568_A
estPtIFG_2781_a
PtIFG_2781_1
PtIFG_616_b
PtIFG_2738_A
PtIFG_616_1
estPtIFG_9113_a
PtIFG_2738_B
PtIFG_2538_5
PtIFG_149_2
LOB
aflp_A91-353
aflp_A11-273
estPpINR_87909_a
aflp_A184-446
aflp_A169-431
estPpINR_28073_a
aflp_91
aflp_53
aflp_109
aflp_87
aflp_A38-289
aflp_A2-253
aflp_A114-365
estPtIFG_8907_a
aflp_113
aflp_214
aflp_A113-375
aflp_A212-463
aflp_118
aflp_3
estPpINR_Pp.ap12_a
aflp_A173-435
aflp_A36-298
aflp_50
aflp_116
aflp_211
aflp_A5-256
aflp_A7-269
aflp_194
aflp_A24-275
estPtIFG_2781_a
aflp_A205-467
aflp_A2-264
aflp_A188-450
MAR
9
estPtIFG_1956_a
PtIFG_2479_1
estPtIFG_9156_a
estPtIFG_9164_a
PtIFG_2323_A
PtIFG_669_b
PtIFG_669_c
PtIFG_669_A
estPtIFG_624_a
PtIFG_1593_21
estPtIFG_2290_a
PtIFG_459_1
estPtIFG_2C11E_a
Cad_1
PtNCS_CAD-06_A
estXUOU_CAD_a
estPtIFG_8415_a
PtIFG_2899_A
PtIFG_3008_4
estPtIFG_8725_a
LOB
aflp_A10_272
aflp_81
aflp_A167_418
aflp_A194_456
aflp_A147_409
aflp_A56_318
aflp_A17_268
aflp_A255_517
aflp_A227_489
aflp_A29_291
aflp_A210_461
aflp_A173_424
aflp_105
aflp_107
aflp_A177_428
aflp_A85_336
aflp_A68_319
estPtNCS_2C11E_a
estPtIFG_8415_a
aflp_A3_254
MAR

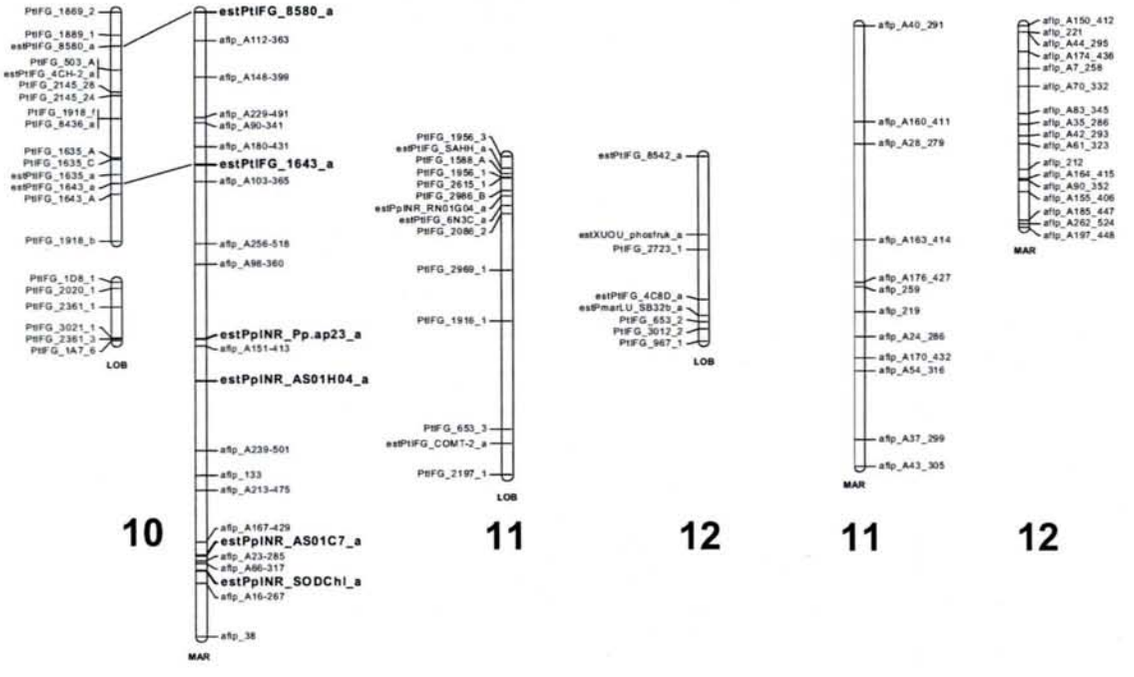
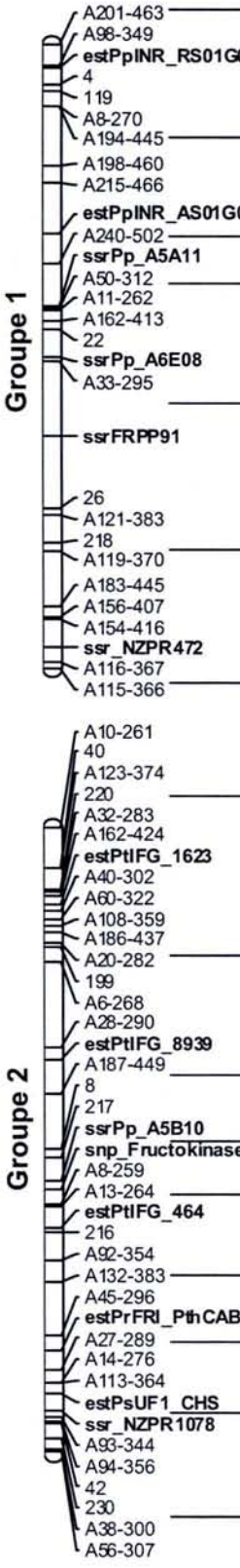
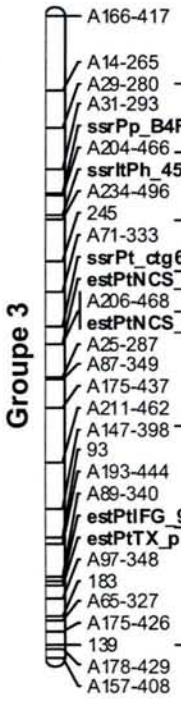


Figure C16 : Alignement entre les cartes génétiques de *Pinus taeda* et de *Pinus pinaster*.

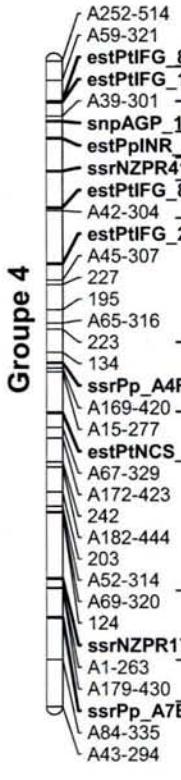
Figure C18 : Localisation des locus microsatellites (ssr_), EST (est_) et SNP (snp_) sur la carte du croisement G2 et informations sur la carte F2, la carte de *P. taeda* et les cartes d'autres espèces de conifères.



Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
x	x	*Pseudotsuga menziesii* : groupe 1
Groupe 4** (paralogue ?)	Groupe 1*	x
Groupe 9	x	x
Groupe 9	x	x
Groupe 9	x	x
x	x	*Pinus radiata* : groupe 10 (homologue au groupe 1 de *P. taeda*)
x	Groupe 2	*Pinus elliottii* : groupe 2
x	Groupe 2	*Pinus elliottii* : groupe 2
Groupe 7	x	x
x	x	*Pinus sylvestris* : groupe 2
x	Groupe 2	*Pinus sylvestris* : groupe 2
x	x	x
x	x	*Pinus sylvestris* : groupe 2 *Pinus radiata* : groupe 3
Groupe 7	x	*Pinus radiata* : groupe 3 (homologue au groupe 2 de *P. taeda*)



	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
Groupe 3	x	x	
Groupe 3	x	x	
Groupe 3	x	x	
x	Groupe 3	*Picea abies* : groupe 3	
x	Groupe 3	x	
x	Groupe 3	*Pinus sylvestris* : groupe 3	
x	Groupe 3	*Pinus sylvestris* : groupe 3	

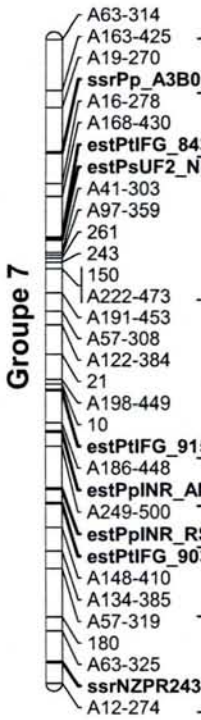


x	Groupe 4	*Pinus sylvestris* : groupe 4 *Pinus elliottii* : groupe 4	
x	Groupe 4		
x	x	x	
x	x	*Picea abies* : groupe 4	
Groupe 8	x	*Pinus radiata* : groupe 4	
x	Groupe 4	*Pinus radiata* : groupe 4	
x	Groupe 4	*Pinus sylvestris* : groupe 4 *Pinus elliottii* : groupe 4	
Non lié	x	x	
x	Groupe 4	x	
x	x	*Pinus radiata* : groupe 4 (homologue au groupe 4 de *P. taeda*)	
Groupe 8	x	x	

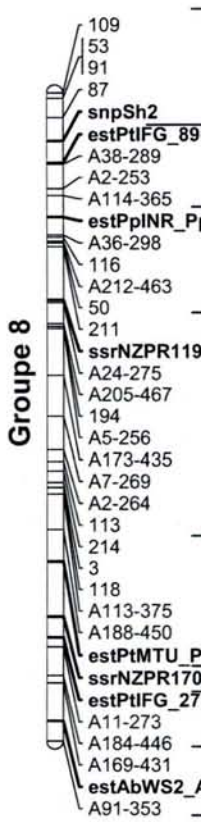
S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Groupe 5	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
A153-415 A34-296 A129-380 A30-281 **ssrPp_A3D04** A146-397	x	x	x
A120-382 **estPtIFG_8779** A246-497 181 A185-436	x	Groupe 5	*Pinus elliottii* : groupe 5
A159-410 A261-523 **estPtIFG_9092**	Groupe 2	Groupe 5	*Pinus radiata* : groupe 11 *Pinus sylvestris* : groupe 5 *Pinus elliottii* : groupe 5
estPtMTU_PtCW1	x	x	x
estPtIFG_893 A35-297	x	Groupe 5	*Pinus radiata* : groupe 11
ssrNZPR823 A83-334 A117-368 A46-297 A64-315 A225-487	Groupe 8 (paralogue !)	Groupe 5	*Pinus radiata* : groupe 11 (homologue au groupe 5 de *P. taeda*)
A184-435 A197-459 A237-499 **ssrRPtEST11** A77-339	Groupe 2	x	x
46 247 A152-414 89 **estPtNCS_PtaAGP6_1**	Non lié	Groupe 5	*Picea abies* : groupe 4 (paralogue !)

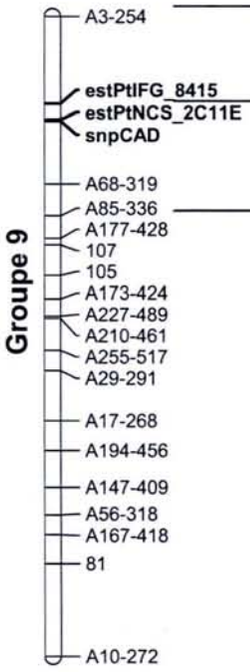
Groupe 6	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
estPtIFG_2358 A95-357 A81-343	x	Groupe 6	*Pinus elliottii* : groupe 6
A87-338 204 **estPrFRI_PrMC2** A41-292	x	x	x
A251-513 **estPtIFG_8972** A158-409 246 A171-422	x	Groupe 6	*Pinus sylvestris* : groupe 6 *Pinus elliottii* : groupe 6
A166-428 52 A195-446 **estPtIFG_1764**	x	Groupe 6	*Pinus sylvestris* : groupe 6 *Pinus elliottii* : groupe 6
snpPAL A96-358 A190-452 A168-419 A5-267 48	Groupe 5 (paralogue ?)	Groupe 6	*Pinus sylvestris* : groupe 6 *Picea abies* : groupe 6
A70-321 **estPpINR_Pp.ap9** A193-455 A224-486	x	x	x
173 A68-330 **estPtIFG_8647** (non lié sur la G2)	Groupe 1	Groupe 6	*Pinus elliottii* : groupe 6
estPtIFG_9044 A44-306 A214-465	x	Groupe 6	*Pinus sylvestris* : groupe 6 *Pinus elliottii* : groupe 6
estPtIFG_8702	x	Groupe 6	
ssrPp_cn524	Groupe 1	x	x
estPtIFG_606 A125-376 A149-411	x	Groupe 6	*Pinus elliottii* : groupe 6
snpCcoaOMT **estPtNCS_CCoAOMT** A32-294	x	Groupe 6	*Pseudotsuga menziesii* : groupe 6
A131-382 **snpCesA01**	x	x	x



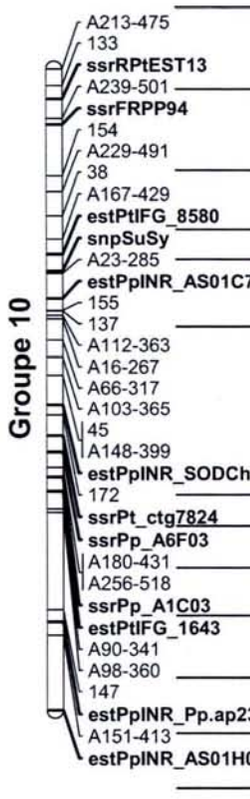
	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
ssrPp_A3B05b	x	x	x
estPtIFG_8436	x	Groupe 10 (Paralogue ?)	*Pinus sylvestris* : groupe 7
estPsUF2_NIR, A41-303, A97-359, 261, 243, 150, A222-473	x	x	*Pinus sylvestris* :groupe E1101a (non homologue)
A191-453, A57-308, A122-384, 21, A198-449, 10, estPtIFG_9151, A186-448	x	Groupe 7	*Pinus elliottii* : groupe 7 *Pinus sylvestris* : groupe 7
estPpINR_AN01E04	x	x	x
A249-500, estPpINR_RS01D05	x	x	x
estPtIFG_9036, A148-410, A134-385	x	Groupe 8 (Paralogue ?)	x
A57-319, 180, A63-325, ssrNZPR243, A12-274	Groupe 4	x	*Pinus radiata* : groupe 6 (homologue au groupe 3 de *P. taeda*, paralogue ?)



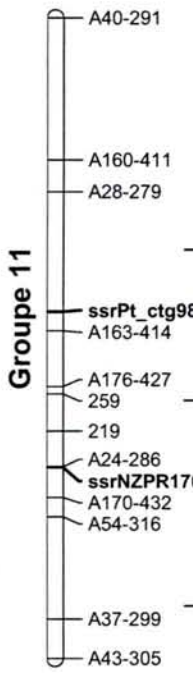
109, 53, 91, 87, snpSh2	x	x	x
estPtIFG_8907, A38-289, A2-253, A114-365	x	Groupe 8	*Pinus elliottii* : groupe 8
estPpINR_Pp.ap12, A36-298, 116, A212-463, 50	x	Groupe 8	x
211, ssrNZPR119, A24-275, A205-467, 194, A5-256, A173-435, A7-269, A2-264, 113	Groupe 10	x	*Pinus radiata* : groupe 15 (non-homologue)
214, 3, 118, A113-375, A188-450, estPtMTU_PtCW2	x	x	x
ssrNZPR1702a	Groupe 10	x	x
estPtIFG_2781, A11-273, A184-446	x	Groupe 8	
A169-431, estAbWS2_AG3.18, A91-353	x	x	x



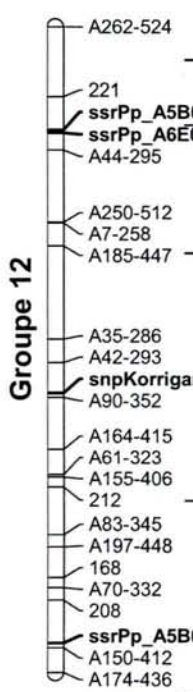
	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
A3-254 / estPtIFG_8415	x	Groupe 9	x
estPtNCS_2C11E / snpCAD / A68-319 / A85-336	x	Groupe 9	*Pinus sylvestris* : groupe 9 *Pseudotsuga menziesii* : groupe 9
A177-428 / 107 / 105 / A173-424 / A227-489 / A210-461 / A255-517 / A29-291 / A17-268 / A194-456 / A147-409 / A56-318 / A167-418 / 81 / A10-272			



A213-475 / 133 / ssrRPtEST13 / A239-501	x	x	x
ssrFRPP94 / 154 / A229-491	Groupe 5	x	x
38 / A167-429 / estPtIFG_8580	Groupe 5	Groupe 10	*Pinus elliottii* : groupe 10
snpSuSy / A23-285	x	x	x
estPpINR_AS01C7 / 155 / 137	Groupe 5	x	x
A112-363 / A16-267 / A66-317 / A103-365 / 45 / A148-399 / estPpINR_SODChl	Groupe 5	Non lié	*Pinus sylvestris* : groupe 10 (locus PsUPS2_PST13)
172 / ssrPt_ctg7824	x	x	x
ssrPp_A6F03 / A180-431	Groupe 5	x	
A256-518 / ssrPp_A1C03	x	x	x
estPtIFG_1643 / A90-341 / A98-360	x	Groupe 10	*Pinus elliottii* : groupe 10
147 / estPpINR_Pp.ap23 / A151-413	x	x	x
estPpINR_AS01H04	x	x	*Picea abies* : groupe 10



	Localisation sur la carte F2	Localisation sur la carte de *Pinus taeda*	Localisation sur les cartes d'autres espèces de conifères
ssrPt_ctg988 – A176-427	Groupe 6	x	x
219 – A54-316	Groupe 6	x	*Pinus radiata* : groupe 8 (homologue au groupe 10 de *P. taeda*, paralogue !)



221 – ssrPp_A5B01_a	x	x	x
ssrPp_A6E05 – A185-447	Groupe 12	x	x
A35-286 – snpKorrigan – A155-406	x	x	x
212 – ssrPp_A5B01_b – A174-436	x	x	x

* : les cartes génétiques de six espèces de conifères présentées dans cette figure ont une numérotation qui s'aligne sur la carte de référence de *Pinus taeda* (Brown et al. 2001, Komulainen et al. 2003, Krutovskii et al. 2004, Troggio et al. 2004, annexe III). Pour *Pinus radiata,* les groupes sont numérotés différemment mais leur homologie a été vérifiée avec ceux de *P. taeda* (P. Wilcox, communication personnelle).

** : pour la correspondance des cartes F2 et G2, se reporter à la figure XX.

Seules deux paires d'amorces sur les 32 cartographiées ont révélé des locus paralogues. Brown et al. (2001) ont montré que les marqueurs orthologues basés sur les EST de *Pinus taeda* étaient des marqueurs de choix pour la cartographie génétique comparée chez les conifères parce qu'ils sont transférables entre espèces et qu'ils fournissent peu d'amplifications non spécifiques à partir du moment où on choisit bien l'endroit où se trouvent les amorces PCR. En effet, les amorces PCR ont été placées dans ou près des régions 3'UTR de manière à être le plus spécifique d'un membre d'une famille multigénique (Temesgen et al. 2001). Les conifères possèdent de nombreuses familles multigéniques et les régions 3'UTR, variables entre les différents membres d'une famille de gènes, garantissent une meilleure spécificité des amorces et une probabilité plus grande d'amplifier un locus orthologue.

C.3.2 Synténie chez les conifères

La figure C18 présente l'ensemble des marqueurs orthologues localisés sur la carte du pedigree G2, et indique les positions dans la carte F2, la carte de *Pinus taeda* et chez d'autres conifères. Il est intéressant de noter que l'homologie des groupes de *P. taeda* et de *P. pinaster* est confirmée avec les autres espèces de la famille des Pinaceae. En effet, d'autres études portant sur la cartographie génétique comparée entre espèces de conifères ont été réalisées et peuvent être confrontées avec les résultats présentés ci-dessus dans le cas du pin maritime et du pin taeda. Le premier exemple d'alignement de deux cartes génétiques de conifères a été publié par Devey et al. (1999) entre *P. radiata* et *P. taeda*. Cette étude utilisait des marqueurs de type RFLP et microsatellites comme points d'ancrage entre les cartes. Les 12 groupes des deux espèces ont pu être alignés avec succès, néanmoins les marqueurs de type RFLP se sont révélés problématiques du fait de la présence de familles multigéniques chez les conifères (Kinlaw et Neale 1997). Elles peuvent en effet provoquer des profils multibandes souvent difficiles à interpréter. En ce qui concerne les microsatellites, j'ai montré précédemment les difficultés qui peuvent être rencontrées lorsque l'on souhaite les utiliser comme marqueurs orthologues. De ce fait, une autre approche consistant à utiliser des marqueurs basés sur l'amplification PCR d'EST a été initiée (Harry et al. 1998, Temesgen et al. 2001). Cette méthode a tout d'abord permis de montrer que les marqueurs basés sur les EST peuvent être hautement transférables entre espèces de la famille des Pinaceae, et a ainsi permis d'aligner les cartes génétiques de *P. taeda* et *P. elliottii* (Brown et al. 2001). D'autres comparaisons

Figure C19a : Conservation de la synténie chez les Pinaceae (groupe 6) : alignement des cartes génétiques de 5 espèces de pin.

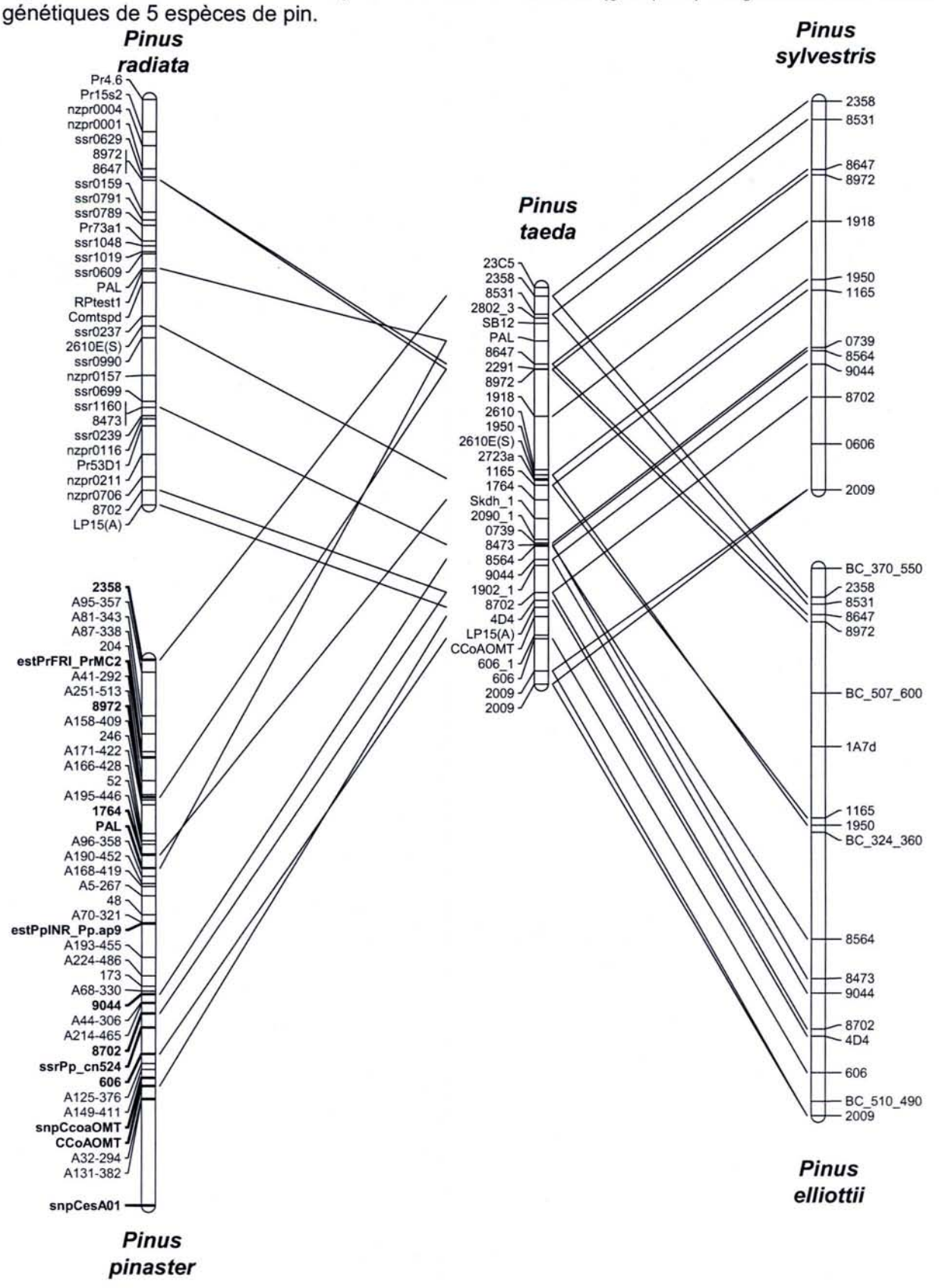
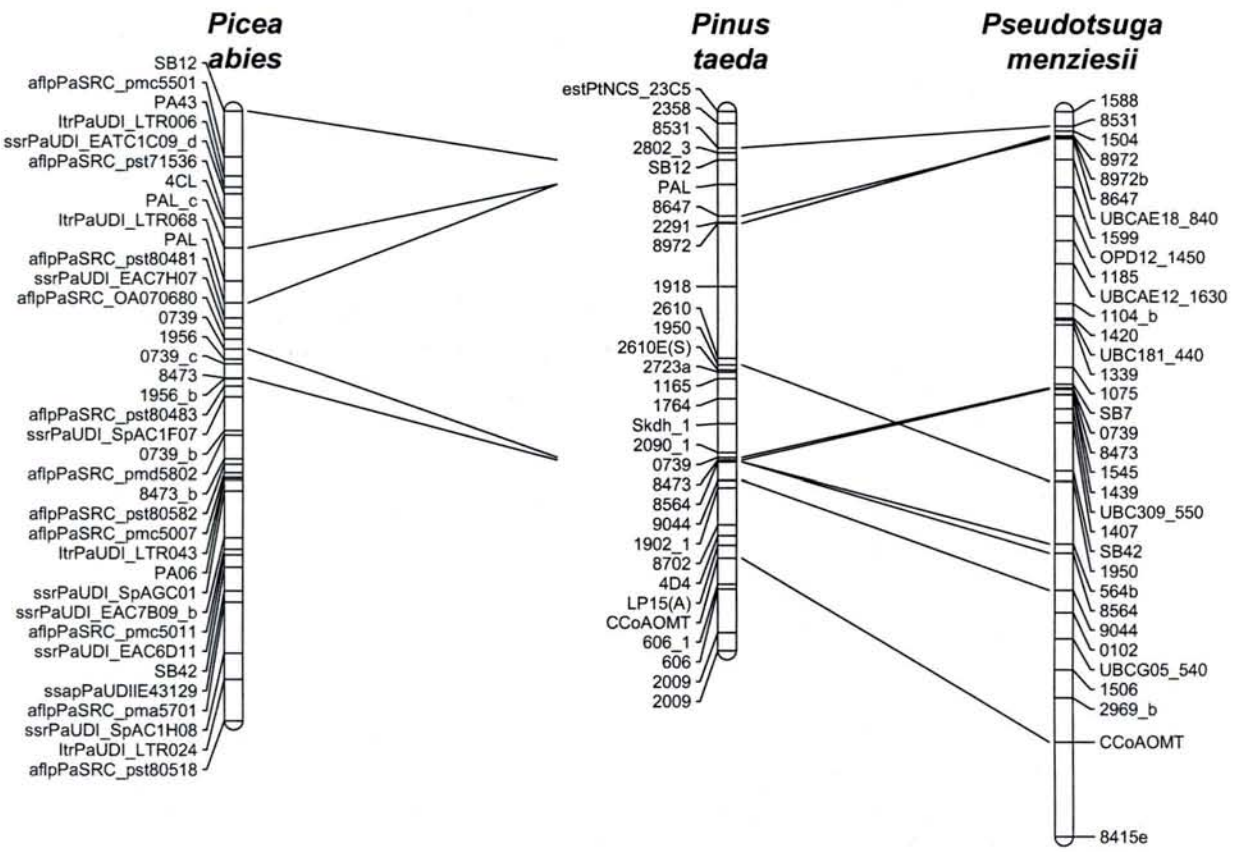


Figure C19b : Conservation de la synténie chez les Pinaceae (groupe 6) : alignement de la carte du pin taeda avec les cartes du sapin de Douglas et de l'épicéa commun.



d'espèces de pins deux à deux ont alors été réalisées, toujours avec *P. taeda* comme espèce de référence (comme par exemple avec le pin sylvestre, Komulainen et al. 2003). Une autre carte que celle de Devey et al. (1999) a été construite pour *P. radiata* (Wilcox et al. 2001), et est en cours de comparaison avec la carte de *P. taeda*. Enfin la comparaison a été étendue à des espèces de la famille des Pinaceae autres que des pins : le sapin de Douglas et l'épicéa commun (Krutovskii et al. 2004, Troggio et al. 2004). Des études de comparaison sont actuellement en cours entre la famille des Pinaceae et des espèces d'une autre famille de gymnospermes (genre *Cryptomeria*, D. Neale, Y. Tsumura, communication personnelle).

Dans chaque cas une bonne conservation de la synténie a été observée entre les différentes cartes alignées avec *P. taeda*. Ainsi, et malgré le fait que ces comparaisons sont souvent de faible résolution (rarement plus de 10 marqueurs communs par groupe), on peut penser que la structure du génome des Pinaceae n'a pas du connaître de grande modification en terme de réarrangements chromosomiques depuis la divergence de ces espèces. Cette hypothèse est d'ailleurs confirmée par la conservation du nombre de chromosomes entre ces espèces (excepté pour le sapin de Douglas, *Pseudotsuga menziesii*, où n = 13).

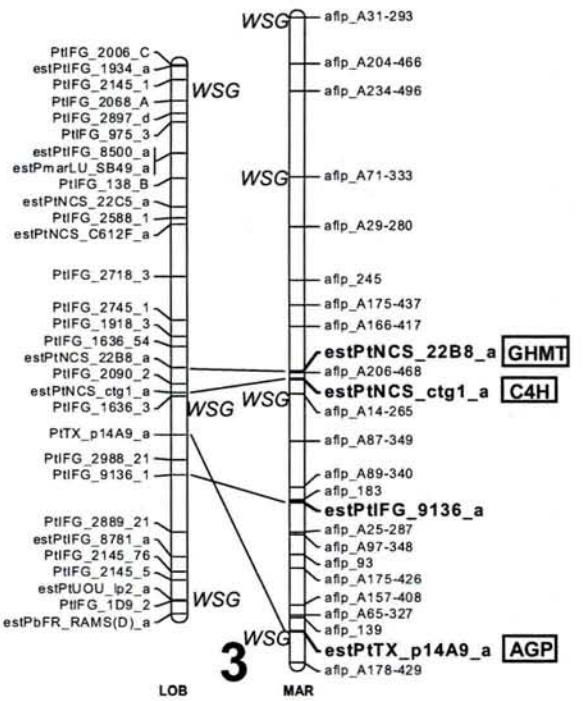
La figure C19 présente l'alignement des cartes génétiques des espèces impliquées dans le projet CCGP au niveau du groupe 6, au sein du genre *Pinus* (figure C19a) et au-delà chez les Pinaceae (figure C19b). Le groupe 6 a été choisi car il possède le plus de marqueurs communs entre toutes les espèces considérées. La même comparaison pourrait être réalisée pour les autres chromosomes mais avec moins de marqueurs. L'ordre des marqueurs dans ce groupe semble relativement bien conservé entre les espèces. Les quelques différences d'ordre observées sont vraisemblablement dues à l'imprécision de l'analyse de liaison génétique plutôt qu'à de réels réarrangements chromosomiques.

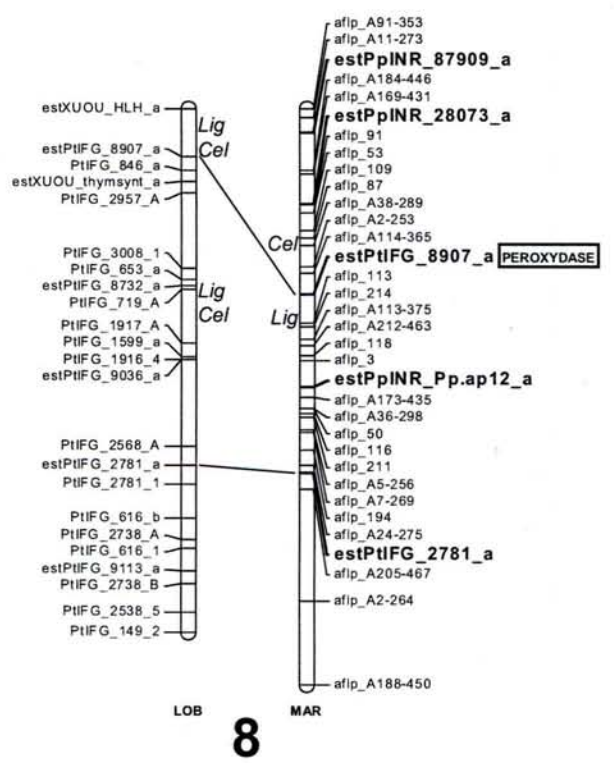
Ces résultats indiquent que le génome des Pinaceae a une structure très conservée, même au-delà d'un genre. Néanmoins, ces résultats sont à prendre avec prudence pour plusieurs raisons. Tout d'abord le nombre de marqueurs communs est faible et ces marqueurs sont espacés le long des chromosomes. Comme je l'ai décrit dans la partie C.2.2, 1 cM correspond à une taille physique énorme chez les pins. De ce fait, des réarrangements de type délétion ou insertion de longs fragments chromosomiques peuvent avoir eu lieu sans avoir été détecté. Le développement d'autres marqueurs orthologues est donc nécessaire afin de confirmer les hypothèses de conservation de la macro-synténie. Ensuite, le lot de marqueurs utilisés a été choisi de manière à ne considérer que les locus simple bande, afin de réduire les chances de se trouver en présence de familles multigéniques (Temesgen et al. 2001). Les informations obtenues ne concernent donc que les parties les moins complexes de la fraction codante du

Figure C20 : Comparaison de la position de QTL pour la qualité du bois sur les groupes homologues 3 et 8 de *P. taeda* et *P. pinaster*.

QTL communs : *WSG* : *Wood Specific Gravity* (Sewell et al. 2000, Pot 2004); *Lig* : taux de lignine; *Cel* : taux de cellulose (Sewell et al. 2002, Pot 2004).

Gènes candidats : GHMT : *Glycine hydroxymethyltransferase*; C4H : *trans cinnamate 4 hydroxylase*; AGP : *Arabinogalactane-like protein*.





génome, ce qui est très restrictif. Une étude de famille multigénique entière, avec pour objectif la localisation de tous les membres de la famille permettrait de mieux comprendre les mécanismes de duplication de gènes qui peuvent avoir eu lieu à l'intérieur ou entre espèces. Les seuls résultats obtenus dans le cadre de cette étude sont la localisation de différents membres de la famille des *Arabino-galactanes* (AGP). Les trois gènes cartographiés chez le pin maritime se trouvent sur trois groupes différents et colocalisent avec les mêmes gènes sur les mêmes chromosomes de la carte de *P. taeda*. Ce résultat préliminaire indique que la duplication de ces gènes a dû avoir lieu avant la divergence des deux espèces. Les autres exemples sont ceux des locus paralogues cartographiés chez le pin maritime et discutés plus haut. Une famille multigénique (*Adh*) a été étudiée chez *Pinus banksiana* (Perry et Furnier 1996) et sept membres différents ont été détectés sur deux groupes.

C.3.3 Application de la cartographie génétique comparée entre espèces de pins

Au-delà de la compréhension de la structure et de l'évolution du génome des conifères, l'alignement des cartes génétiques des conifères peut avoir d'autres applications. Il est en effet possible de trouver des QTL conservés chez des espèces différentes (Paterson et al. 1991). Au niveau intraspécifique, la comparaison des cartes F2 et G2 a permis d'identifier une région chromosomique où des QTL co-localisent avec un gène candidat impliqué dans les processus de réponse aux conditions environnementales et de qualité du bois (gène *Korrigan*). De la même manière, la co-localisation de QTL entre espèces de pins pourrait permettre de vérifier leur position et de fournir éventuellement des gènes candidats positionnels.

Des QTL contrôlant des caractères similaires liés à la qualité du bois ont été localisés sur les cartes alignées de *P. taeda* et de *P. pinaster*. C'est le cas par exemple des QTL liés à la densité (Sewell et al. 2000, Pot 2004) et à la composition chimique du bois (Sewell et al. 2002, Pot 2004). L'alignement des cartes génétiques des deux espèces de pins permet de mettre en évidence la co-localisation de QTL sur les groupes 3 et 8 (figure C20).

Pour le groupe 3, les QTL communs détectés correspondent à des caractères de densité du bois. Des gènes impliqués dans des processus physiologiques liés à la formation du bois ont également été localisés chez les deux espèces au même endroit. Le gène représenté par l'EST PtNCS_C4H-1 a été annoté fonctionnellement comme étant une *trans-cinnamate 4 hydroxylase* (C4H). L'implication de ces enzymes dans la voie de biosynthèse des lignines a été démontrée (Sewalt et al. 1997). L'EST PtTX_p14A9 correspond à un membre de la

famille des *Arabino-galactanes* qui sont impliquées dans la formation du xylème secondaire (Plomion et al. 2001). Enfin, l'EST PtNCS_22B8 est une *glycine hydroxymethyltransferase*. Les *methyltransferases* sont une des familles de gènes les plus exprimées chez les ligneux lors de la formation du bois (Plomion et al. 2001).

Pour ce qui est du groupe 8, les QTL communs identifiés concernent les propriétés chimiques du bois, taux de lignine et d'α-cellulose. Un gène codant pour une Péroxydase co-localise également avec ces QTL chez les deux espèces (estPtIFG_8907). Les péroxydases étaient impliquées dans la polymérisation des sous-unités de lignines (Higushi 1996).

La co-localisation de ces deux clusters de QTL est un résultat intéressant : il indique que les mêmes gènes sont impliqués dans des caractères similaires pour des espèces ayant divergé il y a environ 100 Ma. De plus, cette co-localisation représente une vérification de la position des QTL. Une validation plus stricte de la position de ces QTL, ainsi que de leur association avec les gènes candidats positionnels identifiés, doit être réalisée, par le biais d'une étude d'association sur des individus en population naturelle. Ces études sont en cours à la fois aux Etats-Unis pour *P. taeda* et à l'INRA pour *P. pinaster* (E. Eveno, thèse en cours).

Il n'est pas exclu que ces co-localisations soient simplement dues au hasard. Des modèles statistiques ont été développés pour vérifier la position de QTL communs dans des fonds génétiques différents (Lin et al. 1995). Cette méthode n'a pas pu être appliquée dans le cas présent en raison du nombre restreint de locus communs entre les deux cartes. Cependant, cette co-localisation de QTL avec des gènes candidats fonctionnels (ces gènes appartiennent à des voies de biosynthèses impliquées dans la formation du bois) et expressionnels (ils sont différentiellement exprimés dans des types de bois différents, Le Provost 2003) chez deux espèces différentes peut avoir une signification biologique réelle. Le développement de marqueurs orthologues additionnels s'avère nécessaire pour avoir accès à une comparaison plus fine de ces régions. Les marqueurs SNP devraient permettre d'apporter ce niveau de précision (disponibles en grand nombre, faciles à mettre en œuvre, présents dans des gènes potentiellement orthologues).

CONCLUSION ET PERSPECTIVES

Tableau C1 : Evaluation des différents types de marqueurs pour la construction d'une carte génétique du pin maritime, pour la cartographie génétique comparée chez les conifères, en tant que gènes candidats et pour la détection de QTL.

	Développement	Cartographie génétique	Cartographie comparée	Détection QTL	Gène candidat
AFLP	++	++	--	+/-	--
SSR (transfert)	+	+/-	+/-	+	-
SSR-ADNc	++	+/-	+	+	+
SSR (banque enrichie)	--	+	+/-	++	--
EST (SSCP, DGGE)	+	+	++	+	++
SNP	+/-	+	++	+/-	++

Légende

++	Très intéressant
+	Intéressant
+/-	Bilan mitigé
--	Peu intéressant

Partie D : Conclusion et perspectives

D.1 Des marqueurs moléculaires aux propriétés variables pour des applications diverses (cartographie génétique, cartographie comparée, approche gène candidat)

Le tableau C1 a permis de dresser un bilan sur l'utilité de différents types de marqueurs moléculaires mis en œuvre au cours de cette thèse, pour des applications en cartographie génétique et cartographie comparée. Chaque type de marqueur présente ses avantages et ses inconvénients. D'un point de vue général, le développement de **marqueurs *in silico*** (reposant sur des séquences nucléotidiques disponibles dans les bases de données d'EST essentiellement) est une approche très intéressante car elle fournit des marqueurs polymorphes sans passer par de lourds protocoles techniques. C'est le cas des marqueurs microsatellites présents dans les séquences codantes et des SNP. Il n'est donc pas étonnant de voir ces approches de plus en plus utilisées chez de nombreuses espèces de plantes ou d'animaux, à en juger par le nombre de publications parues récemment qui utilisent cette stratégie (tableau B3).

Parmi les différents marqueurs qui passent par un travail de laboratoire pour leur développement, la technique **AFLP** (de par sa relative simplicité et le nombre important de locus polymorphes qu'elle produit) a été démontré comme permettant de saturer des cartes génétiques de manière très efficace, sans aucune connaissance *a priori* sur le génome considéré. De ce fait, et malgré des inconvénients majeurs (pas transférables entre espèces, dominants et bialléliques), les AFLP restent les marqueurs de choix pour la construction de cartes génétiques, surtout pour des espèces étudiées par une équipe voire un chercheur au niveau international (comme c'est le cas pour de nombreuses essences forestières) ! Une fois la carte génétique établi, on pourra alors envisager d'y apporter des marqueurs plus informatifs pour des applications en détection de QTL ou pour la cartographie génétique comparée. Notons que ces marqueurs sont aussi très utilisés en génétique des populations, pour décrire le niveau et la structure de la diversité des espèces. Mariette et al. (2002) ont récemment montré que leur nombre compensait leur faible niveau d'information.

La cartographie génétique comparée nécessite de mettre en œuvre des marqueurs transférables entre espèces phylogénétiquement proches. L'intérêt de marqueurs microsatellites non codant (**SSR**) et de polymorphismes géniques (**EST-P**) a été évalué pour atteindre cet objectif. Devant la faible transférablité inter-spécifique des SSR chez les pins, nous avons - dans le

cadre d'un consortium international de cartographie comparée des conifères – mis au point des marqueurs orthologues basés sur des gènes en utilisant les techniques SSCP et DGGE pour révéler du polymorphisme nucléotidique. Malgré les inconvénients de ces marqueurs (1/ la nature codante de ces marqueurs peut réduire les chances d'observer du polymorphisme, et 2/ le nombre important de familles multigéniques chez les conifères peut conduire à la détection de locus paralogues), la mutualisation des moyens de plusieurs équipes de recherches a permis d'apporter des résultats intéressant au niveau de la macrosynténie des conifères, même si le nombre de marqueurs communs reste encore limité. L'augmentation du nombre d'EST chez les conifères couplé à la capacité de rechercher rapidement des **SNP** *in silico*, permettra d'augmenter, dans un futur proche, le nombre de sites polymorphes utilisables pour la cartographie comparée des conifères. Loin de dénigrer les marqueurs SSR développé dans le cadre de cette thèse, ils ont quand même servi à aligner les deux cartes génétiques disponibles chez le pin maritime.

Comme cela est évoqué à plusieurs reprises dans cette thèse, la caractérisation biologique des QTL chez les conifères passerait vraisemblablement par une **approche gène candidat**. Ceci tient tout autant à l'impossibilité de mener des approches de clonage positionnel sur le « méga-génome » des conifères, qu'à la taille efficace des populations (Ne = 1000 environ) qui induit des déséquilibres de liaison qu'entre sites polymorphes physiquement proches. Dans ce cadre, les marqueurs géniques fournissent non seulement des points d'ancrages pour la cartographie comparée mais aussi des gènes candidats sur lesquels pourront se greffer des études d'association[26] en population naturelle. Chez l'Homme par exemple, des études d'association pour comprendre les bases génétiques de maladies complexes ont été réalisées grâce à des SNP répartis sur l'ensemble du génome (on parle alors de « *Genome scan* », Lindblad-Toh et al 2000). Il est difficile d'imaginer d'employer des études exhaustives de type *genome scan* chez les conifères du fait de l'absence de carte physique, de la faible résolution des cartes génétiques et du faible nombre de SNP disponibles. Néanmoins l'approche gène candidat, qui est utilisée dans le cadre du programme de génomique du pin maritime, pourrait permettre de choisir un lot représentatif de marqueurs (par rapport à la fonction ciblée) pour réaliser de telles études. Comme je l'ai évoqué auparavant, les données d'études d'expressions de gènes, les connaissances acquises sur les mécanismes physiologiques et les informations contenues dans les bases de données de séquences peuvent permettre de proposer les gènes qui *a priori* impliqués dans un caractère d'intérêt. Dans le cas

[26] Etude d'association : analyse génétique de deux populations (malades et témoins par exemple) permettant d'établir une association entre un gène et un phénotype.

de l'étude des caractères liés à la qualité du bois et aux réponses aux contraintes environnementales chez le pin maritime, des EST ont été produits à partir de racines et de xylème en différenciation, à des stades de développement et dans des conditions environnementales différentes. Des études d'expression sont en cours de réalisation (thèses de C. Dubos 2001, G. Le Provost, 2003 et thèses en cours de J. Paiva, P. Chaumeil) et des gènes candidats « expressionnels » pourront être proposés à l'issu de ces études. Ces gènes pourront alors servir de base à la fois pour des études d'association et de cartographie génétique comparée.

D.2 Des applications futures en génétique des populations

La disponibilité de SNP détectés dans les bases d'EST est une ressource importante pour le développement de marqueurs. Un projet européen (projet TreeSNiPs) ayant pour but l'étude de la diversité nucléotidique d'espèces forestières pour des gènes liés à des caractères d'adaptation aux contraintes environnementales (stress hydrique, phénologie du débourrement) est actuellement en cours et utilise les ressources développées dans cette thèse : SNP mais aussi SSR pour décrire à la fois la diversité adaptative et neutre.

Les locus microsatellites que nous avons mis au point sont actuellement utilisés pour du diagnostic génétique chez le pin maritime, allant de l'identification des grandes régions de provenances, à l'identification des variétés améliorées.

Les trois locus microsatellites décrits à l'annexe II ont été utilisés pour décrire la diversité génétique du pin maritime à l'aide de 47 populations naturelles (Derory et al. 2002). Les résultats obtenus confirment les observations faites sur la structure géographique du pin maritime en trois grands écotypes (figure A13), mais ne permettent pas, du fait du faible nombre de locus utilisés, de définir une stratégie précise de conservation des ressources génétiques de l'espèce. La disponibilité des nouveaux marqueurs microsatellites développés au cours de cette thèse permettra de décrire la diversité génétique manière plus complète.

Le programme d'amélioration génétique du pin maritime utilise l'hybridation entre les provenances corses et landaises pour cumuler les avantages des premiers (rectitude du fût) et des seconds (vigueur et résistance au froid). Des variétés hybrides ont été produites mais leur mise sur le marché fait actuellement l'objet de discussion : l'introduction de matériel non indigène de pin maritime Portugais dans les années 1950 ayant déjà conduit à des catastrophes écologiques et économiques en Aquitaine, suite aux grands froids de l'hiver 1981. Dans l'étude de Derory et al. (2002), le locus FRPP91 a été décrit comme pouvant différencier les

individus landais, des individus corses et des hybrides « corse x landes ». Un tel test diagnostic devrait permettre de rassurer la profession en permettant la traçabilité des variétés hybrides qui seront mises sur le marché.

La plupart des variétés améliorées actuellement disponibles sont issues du croisement entre géniteurs landais dans des vergers à graines de pollinisation libre. Un projet financé par le Ministère de l'Agriculture est en cours et vise à estimer le taux de pollution dans les vergers à graines (quantité de pollen extérieur au verger réduisant les gains génétiques annoncés), et au-delà, de mettre en place un système de traçabilité des lots de graines commercialisés en France (garantie sur l'origine génétique). Des marqueurs microsatellites situés sur des chromosomes différents permettront de donner un génotype unique aux géniteurs G0 (appartenant à la première génération d'amélioration). Un échantillonnage des arbres G1 (issus de croisements contrôlés entre G0) plantés dans les vergers et des graines G2 (issus du croisement panmictique des arbres G1), permettra de suivre la nature et la fréquence des génotypes multilocus au cours des différentes générations et d'apporter des éléments de réponse aux questions précédentes.

D.3 Une suite à la cartographie génétique comparée des conifères

Bien que la comparaison des cartes de *P. taeda* avec les autres Pinaceae reste de faible résolution (en terme de marqueurs communs dans les différents groupes de liaison), les résultats de conservation de la synténie chez les pins, montrent que le génome des conifères n'a pas subi de profond changement structuraux depuis la divergence des espèces. Ainsi, les quelques connaissances obtenues chez des espèces de Pinaceae en matière de structure du génome devraient pouvoir être réunies jusqu'à pouvoir considéré le génome de cette famille comme un système génétique unique. Une telle méta-analyse des données actuellement disponibles a été initiée dans cette thèse (figures C18 et C19).

En comparaison avec d'autres espèces de plantes, les caractéristiques du génome de ces espèces anciennes sont uniques. En effet, la taille physique de leur génome est du même ordre de grandeur que celui de quelques céréales comme le maïs ou le blé. De plus conifères et céréales possèdent les mêmes proportions de séquences répétées. Cependant, alors que les céréales, comme beaucoup d'espèces de plantes, ont connu des remaniements chromosomiques en terme de nombre de chromosomes et de ploïdie, le génome des conifères semble avoir peu évolué entre espèces d'un même genre voire entre genre d'une même

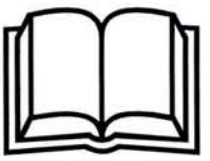
famille. Ceci est particulièrement étonnant si l'on considére leur date d'apparition sur Terre par rapport aux autres plantes !

D'autres particularités biologiques différentient également les conifères des autres plantes. Outre leur caractère pérenne[27], ces organismes sont fortement, voire strictement, allogames et leur diversité génétique a été décrite comme étant très élevée. Est-ce que des mécanismes liés au mode de reproduction, au potentiel adaptatif ou à la diversité de l'espèce, ou bien une combinaison de tous ces facteurs joueraient un rôle sur la structure du génome ?

Ces interrogations montrent que de nombreuses questions restent en suspend en ce qui concerne le génome des conifères, ce qui ouvre la porte vers des études futures. Les perspectives concernant la cartographie génétique comparée des conifères sont-elles de développer encore plus de marqueurs moléculaires et de les tester à nouveau sur les espèces dont les cartes sont déjà alignées ? Aucun projet à ce jour ne concerne cette stratégie, ni même de projet précis de continuation de la cartographie comparée des conifères. Cette stratégie doit s'intégrer aux futurs projets de génomique chez les espèces forestières en prenant pour base les résultats obtenus dans le cadre du projet CCGP (Conifer Comparative Genome Project). Ce dernier a fourni une ressource de départ par le biais d'un jeu de marqueurs disponibles via la base de données Treegenes (http://dendrome.ucdavis.edu/treegenes.html). La poursuite des travaux concernant la synténie des Pinaceae pourra s'opérer selon deux directions : 1) pour les espèces relativement bien étudiées (pour lesquelles une carte génétique et des EST sont disponibles : comme c'est le cas pour *P. pinaster* et *P. taeda*), l'alignement des cartes pourra se poursuivre en se servant directement des SNP détectés *in silico*, 2) dans le cas d'une espèce peu étudiée, les marqueurs obtenus et cartographiés sur le génome des espèces modèles pourront servir de point de départ pour construire une nouvelle carte génétique.

[27] On trouve néanmoins des espèces pérennes chez les angiosperme forestiers et fruitiers.

BIBLIOGRAPHIE



A

Achéré V., Faivre Rampant P., Jeandroz S., Besnard G., Markussen T., Aragones A., Fladung M., Ritter E., Favre J-M. (2004) A full saturated linkage map of *Picea abies* including AFLP, SSR, ESTP, 5S rDNA and morphological markers. Theor Appl Genet, sous presse.

Amarasinghe V., Brown G.R., Mank J.E., Carlson J.E. (2002) Microsatellite DNA loci for Western Hemlock [*Tsuga heterophylla* (Raf.) Sarg]. Mol Ecol 2 : 236-238.

Amarasinghe V., Carlson J.E. (2002) The development of microsatellite DNA markers for genetic analysis in Douglas-fir. Can J For Res 32 : 1904-1915.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408 : 796-815.

Arcade A., Anselin F., Faivre Rampant P., Lesage M.C., Laurans F., Paques L.E., Prat D., (2000) Application of AFLP, RAPD and ISSR markers to genetic mapping of European larch and Japanese larch. Theor. Appl. Genet. 100 : 299-307.

Arumuganathan K., Earle E.D. (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Reporter 9: 208-218.

B

Babula D., Kaczmarck M., Barakat A., Delseny M., Quiros C.F., Sadowski J. (2003) Chromosomal mapping of *Brassica oleracea* based on ESTs from *Arabidopsis thaliana*: complexity of the comparative map. Mol Gen Genomics 268 : 656-665.

Bachem C. W. B., van der Hoeven R. S., de Bruijn S. M., Vreugdenhil D., Zabeau M., Visser R. G. F. (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant Journal 9 : 745-753.

Bahrman N., Zivy M., Damerval C., Baradat P. (1994) Organisation of the variability of abundant proteins in seven geographical origins of maritime pine (*Pinus pinaster* Ait.) Theor Appl Genet 88 : 407-411.

Baradat P., Marpeau A. (1988) Le pin maritime *Pinus pinaster* Ait. : biologie et génétique des terpènes pour la connaissance et l'amélioration de l'espèce. Thèse, Université Bordeaux I.

Barreneche T., Bodénès C., Lexer C., Trontin J.F., Fluchs S., Streiff R., Plomion C., Roussel G., Steinkellner H., Burg K., Favre J-M., Glossl J., Kremer A. (1998) A genetic linkage map of *Quercus robur* (Pedunculate oak) based on RAPD, SCAR, microsatellite, isozyme and rDNA markers. Theor Appl Genet 97 : 1090-1103.

Barreneche T., Casasoli M., Russell K., Akkak A., Meddour H., Plomion C., Villani F., Kremer A. (2004) Comparative mapping between *Quercus* and *Castanea* using simple-sequence repeats (SSRs). Theor Appl Genet, in press.

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

Bebenek K., Abbotts J., Wilson S.H., Kunkel T.A. (1993) Error-prone polymerization by HIV-1 reverse transcriptase. J Biol Chem 268 : 10324-10334.

Bhalerao R., Nilsson O., Sandberg G. (2003) Out of the woods: forest biotechnology enters the genomic era. Current Opinion in Biotechnology 14 : 206-213.

Binelli G., Bucci G. (1994) A genetic linkage map of *Picea abies* Karst., based on RAPD markers, as a tool in population genetics. Theor Appl Genet 88 : 283-288.

Botstein D., White R.L., Skolnick M., Davis R.M. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32 : 314-331.

Bradshaw H.D., Ceulemans R., Davis J, Settler R. (2000) Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. J Plant Growth Regul 19 : 306-313.

Brendel O., Pot D., Plomion C., Rozenberg P., Guehl J-M. (2002) Genetic parameters and QTL analysis of $\delta^{13}C$ and ring width in maritime pine. Plant Cell and Env 25 : 945-953.

Brown G.R., Amarasinghe V., Kiss G., Carlson J.E. (1993) Preliminary karyotype and chromosomal localization of ribosomal DNA sites in white spruce using fluorescence *in situ* hybridization. Genome 36 : 310-316.

Brown G.R., Carlson J.E. (1997) Molecular cytogenetics of genes encoding the 18S-5.8S-26S rDNA and 5S rDNA in two species of spruce (*Picea*). Theor Appl Genet 95 : 1-9.

Brown G.R., Newton C.H., Carlson J.E. (1998) Organization and distribution of a Sau3A tandem repeated DNA sequence in *Picea* (Pinaceae) species. Genome 41 : 560-565.

Brown G.R., Kadel E.E. III, Bassoni D.L., Kiehne K.L., Temesgen B., Van Buijtenen J.P., Sewell M.M., Marshall K.A., Neale D.B. (2001) Anchored reference loci in Loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics 159 : 799-809.

Brown G.R., Bassoni D.L., Gill G.P., Fontana J.R., Wheeler N.C., Megraw R.A., Davis M.F., Sewell M.M., Tuskan G.A., Neale D.B. (2003) Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics 164 : 1537-1546.

Brunner A.M., Busov V., Strauss S.H. (2004) Poplar genome sequence: functional genomics in a ecologically dominant plant species. Trends in Plant Science 9 : 49-56.

Bucci G., Kubisiak T.L., Nance W.L., Menozzi P. (1997) A population 'consensus', partial linkage map of *Picea abies* Karst. based on RAPD markers. Theor Appl Genet 95 : 643-654.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Burban C., Petit R.J. (2003) Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. Molecular Ecology 12 : 1487-1495.

Butland S., Chow M.L., Ellis B.E. (1998) A diverse family of phenylalanine ammonia-lyase genes expressed in pine trees and cell structures. Plant Mol Biol 37 : 15-24.

C

Cantón F.R., Le Provost G., Garcia V., Barré A., Frigério J-M., Paiva J., Fevereiro P., Ávila C., Mouret J-F., de Daruvar A., Cánovas F.M., Plomion C. (2004) Transcriptome analysis of wood formation in maritime pine. In : Sustainable Forestry, Wood products & Biotechnology, BIOFOR proceeding, sous presse.

Cardle L., Ramsay L., Milbourne D., Macaulay M., Marshall D., Waugh R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156 : 847-854.

Cato S.A., Gardner R.C., Kent J., Richardson T.E. (2001) A rapid PCR-based method for genetically mapping ESTs. Theor Appl Genet 102 : 296-306.

Cavell A.C., Lydiate D.J., Parkin I.A.P., Dean C., Trick M. (1998) Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. Genome 41 : 62-69.

Cervera M.T., Storme V., Ivens B., Gusmao J., Liu B.H.., Hostyn V., Van Slycken J., Van Montagu M., Boerjan W. (2001) Dense genetic linkage maps of three populus species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. Genetics 158 : 787-809.

Chen X., Cho Y.G., McCouch S.R. (2002) Sequence divergence of rice microsatellites in *Oryza* and other plant species. Mol Gen Genomics 268 : 331-343.

Cho Y.G., Ishii T., Temnykh S., Chen X., Lipovich L., McCough S.R., Park W.D., Ayer N., Cartinhour S. (2000) Diversity of microsatellites derived from genomic libraries and Genbank sequences in rice (*Oriza sativa*). Theor Appl Genet 100 : 713-722.

Chou Q., Russel M., Birch M., Raymond J., Bloch W. (1992) Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Nuc Ac Res 20 : 1717-1723.

Christoffels A., van Gelder A., Greyling G., Miller R., Hide T., Hide W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nuc Ac Res 29(1) : 238-238.

Cliff A.D.,Ord J.K. (1973) Spatial autocorrelation. Pion, London.

Cordeiro G.M., Casu R., McIntyre C.L., Manners J.M., Henry R.J. (2001) Microsatellite markers from sugarcane (*Saccharum spp.*) ESTs cross transferable to erianthus and sorghum. Plant Sci 160 : 1115-1123.

Costa P. (1999) Réponse moléculaire, physiologique et génétique du pin maritime à une contrainte hydrique. Thèse Université Henri Poincaré, Nancy I.

Costa P., Pot D., Dubos C., Frigerio J-M., Pionneau C., Bodénès C., Bertocchi E., Cervera M.T., Remington D.L., Plomion C. (2000) A genetic map of maritime pine based on AFLP, RAPD and protein markers. Theor Appl Genet 100 : 39-48.

Cox, D. R.; Burmeister, M.; Price, E. R.; Kim, S.; Myers, R. M. (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250 : 245-250.

D

Derory J., Mariette S., Gonzalez-Martinez S.C., Chagné D., Madur D., Gerber S., Brach J., Persyn F., Ribeiro M.M., Plomion C. (2002) What can nuclear microsatellites tell us about maritime pine genetic resources conservation and provenance certification strategies ? Ann For Sci 59 : 699-708.

Devey M.E., Jermstad K.D., Tauer C.G., Neale D.B. (1991) Inheritance of RFLP loci in a loblolly pine three-generation pedigree. Theor Appl Genet 83 : 238-242.

Devey M.E., Fiddler T.A., Liu B.H., Knapp S. J., Neale D.B. (1994) An RFLP linkage map for loblolly pine based on a three-generation outbred pedigree. Theor Appl Genet 88 : 273-278.

Devey M.E., Bell J.C., Smith D.N., Neale D.B., Moran G.F. (1996) A genetic linkage map for *Pinus radiata* based on RFLP, RAPD, and microsatellite markers. Theor Appl Genet 92 : 673-679.

Devey M.D., Sewell M.M., Uren T.L., Neale D.B. (1999) Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theor Appl Genet 99 : 656-662.

Devos K.M., Gale M.D. (2000) Genome relationships: the grass model in current research. Plant Cell. 12 : 637-646.

Dirlewanger E., Cosson P., Tavaud M., Aranzana M.J., Poizat C., Zanetto A., Arus P., Laigret F. (2002) Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). Theor Appl Genet 105 : 127-138.

Doganlar S., Frary A., Daunay M.C., Lester R.N., Tanksley S.D. (2002a) A comparative gentic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the Solanaceae. Genetics 161 : 1697-1711.

Doganlar S., Frary A., Daunay M.C., Lester R.N., Tanksley S.D. (2002b) Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics 161 : 1713-1726.

Dubos C. (2001) Réponse moléculaire de jeunes plants de pin maritime soumis à un stress hydrique en milieu hydroponique. Thèse Université Henri Poincaré, Nancy I.

Dubos C., Plomion C. (2003) Identification of water-deficit responsive genes in maritime pine (*Pinus pinaster* Ait.) roots. Plant Mol Biol 51 : 249-262.

E

Echt C.S., May-Marquardt P., Hseih M., Zahorchak R. (1996) Characterization of microsatellite markers in eastern white pine. Genome 39 : 1002-1008.

Echt C.S., May-Marquardt P. (1997) Survey of microsatellite in pine. Genome 40 : 9-17.

Echt C.S., Vendramin G.G., Nelson C.D., May-Marquardt P. (1999) Microsatellite DNA as shared genetic markers among conifer species. Can J For Res 29 : 365-371.

Elsik C.G., Williams C.G. (2000) Retroelements contribute to the excess low-copy number DNA in pine. Mol Gen Genet 264 : 47-55.

Elsik C.G., Minihan V.T., Hall S.E., Scarpa A.M., Williams C.G. (2000b) Low-copy microsatellite markers for *Pinus taeda* L. Genome 43 : 550-555.

Elsik C.G., Williams C.G. (2001) Low-copy microsatellite recovery from a conifer genome. Theor Appl Genet 103 : 1189-1195.

Eujayl I., Sorrells M.E., Baum M., Wolters P., Powell W. (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104 : 399-407.

Eujayl I., Sledge M.K., Wang L., May G.D., Chekhovskiy K., Zwonitzer J.C., Mian M.A.R. (2003) *Medicago trunculata* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. Theor Appl Genet, in press.

Ewing B., Green P. (1998) Base calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res 8 : 186-194.

Ewing B., Hiller L.D., Wendl M.C., Green P. (1998) Base calling of automated sequencer traces using *Phred*. II. Accuracy assessment. Genome Res 8 : 175-185.

F

Farjon A. (1984) Pines: drawings and descriptions of the genus *Pinus*. E.J. Brill, Leiden.

Ferguson M. C. (1901) The development of the egg and fertilization in *Pinus strobus*. Annals of Botany 15 : 435-479.

Feuillet C., Keller B. (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. Annals of Botany 89 : 3-10.

Fisher P.J., Richardson T.E., Gardner R.C. (1998) Characteristics of single- and multi-copy microsatellites from *Pinus radiata*. Theor Appl Genet 96 : 969-979.

Foolad M.R., Arulsekar S., Becerra V., Bliss F.A. (1995) A genetic map of *Prunus* based on an interspecific cross between peach and almond. Theor Appl Genet 91 : 262-269.

Foulongne M., Pascal T., Arus P., Kervella J. (2003) The potential of *Prunus davidiana* for introgression into peach [*Prunus persica* (L. Batsch] assessed by comparative mapping. Theor Appl Genet 107 : 227-238.

Frary A., Nesbitt T.C., Frary A., Grandillo S., van der Knapp E., Cong B., Liu J., Meller J., Elber R., Alpert K.B., Tanksley S.D. (2000) *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science 289 : 85-88.

Freeling M. (2001) Grasses as a single genetic system. Reassessment 2001. Plant Physiol 125 : 1191-1197.

Fu H., Zheng Z., Dooner H.K. (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99 : 1082-1087.

G

Gale M.D., Devos K.M. (1998) Comparative genetics in the grasses. Proc Natl Acad Sci 95 : 1971-1974.

Gao L., Tang J., Li H., Jia J. (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed 12 : 245-261.

Giordano M., Mellai M., Hoogendoorn B., Momigliano-Richiardi P. (2001) Determination of SNP allele frequencies in pooled DNAs by primer extension genotyping and denaturing high-performance liquid chromatography. Journal of Biochemical and Biophysical Methods 47 : 101-110.

Gocmen B., Jermstad K.D., Neale D.B., Kaya Z. (1996) Development of random amplified polymorphic DNA markers for genetic mapping in Pacific yew (*Taxus brevifolia*). Can J For Res 26 : 497-503.

Goldstein D.B., Schlotterer C. (1999) Microsatellites. Evolution and applications. Goldstein and Schlotterer eds, Oxford University Press.

Gonzalez-Martinez S.C., Agundez D., Alia R, Salvador L., Gil L. (2001) Geographical variation of gene diversity of *Pinus pinaster* Ait. in Iberian peninsula. In: Genetic response of forest ecosystems to changing environmental condition (ed. Müller-Starck G.), pp. 161-171. Kluwer Academic Press, Dordrecht.

Gonzalez-Martinez S.C., Robledo-Arnuncio J.J., Collada C., Diaz A., Williams C.G., Alia R., Cervera M.T. (2003) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theor Appl Genet, in press.

Gosselin I., Zhou Y., Bousquet J., Isabel N. (2002) Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD, SCAR and ESTP markers. Theor Appl Genet 104 : 987-997.

Grattapaglia D., Sederoff R. (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus europhylla* using a pseudo testcross: mapping strategy and RAPD markers. Genetics 137 : 1121-1137.

Groover A., Devey M., Fiddler T., Lee J., Megraw R., Mitchel-Olds T., Sherman B., Vujcic S., Williams C., Neale D. (1994) Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. Genetics 138 : 1293-1300.

Guldberg P., Guttler F. (1993) A simple method for identification of point mutations using denaturing gradient gel electrophoresis. Nucleic Acids Research 21 : 2261-2262.

Gupta P.K., Rustgi S., Sharma S., Singh R., Kumar N., Balyan H.S. (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Gen 270 : 315-323.

H

Haldane J. (1919) The combination of linkage values and the calculation of distance between loci of linked factors. J Genet 8 : 299-309.

Harfouche A., Baradat P., Kremer A. (1995) Variabilité intraspécifique chez le pin maritime (*Pinus pinaster* Ait.) dans le Sud-est de la France. I. Variabilité des populations autochtones et de l'ensemble de l'aire de l'espèce. Ann For Sci 52 : 307-328.

Harry D.E., Temesgen B., Neale D.B. (1998) Codominant PCR-based markers for *Pinus taeda* developed from mapped cDNA clones. Theor Appl Genet 97 : 327-336.

Hawken R., Murtaugh J., Flickinger G.H., Yerle M., Robic A., Milan D., Gelin J., Beattie G.W., Schook L.B., Alexabnder L. (1999) A first generation porcine whole genome radiation hybrid map. Mammalian Genome 10 : 824-830.

Hayashi E., Kondo T., Terada K., Kuramoto N., Goto Y., Okamura M,. Kawasaki (2001) Linkage map of Japanese black pine based on AFLP and RAPD markers including markers linked to resistance against the pine needle gall midge. Theor Appl Genet 102 : 871-875.

Hayashi, E.; Kondo, T.; Terada, K.; Kuramoto, N.; Goto, Y.; Okamura, M.; Kawasaki, H (2002) Linkage map based on AFLP and RAPD markers in Japanese black pine. Bulletin of the National Forest Tree Breeding Center (No.18) : 49-57

Hazen S.P., Kay S.A. (2003) Gene arrays are not just for measuring gene expression. Trends in Plant Science 8 : 413-416.

Hemmat M., Weeden N.F., Manganaris A.G., Lawson D.M. (1994) Molecular marker linkage map for apple. J Heredity 85 : 4-11.

Heslop-Harrison J.S. (2000) Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. Plant Cell 12 :617-635.

Hicks M., Adams D., O'Keefe S., MacDonald E., Hodgetts R. (1998) The development of RAPD and microsatellite markers in lodgepole pine (*Pinus contorta* var. *latifolia*). Genome 41 : 797-805.

Higushi T. (1996) Biochemistry and molecular biology of wood. pp 169-177. T.E. Timell eds, Springer, New York.

Hodgetts R.B., Aleksiuk M.A., Brown A., Clarke C., MacDonald E., Nadeem S., Khasa (2001) Develoment of microsatellite markers for white spruce (*Picea glauca*) and related species. Theor Appl Genet 102 : 1252-1258.

Hodgkin J., Horvitz H.R., Jasny B.R., Kimble J. (1998) *C. elegans*: sequence to biology. Science 282 : 2011.

Holton T.A., Christophe J.T., McClure L., Harker N., Henry R.J. (2002) Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat. Mol Breed 9 : 63-71.

Huang S., Sirikhachornkit A., Faris J.D., Su X., Gill S.B., Haselkorn R., Gornicki P. (2002) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. Plant Mol Biol 48 : 805-820.

Hudson T.J., Church D.M., Greenaway S., Nguyen H., Cook A., Steen R.G., Van Etten W.J., Castle A.B., Strivens M.A., Trickett P., Heuston C., Davison C., Southwell A., Hardisty R., Varela-Carver A., Haynes A.R., Rodriguez-Tome P., Doi H., Ko M.S., Pontius J., Schriml L., Wagner L., Maglott D., Brown S.D., Lander E.S., Schuler G., Denny P. (2001) A radiation hybrid map of mouse genes. Nat Genet 29 : 201-5.

Humphry M.E., Konduri V., Lambrides C.J., Magner T., McIntyre C.L., Aitken E.A.B., Liu C.J. (2002) Development of a mungbean (*Vigna radiata*) RFLP linkage map and its comparison with lablab (*Lablab purpureus*) reveals a high level of colinearity between the two genomes. Theor Appl Genet 105 : 160-166.

I

Illy (1966) Recherches sur l'amélioration génétique du pin maritime. Ann For Sci 23 : 758-948.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 460 : 860-921.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Iwata H., Ujino-Ihara T., Yoshimura K., Nagasaka K., Mukai Y., Tsumura Y. (2001) Cleaved amplified polymorphic markers in sugi, *Cryptomeria japonica* D. Don, and their locations on linkage map. Theor Appl Genet 103 : 881-895.

J

Jenkins S., Gibson N. (2002) High-troughput SNP genotyping. Comp Funct Genom 3 : 57-66.

Jermstad K.D., Bassoni D.L., Wheeler N.C., Neale D.B. (1998) A sex-averaged genetic linkage map in coastal Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco var *'menziesii'*) based on RFLP and RAPD markers. Theor Appl Genet 97 : 762-770.

K

Kalendar R., Grob T., Regina M., Suoniemi A., Schulman A. (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor Appl Genet 98 : 704-711.

Kamm A., Doudrick R.L., Heslop-Harrison J.S., Schmidt T. (1996) The genomic and physical organization of *Ty*1-*copia*-like sequences as a component of large genomes in *Pinus elliottii* var. *elliottii* and other gymnosperms. Proc Natl Acad Sci USA 93 : 2708-2713.

Kantety R.V., LaRota M., Matthews D.E., Sorrells M.E. (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum wheat. Plant Mol Biol 48 : 501-510.

Karhu A., Dieterich J.H., Savolainen O. (2000) Rapid expansion of microsatellite sequences in pines. Mol Bio Evol 17 : 259-265.

Kaya Z., Neale D.B. (1995) Utility of random amplified polymorphic DNA (RAPD) markers for linkage mapping in Turkish red pine (*Pinus brutia* Ten.). Silvae Genet 44 : 110-116.

Keller B., Feuillet C. (2000) Colinearity and gene density in grass genomes. Trends Plant Sci 5 : 246-251.

Keys R.N., Autino A., Edwards K.J., Fady B., Pichot C., Vendramin G.G. (2000) Characterization of nuclear microsatellites in *Pinus halepensis* Mill. and their inheritance in *P. halepensis* and *Pinus brutia* Ten. Mol Ecol 9 : 2155-2234.

Kinlaw C.G., Neale D.B. (1997) Complex gene families in pine genomes. Trends Plant Sci 2 : 356-359.

Kirst M., Johnston A.F., Baucom C., Ulrich E., Hubbard K., Staggs R., Paule C., Retzel E., Whetten R., Sederoff R. (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc Natl Acad Sci USA 100: 7383-7388.

Komulainen P., Brown G.R., Mikkonen M., Karhu A., Garcia-Gil M.R., O'Malley D., Lee B., Neale D.B., Savolainen O. (2003) Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*. Theor Appl Genet 107 : 667-678.

Kosambi D.D. (1944) The estimation of map distance from recombination values. Ann Eugen 12 : 172-175.

Kossack D.S., Kinlaw C.S. (1999) *IFG*, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive hystory in pines. Plant Mol Biol 39 : 417-426.

Kostia S., Varvio S.L., Vakkari P., Pulkkinen P. (1995) Microsatellite sequences in a conifer, *Pinus sylvestris*. Genome 38 : 244-1248.

Kowalski S.P., Lan T.H., Feldmann K.A., Paterson A.H. (1994) Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals island of conserved organization. Genetics. 138 : 449-510.

Kriebel H.B. (1985) DNA sequence componenets of *Pinus strobus* nuclear genome. Can J For Res 15 : 1-4.

Krupkin A.B., Liston A., Strauss S.H. (1996) Phylogenetic analysis of hard pine (*Pinus* subgenus *Pinus*, Pinaceae) from chloroplast DNA restriction site analysis. American Journal of Botany 83 : 489-498.

Krutovskii K.V., Vollmer S.S., Sorensen F.C., Adams W.T., Knapp S.J., Strauss S.H. (1998) RAPD genome maps of douglas fir. J Hered 89 : 197-205.

Krutovskii K.V., Troggio M. , Brown G.R., Jermsatd K.D., Neale D.B. (2004) Comparative mapping in the Pinaceae. Genetics, in press.

Kubisiak T.L., Nelson C.D., Nance W.L., Stine M. (1995) RAPD linkage mapping in a longleaf pine x slash pine F1 family. Theor Appl Genet 90 : 1110-1127.

Kuppuswamy M.H., Hoffman J.W., Kasper C.K., Spitzer S.G., Groce S.L., Bajaj S.P. (1991) Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. Proc Natl Acad Sci USA 88 : 1143-1147.

Kutil B.L. and Williams C.G. (2001) Triplet-repeat microsatellites shared among hard and soft pines. J Hered 92 : 327-332.

Kwok P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet 2 : 235-258.

L

Lagercrantz U., Lydiate D.J. (1996) Comparative genome mapping in *Brassica*. Genetics 144 : 1903-1910.

Lan T.H., Delmonte T.A., Reischmann K.P., Hyman J., Kowalski S.P., McFerson J., Kresovich S., Paterson A.H. (2000) An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res 10 : 776-788.

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lilncoln S.E., Newberg L. (1987) MAPMAKER : an interactive computer package for constructing primary genetic maps of experimental and natural populations. Genomics 1 : 174-181.

Le Provost G. (2003) Effet de la saison et d'un stress mécanique sur la variation du transcriptome dans le xylème en formation chez le pin maritime (Pinus pinaster Ait.). Thèse Université Henri Poincaré, Nancy I.

Le Provost G., Paiva J., Pot D., Brach J., Plomion C. (2003) Seasonal variation in transcript accumulation in wood forming tissues of maritime pine (*Pinus pinaster* Ait.) with emphasis on a Cell wall Glycin Rich Protein. Planta. 217 : 820-830.

Lerceteau E., Plomion C., Andersson B. (2000) AFLP mapping and detection of quantitative trait loci (QTLs) for economically important traits in *Pinus sylvestris*: a preliminary study. Molecular Breeding 6 : 451-458.

Lian C., Miwa M., Hogetsu T. (2000) Isolation and characterization of microsatellite loci from the Japanese red pine, *Pinus densiflora*. Mol Ecol 9 : 1171-1193.

Liebhard R., Koller B., Gianfranceschi L., Gessler C. (2003) Creating a saturated reference map for the apple (*Malus* x *domestica* Borkh.) genome. Theor Appl Genet 106 : 1497-1508.

Lin Y.R., Schertz K.F., Paterson A.H. (1995) Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics 141 : 391-411.

Liston A., Robinson W.A., Pinero D, Alvarez-Buylla E.R. (1999) Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Molecular Phylogenetics and Evolution 11 : 95-109.

Little E.L. Jr, Critchfield W.B. (1969) Subdivisions of the genus *Pinus* (Pines). USDA Forest Service, Washington, DC. Miscellaneous publication 1144.

Liu B.H. (1998) Statistical genomics – Linkage, Mapping and QTL analysis. CRC Press, New York.

Liu MengJun (1998) Genome map and its application in fruit tree. Advances in Horticulture 2 : 1-7.

Livingstone K.D., Lackney V.K., Blauth J.R., Van Wijk R., Jahn M.K. (1999) Genome mapping in *Capsicum* and the evolution of genome structure in the Solanaceae. Genetics 152 : 1183-1202.

M

Mariette S., Chagné D., Decroocq S., Vendramin G.G., Lalanne C., Madur D., Plomion C. (2001) Microsatellite markers for *Pinus pinaster* Ait. Ann For Sci 58: 203-206.

Mariette S., Chagné D., Lézier C., Pastuszka P., Raffin A., Plomion C., Kremer A. (2002) Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. Heredity 86 : 469-479.

Markussen T., Fladung M., Achere V., Favre J-M., Faivre-Rampant P., Aragones A., Da Silva P., Harvengt L., Ritter E. (2003) Identification of QTLs controlling growth, chemical and physical wood property traits in *Pinus pinaster* Ait. *Sylvae Genet* 52 : 8-15.

Marques C.M., Araujo J.A., Ferreira J.G., Whetten R., O'Malley D.M., Liu B.H., Sederoff R. (1998) AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*. Theor Appl Genet 96 : 727-737.

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P., Gish W.R. (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23 : 452-456.

Menancio-Hautea D., Fatokun C.A., Kumar L., Danesh D., Young N.D. (1993) Comparative genome analysis of mungbean (*Vigna radiata* L. Wilczek) and cowpea (*V. unguiculata* L. Walpers) using RFLP mapping data. Theor Appl Genet 86 : 797-810.

Metzgar D., Bytof J., Wills C. (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10 : 72-80.

Miksche J.P. (1967) Variation in DNA content of several gymnosperms. Canadian Journal of Genetics and Cytology 9 : 717-722.

Mirov N.T. (1967) The Genus *Pinus*. New York: Ronald press.

Morgan T.H. (1911) Random segregation *versus* coupling in mendelian inheritance. Science 36 : 718-719.

Morgante M., Hanafey M., Powell W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30 : 194-200.

Moore G., Devos K.M., Wang Z., Gale M.D. (1995) Grasses, line up and form a circle. Current Biology 5 : 737-739.

Moriguchi Y., Iwata H., Ujino-Ihara T., Yoshimura K., Taira H., Tsumura Y. (2003) Development and characterization of microsatellite markers for *Cryptomeria japonica* D. Don. Theor Appl Genet 106 : 751-758.

Morton N.E. (1955) Sequential tests for the detection of linkage. Am J Hum Gen 7 : 277-318.

Murray B.G., Friesen N., Heslop-Harrison J.S. (2002) Molecular cytogenetic analysis of *Podocarpus* and comparison with other gymnosperm species. 89 : 483-489.

Myburg A.A., Griffin A.R., Sederoff R.R., Whetten R.W. (2003) Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F1 hybrid based on a double pseudo-backcross mapping approach. Theor Appl Genet 107 : 1028-1042.

Myers R.M., Maniatis T., Lerman L.S. (1987) Detection and localization of single base changes by denaturing gradient gel electrophoresis. Methods Enzymol. 155: 501-527

N

Neale D.B., Williams C.G. (1991) Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J For Res 21 : 545-554.

Nei M. (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70 : 3321-3323.

Nelson C.D. Nance, W.L. Doudrick, R.L. (1993) A partial genetic linkage map of Slash pine (*Pinus elliotti* Englem var *elliottii*) based on random amplified polymorphic DNA's. Theor Appl Genet 87 : 145-151.

Nicol F., His I., Jauneau A., Verhnettes S., Canut H., Hofte H. (1998) A plasma membrane bound putative endo-1,4-B-D-glucanase is required for normal wall assembly and cell elongation in Arabidopsis. EMBO J 17 : 5563-5576.

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258 : 67-86 et 148-162.

O

Ogata H., Audic S., Renesto-Audiffren P., Fournier P-E., Barbe V., Samson D., Roux V., Cossart P., Weissenbach J., Claverie J-M., Raoult D. (2001) Mechanisms of evolution in *Rickettsia conori* and *R. prowazekii*. Science 293 : 2093-2098.

Ohri D., Khoshoo T.N. (1986) Genome size in gymnosperms. Pl Syst Evol 153 : 119-132.

Orita M., Iwahana H., Kanazawa H., Hayashi K., Sekiya T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand cinformation polymorphisms. Proc Natl Acad Sci USA 86 : 2766-2770.

P

Paglia G.P., Olivieri A.M., Morgante M. (1998) Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol Gen Genet 258 : 466-478.

Pastuszka P., Raffin A., Alazard P. (2002) Aperçu historique du programme d'amélioration du pin maritime. Le progrès génétique en forêt. p 7-12.

Paterson A.H., Damon S., Hewitt J.D., Zamir D., Rabinowitch H.D., Lincoln S.E., Lander E.S., Tanksley S.D. (1991) Mendelian factor underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127 : 181-197.

Peakall R., Gilmore S., Keys W., Morgante M., Rafalski A. (1998) Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15 : 1275-1287.

Pelgas B., Isabel N., Bousquet J. (2004) Efficient screening for expressed sequence tag polymorphisms (ESTPs) by DNA pool sequencing and denaturing gradient gel electrophoresis (DGGE) in spruces. Mol Breed, in press.

Perry D.J., Furnier G.R. (1996) *Pinus banksiana* has at least seven alcohol dehydrogenase genes in two linked groups. Proc Natl Acad Sci USA 93 : 13020-13023.

Peterson D.G., Schulze S.R., Sciara E.B., Lee S.A., Bowers J.E., Nagel A., Jiang N., Tibbitts D.C., Wessler S.R., Paterson A.H. (2002) Integration of Cot Analysis, DNA Cloning, and High-Throughput Sequencing Facilitates Genome Characterization and Gene Discovery. Genome Res 12 : 795-807.

Petit R.J., Barhman N., Baradat P. (1995) Comparison of genetic differentiation in maritime pine (*Pinus pinaster* Ait.) estimated using isozymes, total proteins and terpenic loci. Heredity 75 : 382-389.

Pfeiffer A., Olivieri A.M., Morgante M. (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome 40 : 411-419.

Pflieger S., Lefebvre V., Causse M. (2001) The candidate gene approach in plant genetics a review. Mol Breed 7 : 275-291.

Plomion C., Bahrman N., Durel C-E., O'Malley D.M. (1995a) Genomic mapping in *Pinus pinaster* (maritime pine) using RAPD and protein markers. Heredity 74 : 661-668.

Plomion C., O'Malley D.M. (1996) Recombination rate differences for pollen parents and seed parents in *Pinus pinaster*. Heredity 77 : 341-350.

Plomion C., Hurme P., Frigerio J-M., Ridolfi M., Pot D., Pionneau C., Avila C., Gallardo F., David H., Neutelings G., Campbell M., Canovas F.M., Savolainen O., Bodenes C., Kremer A. (1999) Developing SSCP markers in two *Pinus* species. Molecular Breeding 5 : 21-31.

Plomion C., Le Provost G., Stokes A. (2001) Wood formation in trees. Plant Physiol 127 : 1513-1523.

Pot D. (2004) Déterminisme génétique de la qualité du bois chez le pin maritime : du phénotype aux gènes. Thèse ENSA de Rennes.

Price R.A., Liston A., Strauss S.H. (1998) Phylogeny and systematics of *Pinus*. In Richardson D.M. (ed) Ecology and biogeography of *Pinus*. Cambridge University Press, pp. 49-68.

Q

Qi X., Stam P., Lindhout P. (1996) Comparison and integration of four barley genetic maps. Genome 39 : 379-394.

R

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

Rajora O.P., Rahman M.H., Dayanandan S., Mosseler A. (2001) Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. Mol Gen Genet 264 : 871-882.

Rake A.V., Miksche J.P., Hall R.B., Hansen K.M. (1980) DNA reassociation kinetics of four conifers. Canadian Journal of Genetics and Cytology 22 : 69-79.

Remington D.L. Whetten R.W. Liu B.H. O'Malley D.M. (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor. Appl. Genet. 98 : 1279-1292.

Ribeiro M., Plomion C., Petit R., Vendramin G.G., Szmidt A.E. (2001) Variation of chloroplast single-sequence repeats in Portuguese maritime pine (*Pinus pinaster* Ait.) Theor Appl Genet 102 : 97-103.

Rico C., Rico I., Hewitt G. (1996) 470 million years of conservation of microsatellite loci among fish species. Proc R Soc Lond B Biol Sci 263 : 549-557.

Ruiz C., Asins M.J. (2003) Comparison between *Poncirus* and *Citrus* genetic linkage maps. Theor Appl Genet 106 : 826-836.

S

Saintagne C. (2003) Distribution des régions génomiques différenciant deux espèces proches : le chêne sessile (*Quercus petraea*) et le chêne pédonculé (*Q. robur*). Thèse Université Henri Poincaré, Nancy I.

Salvador L., Alia R, Agundez D., Gil L. (2000) Genetic variation and migration pathways of maritime pine (*Pinus pinaster* Ait.) in the Iberian peninsula. Theor Appl Genet 100 : 89-95.

Salzberg S.L., Delcher A.L., Kasif F., White O. (1998) Microbial gene identification using Markov interpolated models. Nuc Ac Res 26 : 544-548.

Saylor L.C. (1961) A karyotype analysis of selected species of *Pinus*. *Sylvae Genetica* 10 : 77-84.

Sax K., Sax H.J. (1933) Chromosome number and morphology in the conifers. Jour Arnold Arboretum 14 : 356-375.

Schiex T., Gaspin C. (1997) Cartagene: constructing and joining maximum likelihood genetic maps. Proc. of ISMB'97, Halkidiki, Greece, June 1997.

Schiex T., Gouzy J., Moisan A., de Oliveira Y. (2003) FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nuc Ac Res 31 : 3738-3741.

Schneider S., Roessli D., Excoffier L. (2000) ARLEQUIN v.2000 (Genetics and Biometry Laboratory, University of Geneva, Switzerland).

Scott K.D., Eggler P., Seaton G., Rossetto M., Ablett E.M., Lee L.S., Henry R.J. (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100 : 723-726.

Scott L.J., Shepherd M., Henry R.J. (2003) Characterization of highly conserved microsatellite loci in *Araucaria cunninghamii* and related species. Plant Syst Evol 236 : 115-123.

Scotti I., Magni F., Fink R., Powell W., Binelli G., Hedley P.E. (2000) Microsatellite repeats are not randomly distributed within Norway spruce (*Picea abies* K.) expressed sequences. Genome 43 : 41-46.

Scotti I., Magni F., Paglia G.P., Morgante M. (2002a) Trinucleotide microsatellites in Norway spruce (*Picea abies*): their features and the development of molecular markers. Theor Appl Genet 106 : 40-50.

Scotti I., Paglia G.P., Magni F., Morgante M. (2002b) Efficient Development of Dinucleotide microsatellite markers in Norway spruce (*Picea Abies* Karst.) through dot-blot selection. Theor Appl Genet 104 : 1035-1041.

Schmidt A., Doudrick R.L., Heslop-Harrison J.S., Schmidt D.T. (2000) The contribution of short repeats of low sequence complexity to large conifer genomes. Theor Appl Genet 101 : 7-14.

Sewalt V.J.H., Ni W., Blount J.W., Jung H.G., Massoud S.A., Howles P.A., Lamb C., Dixon R.A (1997) Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of L-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase. Plant physiol 115 : 41-50.

Sewell M.M., Sherman B.K., Neale D.B. (1999) A consensus map for Loblolly Pine (*Pinus taeda* L.) I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees. Genetics 151 321-330.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Sewell M.M., Bassoni D.L., Megraw R.A., Wheeler N.C., Neale D.B. (2000) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. Theor Appl Genet 101 : 1273-1281.

Sewell M.M., Davis M.F., Tuskan G.A., Wheeler N.C., Elam C.C., Bassoni D.L., Neale D.B. (2002) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). II. Chemical wood properties. Theor Appl Genet 104 : 214-222.

Sharma S., Balyan H.S., Kulwal P. L., Kumar N., Varshney R.K., Prasad M., Gupta P.K. (2002) Study of interspecific SSR polymorphism among 14 species from *Triticum-Aegilops* group. Wheat Information Service 95 : 23-28.

Shepherd M., Cross M., Maguire T.L., Dieters M.J., Williams C.G., Henry R.J. (2002) Transpecific microsatellites for hard pines. Theor Appl Genet 104 : 819-827.

Smith D.N. and Devey M.E. (1994) Occurrence and inheritance of microsatellites in *Pinus radiata*. Genome 37 : 977-983.

Song W.Y., Wang G.L., Chen L.L., Kim H.S., Pi L.Y., Holsten T., Gardner J., Wang B., Zhai W.X., Zhu L.H., Fauquet C., Ronald P. (1995) A receptor kinase like protein encoded by the rice disease resistance gene, XA21. Science 270 : 1856-1859.

Soranzo N., Provan J., Powell W. (1998) Characterization of microsatellite loci in *Pinus sylvestris* L. Mol Ecol 7 : 1247-1263.

Sosinski B., Gannavarapu M., Hager L.D., Beck L.E., King G.J., Ryder C.D., Rajapakse S., Baird W.V., Ballard R.E., Abbott A.G. (2000) Characterisation of microsatellite markers in peach [*Prunus persica* (L.) Batsch]. Theor Appl Genet 101 : 421-428.

Stam P. (1993) Construction of integrated linkage maps by means of a new computer package: JOINMAP. Plant J. 3 : 739-744.

Suárez M.C., Bernal A., Gutiérrez J., Tohme J., Fregene M. (2000) Developing expressed sequence tags (ESTs) from polymorphic transcript-derived fragments (TDFs) in cassava (*Manihot esculenta* Crantz). Genome 43 : 62-67.

Sutton T., Whitford R., Baumann U., Dong C., Able J.A., Langridge P. (2003) The *Ph2* pairing homoeologous locus of wheat (*Triticum aestivum*): identification of candidate meiotic genes using a comparative genetics approach. Plant Journal 36 : 443-456.

Syvanen A.C. (1999) From gels to chips: 'minisequencing' primer extension for analysis of point mutations and single nucleotide polymorphisms. Hum Mutat 13 : 1-10.

T

Tanksley S.D., Bernatzky R., Lapitan N.L., Prince J.P. (1988) Conservation of gene repertoire but not gene order in pepper and tomato. Proc Natl Acad Sci 83 : 6419-6423.

Tanksley S.D., Ganal M.W., Prince J.P., de Vicente M.C., Bonierbale M.W., Broun P., Fulton T.M., Giovannoni J.J., Grandillo S., Martin G.B., Messeguer R., Miller J.C., Miller L., Paterson A.H., Pineda O., Roder M.S., Wing R.A., Wu W., Young N.D. (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics 132 : 1141-1160.

Temesgen B., Neale D.B., Harry D.E. (2000) Use of haploid mixtures and heteroduplexe analysis enhance polymorphism revealed by Denaturing Gradient Gel Electrophoresis. BioTechniques 28 : 114-122.

Temesgen B., Brown G.R., Harry D.E., Kinlaw C.S., Sewell M.M., Neale D.B. (2001) Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (Pinus taeda L.). Theor Appl Genet 102 : 664-675.

Temnykh S., Park W.D., Ayres N., Cartinhour S., Hauck N., Lipovich L., Cho Y.G., Ishii T., McCough S.R. (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100 : 697-712.

Temnykh S., DeClerck G., Lukashova A., Lipovich L., Cartinhour S., McCouch S. (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Gen Res 11 : 1441-1452.

Thiel T., Michalek W., Varshney R.K., Graner A. (2003) Exploiting EST database for the development and characterization of gene derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106 : 411-422.

Toth G., Gaspari Z., Jurka J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Gen Res 10 : 967-981.

Travis S.E., Ritland K., Whitman T.G., Keim P. (1998) A genetic linkage map of Pinyon pine (*Pinus edulis*) based on amplified fragment length polymorphism. Theor Appl Genet 97 : 871-880.

Troggio M., Kubisiak T.L., Bucci G., Menozzi P. (2001) Randomly amplified polymorphic DNA linkage relationships in different Norway spruce populations. Can J Forest Res 31 : 1456-1461.

Troggio M., Brown G.R., Chagné D., Krutovskii K.V., Menozzi P., Scotti I., Plomion C., Neale D.B. (2004) Comparative EST and genetic mapping in conifers: Norway spruce and loblolly pine. Manuscrit en preparation.

Tulsieram L.K., Glaubitz J.C., Kiss G., Carlson J.E. (1992) Single tree genetic linkage analysis in conifers using haploid DNA from megagametophytes. BioTechnology 10 :686-690.

V

Van der Burgh J. (1973) Hölzer der niederrheinischen Braunkohlenformation, 2. Hölzer der Braunkohlengruben "Maria Theresa" zu Herzogenrath, "Zukunft West" zu Eschweiler

und "Victor" (Zülpich Mitte) zu Zülpich. Nebst einer systematisch-anatomischen Bearbeitung der Gattung *Pinus* L. Rev Paleobot Palynol 15 : 73-275.

Van Ooijen J.W. and Voorrips R.E. (2001) Joinmap 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands

Van de ven W.T.G., Mac Nicol R.J. (1996) Microsatellites as DNA markers in Sitka spruce. Theor Appl Genet 93 : 613-617.

Varshney R.K., Thiel T., Stein N., Langridge P., Graner A. (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol 7 : 537-546.

Vendramin G.G., Anzidei M., Madaghiele A., Bucci G. (1998) Distribution of genetic diversity in *Pinus pinaster* Ait. as revealed by chloroplast microsatellites. Theor Appl Genet 97 : 450-463.

Verhaegen D., Plomion C. (1996) Genetic mapping in *Eucalyptus europhylla* and *Eucalyptus grandis* using RAPD markers. Genome 39 : 1051-1061.

de Vienne D. (1998) Les marqueurs moléculaires en génétique et biotechnologies végétales. INRA éditions. 200p.

Voorrips R.E. (2001) MapChart version 2.0: Windows software for the graphical presentation of linkage maps and QTLs. Plant Research International, Wageningen, The Netherlands

Vos P., Hogers R., Bleeker M., Reijans M., van der Lee T., Hornes M., Frijters A., Pot J., Pelemam J., Kuiper M., Zabeau M. (1995) AFLP : a new technique for DNA fingerprinting. Nucleic Acids Res. 23 : 4407-4414.

Vuylsteke M., Mank R., Antonise R., Bastiaans E., Senior M.L., Stuber C.W., Melchinger A.E., Lübberstedt T., Xia X.C., Stam P., Zabeau M.? Kuiper M. (1999) Two high-density AFLP linkage maps of *Zea mays* L;: analysis of distribution of AFLP markers.

W

Wakamiya I., Newton R. Johnston J.S., Price H.J. (1993) Genome size and environmental factors in the genus *Pinus*. Am. J. Bot. 80 : 1235-1241.

Wang X-R., Tsumura Y., Yoshimaru H., Nagasaka K., Szmidt A.E. (1999) Phylogenetic relationships of eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer and *trn*V intron sequences. American Journal of Botany 86 : 1742-1753.

Wang X-Q., Tank D.C., Sang T. (2000) Phylogeny and divergence times in Pinaceae: evidence from three genomes. Mol Biol Evol 17 : 773-781.

Ware D.H., Jaiswal P., Ni J., Yap I.V., Pan X., Clark K.Y., Teytelman L., Schmidt S.C., Zhao W., Chang K., Cartinhour S., Stein L.D., McCouch S.R. (2002) Gramene, a tool for grass genomics. Plant Physiol 130 : 1606-1613.

Watson J.D., Crick F.H.C (1953) Molecular structure of nucleic acids. Nature 171 : 737-738.

Waugh R., McLean K., Flavell A.J., Pearce S.R., Thomas B.B.T., Powell W. (1997) Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specifis amplification polymorphisms (S-SAP) Mol Gen Genet 253 : 687-694.

Weeden N.F., Muehlbauer F.J., Ladizinsky G. (1992) Extensive conservation of linkage relationships betweenpea and lentil genetic maps. J Hered 83 : 123-129.

Wilcox P.L., Richardson T.E., Corbett G.E., Ball R.D., Lee J.R., Djorovic A., Carson S.D. (2001) Framework linkage maps of *Pinus radiata* D. Don based on pseudotestcross markers. For Genet : 109-117.

Williams J.G.K., Kubelik A.R., Livak-Kenneth J., Rafalski J., Antoni J., Scott V. (1990) DNA polymorphism amplified by arbitrary primers are useful as gentic markers. Nuc. Ac. Res. 18 : 6531-6535.

X

Xu L.A. (2004) Diversité de l'ADN chloroplastique et relations phylogénétiques au sein des Fagacées et du genre Quercus. Thèse Université Henri Poincaré, Nancy I.

Y

Yazdani R., Yeh F., Rimsha J. (1995) Genomic mapping of *Pinus sylvestris* (L) using random amplified polymorphic DNA markers. For Genet 2 : 109-116.

Yin T. M., Wang X. R., Andersson B., Lerceteau-Kohler E. (2003) Nearly complete genetic maps of *Pinus sylvestris* L. (Scots pine) constructed by AFLP marker analysis in a full-sib family. Theoretical and Applied Genetics 106 (6) : 1075-1083.

Yu J., Hu S.N., Wang J., Li S.G., Wong G., Liu B., Deng Y.J., Dai L., Zhou Y., Zhang X.Q., Cao M.L., Liu J., Sun J.D., Tang J.B., Chen Y.J., Huang X.B., Lin W., Ye C., Tong W., Cong L.J., Geng J.N., Han Y.J., Li L., Li W., Hu G.Q., Huang X.G., Li W.J., Li J., Liu Z.W., Liu J.P., Qi Q.H., Liu J.S., Wang X.G., Lu H., Wu T.T., Zhu M., Ni P.X., Han H., Dong W., Ren X.Y., Feng X.L., Cui P., Li X.R., Wang H., Xu X., Zhai W.X., Xu Z., Zhang J.S., He S.J., Zhang J.G., Xu J.C., Zhang K.L., Zheng X.W., Dong J.H., Zeng W.Y., Tao L., Chen X.W., He J., Liu D.F., Tian W., Tian C.G., Xia H.A., Li G., Gao H., Li P., Chen W., Wang X.D., Zhang Y., Hu J.F., Liu S., Yang J., Zhang G.Y., Xiong Y.Q., Li Z.J., Mao L., Zhou C.S., Zhu Z., Chen R.S., Hao B.L., Zheng W.M., Chen S.Y., Guo W., Li G.J., Liu S.Q., Huang G.Y., Tao M., Zhu L.H., Yuan L.P., Yang H.M. (2001) A draft sequence of the rice (*Oryza sativa* ssp. *indica*) genome. Chinese Science Bulletin 46 : 1937-1942.

Z

Zhou Y., Bui T., Auckland L.D., Williams C.G. (2002) Undermethylated DNA as a source of microsatellite from a conifer genome. Genome 45: 91-99.

ANNEXES

Liste des annexes et contribution aux articles présentés

Annexe I : **Chagné D.**, Lalanne C., Madur D., Kumar S., Frigerio J-M., Krier C., Decroocq S., Savouré A., Bou-Dagher-Kharrat M., Bertocchi E., Brach J., Plomion C. (2002) A high density genetic map of maritime pine based on AFLPs. Annals of Forest Science 59 : 627-636.

Annexe II : Mariette S., **Chagné D.**, Decroocq S., Vendramin G.G., Lalanne C., Madur D., Plomion C. (2001) Microsatellite markers for *Pinus pinaster* Ait. Annals of Forest Science 58 : 203-206.
Contribution :
1) mise au point des conditions d'amplification chez le pin maritime;
2) cartographie des locus polymorphes sur la carte génétique décrite par Plomion et al. (1995a);
3) génotypage des locus polymorphes sur des peuplements naturels de pin maritime.

Annexe III : **Chagné D.**, Brown G.R., Lalanne C., Madur D., Pot D., Neale D.B., Plomion C. (2003) Comparative genome and QTL mapping between maritime and loblolly pines. Molecular Breeding 12 : 185-195.

Annexe IV : **Chagné D.**, Chaumeil P., Ramboer A., Collada C., Guevara A., Cervera M-T., Vendramin G.G., Garcia V., Frigerio J-M., Echt C., Richardson T., Plomion C. Cross species transferability and mapping of genomic and cDNA SSRs in pines. Soumis à Theoretical and Applied Genetics (décembre 2003).

Annexe V : Le Dantec L., **Chagné D.**, Pot D., Cantin O., Garnier-Géré P., Bedon F., Frigerio J-M., Chaumeil P., Léger P., Garcia V., Laigret F., de Daruvar A., Plomion C. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Soumis à Plant Molecular Biology (janvier 2004).
Contribution :
1) mise au point des paramètres de détection du pipeline à l'aide du jeu de SNP de référence;
2) analyse des résultats de détection des SNP sur les EST de pin maritime.

Annexe VI : Scotti I., Burelli A., Cattonaro F., **Chagné D.**, Fuller J., Hedley P.E., Jansson G., Neale D.B., Plomion C., Powell W., Troggio M., Morgante M. Analysis of the distribution of marker classes in a genetic linkage map of Norway spruce (*Picea abies* Karst.). Manuscrit en préparation.
Contribution :
1) génotypage des marqueurs EST sur la population de cartographie;
2) construction de la carte génétique.

ANNEXE I

A high density genetic map of maritime pine based on AFLPs

Ann. For. Sci. 59 (2002) 627–636
© INRA, EDP Sciences, 2002
DOI: 10.1051/forest:2002048

Original article

A high density genetic map of maritime pine based on AFLPs

David Chagné[a], Céline Lalanne[a], Delphine Madur[a], Satish Kumar[b], Jean-Marc Frigério[a], Catherine Krier[a], Stéphane Decroocq[a], Arnould Savouré[c], Magida Bou-Dagher-Kharrat[c], Evangelista Bertocchi[a], Jean Brach[a] and Christophe Plomion[a*]

[a] INRA, Équipe de Génétique et Amélioration des Arbres Forestiers, 69 route d'Arcachon, 33612 Cestas Cedex, France
[b] Forest Research, Applications of Genomic Science, Sala Street, Rotorua, 3021, New Zealand
[c] Physiologie Cellulaire et Moléculaire des Plantes, UMR 7632 CNRS, Université Pierre et Marie Curie, case 156, 4 Place Jussieu, 75252 Paris Cedex 05, France

(Received 16 August 2001; accepted 13 February 2002)

Abstract – We constructed a high-density linkage map of maritime pine (*Pinus pinaster* Ait.) based on AFLP (Amplified Fragment Length Polymorphism) markers using a three-generation outbred pedigree. In a first step, male and female maps were established independently with test-cross markers segregating 1:1 (presence:absence of the amplified fragment in the full-sib progeny). In a second step, both maps were merged using intercross markers segregating 3:1 in the progeny. A combination of MAPMAKER and JOINMAP softwares was used for the mapping process. A consensus map was obtained and is available at URL http://www.pierroton.inra.fr/genetics/pinus/Map3/index.html. It covers 1441 cM and comprises a total of 620 AFLP markers on 12 linkage groups. The physical size of the maritime pine genome (51.5 pg/2C) was measured by flow cytometry, providing a physical/genetic size ratio of 13.78 Mb/cM. This map will be used to dissect the genetic architecture of economically (growth, wood quality) and ecologically (water-use efficiency) important traits into mendelian inherited components (QTLs: Quantitative Trait Loci). It will also provide a framework to localize more informative markers (ESTs: Expressed Sequence Tags) to be used as candidate genes in QTL detection experiments. The location of orthologous markers (ESTs and SSRs: Simple Sequence Repeats) will also allow the study of the genome structure of closely related conifer species using a comparative genome mapping approach.

***Pinus pinaster* / genetic linkage map / AFLP / double pseudo-testcross / physical size**

Résumé – Établissement d'une carte génétique à haute densité du pin maritime à partir de marqueurs AFLP. Nous avons construit une carte génétique du pin maritime (*Pinus pinaster* Ait.) en génotypant une famille de plein-frères appartenant à la troisième génération du programme d'amélioration, avec des marqueurs AFLP. Dans un premier temps, les cartes des parents mâle et femelle ont été établies indépendamment avec des marqueurs de type « test-cross » ségréguant dans les proportions 1:1 (présence:absence du fragment amplifié dans la famille de plein-frères). Dans un second temps ces deux cartes ont été fusionnées à l'aide de marqueurs de type « intercross », ségréguant dans les proportions 3:1. La construction des cartes a été réalisée à l'aide des logiciels de cartographie génétique JOINMAP et MAPMAKER. Une carte génétique consensus des deux parents comprenant 12 groupes de liaison a finalement été obtenue et est accessible à l'URL suivante : http://www.pierroton.inra.fr/genetics/pinus/Map3/index.html. Elle couvre 1441 cM et comprend 620 marqueurs. Par ailleurs, la taille physique du génome du pin maritime a été estimée par cytométrie de flux à 51.5 pg/2C, donnant un rapport taille physique/taille génétique de 13.78 Mb/cM. Cette carte sera maintenant utilisée pour étudier l'architecture génétique de caractères d'intérêt économique (croissance, qualité du bois) et écologique (efficience d'utilisation de l'eau). Il s'agira de localiser les zones du génome (QTL, Quantitative Trait Loci) impliquées dans le contrôle génétique de ces caractères complexes. La carte génétique fournira aussi un support pour localiser d'autres types de marqueurs, tels que des gènes (EST, Expressed Sequence Tags) qui seront utilisés comme marqueurs candidats pouvant correspondre aux QTL. La localisation de marqueurs orthologues (EST et SSR, Simple Sequence Repeats) permettra d'étudier en outre la structure du génome des conifères en utilisant une approche par cartographie comparée.

pin maritime / carte génétique / AFLP / double pseudo-testcross / taille physique

* Correspondence and reprints
Tel.: +33 5 57 12 28 38; fax: +33 5 57 12 28 81; e-mail: plomion@pierroton.inra.fr

1. INTRODUCTION

Maritime pine (*Pinus pinaster* Ait.) is the most economically and ecologically important conifer species in the southwestern Europe, where it covers about 4 millions hectares. In France, INRA (Institut National de la Recherche Agronomique) started a breeding programme of maritime pine in the early sixties to provide foresters with improved varieties for growth and straightness. This program has now reached its third generation. Although positive genetic gains are obtained through classical breeding strategies [5], there is a great need to improve selection efficiency. Indeed, forest tree selection faces three major stumbling blocks: (i) late selection age (12 years of age for maritime pine, [32]), (ii) complex traits with low to medium heritabilities [17, 31, 48], (iii) and late flowering (8 years of age for maritime pine). The development of molecular marker techniques provides new tools to detect the genomic regions involved in the genetic control of quantitative traits (QTLs, Quantitative Trait Loci, [59]), which, in turn, will improve selection efficiency and will increase genetic gains per unit of time. A prerequisite of this strategy is the availability of a saturated genetic linkage map for the studied species.

Previous reviews have described the specificity of the different mapping strategies used in forest trees [14, 42]. A comprehensive review of inheritance and mapping studies in conifers, indicating the type of pedigree and marker techniques used, is also available at: http: //www.pierroton.inra.fr/genetics/labo/mapreview.html. Chronologically, inheritance and mapping studies were performed using the megagametophyte, a nutritive haploid tissue surrounding the embryo of gymnosperm seeds and corresponding to the female inheritance transmitted to the embryo [63]. Markers used by the forest tree geneticists in the 70's and 80's were isozymes [1]. However, a large proportion of the genome could not be covered by a too few number of loci. The use of this haploid tissue climaxed in the mid-90's, when randomly amplified polymorphic DNA (RAPD, [68]) became the most popular marker technique to produce genetic maps for plant species. In particular, the haploid megagametophyte of conifer seeds avoided the drawback of the dominant nature of RAPDs. The "megagametophyte-RAPD" strategy was used in several conifer species, including *P. pinaster* [44], from which the first conifer saturated map was published. In the late 90's, RAPDs were progressively abandoned with the availability of a more reliable technique: Amplified Fragment Length Polymorphism (AFLP, [66]), which was used in several conifer species such as pinyon pine [62], loblolly pine [51] and maritime pine [18, 53]. Although very popular in the forest geneticist community, the megagametophyte approach faces two major limitations. First, it requires the development of specific populations and is not applicable to QTL analysis for mature traits in existing plantations. Indeed, the megagametophyte is a temporary tissue that can only be collected from the seedling stage during the germination of the embryo. Therefore, the dissection of the genetic architecture of adult trait would require several years to start. In addition, only the maternal effect of QTL can be estimated [45, 46]. Second, the haploid progeny cannot be considered as a "perpetual" mapping population, because of the relatively low amount of DNA that can be extracted from this tissue. Consequently, it will prevent a high number of markers, as well as markers requiring a high amount of DNA such as RFLPs (Restriction Fragment Length Polymorphisms, [8]), from being mapped over time.

Conversely, adult trees can be grafted and propagated by cuttings, and diploid progenies can constitute "perpetual" population, analogous to Recombinant Inbred Lines in crop plants. Carlson et al. [11] were the first to show that RAPD primers could be screened for informative markers segregating in a 1:1 ratio in diploid tissue of full-sib progenies. This idea was extended by Grattapaglia and Sederoff [24] to construct parental maps of an interspecific eucalyptus hybrid family, in a mapping strategy named "two-way pseudo-testcross". It was further used in conifers [3, 33] with RAPDs and AFLPs.

Although dominant biallelic markers (RAPD and AFLP) continue to be the most easy-to-use technique, they present major limitation since they cannot capture the multiallelic nature of QTLs. Alternatively, other research groups started to use codominant markers such as RFLPs [10, 19, 54], PCR-RFLP [26], ESTs (Expressed Sequence Tags) [12, 47, 61] and more recently SSRs (Simple Sequence Repeats) [19, 22, 43], allowing gene action to be precisely defined (estimation of additive and dominant effects of QTLs, [55]) and providing anchor points in comparative mapping experiments [39].

This brief review of the history of molecular marker development can give us insights on how to proceed in the development of a molecular genetics project in maritime pine. In a first step we identified a three-generation outbred pedigree comprising 202 individuals and segregating for traits of interest. Second, we quickly established a fully saturated map based on AFLP markers. Third, we are now mapping QTL for traits of interest and developing SSRs and ESTPs (EST Polymorphisms) to provide more informative markers which should be easily transferred to other pedigrees of maritime pine and other pine species, with the main objective of QTL validation [39]. The main goal of this paper is to present a saturated map of maritime pine which corresponds to the second step of this strategy.

2. MATERIALS AND METHODS

2.1. Mapping population

A three-generation outbred pedigree (9.103.3 × 10.159.3) was used to construct the genetic map (*figure 1*). The two parental trees were mated in 1980 and seeds from the controlled cross were sown in spring 1982. They produced 202 progeny seedlings that were

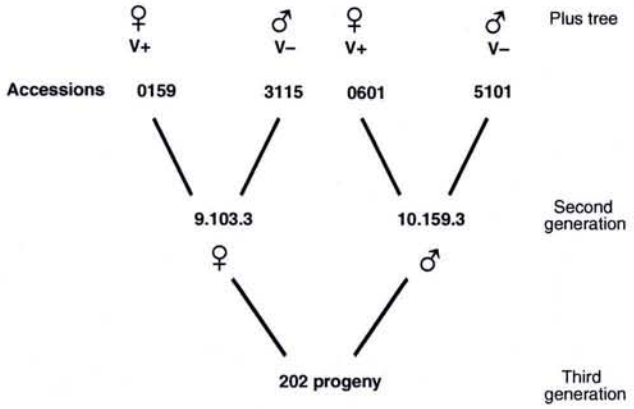


Figure 1. Mapping pedigree: 202 full-sibs belonging to the third generation of the maritime pine breeding programme (V+: vigorous trees, V−: less vigorous trees).

planted in autumn 1982. The four grand parents were "plus trees" phenotypically selected for stem growth and straightness in the local provenance of the Landes de Gascogne, and grafted in clonal orchards. These grand parents were tested in a polycross progeny test and classified according to their breeding value as Vigor "+" (for vigorous trees) and Vigor "−" (for less vigorous trees). The progeny was located in Malente (Gironde, France) on a semi-humid podzolic soil. Spacing was 4 m between rows and 1.1 m between individual trees, i.e. 2 272 trees ha^{-1}. This full-sib family belongs to a progeny test of the third generation breeding population.

2.2. AFLP assay and gel electrophoresis

Genomic DNA was extracted as described by Doyle and Doyle [21]. AFLP markers were obtained following the protocol of Costa et al. [18] with slight modifications: the *Eco*RI primers used for selective amplification were radio labelled for 1.0 h at 37 °C in 1 × OPA buffer (Pharmacia), 9.5 U of T4 kinase (Pharmacia), 100 µM of primer and 10 µCi of γ^{33}P-ATP. The reaction was stopped by incubating the reaction mix for 10 min at 80 °C. After selective amplification, 4 µl of denaturated template was loaded, after one hour of pre-run, on to 52 cm gels composed of 4% 19:1 acrylamide: bis-acrylamide, 7 M urea and 1 × TBE. The run was performed at 80 W for 150 min or more, depending on the primer combination. The gel was fixed after running in 10% acetic acid for 20 min, rinsed in distilled water and dried overnight at 50 °C. Finally, gels were exposed on Konica AX autoradiographic film for about 8 days.

Fifty-two primer-enzyme combinations (PEC, see *table I*) were chosen on the basis of their repeatability, pattern (i.e. ease of scoring) and level of polymorphism. Presence or absence of AFLP fragments was directly scored on the gel image (*figure 2*).

Polymorphic AFLP fragments were named considering (1) the PEC used; (2) the fragment length, and (3) the quality of the scoring: "a" for intense bands, "b" for weak bands, and "c" for the bands that were difficult to score. A table of correspondence between the locus ID and the PEC used is available online at URL: http://www.pierroton.inra.fr/genetics/pinus/Map3/marker_table.html.

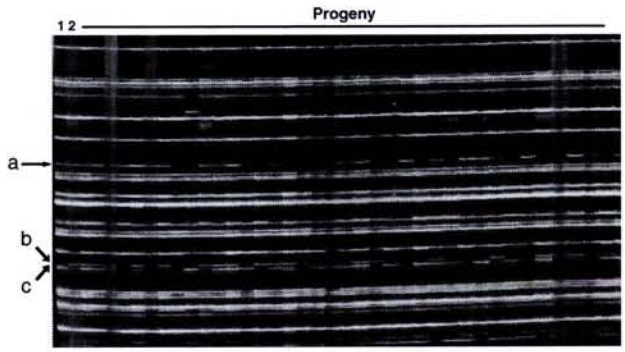


Figure 2. Example of AFLP profile showing the three types of segregation. Lanes 1 and 2 correspond to the parents (female and male) and other lanes correspond to the full-sib progeny. (A) Inter-cross marker, heterozygous in both parents and segregating 3:1 in the progeny; (B) Test-cross marker, heterozygous in the male and absent in the female, and segregating 1:1 in the progeny; (C) Test-cross marker, heterozygous in the female and absent in the male, and segregating 1:1 in the progeny.

2.3. Mapping procedure

We used the two-way pseudo-test cross mapping strategy to construct the linkage maps [24]. Markers were subdivided into two groups considering their segregation patterns. The first group comprised markers in the testcross configuration between the parents (heterozygous in one parent and homozygous null in the other), which presented a 1:1 segregation ratio in the progeny. The second group concerned markers heterozygous in both parents, and therefore segregating in a 3:1 ratio in the progeny. Mendelian segregation of the markers was tested by chi-square tests ($P > 0.01$). The few distorted 1:1 and 3:1 markers were discarded from further analysis. They generally belonged to the "c" quality score category.

Because of the low information content between pairs of markers segregating in the 1:1 and 3:1 configuration [52], a preliminary grouping of the 1:1 markers only was performed for each parent using MAPMAKER software [34] with a LOD threshold of 6. Our objective was to construct precise parental maps with 1:1 markers to compare with the results obtained later with JOINMAP. The two parental maps based on 1:1 and 3:1 markers were built using JOINMAP v1.4 software [57] with a minimum LOD of 3 used as grouping criterion and then aligned based on 3:1 markers. Whenever the ordering of 3:1 markers was disturbed, the corresponding markers were discarded until a good ordering was obtained. A consensus map was finally built using all 1:1 and selected 3:1 markers using JOINMAP. Linkage groups were drawn using MAPCHART [65]. Recombination rates were converted to map distances in centiMorgans (cM) using the Kosambi mapping function.

2.4. Physical size measurement

DNA content of embryos or megagametophytes was assessed by flow cytometry. Ten seeds were first imbibed overnight and then dissected to separate the megagametophyte from the embryo. *Triticum aestivum* (2C = 30.9 pg, [35]) was used as an internal standard. *Pinus* tissues and hexaploid wheat leaf were chopped together with a razor blade in Galbraith buffer [23] slightly modified by the addition of 10 mM metabisulfite, 1% (w/v), Triton X-100 and

Table I. List of AFLP primer pairs used to construct the maritime pine genetic map and number of polymorphic fragments.

PEC	Number of amplified fragments	Number of markers segregating 1:1 (1)	Number of markers segregating 3:1 (2)	Total (1)+(2)
1 E+ACA/M+CCAG	140	10	7	17
2 E+ACA/M+CCGA	130	9	7	16
3 E+ACG/M+CCGC	108	8	1	9
4 E+ACG/M+CCAG	60	9	5	14
5 E+ACG/M+CCGT	95	10	1	11
6 E+ACG/M+CCTA	56	4	4	8
7 E+ACG/M+CCCA	112	5	3	8
8 E+ACG/M+CCAA	118	12	1	13
9 E+ACG/M+CCTG	62	26	3	29
10 E+ACC/M+CCAG	140	13	5	18
11 E+ACC/M+CCTG	126	9	5	14
12 E+ACC/M+CCGT	70	7	1	8
13 E+ACC/M+CCTA	145	10	4	14
14 E+ACC/M+CCGA	110	8	2	10
15 E+ACT/M+CCGC	130	9	2	11
16 E+ACT/M+CCAG	110	11	1	12
17 E+ACT/M+CCTG	120	14	7	21
18 E+ACT/M+CCGT	130	7	5	12
19 E+ACT/M+CCCA	136	4	4	8
20 E+ACT/M+CCTA	105	9	6	15
21 E+ACAA/M+CCTA	95	6	9	15
22 E+ACAA/M+CCAC	100	6	5	11
23 E+ACAA/M+CCGC	140	9	3	12
24 E+ACAA/M+CCCA	140	12	5	17
25 E+ACAA/M+CCGA	110	4	5	9
26 E+ACAA/M+CCTT	136	17	4	21
27 E+ACAA/M+CCTG	70	9	4	13
28 E+ACAA/M+CCAG	75	10	4	14
29 E+ACAA/M+CCAT	110	12	4	16
30 E+ACAC/M+CCAA	100	9	2	11
31 E+ACAC/M+CCAT	130	16	10	26
32 E+ACAC/M+CCTA	100	7	2	9
33 E+ACAC/M+CCTT	100	12	8	20
34 E+ACAC/M+CCTC	90	9	3	12
35 E+ACAC/M+CCAG	123	9	3	12
36 E+ACAC/M+CCAC	100	13	4	17
37 E+ACAG/M+CCTG	114	15	4	19
38 E+ACAG/M+CCTA	107	9	3	12
39 E+ACAG/M+CCAT	104	18	9	27
40 E+ACAG/M+CCAA	99	14	3	17
41 E+ACAG/M+CCGA	120	6	6	12
42 E+ACAG/M+CCTC	110	5	7	12
43 E+ACAG/M+CCGT	130	3	6	9
44 E+ACAG/M+CCGC	145	9	5	14
45 E+ACAT/M+CCAG	115	10	7	17
46 E+ACAT/M+CCTA	110	15	9	24
47 E+ACAT/M+CCAT	132	7	6	13
48 E+ACAT/M+CCTC	143	20	9	29
49 E+ACAT/M+CCTG	105	2	6	8
50 E+ACAT/M+CCAC	138	13	9	22
51 E+ACAT/M+CCCA	145	6	8	14
52 E+ACAT/M+CCGA	115	7	7	14
TOTAL	5854	513	253	766

1% (w/v) polyethylene glycol (PEG) 8000. After addition of 5 units mL^{-1} RNase A (Roche, France) and 50 µg mL^{-1} propidium iodide (Sigma-Aldrich, France), nuclei were filtered through a 75 µm nylon filter in order to eliminate cell debris. Samples were left 30 min on ice before measurements.

Assuming a linear relationship between fluorescence ratio and amount of DNA, total 2C DNA content was evaluated using the leaf 2C DNA value of hexaploid wheat. For each sample, measurements were made on 2 500 nuclei with duplication. Fluorescence analysis of the stained nuclei was performed on an Epics V cytometer (Beckman-Coulter, Roissy, France) with an argon laser at 488 nm for propidium iodide. The cytometer linearity was checked and adjusted before each set of run.

3. RESULTS AND DISCUSSION

3.1. AFLP markers

The 52 PECs used in this study provided 766 non-distorted AFLP markers. The number of polymorphic fragments ranged from 8 to 29 with an average of 15 polymorphic markers per combination. 253 (33%) markers segregated in the 3:1 ratio and 513 (66%) in the 1:1 ratio. A total of 251 (32.8%) of these 513 markers were heterozygous for the male parent and 262 (34.2%) for the female parent.

In a short time, and for a rather low cost, the AFLP method provided a sufficient amount of polymorphic markers to saturate the genome of maritime pine. In spite of its large genome size, the use of appropriate PECs allowed the production of easy-to-score AFLP gels. The use of *Pst-Mse* PECs has been reported to provide less complex gel patterns but also yields non-randomly distributed markers in conifers [43]. *Pst*I is sensitive to methylation and the use of this endonuclease may target low-copy clustered regions. To avoid this problem and ensure full genome coverage, we used *Eco – Mse* PECs. By using two selective nucleotides in the pre-amplification step (*Eco*RI + 2, *Mse*I + 2), and three to four nucleotides in the selective amplification step (*Eco*RI + 3 / +4 *Mse*I + 4), we could circumvent the complexity of the pine genome to produce clear AFLP patterns [15, 25].

Remington et al. [51] reported a significant effect of the composition of the selective extensions. They showed that the amount of CpG was negatively correlated with the number of polymorphic fragments. In this study, although a slight decrease was also observed, an analysis of variance (not shown) test showed that there were no significant relationship between the number of polymorphic bands and the CpG content in both *Eco*RI and *Mse*I primers (P-value = 0.21).

3.2. Linkage map

Some polymorphic markers were discarded from the linkage analysis because they were distorted ($P < 0.01$). It should be noticed that the observed level of distorsion was not significantly greater than that expected by chance alone. In

respect to the 3:1 markers, only a subset (42%) that showed the same order in the parental maps were kept. Six hundred and twenty markers were finally used to construct the consensus linkage map (*figure 3*). The map consisted of 12 linkage groups, corresponding to the 12 haploid chromosomes of *P. pinaster*.

The total lengths obtained for the female, the male and the consensus maps using JOINMAP and MAPMAKER softwares are presented in *table II*. The total genetic length calculated using MAPMAKER software on the female map (1 807 cM) is not significantly different from those described by Plomion et al. [44] and Costa et al. [18] on the same species (1 860 cM and 1 873 cM, respectively). On the other hand, the comparison between the total genetic lengths obtained with JOINMAP or MAPMAKER are different, even if the same mapping function (Kosambi) was used in both software. Qi et al. [49] in barley and Sewell et al. [54] in loblolly pine reported the same phenomenon, which can be attributed to how the software packages calculate the genetic distances: in any case the assumed level of interference differs slightly from the true interference.

3.3. Physical versus genetic size

Improvements of the extraction buffer allowed analysis of fair quality with a highly reproducible fluorescence index ($2C_{Pinus}/2C_{standard}$). Analysis of *P. pinaster* embryo tissues yielded DNA histograms with coefficients of variation in the 2C peaks ranging from 2 to 4%. Hexaploid wheat was used as an internal standard because its genome size is relatively high and thus more convenient in the assessment of large genome. The mean DNA value (2C) for *P. pinaster* was 51.49 ± 0.51 pg. The ratio between the fluorescence peak of nuclei isolated from the diploid *P. pinaster* embryos and the corresponding megagametophyte haploid tissue was equal to 1.92.

Table II. Total genetic lengths and number of linkage group (LG) obtained for female, male and consensus maps using two different mapping softwares.

	JOINMAP	MAPMAKER
Female	1218 cM (12 LG)	1807 cM (12 LG)
Male	1297 cM (15 LG)	1541 cM (16 LG)
Consensus	1407 cM (12 LG)	–

The *Pinaceae* presents the widest range and diversity of DNA contents in all gymnosperm families [30, 37, 40, 41]. *P. pinaster*, with a 2C DNA value of 51.49 pg/2C (25.7×10^9 base pair per 1C) is close to most of the *Pinus* species. The highest DNA reported in *Pinus* genus and also in gymnosperm is 63.5 pg/2C in *Pinus lambertiana* [37]. For the moment, it is not clear if the large diversity of the *Pinus* genome sizes procures an advantage to environmental conditions as hypothesised by Wakamiya et al. [67].

Table III compares the genetic and physical size of maritime pine and several other plant genome, including forest trees belonging to angiosperms (oak [6], poplar [16], eucalyptus [36]) and gymnosperms (Loblolly pine [54]). The two pine species show higher physical lengths compared to the other species, which translates into a much larger physi cal/genetic size ratio (e.g.: 13.78 Mbp/cM in *P. pinaster* versus 0.22 Mbp/cM in *Arabidopsis thaliana*). *Figure 4* shows the relationship between the number of crossing-over and the mean physical size of a chromosome. The number of crossing-over is highly negatively correlated with chromosome size ($R = -0.88$, $P < 0.01$). As the number of crossing-over occurring during the meiosis does not differ strongly between

Table III. Genome characteristics of 15 plant species.

	Species	Physical size (Mb)	Genetic length (cM) (MAPMAKER estimates)	Chromosome number	Mean genetic size per chromosome (cM)	Physical/genetic size ratio (Mb/cM)
1	*Arabidopsis thaliana*	150[4]	675[50]	5	135	0.22
2	*Prunus persica*	300[4]	712[20]	8	90	0.42
3	*Oryza sativa*	450[4]	1490[2]	12	125	0.3
4	*Populus deltoides*	550[4]	2300[16]	19	121	0.23
5	*Eucalyptus grandis*	600[4]	1370[64]	11	125	0.43
6	*Brassica rapa*	650[4]	1850[56]	10	185	0.35
7	*Quercus robur*	900[4]	1200[6]	12	100	0.75
8	*Lycopersicon esculentum*	980[4]	1280[57]	12	107	0.76
9	*Solanum tuberosum*	1540[4]	1120[58]	12	93	1.37
10	*Zea mays*	2500[4]	1860[13]	10	186	1.34
11	*Lactuca sativa*	2730[4]	1950[27]	9	217	1.4
12	*Triticum tauschii*	4200[7]	1330[38]	7	190	3.15
13	*Hordeum vulgare*	5500[7]	1250[29]	7	178	4.4
14	*Pinus taeda*	21000[4]	1700[19]	12	141	12.35
15	*Pinus pinaster*	25700	1850[18]	12	154	13.78

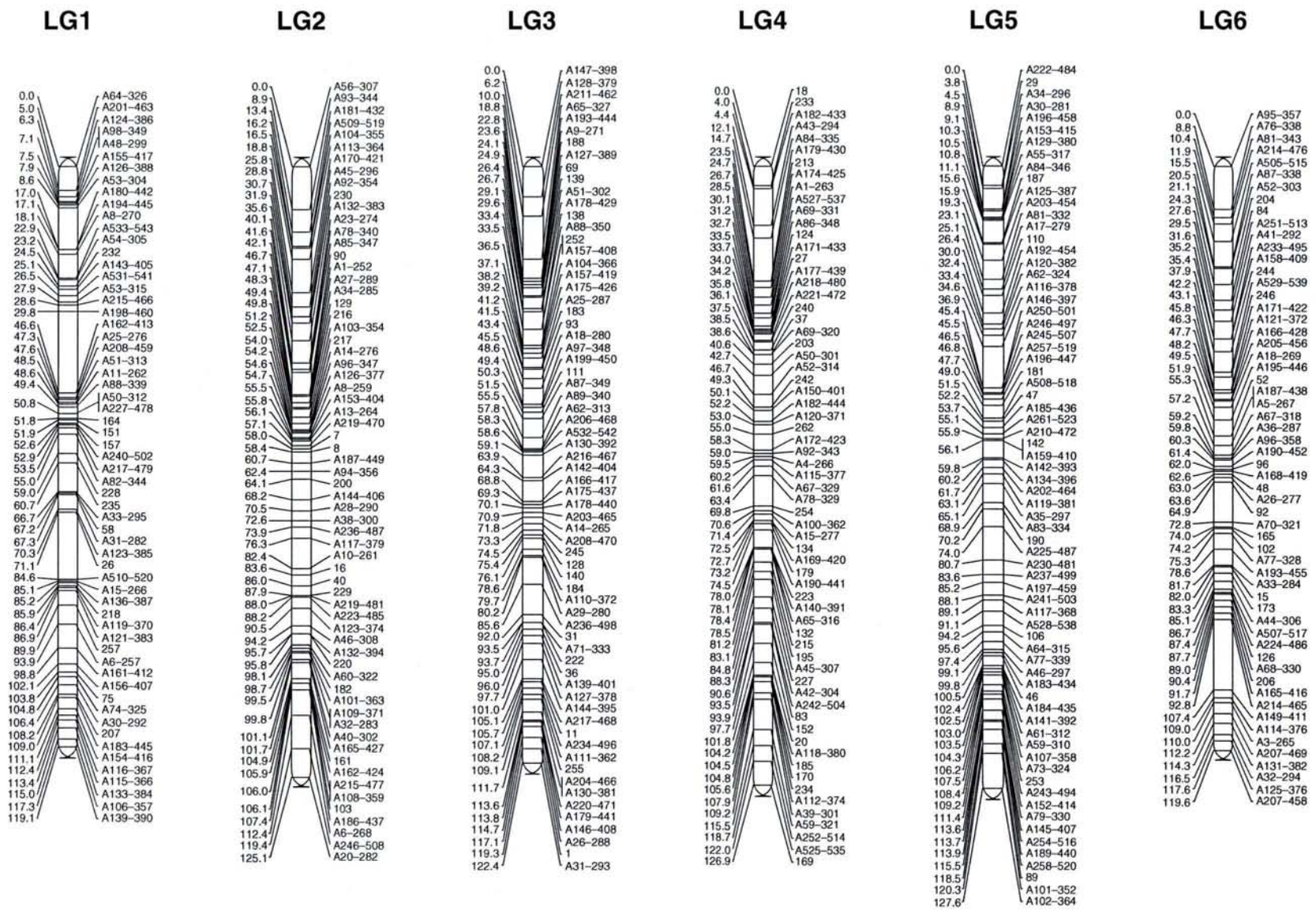


Figure 3. Consensus map based on 620 AFLP markers. Correspondence between marker ID and PEC is available at: http://www.pierroton.inra.fr/genetics/pinus/Map3/marker_table.html. Male and female maps are also available at the same URL.

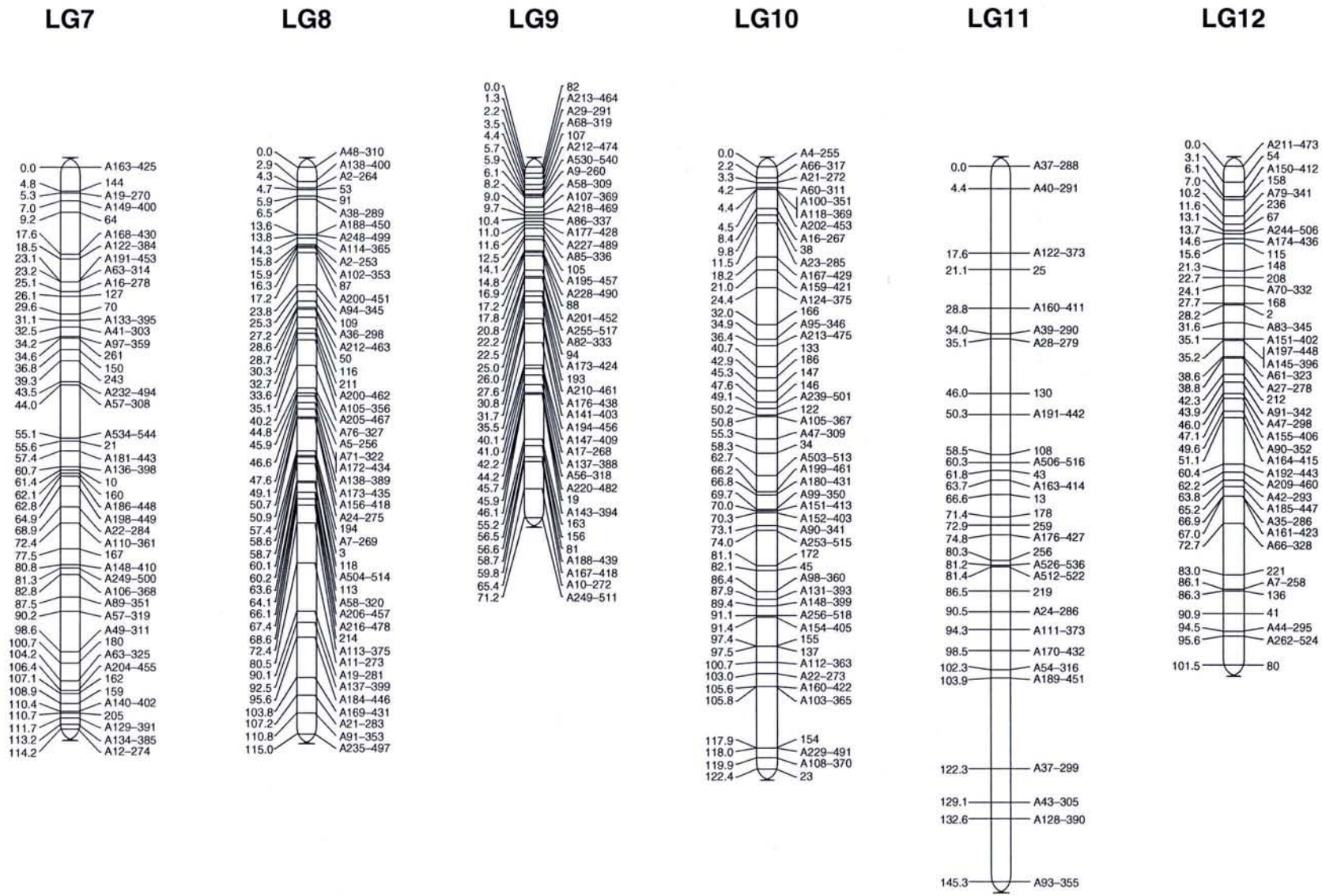


Figure 3. Continued.

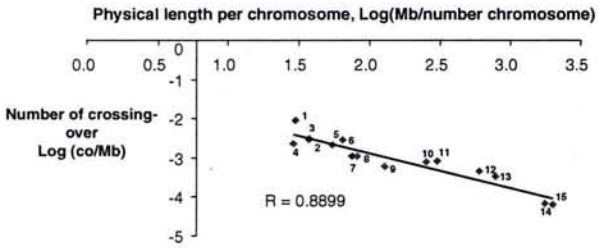


Figure 4. Relationship between the number of crossing-over and the mean physical size per chromosome in 15 plant species. The correspondence between the number and the 15 species is presented in *table III*.

species (*table III*), species with small chromosomes will present a larger amount of recombination per unit of physical size.

4. PERSPECTIVES

The new maritime pine genetic map provides a very useful tool for further genetic analysis. First, this map will serve as a framework to locate comparative anchor tags for comparative genomics. Although AFLP markers have been shown to be poorly transferable between pine species, orthologous markers such as RFLPs, ESTPs [61] or SSRs can be used as anchor-points between the different maps already available for conifer species. ESTs which have been mapped in *Pinus taeda* [26, 61] and ESTs from *P. pinaster* cDNA libraries are currently being located in the AFLP map of maritime pine as part of the Conifer Comparative Genome Project (CCGP; http://dendrome.ucdavis.edu/Synteny/index.html). The aim of CCGP is to compare conifer genetic maps with the *P. taeda* reference map by providing orthologous markers. A hierarchical approach based on different PCR-based methods is used to detect polymorphism in ESTs: PCR fragment length and conformation in denaturing or non-denaturing gel conditions (SSCP [47] and DGGE [60]) are first used because of their low or medium cost and time efficiency. More powerful methods such as point mutation detection by systematic sequencing, or such as the prospecting of variation in the non-coding regions flanking the ESTs [12], will also be used to increase polymorphism rate.

As for the "intraspecific mapping comparison", some of the AFLP markers will be transferable between pedigrees of maritime pine, but to compare maps constructed based on different genetic backgrounds (e.g.: using experimental design such as factorial and diallel), SSRs will be the marker of choice. Their multiallelic nature will also allow tagging multiple alleles at QTLs. Development of a battery of SSRs for maritime pine is therefore a priority.

Secondly, genomic regions controlling adaptive and economically important traits are currently being studied in maritime pine. These include QTLs for growth, wood quality, end-uses properties and water use efficiency [9]. These studies are based on a skeleton map based on evenly spaced AFLP markers genotyped on the whole mapping pedigree (202 full-sibs; Pot, unpublished). The ESTs described in the previous paragraph will also provide positional candidate genes, i.e. whose position coincides with mapped QTLs. However, because of the high physical/genetic size ratio in conifers, it will be of great importance to find the actual genes underlying QTLs of interest, before any attempt of using this information in Marker-Assisted Breeding Program. The location of candidate genes will also contribute to the establishment of a "functional" genetic map.

In an integrative study, it will be essential to use the same markers (ESTs) for comparative mapping and the candidate gene approach, in order to validate the candidate gene-QTL co-locations between phylogenetically related species [39].

Acknowledgements: The authors are grateful to the reviewers for comments on the manuscript. This work was supported by funding from the European Union (ANACONGEN, BI04-CT97-2125) and the French Ministry of Research (BIOTECH, décision n° 98C0204).

REFERENCES

[1] Adams W.T., Joly R.J., Linkage relationships among twelve allozyme loci in loblolly pine, J. Hered. 71 (1980) 199–202.

[2] Ahn S., Tanskley S., Comparative linkage maps of the rice and maize genomes, Proc. Natl. Acad. Sci. USA 90 (1993) 7980–7984.

[3] Arcade A., Anselin F., Faivre Rampant P., Lesage M.C., Laurans F., Paques L.E., Prat D., Application of AFLP, RAPD and ISSR markers to genetic mapping of European larch and Japanese larch, Theor. Appl. Genet. 100 (2000) 299–307.

[4] Arumuganathan K., Earle E.D., Nuclear DNA content of some important plant species, Plant Mol. Biol. Rep. 9 (1991) 208–218.

[5] Baradat P., Pastuszka P., Le pin maritime, in: Gallais A., Bannerot H. (Éds.), Amélioration des espèces végétales cultivées, INRA édition, Paris, 1992, pp. 695–709.

[6] Barreneche T., Bodénès C., Lexer C., Trontin J.F., Fluchs S., Streiff R., Plomion C., Roussel G., Steinkellner H., Burg K., Favre J.-M., Glossl J., Kremer A., A genetic linkage map of *Quercus robur* (Pedunculate oak) based on RAPD, SCAR, microsatellite, isozyme and rDNA markers, Theor. Appl. Genet. 97 (1998) 1090–1103.

[7] Bennett M.D., Smith J.B., Nuclear DNA amount in angiosperms, Phil. Trans. R. Soc. Lond. B 274 (1976) 227–274.

[8] Botstein D., White R., Skolnick M., Davis R., Construction of genetic linkage map in man using restriction fragment length polymorphism, Am. J. Hum. Genet. 32 (1980) 641–656.

[9] Brendel O., Pot D., Plomion C., Rozenberg P., Guelh J.-M., Genetic parameters and QTL analysis of $\delta^{13}C$ and ring width in maritime pine, Plant Cell Env. 25 (2002) 945–953.

[10] Byrne M., Murell J.C., Allen B., Moran G.F., An integrated genetic linkage map for eucalypts using RFLP, RAPD and isozyme markers, Theor. Appl. Genet. 91 (1995) 869–875.

[11] Carlson J.E., Tulsieram L.K., Glaubitz J.C., Luk V.W.K., Kauffeld C., Rutledge R., Segregation of random amplified DNA markers in F1 progeny of conifers, Theor. Appl. Genet. 83 (1991) 194–200.

[12] Cato S.A., Gardner R.C., Kent J., Richardson T.E., A rapid PCR-based method for genetically mapping ESTs, Theor. Appl. Genet. 102 (2001) 296–306.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

[13] Causse M., Santoni S., Damerval C., Maurive A., Charcosset A., Deatrick J., de Vienne D., A composite map of expressed sequences in maize, Genome 39 (1996) 418–432.

[14] Cervera M.T., Plomion C., Malpica C., Molecular markers and genome mapping in woody plants, in: Jain S.M., Minocha S.C. (Eds.), Molecular Biology of Woody Plants, Vol. I, 2000, pp. 375–394.

[15] Cervera M.T., Remington D., Frigerio J.M., Storme V., Ivens B., Boerjan W., Plomion C., Improved AFLP analysis of tree species, Can. J. For. Res. 30 (2000) 1608–1616.

[16] Cervera M.T., Storme V., Ivens B., Gusmao J., Liu B.H., Hostyn V., Van Slycken J., Van Montagu M., Boerjan W., Dense genetic linkage maps of three populus species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers, Genetics 158 (2001) 787–809.

[17] Costa P., Durel C.E., Time trends in genetic control over height and diameter in maritime pine, Can. J. For. Res. 26 (1996) 1209–1217.

[18] Costa P., Pot D., Dubos C., Frigerio J.-M., Pionneau C., Bodénès C., Bertocchi E., Cervera M.T., Remington D.L., Plomion C., A genetic map of maritime pine based on AFLP, RAPD and protein markers, Theor. Appl. Genet. 100 (2000) 39–48.

[19] Devey M.E., Bell J.C., Smith D.N., Neale D.B., Moran G.F., A genetic map for *Pinus radiata* based on RFLP, RAPD and microsatellite markers, Theor. Appl. Genet. 92 (1996) 673–679.

[20] Dirlewanger E., Pronier V., Parvery C., Rothan C., Guye A., Monet R., Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers, Theor. Appl. Genet. 97 (1998) 888–895.

[21] Doyle J.J., Doyle J.L., Isolation of plant DNA from fresh tissue, Focus 12 (1990) 13–15.

[22] Echt C.S., Nelson C.D., Linkage mapping and genome length in eastern white pine (*Pinus strobus* L.), Theor. Appl. Genet. 94 (1997) 1031–1037.

[23] Galbraith D., Harkins K., Maddox J., Ayres N., Sharma D., Firoozabady E., Rapid flow cytometric analysis of the cell cycle in intact plant tissues, Science 220 (1983) 1049–1051.

[24] Grattapaglia D., Sederoff R., Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus europhylla* using a pseudo testcross: mapping strategy and RAPD markers, Genetics 137 (1994) 1121–1137.

[25] Han T.H., van Eck H.J., de Jeu M.J., Jacobsen E., Optimization of AFLP fingerprinting of organisms with large-sized genome: a study on *Alstoemeria* spp, Theor. Appl. Genet. 98 (1999) 465–471.

[26] Harry D.E., Temesgen B., Neale D.B., Codominant PCR-based markers for *Pinus taeda* developed from mapped cDNA clones, Theor. Appl. Genet. 97 (1998) 327–336.

[27] Hulbert S.H., Hott T.W., Legg E.J., Lincoln S.E., Lander E.S., Michelmore R.W., Genetic analysis of the fungus, *Bremia lactucae*, using restriction length polymorphism, Genetics 120 (1988) 947–958.

[28] Jacobs J.M.E., van Eck H.J., Arens P., Verkerk-Bakker B., te Lintel Hekkert B., Bastiaansen H.J.M., El-Karbotly A., Pereira A., Jacobsen E., Stiekema W.J., A genetic map of potato (*Solanum tuberosum*) integrating molecular markers, including transposons and classical markers, Theor. Appl. Genet. 91 (1995) 239–300.

[29] Kleinhofs A., Kilian A., Saghai Maroof M.A., Biyashev R.M., Hayes P., Chen F.Q., Lapitan N., Fenwick A., Blake T.K., Kanazin V., Ananiev E., Dahleen L., Kudrna D., Bollinger J., Knapp S.J., Liu B., Sorrells M., Heun M., Franckowiack J.D., Hoffmann D., Skadsen R., Steffenson B.J., A molecular, isozyme and morphological map of barley (*Hordeum vulgare*) genome, Theor. Appl. Genet. 86 (1993) 705–712.

[30] Kohler B., Guttenberger H., Borzan Z.G., C-banding pattern of the chromosomes in the macrogametophyte of Norway spruce, Silvae Genet. 45 (1996) 16–21.

[31] Kremer A., Lascoux D.M., Genetic architecture of height growth in maritime pine (*Pinus pinaster* Ait.), Silvae Genet. 37 (1988) 1–8.

[32] Kremer A., Lascoux M., Nguyen A., Morphogenetic subdivision of height growth and early selection in maritime pine, Proceedings of the 21st Southern Forest Tree Improvement Conference, 1991, pp. 203–221.

[33] Kubisiak T.L., Nelson C.D., Nance W.L., Stine M., RAPD linkage mapping in a longleaf pine × slash pine F1 family, Theor. Appl. Genet. 90 (1996) 1110–1127.

[34] Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E., Newberg L., MAPMAKER: an interactive computer package for constructing primary genetic maps of experimental and natural populations, Genomics 1 (1987) 174–181.

[35] Marie D., Brown S., A cytometric exercise in plant DNA histograms, with 2C values for 70 species, Bio. Cell. 78 (1993) 41–51.

[36] Marques C.M., Araujo J.A., Ferreira J.G., Whetten R., O'Malley D.M., Liu B.H., Sederoff R., AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*, Theor. Appl. Genet. 96 (1998) 727–737.

[37] Murray B., Nuclear DNA amounts in gymnosperms, Ann. Bot. 83 (1998) 3–15.

[38] Namuth D.M., Lapitan N.L.V., Gill K.S., Gill B.S., Comparative RFLP mapping of *Hordeum vulgare* and *Triticum tauschii*, Theor. Appl. Genet. 89 (1994) 865–872.

[39] Neale D.B., Sewell M.M., Brown G.R., Molecular dissection of the quantitative inheritance of wood property traits in loblolly pine, Ann. For. Sci. 59 (2002) 595–605.

[40] O'Brien I., Smith D., Gardner R., Murray B., Flow cytometric determination of genome size in *Pinus*, Plant Sci. 115 (1996) 91–99.

[41] Ohri D., Khoshoo T.N., Genome size in gymnosperms, Pl. Syst. Evol. 153 (1986) 119–132.

[42] O'Malley D.M., Whetten R., Moleculars markers and forest trees, in: Caetano-Anollès G., Gresshoff P.M. (Eds.), DNA markers: Protocols, Applications and Overviews, 1997, pp. 237–257.

[43] Paglia G.P., Olivieri A.M., Morgante M., Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.), Mol. Gen. Genet. 258 (1998) 466–478.

[44] Plomion C., Bahrman N., Durel C.E., O'Malley D.M., Genomic mapping in *Pinus pinaster* (maritime pine) using RAPD and protein markers, Heredity 74 (1995) 661–668.

[45] Plomion C., Durel C.E., Estimation of the average effects of specific alleles detected by the pseudo-testcross QTL mapping strategy, Genet. Sel. Evol. 28 (1996) 223–235.

[46] Plomion C., Durel C.E., O'Malley D.M., Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions, Theor. Appl. Genet. 93 (1996) 849–858.

[47] Plomion C., Hurme P., Frigerio J.-M., Ridolfi M., Pot D., Pionneau C., Avila C., Gallardo F., David H., Neutelings G., Campbell M., Canovas F.M., Savolainen O., Bodénès C., Kremer A., Developing SSCP markers in two *Pinus* species, Mol. Breed. 5 (1999) 21–31.

[48] Pot D., Chantre G., Rozenberg P., Rodrigues J.C., Jones G.L., Pereira H., Hannrup B., Cahalan C., Plomion C., Genetic control of pulp and timber properties in maritime pine (*Pinus pinaster* Ait.), Ann. For. Sci. 59 (2002) 563–575.

[49] Qi X., Stam P., Lindhout P., Comparison and integration of four barley genetic maps, Genome 39 (1996) 379–394.

[50] Reiter R.S., Williams J., Feldman K., Rafalski J.A., Tingey S.V., Scolnik P.A., Global and local genome mapping in *Arabidopsis thaliana* recombinant inbred lines and random amplified polymorphic DNAs, Proc. Natl. Acad. Sci. USA 89 (1992) 1477–1481.

[51] Remington D.L., Whetten R.W., Liu B.H., O'Malley D.M., Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*, Theor. Appl. Genet. 98 (1999) 1279–1292.

[52] Ritter E., Gebhardt C., Salamini F., Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents, Genetics 125 (1990) 645–654.

[53] Ritter E., Aragonés A., Markussen T., Acheré V., Espinel S., Fladung M., Wrobel S., Faivre-Rampant P., Jeandroz S., Favre J.-M., Towards construction of an ultra high density linkage map for *Pinus pinaster*, Ann. For. Sci. 59 (2002) 637–643.

[54] Sewell M.M., Sherman B.K., Neale D.B., A consensus map for loblolly Pine (*Pinus taeda* L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees, Genetics 151 (1999) 321–330.

[55] Sewell M.M., Bassoni D.L., Megraw R.A., Wheeler N.C., Neale D.B., Identification of QTLs influencing wood property traits in loblolly pine

(*Pinus taeda* L.). I. Physical wood properties, Theor. Appl. Genet. 101 (2000) 1273–1281.

[56] Song K.M., Suzuki J.Y., Slocum M.K., Williams P.H., Osnorn T.C., A linkage of *Brassica rapa* (*syn. Campestris*) based on restriction fragment length polymorphism loci, Theor. Appl. Genet. 82 (1991) 296–304.

[57] Stam P., Construction of integrated linkage maps by means of a new computer package: JOINMAP, Plant J. 3 (1993) 739–744.

[58] Tanksley S.D., Ganal M.W., Prince J.P., de Vicente M.C., Bonerbiale M.W., Broun P., Fulton T.M., Giovannoni J.J., Grandillo S., Martin G.B., Messeguer R., Miller J.C., Miller L., Paterson A.H., Pineda O., Roder M.S., Wing R.A., Wu W., Young N.D., High dendity molecular linkage maps of the tomato and potato genomes, Genetics 132 (1992) 1141–1160.

[59] Tanksley S.D., Mapping Polygenes, Ann. Rev. Genet. 27 (1993) 205–233.

[60] Temesgen B., Neale D.B., Harry D.E., Use of haploid mixtures and heteroduplex analysis enhance polymorphism revealed by Denaturing Gradient Gel Electrophoresis, BioTechniques 20 (2000) 114–122.

[61] Temesgen B., Brown G.R., Harry D.E., Kinlaw C.S., Sewell M.M., Neale D.B., Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.), Theor. Appl. Genet. 102 (2001) 664–675.

[62] Travis S.E., Ritland K., Whitman T.G., Keim P., A genetic linkage map of Pinyon pine (*Pinus edulis*) based on amplified fragment length polymorphism, Theor. Appl. Genet. 97 (1998) 871–880.

[63] Tulsieram L.K., Glaubitz J.C., Kiss G., Carlson J.E., Single tree genetic linkage analysis in conifers using haploid DNA from megagametophytes, BioTechnology 10 (1992) 686–690.

[64] Verhaegen D., Plomion C., Genetic mapping in *Eucalyptus urophylla* and *E. grandis* using RAPD markers, Genome 39 (1996) 1051–1061.

[65] Voorips R.E., Mapchart version 2.0: Windows software for the graphical presentation of linkage maps and QTLs, Plant Research International, Wageningen, The Netherlands, 2001.

[66] Vos P., Hogers R., Bleeker M., Reijans M., van der Lee T., Hornes M., Frijters A., Pot J., Pelemam J., Kuiper M., Zabeau M., AFLP: a new technique for DNA fingerprinting, Nucleic Acids Res. 23 (1995) 4407–4414.

[67] Wakamiya I., Newton R. Johnston J.S., Price H.J., Genome size and environmental factors in the genus *Pinus*, Am. J. Bot. 80 (1993) 1235–1241.

[68] Williams J.G.K., Kubelik A.R., Livak K.J., Rafalski J.A., Tingey S.V., DNA polymorphisms amplified by arbitrary primers are useful as genetic markers, Nucleic Acids Res. 18 (1990) 6531–6535.

To access this journal online:
www.edpsciences.org

ANNEXE II

Microsatellite markers for *Pinus pinaster* Ait

Ann. For. Sci. 58 (2001) 203–206
© INRA, EDP Sciences, 2001

Note

Microsatellite markers for *Pinus pinaster* Ait.

Stéphanie Mariette[a,#], David Chagné[a,#], Stéphane Decroocq[a], Giuseppe Giovanni Vendramin[b], Céline Lalanne[a], Delphine Madur[a] and Christophe Plomion[a,*]

[1] INRA, BP45, Laboratoire de Génétique et Amélioration des Arbres Forestiers, 33610 Cestas, France
[2] Istituto Miglioramento Genetico Piante Forestali, CNR, Via A. Vannucci 13, 50134 Firenze, Italy

(Received 26 January 2000; accepted 13 June 2000)

Abstract – Simple sequence repeats (SSRs) or microsatellites are valuable tools for genome mapping and population genetic studies for as they are codominant and highly polymorphic markers. Seventy-six SSR primer pairs from four *Pinus* species were tested to amplify microsatellites in *Pinus pinaster*. Twenty-six primer pairs were stemmed from a microsatellite library on *P. pinaster* and the other primer pairs were obtained in other species of the same genus (*P. radiata*, *P. strobus* and *P. halepensis*). Only three out of the 76 SSR primer pairs amplified at a single polymorphic locus in *P. pinaster*. The Mendelian inheritance of those three primer pairs was studied and their genetic map position was determined. The number of alleles and the level of heterozygosity were assessed in an analysis of a sample of 196 trees. The development of microsatellites in *Pinus* species has been reported to be a difficult task because of the size and complexity of their genome. The results of this study showed that cross-species amplification was quite unsuccessful.

***Pinus pinaster* / genetic variability / genetic mapping / microsatellite / cross-species amplification**

Résumé – Marqueurs microsatellites chez *Pinus pinaster* Ait. Les microsatellites (SSRs) sont des outils de choix pour la cartographie génétique et les études de génétique des populations parce qu'ils sont des marqueurs codominants et très polymorphes. Soixante-seize paires d'amorces de quatre espèces de pin ont été utilisées afin d'amplifier des microsatellites chez *Pinus pinaster*. Vingt-neuf paires d'amorces étaient issues d'une banque enrichie en microsatellites sur *P. pinaster* et les autres paires d'amorces avaient été obtenues sur d'autres espèces du même genre (*P. radiata*, *P. strobus* et *P. halepensis*). Sur un total de 76 paires d'amorces, seulement trois ont amplifié un seul locus microsatellite polymorphe chez *P. pinaster*. Leur ségrégation mendélienne a été étudiée et chaque locus a été localisé sur une carte génétique. Le nombre d'allèles et l'hétérozygotie ont été ensuite évalués en analysant un échantillon de 196 arbres. Le développement de microsatellites chez les espèces du genre *Pinus* s'est révélée difficile en raison de la taille et de la complexité de leur génome. Les résultats de cette étude ont montré que l'amplification inter-espèces n'a rencontré que peu de succès.

***Pinus pinaster* / variabilité génétique / cartographie génétique / microsatellite / amplification inter-spécifique**

Pinus pinaster Ait. is one of the most abundant conifer in South-Western Europe. It is an important species from an ecological (swamp draining and dunes protection) and economical (wood production, pulp and paper making industry) point of view. In France, it covers 1.4 M hectares which represents 10% of the forest surface. A breeding programme for *P. pinaster* was started in the sixties. It has now reached its third generation and has allowed the deployment of improved varieties. Genetic diversity studies were performed using terpenic,

* Correspondence and reprints
Tel. (33) 05 57 97 90 76; Fax. (33) 05 57 97 90 88; e-mail: plomion@pierroton.inra.fr

[#] These authors have contributed equally to this work.

protein, allozymic and chloroplast microsatellite markers throughout the natural range of this species (Baradat et al., 1991 [1]; Barhman et al., 1992 [2]; Petit et al., 1995 [14]; Vendramin et al., 1998 [19]). Nuclear microsatellites are valuable codominant multiallelic DNA markers but not yet available in *P. pinaster* for testing the validity of controlled crosses, for fingerprinting clones and for studying the genetic diversity of the provenances used in the breeding programme. In this study, our aim was to test 76 primer pairs from four *Pinus* species to amplify microsatellite loci in *P. pinaster*.

We adopted two strategies to amplify microsatellites in *P. pinaster*. First, we tested 47 existing primer pairs, as described in the literature or by personal communication, from three other *Pinus* species (*table I*). Second, we constructed an enriched microsatellite library, from which we designed and tested 29 primer pairs.

The microsatellite library, enriched with CA and GA repeats, was constructed from *P. pinaster* genomic DNA, as described by Edwards et al. (1996) [8]. The protocol was modified for hybridisation and washing as followed. Nylon membranes were prehybridised in 6X SSC 5X Denhardt's 1% SDS at 65 °C for 48 h renewing the solution after 24 h. Hybridisation was performed in the same conditions for 20 h. Washing was performed three times in 2X SSC, 1% SDS 65 °C for 15 min, then 1X SSC, 1% SDS for 10 min at 65 °C. A first PCR was performed on DNA that was bound and then eluted from the membrane. PCR products were used for a second round of enrichment. After this second step of enrichment, PCR products were cloned according to the protocol outlined in the Topo TA Cloning kit (Invitrogen, The Netherlands). They were sequenced using LI-COR automatic sequencers 4000 and 4000L (LI-COR Inc., Nebraska, USA). A total of 65 clones containing a microsatellite were detected from 80 clones randomly chosen from the library. Primers were designed for 29 SSRs using the primer software (version 5.0, Whitehead Institute for Biomedical Research, 1991).

The extraction of DNA and the amplification of microsatellites were performed as followed. Genomic DNA was extracted from needles as described by Doyle and Doyle (1991) [5]. The PCR was carried out in a Thermal Cycler Perkin Elmer GeneAmp PCR system 9600, using 0.4 units of Gibco BRL *Taq* Polymerase (Life Technologies, Inc. Gaithersburg MD, USA), and approximately 6 ng of genomic DNA in a total volume of 10 µl containing 200 µM of each nucleotide and 0.2 µM of each primer. Optimized $MgCl_2$ concentrations are indicated in table I. Each forward primer was labelled with the infra-red fluorescent dye IR800 (purchased at MWG Biotech). After a preliminary denaturation step at 94 °C for 4 min, PCR amplifications were performed for 35 cycles under the following conditions: 30 s at 94 °C, 30 s at the annealing temperature (see table I), and 45 s at 72 °C, with a final extension step of 10 minutes at 72 °C. After the amplification, 2 µl of PCR product were mixed with 7 µl of loading buffer (78% formamide, 10 mM EDTA pH 7.6, 0.1% bromophenol blue and 0.1% xylene cyanol), heated for 5 min at 75 °C and quickly cooled on ice. Afterwards 1 µl of denatured SSR fragments was loaded into a 25 cm long denaturating gel containing 8% acrylamide/bisacrylamide (19:1), 6 M urea and 0.4X TBE (134 mM TRIS, 45 mM boric acid, 2.5 mM EDTA). Electrophoresis was performed in the LI-COR automated sequencers using a 1X TBE running buffer at 1500 V, 40 mA and 45 °C of plate temperature. The RFLPscan version 3.0 (Scanalytics) software was used to score the SSR fragments.

Only three (one from *P. halepensis* and two from *P. pinaster*) out of the 76 primer pairs screened amplified at a single highly polymorphic locus in *P. pinaster*

Table I. Amplification of *Pinus* microsatellites in *Pinus pinaster*.

Species (number of primer pairs tested)	Sub-section	Amplification[d]		Banding pattern[e]		
		+	–	SML	SPL	C
Pinus radiata (n = 11)[a]	Attenuatae	73%	27%	12%	0%	88%
Pinus strobus (n = 11)[b]	Strobi	100%	0%	73%	0%	27%
Pinus halepensis (n = 25)[c]	Halepenses	68%	32%	23%	6%	71%
Pinus pinaster (n = 29)	Australes	79%	21%	34%	9%	57%

[a] 7 pairs from G.F. Moran (unpublished results), 2 pairs from [17] Smith and Devey (1994), 2 pairs from [10] Fisher et al. (1998).
[b] 4 pairs from [6] Echt et al. (1996) and 7 pairs available at URL http:/www.resgen.com.
[c] 25 pairs from G.G. Vendramin (unpublished).
[d] + : amplification; - : no amplification.
[e] SML: single monomorphic locus; SPL: single polymorphic locus; C: complex banding pattern.

Table II. Characteristics of three primer pairs for amplifying maritime pine microsatellite loci, with expected (H_E) and observed (H_0) levels of heterozygosity based on 196 individuals.

Locus	Primers (5'→3')	Micro-satellite sequence	Length of PCR product	T_a (°C)[a]	$MgCl_2$ (mM)	Number of alleles	H_E	H_O	Map location[b]	EMBL Accession number
FRPP91	F:GTACTCCCACATAAAATGAGACTT R:CCGAAATACATTGCAGGTTA	$(CT)_{20}$	168	61	2.25	25	0.862	0.684	9	AJ012085
FRPP94	F:GGCAAACCTCTTTTAGAGTGC R:TTTGTCGATTTTTCTTGAAATCTAA	$(CT)_{22}$	162	60	2.5	17	0.726	0.571	5	AJ012086
ITPH4516	F:TGATGCAAACAAGTTCCATG R:AGCACTCGCTAAACTATGAAGG	$(CT)_{27}$	159	61	2.25	20	0.894	0.684	3	AJ012087

[a] T_a, annealing temperature; [b] Linkage group according to the genetic map of Costa et al. (2000).
See also URL http://www.pierroton.inra.fr/genetics/pinus/map2.html

genomic DNA (*table I*). Their Mendelian pattern of inheritance was tested and their allelic variations were examined for 196 individuals, that belonged to eight different populations from southwest France. The characteristics of these three SSRs are summarized in *table II*. Each microsatellite locus revealed a high amount of polymorphism (mean number of alleles = 20.7). The average observed heterozygosity was 0.65. We used a haploid (megagametophyte) mapping pedigree to show that they exhibit a Mendelian pattern of inheritance and we could position them in a previously constructed linkage map (Costa et al., 2000 [4]). About a third of the primer pairs analyzed in this study resulted in single locus-specific amplification. Among these loci, 87.5% were monomorphic and 12.5% were polymorphic. The majority of the remaining primer pairs gave either no amplification (22.4%) or produced multiband patterns (46%) (*table I*). The difficulty of developing informative (single polymorphic locus) microsatellites has already been reported in other conifer species (Echt et al., 1996 [6]; Pfeiffer et al., 1997 [15]) and can be attributed to their large genome size and complexity (Wakamiya et al., 1993 [20]; Kinlaw and Neale, 1997 [12]).

In this study, we showed that only one primer pair from other *Pinus* species (*P. halepensis*) could be transferred to *P. pinaster*. According to Farjon (1984) [9], *P. radiata* belongs to the same section as *P. pinaster* whereas *P. strobus* belongs to the section *Strobus* and *P. halepensis* to the section *Pinea*. However, a natural hybrid between *P. halepensis* and *P. pinaster* was mentioned by Schütt (1959) [16], which may explain our result. As reported by Echt and May-Marquardt (1997) [7], we also found that SSR information do not transfer across *Pinus* species. However, ten polymorphic SSRs markers developed in *P. halepensis* produced single variable bands segregating in a Mendelian manner in the species *P. brutia* (G.G. Vendramin, personal communication). In this case, cross-species amplification seemed to be easier because these *Pinus* species show a low degree of divergence (Bucci et al., 1998 [3]). Similarly, some studies have shown that SSRs isolated from several species amplify the corresponding and polymorphic PCR products in closely related species. Kijas et al. (1995) [11] tested two primer sets in 10 different *Citrus* species and two related genera and found conservation of the sequences. Using 17 sets of primers developed from sessile oak, *Quercus petraea*, Steinkellner et al. (1997) [18] found that two of the loci were polymorphic in all the *Quercus* species tested. In general, the success of the amplification diminishes with increasing species divergence (Steinkellner et al., 1997 [18]; Whitton et al., 1997 [21]; Lefort et al., 1999 [13]). Further development of *P. pinaster* microsatellites will be focused on an enriched cDNA library.

Acknowledgements: We thank Cécile Cabrero and Audrey Lartigue for their partnership in this work, Gavin Moran for providing unpublished *Pinus radiata* primer pairs. We thank two anonymous reviewers for their useful remarks on a previous version of the manuscript. This work was supported by grants from France (Ministère de l'Agriculture et de la Pêche-DERF n° 61.21.04/98), and the European Union (INCO, ERBIC-08CT-970200).

REFERENCES

[1] Baradat P., Marpeau A., Walter J., Terpene markers, in: Müller-Starck G., Ziehe M. (Eds.), Genetic variation in European populations of forest trees, Sauerländer, Francfort, 1991, pp. 40–66.

[2] Barhman N., Baradat P., Petit R.J., Structuration de la variabilité génétique du Pin maritime dans l'ensemble de son aire naturelle, in : Colloque international en hommage à Jean Pernès, Publications du Bureau des Ressources Génétiques, 1992, pp. 352–368.

[3] Bucci G., Anzidei M., Madaghiele A., Vendramin G.G., Detection of haplotypic variation and natural hybridization in *halepensis*-complex pine species usng chloroplast simple sequence repeat (SSR) markers, Mol. Ecol. 7 (1998) 1633–1643.

[4] Costa P., Pot D., Dubos C., Frigerio J.M., Pionneau C., Bodénès C., Bertocchi E., Cervera M.T., Remington D.L., Plomion C., A genetic map of maritime pine based on AFLP, RAPD and protein markers, Theor. Appl. Genet. 100 (2000) 39–48.

[5] Doyle J.J., Doyle J.L., Isolation of plant DNA from fresh tissue, Focus 12 (1990) 13–15.

[6] Echt C.S., May-Marquardt P., Hseih M., Zahorchak R., Characterization of microsatellite markers in Eastern white pine, Genome 39 (1996) 1102–1108.

[7] Echt C.S., May-Marquardt P., Survey of microsatellite DNA in pine, Genome 40 (1997) 9–17.

[8] Edwards K.J., Barker J.H.A., Daly A., Jones C., Karp A., Microsatellite librairies enriched for several microsatellite sequences in plants, BioTechniques 20 (1994) 758–760.

[9] Farjon A., Pines: drawings and descriptions of the genus Pinus, Brill E.J. (Ed.), Leiden, 1984.

[10] Fisher P.J., Richardson T.E., Gardner R.C., Characteristics of single-and multi-copy microsatellites from Pinus radiata, Theor. Appl. Genet. 96 (1998) 969–979.

[11] Kijas J.M.H., Fowler J.C.S., Thomas M.R., An evolution of sequence tagged microsatellite site markers for genetic analysis within citrus and related species, Genome 38 (1995) 349–355.

[12] Kinlaw C.S., Neale D.B., Complex gene families in pine genomes, Trends Plant Sci. 2 (1997) 356–359.

[13] Lefort F., Brachet S., Frascaria-Lacoste N., Edwards K.J., Douglas G.C., Identification and characterization of microsatellite loci in ash (*Fraxinus excelsior* L.) and their conservation in the olive family (Oleaceae), Mol. Ecol. 8 (1999) 1088–1090.

[14] Petit R.J., Barhman N., Baradat P.H., Comparison of genetic differentiation in maritime pine (*Pinus pinaster* Ait.) estimated using isozyme, total proteine and terpenic loci, Heredity 75 (1995) 382–389.

[15] Pfeiffer A., Olivieri A.M., Morgante M., Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome 40 (1997) 1–9.

[16] Schütt P., Züchtung mit Kiefern Teil 2. Kreuzungen, Resistenzzüchtung und Zytologie. Mitt. Bundesforsch. Forst- und Holzwirt. Forstgenetik und Forstpflanzenzüchtung, 42 (1959) 1–40.

[17] Smith D.N., Devey M.E., Occurence and inheritance of microsatellites in *Pinus radiata*, Genome 37 (1994) 977–983.

[18] Steinkellner H., Lexer C., Turetschek E., Glössl J., Conservation of $(GA)_n$ microsatellite loci between *Quercus* species, Mol. Ecol. 6 (1997) 1189–1194.

[19] Vendramin G.G., Anzidei M., Madaghiele A., Bucci G., Distribution of genetic diversity in *Pinus pinaster* Ait. as revealed by chloroplast microsatellites, Theor. Appl. Gen. 97 (1998) 456–463.

[20] Wakamiya I., Newton R.J., Johnston J.S., Price H.J., Genome size and environmental factors in the genus *Pinus*, Am. J. Bot. 80 (1993) 1235–1241.

To access this journal online:
www.edpsciences.org

ANNEXE III

Comparative genome and QTL mapping between maritime and loblolly pines

Molecular Breeding **12:** 185–195, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Comparative genome and QTL mapping between maritime and loblolly pines

David Chagné[1], Garth Brown[2,3], Céline Lalanne[1], Delphine Madur[1], David Pot[1], David Neale[2,3] and Christophe Plomion[1,*]

*[1]Equipe de génétique et amélioration des arbres forestiers, INRA – Pierroton, 69 route d'Arcachon, Cestas Cedex, 33612, France; [2]Institute of Forest Genetics, Pacific Southwest Research Station, US Department of Agriculture Forest Service, Davis, CA 95616, USA; [3]Department of Environmental Horticulture, University of California, Davis, CA 95616, USA; *Author for correspondence (e-mail: plomion@pierroton.inra.fr; phone: 33 5 57 12 28 38; fax: 33 5 57 12 28 81)*

Received 2 August 2002; accepted in revised form 21 March 2003

Key words: Comparative mapping, EST polymorphism, *Pinus pinaster*, *Pinus taeda*, Wood quality QTL

Abstract

Genetic markers developed from expressed sequence tags (ESTs) were used as orthologous loci for comparative genome studies in the genus *Pinus*. A total of 309 ESTs derived from conifer gene sequences were tested for amplification and polymorphism in maritime pine (*Pinus pinaster* Ait.). Electrophoresis-based techniques made it possible to map 50 expressed sequence tag polymorphisms (ESTPs). The map positions of 32 markers were compared to putative orthologous loci on the loblolly pine (*Pinus taeda* L.) linkage map, which is the reference map of the conifer genetic mapping community. Overall, synteny was maintained between the two species. This report agrees with other pairwise genome comparisons in pine and supports the cytogenetic evidence that chromosome evolution in the genus is conservative. The alignment of homologous linkage groups allowed, for the first time in conifers, the comparison of QTL location. The position of two QTLs controlling wood density and cell wall components were found to be conserved between the two species.

Introduction

Comparative genome mapping consists of studying the conservation of gene content (synteny) and order (colinearity) in the genome of closely related species. One outcome of comparative mapping is the knowledge of genome evolution of the species (or genera) considered. The genetic maps of important plant species have been compared. The results (Gale and Devos 1998 and references therein) showed that gene content and order can be conserved, and that the differences in genome structure are mainly due to chromosome rearrangements (inversion, duplication, translocation, deletion) or polyploidization. A second goal of comparative mapping is the transfer of genetic information between species. The ability to consider a number of species, such as the grasses, as a single genetic system has profound applications, including the prediction of the location of markers, candidate genes, and quantitative trait loci (QTL) in non-model species. For example, Chen et al. (1999) validated QTLs controlling fruit traits by comparative QTL mapping in several tomato species.

There is no known conifer species with a small genome to serve as a model species in gymnosperms, as *Arabidopsis thaliana* serves for angiosperms. However, comparative mapping helps to synthesize genetic map data obtained from a wide variety of species. The Conifer Comparative Genome Project (CCGP, http://dendrome.ucdavis.edu/Synteny/) was established to provide the forest genetics community with orthologous markers and bioinformatic tools to integrate the genetic maps of important conifer species belonging to the *Pinaceae* family. Little is known about the genome organization of pines. Interestingly, the karyotype (2n = 2x = 24) and the centromere po-

sitions are similar for most of the members of the *Pinus* genus, indicating that no major chromosomal rearrangement occurred since the divergence of pines (Saylor 1971; Prager et al. 1976). However, the pine genome is characterized by a large physical size (25,000–30,000 Mb 1C; Wakayima et al. 1993) and by a large low-copy fraction (15–24%; Kriebel 1985; Elsik and Williams 2000).

Ahuja et al. (1994) first showed that cDNA probes cross-hybridized between conifer species and concluded that comparative genome mapping was feasible in conifers. A first pairwise comparison was performed between two pine species, *Pinus taeda* L. (loblolly pine) and *Pinus radiata* D. Don (Devey et al. 1999), using restriction fragment length polymorphisms (RFLPs) and microsatellite markers, and revealed a significant colinearity between their genomes. The ongoing sequencing of cDNA libraries in pines (Whetten et al. 2001) provides an alternative source of markers for comparative mapping. Expressed sequence tag polymorphism (ESTP) markers can be easily assayed by the polymerase chain reaction (PCR) and are known to be more conserved and thus more easily transferable between species than microsatellites (Echt and May-Marquardt 1997) or anonymous markers like RAPDs (randomly amplified polymorphic DNA) or AFLPs (amplified fragment length polymorphisms). The polymorphism detected in ESTPs may correspond to either insertion-deletions (Indels) or single nucleotide polymorphisms (SNPs). A set of ESTPs was developed in the framework of the CCGP to serve as anchor loci for interspecific amplification and mapping (Temesgen et al. 2001; Brown et al. 2001). The results obtained by the comparative mapping analysis between *P. taeda* and *P. elliottii* (Brown et al. 2001) using ESTPs showed significant conservation of gene content and order.

The construction of genetic maps and the expansion of QTL detection studies in the last ten years has improved the understanding of the genetic control of economic traits (Paterson 1995). In the near future, these approaches could permit the development of genetically improved trees using marker-assisted selection (MAS). However, considering some of the characteristics of forest trees: long-lived, highly diverse organisms growing in diverse environments, it is a prerequisite to gather information regarding the stability of QTLs in different genetic backgrounds, at different stages of maturation and on different sites. Plomion et al. (1996) and Verhaegen et al. (1997) adressed the maturation effects on QTLs detected for growth and wood density. Neale et al. (2002) reported the effect of genetic background on wood quality QTLs using several pedigrees of varying coancestry in loblolly pine. The alignment of genetic maps between both species would serve to validate wood density and cell wall chemistry QTLs and to co-localize positional candidate genes controlling these traits.

In the present article, we compared the genetic maps of maritime pine (*Pinus pinaster* Ait.) and loblolly pine (*Pinus taeda* L.). According to Price et al. (1998), these two species belong to the same subgenus (Hard pines, subgenus *Pinus*), but to two distinct subsections (*Australes* for *P. taeda* and *Pinus* for *P. pinaster*). Maritime pine is the most abundant forest tree species in Southwestern Europe (France, Portugal, Spain) and loblolly pine is widely used for reforestation in Northern America. Beyond the ecological importance of these species, they have a large economic value for wood production. They are also both subjected to an intensive breeding program in which growth, stem form and wood quality represent the main target traits.

Materials and methods

Plant material

A three generation outbred pedigree belonging to the maritime pine breeding program was used.

Orthologous markers

A total of 309 primer pairs developed from conifer gene sequences were tested for amplification in *P. pinaster* (Table 1). Set #1 and set #2 corresponded to the putative orthologous anchor loci developed by the CCGP. These ESTP markers were chosen following several criteria described by Temesgen et al. (2001) and Brown et al. (2001). PCR amplifications were performed following the same conditions as described by Temesgen et al. (2001). Set #3 includes potential candidate genes for wood quality traits in loblolly pine (G. Gill and D. Neale, unpublished). The pine EST data (http://www.cbc.umn.edu/ResearchProjects/Pine/DOE.pine/index.html) available for each of these genes were aligned and variable regions were detected *in silico* using Sequencher (Gene Codes Corp., Ann Arbor, MI). PCR primers were designed to amplify regions encompassing single nucleotide polymorphisms. The PCR conditions were the same

Table 1. Description of the PCR primer sets: amplification and polymorphism rates of the EST used for mapping in *Pinus pinaster*. The five sets correspond to different sources of ESTPs

Primer set	Number of markers tested	Amplification rate	Polymorphism rate	References for primer pairs
Set #1	56	96.4% (54/56)	38.8% (21/54)	Temesgen et al. 2001
Set #2	38	71.0% (27/38)	14.8% (4/27)	Brown et al. 2001
Set #3	65	81.5% (53/65)	22.6% (12/53)	G. Gill and D. Neale, personal communication
Set #4	90	52.2% (47/90)	31.9% (15/47)	http://www.pierroton.inra.fr/genetics/pinus/primers.html
Set #5	60	63.3% (38/60)	5.2% (2/38)	Cato et al. 2001
Total	309	70 % (217/309)	24 % (54/217)	–

Table 2. Description of the ESTPs developed in maritime pine (set #4): accession, marker ID and primer pairs. The other ESTs belonging to set #4 are described in Plomion et al. (1999)

Marker ID	Accession	Primer pair	
PpINR_AN01E4	AL749565	F: TGATGTTATTGAGGGGTTGA	R: TCGGTTAGTTTTGTGTTGGT
PpINR_AS01C7	AL749806	F: ACGCAGAGTAGAAACCAACA	R: TCCCAGACAAGACAAACAAT
PpINR_AS01F3	AL749831	F: GAGAACCGAACAGCAGGAAT	R: GCATGAACTCAGGGGACC
PpINR_AS01G01	AL749839	F: GAGAATTGGGTGTGTTGTTA	R: TCGCAGTTGTGTGAAGAT
PpINR_AS01H04	AL749850	F: TAGCTGCTCCCCTCAAGACC	R: GGCACCCACTTGTTCCTCA
PpINR_Pp.ap12	AJ309112	F: ATTAGCAGGGCATCTGTCTG	R: CACGCCTCTCATTTTCATC
PpINR_Pp.ap23	AJ309109	F: CTGGGATGAGATTGAAGA	R: GTGACTTGGACGAAAATAAT
PpINR_Pp.ap9	AJ309115	F: GCAGCGTTCGTCTTCATAAT	R: GCGGTCACATGGAAAAACT
PpINR_RS01D05	AL750878	F: ACGGGCCGAGGAACTGGACC	R: GCAGGGCTCGGGCAATCGTT
PpINR_RS01G05	AL750905	F: AAAGCGTTTCTGGAACRKKTTA	R: GTTTTGGTCAAGGCAATCC

as for sets #1 and #2, except that a 60 °C annealing temperature was added. Set #4 was derived from randomly chosen conifer cDNA sequences (Table 2). PCR primers were designed to amplify 5′ or 3′ untranslated regions and PCR was performed as described by Plomion et al. (1999). Set #5 consisted of 60 markers developed by Cato et al. (2001) in *P. radiata* based on the amplification of the flanking regions upstream or downstream of a coding region.

Amplification products were loaded on 1.5% agarose gels for all the 309 markers. Gels were checked for the amplification of a single band. Complex patterns were eliminated from further analysis since they could result from the amplification of paralogous loci or non-specific amplification.

Polymorphism detection

For each primer set, the two parents and four progeny of the maritime pine pedigree were amplified and tested for polymorphism. A *P. taeda* DNA sample was used as a control for set #1, set #2 and set #3 and a *P. radiata* DNA sample for set #5. Sets #4 and #5 were not tested for amplification or polymorphism in *P. taeda*. Different electrophoretic methods (see Table 3) were used to detect polymorphism depending on the primer set. Polyacrylamide gel electrophoresis (PAGE) was used for length polymorphism detection for all primer sets. PCR products were run on 4–10% non-denaturing gels and stained with ethidium bromide. Denaturing gradient gel electrophoresis (DGGE; Myers et al. 1987) was used for set #1, set #2 and set #3. Gel conditions and gradient composition were exactly the same as described by Temesgen et al. (2001) and Brown et al. (2001) for reproducibility and transferability of the markers. Single strand conformation polymorphism (SSCP; Orita et al. 1989) was used for set #4 with optimizations as described by Plomion et al. (1999). Set #5 was assayed according to methods of Cato et al. (2001).

Linkage mapping

A saturated AFLP map of maritime pine was previously described by Chagné et al. (2002) using 620 markers genotyped on a subset of 90 progeny. A sub-

Table 3. Description of the ESTPs mapped in maritime pine: Best identity, electrophoresis conditions and map position

Set	Primer pair	Best identity	Method used for polymorphism detection	Polyacrylamide concentration	Gel conditions	Segregation ratio	Segregation distorsion	Linkage group
Set 1	PtIFG_1623	NS1-associated protein	PAGE	4%	ND	1:1	–	2
Set 1	PtIFG_1643	ABI1 gene product (phosphatase)	DGGE	6%	15–45% UF	1:1:1:1	*	10
Set 1	PtIFG_1764	–	DGGE	6%	15–45% UF	1:1	–	6
Set 1	PtIFG_2274	Adenylyl cyclase	DGGE	6%	15–45% UF	1:1	–	4
Set 1	PtIFG_2358	Phenylalanine tRNA synthetase	DGGE	10%	15–45% UF	1:1	–	6
Set 1	PtIFG_2781	Glucose-induced repressor	PAGE	10%	ND	1:1:1:1	*	8
Set 1	PtIFG_464	Aquaporin	PAGE	4%	ND	1:1	–	2
Set 1	PtIFG_606	*entE* gene	DGGE	6%	15–45% UF	1:1	–	6
Set 1	PtIFG_8415	–	DGGE	10%	15–45% UF	1:1	–	9
Set 1	PtIFG_8436	Ribosomal protein 40S S3A	DGGE	10%	15–45% UF	1:1	–	7
Set 1	PtIFG_8702	Thioredoxin	DGGE	10%	15–45% UF	1:1	–	6
Set 1	PtIFG_8898	Testis mitotic checkpoint	DGGE	6%	15–45% UF	1:1	–	4
Set 1	PtIFG_8907	Peroxidase cationic	DGGE	10%	15–45% UF	1:1:1:1	–	8
Set 1	PtIFG_893	Nonspecific lipid transfer protein	PAGE	6%	ND	1:2:1	–	5
Set 1	PtIFG_8939	Ribosomal protein 40S S16	DGGE	10%	10–30% UF	1:1:1:1	–	2
Set 1	PtIFG_8972	Plasma membrane protein	DGGE	10%	15–45% UF	1:1	–	6
Set 1	PtIFG_9036	Ribosomal protein L37	DGGE	10%	15–45% UF	1:1	–	7
Set 1	PtIFG_9044	Ribosomal protein 40S S27	DGGE	10%	15–45% UF	1:1	*	6
Set 1	PtIFG_9092	Nonspecific lipid transfer protein	DGGE	10%	15–45% UF	1:1	–	5
Set 1	PtIFG_9151	Cucumber basic protein	DGGE	10%	15–45% UF	1:1	**	7
Set 2	PtIFG_8429	3-DMOcytidylyltransferase	DGGE	10%	15–45% UF	1:1	–	4
Set 2	PtIFG_8656	GF6P aminotransferase	DGGE	10%	15–45% UF	1:1	–	U
Set 2	PtIFG_8779	Histone H3	DGGE	10%	15–45% UF	1:2:1	–	5
Set 2	PtIFG_9136	Ribosomal protein S11	DGGE	10%	15–45% UF	1:1:1:1	–	3
Set 4	PtIFG_1584	Deoxychalcone synthase	PAGE	8%	ND	1:1	–	4
Set 4	PtIFG_8580	Embryogenesis abundant protein	PAGE	8%	ND	1:2:1	–	10
Set 3	PtNCS_17G4	Endoglucanase	DGGE	10%	15–45% UF	1:1	–	U
Set 3	PtNCS_1CAB7E	ACC oxidase	PAGE	10%	ND	1:1	–	U
Set 3	PtNCS_22B8	Glycine hydroxymethyltransferase	DGGE	10%	15–45% UF	1:1	–	3
Set 3	PtNCS_2C11E	Cinnamyl alcohol dehydrogenase	DGGE	10%	15–45% UF	1:1	–	9
Set 3	PtNCS_C4H-1	trans-cinnamate 4-hydroxylase	DGGE	10%	15–45% UF	1:1:1:1	–	3
Set 3	PtNCS_6C5A	Isoflavone reductase homolog	DGGE	10%	15–45% UF	1:1	–	4

Table 3. Continued

Set	Primer pair	Best identity	Method used for polymorphism detection	Polyacrylamide concentration	Gel conditions	Segregation ratio	Segregation distorsion	Linkage group
Set 3	PtNCS_7C4C	30S ribosomal protein	PAGE	10%	ND	1:1	–	U
Set 3	PtTX_p14A9	Arabinogalactane-like protein	DGGE	10%	15–45% UF	1:1	–	3
Set 3	PtNCS_PtaAGP6	Arabinogalactane-like protein	DGGE	6%	25–55% UF	1:1:1:1	–	5
Set 3	PtNCS_ctg3	Caffeoyl CoA O-methyltransferase	DGGE	10%	15–45% UF	1:1	–	6
Set 3	PpINR_CHS	Chalcone synthase	DGGE	10%	15–45% UF	1:1	–	2
Set 3	PsUF1_NIR	Nitrite reductase	DGGE	10%	15–45% UF	1:1	–	7
Set 4	PpINR_AN01E4	Glycin decarboxylase	SSCP	8%	4 °C RT	1:1	**	7
Set 4	PpINR_AS01C7	Heat Shock Protein (HSP)	PAGE	8%	ND	1:2:1	–	10
Set 4	PpINR_AS01F3	50S ribosomal protein	SSCP	8%	4 °C RT	1:2:1	–	4
Set 4	PpINR_AS01G01	Initiation factor	SSCP	8%	4 °C RT	3:1	–	1
Set 4	PpINR_AS01H04	RuBP carboxylase	PAGE	8%	ND	1:1	–	10
Set 4	PtMTU_PtCW1	Ferritin	SSCP	8%	4 °C RT	1:1	*	5
Set 4	PtMTU_PtCW2	Peroxidase	Agarose	2%	ND	1:1	–	8
Set 4	AbWS2_AG3.18	Pinene synthase	Agarose	2%	ND	1:1	–	8
Set 4	PpINR_Pp.ap12	–	SSCP	8%	15 °C RT	3:1	–	8
Set 4	PpINR_Pp.ap23	–	Agarose	2%	ND	1:1	–	10
Set 4	PpINR_Pp.ap9	Hypothetical zinc finger protein	SSCP	8%	4 °C RT	1:1	–	6
Set 4	PpINR_RS01D05	MYB-related protein	PAGE	8%	ND	1:1	–	7
Set 4	PpINR_RS01G05	Unknown protein	Agarose	2%	ND	1:1	–	1
Set 4	PpINR_SODChl	Superoxide dismutase	SSCP	8%	4 °C RT	1:2:1	–	10
Set 5	PrFRI_PrMC2	Male cone protein 1 precursor	PAGE	8%	ND	1:1	–	6
Set 5	PrFRI_PthCAB	Chlorophyll a/b binding protein	PAGE	8%	ND	1:1	–	2

DGGE: Denaturing Gradient Gel Electrophoresis; SSCP: Single Strand Conformation Polymorphism; PAGE: Polyacrylamide Gel Electrophoresis, UF: Urea-Formamide denaturing gradient; RT: Running Temperature, ND: Non-Denaturing gel; U: Unlinked marker. *: $0.01 \leq p \leq 0.05$; **: $p > 0.05$; –: non significantly distorted.

set of evenly spaced AFLP markers was then selected and genotyped on 202 individuals of the same family to improve the statistical support of marker ordering and detect QTLs for wood properties with more power. The parental framework maps were built following the strategy described by Chagné et al. (2002) using testcross AFLP markers (segregating in the 1:1 ratio) with MapMaker v2.0 for MacIntosh (Lander et al. 1987) and the ordering of markers was optimized using the *ripple* function. The order was compared to those obtained with Carthagene v0.5 (Schiex et al. 2001) and Joinmap v3.0 (Van Ooijen and Voorrips 2001) mapping software. The *sem* command of Carthagene was used to calculate the LOG10-likelihood of the orders obtained with the three packages. The *nicemapl*, *nicemapd*, *build* and *annealing* algorithms of Carthagene were used to improve the order. When a better order was found, it was validated by the *flips* method (the equivalent of the *ripple* function of MapMaker) in a window of three markers.

The segregation data obtained from the ESTPs were scored visually, checked for segregation distortion ($P > 0.01$) and added to the framework marker data. A LOD threshold of 5.0 was used for mapping these new markers using Joinmap. When the ordering of the AFLP framework markers was modified compared with the QTL detection maps, the *fixed order* function of Joinmap was used. Once the ESTPs were linked to these maps, a consensus map was built using intercross AFLP markers (segregating in the 3:1 ratio). Because of the weak linkage information between pairs of markers segregating 1:1 and 3:1 (Plomion et al. 1997), only those that did not disturb the order used for the QTL analysis were retained. The genetic consensus map was finally drawn using MapChart v2.1 software (Voorrips 2001) and is available at http://www.pierroton.inra.fr/genetics/pinus/Map3/index.html.

Results and discussion

Amplification of orthologous loci in maritime pine

Two hundred and seventeen out of 309 primer pairs amplified in *P. pinaster* (Table 1). The highest rate was for set #1 (96%) and the lowest was for set #5 (63%) designed from *P. radiata* sequences. As *P. radiata* belongs to a different subsection, this result showed that the amplification rate decreased notably with increasing phylogenetic distance (Price et al. 1998), consistent with previous results in pines (Brown et al. 2001) and in spruces (Perry and Bousquet 1998). Most of the markers showed a single band of similar length on agarose gels between *P. pinaster* and *P. taeda* or *P. radiata* controls. This was a first indication that the amplified fragments obtained in maritime pine could be orthologous to those in loblolly and radiata pines. Nevertheless, the primer pair PtIFG_9036 amplified two bands of 280 bp and 244 bp in maritime pine, whereas a single band of 280 bp was amplified in loblolly pine.

Polymorphism rate

Fifty-four out of 217 (24.8%) amplified ESTPs were polymorphic in the maritime pine mapping pedigree (Table 1). Set #5 was the least polymorphic (5.2%) for which a great proportion of the amplified loci was difficult to score or did not segregate in a Mendelian manner. The rates obtained with DGGE, SSCP or PAGE for the other sets ranged from 14.8% to 38.8% and were twice as low as the rates obtained for other pedigrees included in the CCGP (i.e., 52%, Brown et al. 2001).

Mapping ESTPs

Fifty ESTPs were located on the maritime pine genetic map (Figure 1). The map consists of 326 markers distributed in 12 linkage groups with a total length of 1638.5 cM. The introduction of the ESTP markers did not change the ordering of the framework AFLP markers used for QTL detection and all software gave identical maps. This map had very few differences to the map previously described by Chagné et al. (2002). Sixty intercross common AFLP markers (segregating 3:1) were included and made it possible to merge both parental maps. Four ESTPs were not linked. Interestingly, two of them (estPtIFG_8656_a and estPtNCS_7C4C_a) were also unlinked in loblolly pine and may be located in telomeric regions not covered by either genetic map.

Comparative mapping

Based on the positions of 32 ESTPs common to both the maritime pine and loblolly pine maps, ten homologous linkage groups could be identified. Between one and seven markers were common for each of the ten linkage groups. Synteny and colinearity were largely conserved between loblolly and maritime

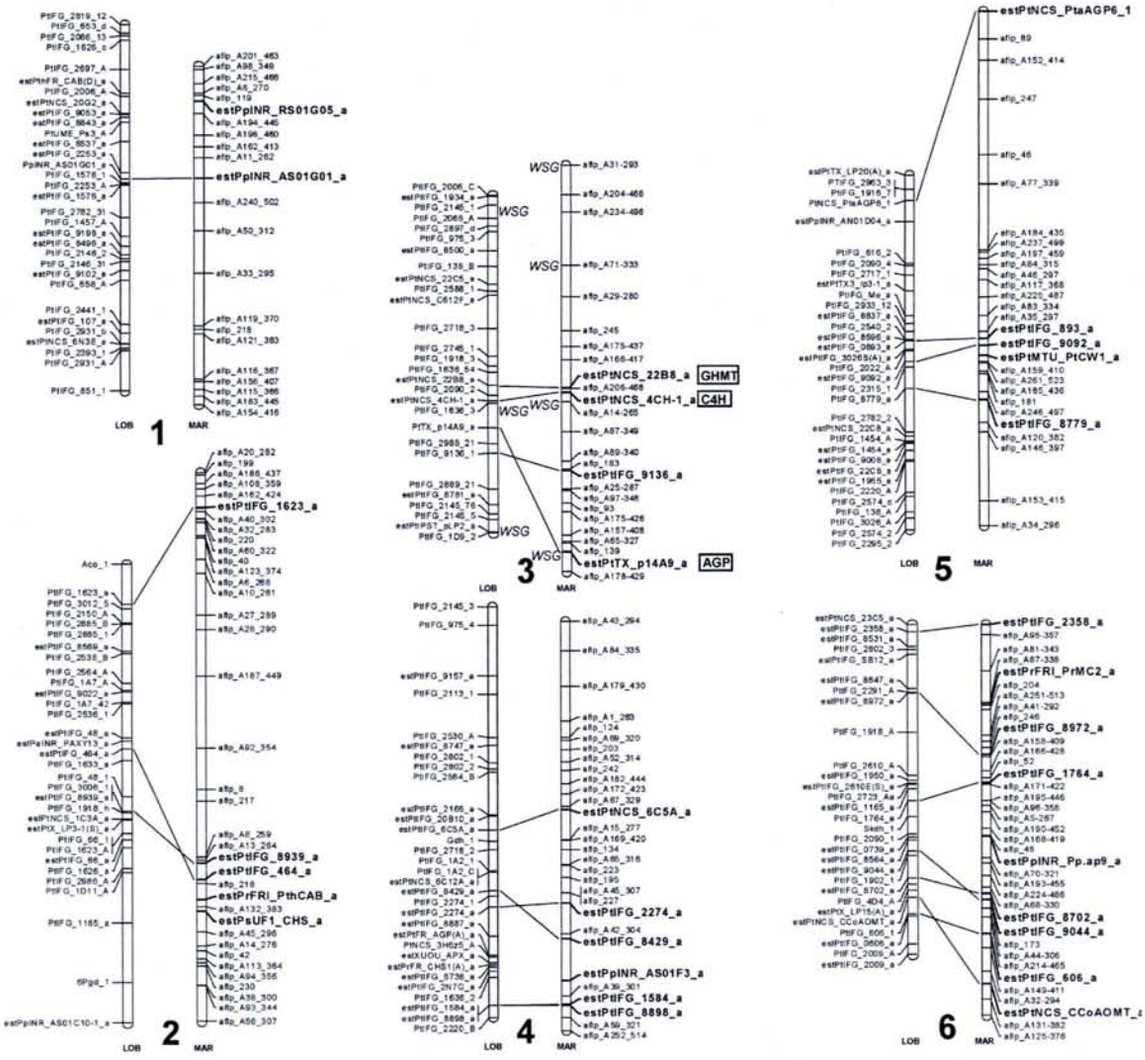


Figure 1. Comparison between loblolly (LOB) and maritime pine (MAR) maps. The naming of genetic makers follows the locus nomenclature of the Treegenes database (http://dendrome.ucdavis.edu/TreeGenes). The Experiment field for markers mapped in maritime pine and loblolly pine are INRCOM and IFGTXP, respectively, but have been omitted in this figure for simplicity. The ESTPs mapped in maritime pine are presented in bold. Homology could be determined for ten out twelve linkage groups using 32 common ESTPs. Two markers (estPtIFG_9036_b and estPtIFG_8436_b) were mapped in different groups between the two species (respectively LG #8 and #10 in loblolly pine and LG 7 in maritime pine). Common wood quality QTLs and related candidate genes are presented for linkage groups #3 and #8. *Lig*: lignin content QTL; *Cel*: α-cellulose content QTL; *WSG*: wood specific gravity QTL (Sewell et al. 2000; 2002; Pot *et al.*, unpublished). GHMT: glycine hydroxymethyltransferase; C4H: trans cinnamate 4 hydroxylase; AGP: arabinogalactane-like protein.

pines although a few differences in gene order were observed among tightly linked markers. These may have resulted from mapping errors rather than actual chromosomal rearrangement. As for the two linkage groups (#1 and #7) harboring a single comparative anchor locus, another pairwise comparison between maritime and radiata pines has confirmed this homology using microsatellite markers (D. Chagné and C. Echt, unpublished). It should also be noticed that two ESTPs did not map to the homologous linkage groups.

Although the present analysis is of low-resolution, it has clearly identified homologous linkage groups between the two species. These results agree with early work on isozymes (Conkle 1981) and more recent genome-wide comparisons between different

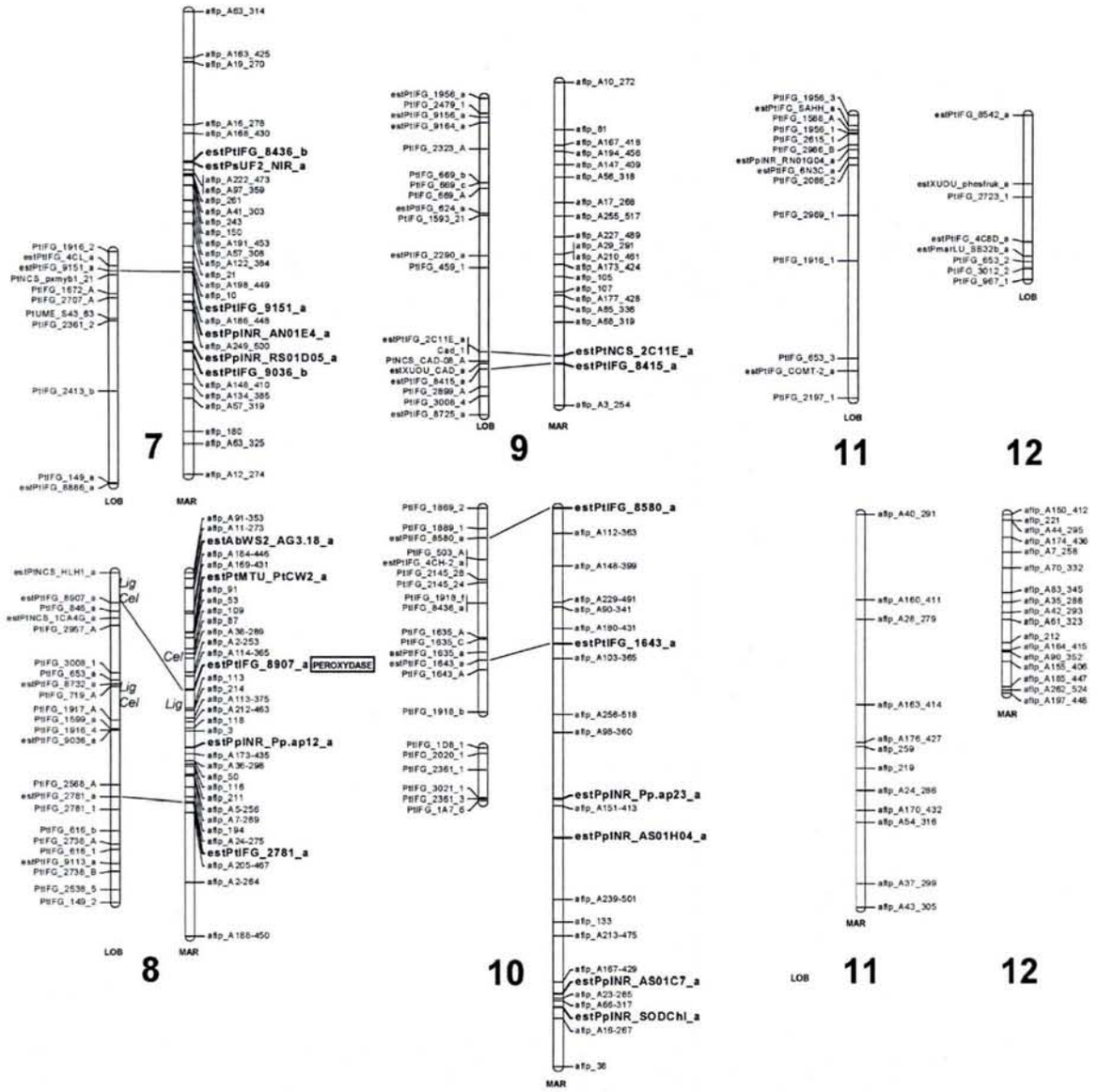


Figure 1. Continued.

subsections of pines: *P. taeda* and *P. radiata* (Devey et al. 1999), *P. taeda* and *P. elliottii* (Brown et al. 2001) and *P. taeda* and *P. sylvestris* (Komulainen et al. in prep), Therefore, the homologous linkage groups between these four pine species are now known and the linkage group numbering system of the loblolly pine reference map can be employed by scientists belonging to the conifer mapping community. Practically, if substantial synteny and colinearity is the rule in pines, predicting the position of genetic information (markers, genes, and QTLs) from one species to another will be possible. It will also provide less studied pine species with markers to quickly generate a genetic map.

PCR products of four maritime pine ESTPs (PtTX_p14A9, PtNCS_C4H-1, PtNCS_22B8, and PtIFG_9151) that mapped on homologous linkage groups, were cloned and sequenced as described by Dubos and Plomion (2002) to check their similarity with corresponding loblolly pine sequences. The re-

Table 4. Sequence similarity between *P. taeda* and *P. pinaster* ESTPs. The fragments from 6 primer pairs tested in maritime pine were sequenced and compared with the reference sequences from loblolly pine. For the primer pair estPtIFG_9036 the two gene family members present in maritime pine were sequenced (INRCOM_estPtIFG_9036_a and INRCOM_estPtIFG_9036_b)

Locus	Nucleotide identity
estPtIFG_8436_b	235/256 (91.8%)
estPtIFG_9036_a	274/280 (97.8%)
estPtIFG_9036_b	227/280 (81.1%)
estPtIFG_9151	272/291 (93.4%)
estPtNCS_C4H-1	481/489 (98.3%)
estPtTX_p14A9	507/518 (97.8%)
estPtNCS_22B8	225/228 (98.6%)

sults presented in Table 4 show nucleotide identity rates ranging from 93.4% to 98.6%. High identities were also reported by Brown et al. (2001) who compared the sequences of ten loci between *P. elliottii* and *P. taeda* and found 99.4% nucleotide identity on average. The sequence similarity obtained between *P. taeda* and *P. elliottii* is higher than the rate obtained between *P. taeda* and *P. pinaster* because *P. taeda* and *P. elliotti* are phylogenetically closer from each compared to *P. taeda* and *P. pinaster* (Price et al. 1998). Paralogous loci are genes or clusters of genes at different chromosomal locations in the same organism that have structural similarities indicating that they derived from a common ancestral gene. Although the presence of paralogous loci cannot be excluded, similar map position and high nucleotide identity make a strong case for locus orthology.

Sequencing was also performed for the two ESTPs that did not map to homologous linkage groups. The fragment corresponding to locus PtIFG_9036_b (informative in *P. pinaster* and missing in *P. taeda*) showed 81.1% identity with the loblolly pine reference sequence, whereas the fragment corresponding to locus PtIFG_9036_a (informative in *P. taeda* and monomorphic in *P. pinaster*) was 98.4% identical to the same loblolly pine sequence. These differences were mainly due to the presence of two insertion/deletion polymorphisms within the sequences (Figure 2). The lower similarity between PtIFG_9036_b and PtIFG_9036_a is an indication of the presence of at least two gene family members in maritime pine.

For primer pair PtIFG_8436, only one band was amplified in both species. However, this locus also mapped to a non-homologous linkage group in the two species (linkage group 10 in loblolly pine and linkage group 7 in maritime pine) and the sequencing of the PtIFG_8436_b in maritime pine showed a lower similarity (91.8%) than the other loci. This indicates that estPtIFG_8436_b could also correspond to a paralogous gene. Even if the markers were made to avoid the amplification of paralogous loci by designing one of the PCR primers in the 3'UTR region (Temesgen et al. 2001; Brown et al. 2001), this kind of phenomenon is not surprising given the complexity of the pine genome (Kinlaw and Neale 1997).

Comparative QTL and candidate gene mapping

The alignment of the loblolly and maritime pine maps made it also possible to compare the location of QTLs influencing similar traits (Figure 1). Sewell et al. (2000, 2002) described the positions and effects of QTLs controlling wood properties in loblolly and Pot et al. (unpublished, cited in Plomion et al. 2003) reported similar findings for QTLs in maritime pine. The comparison of the positions of these QTLs shows a similar locations for QTLs influencing wood properties on two linkage groups. QTLs controlling wood density and related traits in different annual growth rings (wood specific gravity, density homogeneity, mass) were detected in linkage group #3 and putative orthologous QTLs for lignin and α-cellulose contents were observed on linkage group #8.

In addition, putative candidate genes belonging to physiological pathways related to wood formation (PtNCS_C4H-1, PtTX_p14A9 and PtNCS_22B8) co-localized with the physical wood properties QTLs in linkage group #3. Locus PtNCS_C4H-1 corresponds to *trans-cinnamate 4 hydroxylase* (C4H), an enzyme involved in the lignin biosynthesis pathway (Sewalt et al. 1997). PtTX_p14A9 belongs to the *arabinogalactane like protein* (AGP) gene family known to be involved in plant development and in xylem formation in particular (Plomion et al. 2001). PtNCS_22B8 is a *glycine hydroxymethyltranferase* involved in methyl-transfer reactions, whose related genes have been described as the most highly expressed genes in wood-forming tissues in poplar and pine (Plomion et al. 2001). Similarly, one peroxidase (PtIFG_8907) was found to co-localize with chemical properties QTLs in linkage group #8. Peroxidases are involved in lignin sub-units polymerization (Higushi 1996).

```
IFGTXP_PtIFG_9036_a   CAGGACGAAT GAGATACCTG CGTCACATGC CTCGTCGTTT CAAGAATAAT
INRCOM_PtIFG_9036_a   .......... .......... .......... .......... ..........
INRCOM_PtIFG_9036_b   .......... .......... .......... ....G..... G.........

IFGTXP_PtIFG_9036_a   TTCCGGGAGG GGACCCAGGC AACCCCCAAG AAGAAGGCTA CTGCAACCAC
INRCOM_PtIFG_9036_a   .......... .......... .......... .......... ..........
INRCOM_PtIFG_9036_b   ...T...... ..-------. ...G.A.GN- .......... ..........
                                  Indel_1
IFGTXP_PtIFG_9036_a   ATTTTAAATG ATTTAAGAAT GACCCTGAAA TGTTTTGTAG AGTTTTTAAC
INRCOM_PtIFG_9036_a   .......... .......... .......... .......... ..C.......
INRCOM_PtIFG_9036_b   .......... .......... .......... ..C...A... ..G.......

IFGTXP_PtIFG_9036_a   CACGTTTGTG AGCTGGCACA ACTTGCATTA TTTTCATATG CTGTCATAGA
INRCOM_PtIFG_9036_a   ...A...... .......... .......... .......... ..........
INRCOM_PtIFG_9036_b   .......... .......... .......... .--------- ----------
                                                                   Indel_2
IFGTXP_PtIFG_9036_a   CAGTTTTATT TTCTTTCACA ANTCGTAGTT TTCAAGTAAA ACTCATTTCC
INRCOM_PtIFG_9036_a   .......... .......... .C..A..... .......... .ACT-C....
INRCOM_PtIFG_9036_b   ---------- .......... .C..A....N .N........ ....--..T.

IFGTXP_PtIFG_9036_a   ATATTTTTGA TTGAGGTTGT ATCGGATGAC
INRCOM_PtIFG_9036_a   .......... .......... ..........
INRCOM_PtIFG_9036_b   .........- C......-.. ....-.....
```

Figure 2. Sequence alignment for amplification products of primer pair PtIFG_9036. The sequence similarity between the putative orthologous loci IFGTXP_estPtIFG_9036_a and INRCOM_estPtIFG_9036_a are characterized by few single base pair differences. The paralogous loci INRCOM_estPtIFG_9036_a and INRCOM_estPtIFG_9036_b differ by two insertions/deletions (Indel_1 and Indel_2).

Perspectives of comparative mapping in pines

Comparative genome mapping in pines suggests the genus may be considered a single genetic system. Compared to annual crops (e.g., *Triticaceae*, *Cruciferaceae*) the genome structure of Pinaceae has not evolved considerably since its divergence 100-70 million years ago. These results have to be confirmed by developing more orthologous markers, in a broader range of *Pinus* species belonging to the two subgenera of hard and soft pines, and also to species belonging to other conifer genera. For example, EST markers are currently developed in *Picea abies* and *Pseudotsuga menziesii* and will be tested in the pine species involved in the CCGP (Krutovskii et al. personal communication). A meta-analysis of all pairwise comparisons will give more support to this hypothesis.

The coincidence of candidate genes with wood density and cell wall chemistry QTLs in the maritime and loblolly pine maps is a first step towards the application of comparative genome mapping to help QTL verification. Such 'positional candidate genes' should also be validated by association studies in natural populations in both species and/or the phenotypic analysis of genetically modified pines. The ultimate goal would be to use these candidate genes as diagnostic tools to select trees presenting valuable characteristics.

Acknowledgements

We thank Sheree Cato from Forest Research (New Zealand) for providing us with her protocol on the ESTP mapping method, before it was published. We also thank Pauline Géré-Garnier from INRA Pierroton (France) for helpful comments on the manuscript. Funding was provided by the European Union (ANACONGEN: BIO4-CT97-2125, GEMINI QLRT-1999-00942), and by Plant Genome National Research Initiative (USDA NRI grant 95-37300-1632).

References

Ahuja M.R., Devey M.E., Groover A.T., Jermstad K.D. and Neale D.B. 1994. Mapped DNA probes from loblolly pine can be used for restriction fragment length polymorphism mapping in other conifers. Theor. Appl. Genet. 88: 279–282.

Brown G.R., Kadel E.E. III, Bassoni D.L., Kiehne K.L., Temesgen B., Van Buijtenen J.P. et al. 2001. Anchored reference loci in Loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics 159: 799–809.

Cato S.A., Gardner R.C., Kent J. and Richardson T.E. 2001. A rapid PCR-based method for genetically mapping ESTs. Theor. Appl. Genet. 102: 296–306.

Chagné D., Lalanne C., Madur D., Kumar S., Frigerio J.M., Krier C. et al. 2002. A high density genetic map of maritime pine based on AFLPs. Ann. For. Sci. 59: 627–636.

Chen F.Q., Foolad M.R., Hyman J., StClair D.A. and Beelaman R.B. 1999. Mapping of QTLs for lycopene and other fruit traits

in a *Lycopersicon esculentum* × *L. pimpinellifolium* cross and comparison of QTLs across tomato species. Mol. Breed. 5: 283–299.

Conkle M.T. 1981. Isozyme variation and linkage in six conifer species. In: Conkle M.T. (ed.), Proceedings of the symposium on isozymes of North American forest trees and insects. USDA Forest Service, Berkeley, CA, USA, pp. 11–17.

Devey M.D., Sewell M.M., Uren T.L. and Neale D.B. 1999. Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theor. Appl. Genet. 99: 656–662.

Dubos C. and Plomion C. 2002. Identification of water-deficit responsive genes in maritime pine (*Pinus pinaster* Ait.) roots. Plant Mol. Biol. (in press).

Echt C.S. and May-Marquardt P. 1997. Survey of microsatellite DNA in pine. Genome 39: 1102–1108.

Elsik C.G. and Williams C.G. 2000. Retroelements contribute to the excess low-copy-number DNA in pine. Mol. Gen. Genet. 264: 47–55.

Fisher P.J., Richardson T.E. and Gardner R.C. 1998. Characteristics of single- and multi-copy microsatellite from *Pinus radiata*. Theor. Appl. Genet. 96: 969–979.

Gale M.D. and Devos K.M. 1998. Comparative genetics in the grasses. Proc. Natl. Acad. Sci. USA 95: 1971–1974.

Higushi T. 1996. Biochemistry and molecular biology of wood. In: Timell T.E. (ed.), Springer, New York, pp. 169–177.

Kinlaw C.S. and Neale D.B. 1997. Complex gene families in pine genomes. Trends Plant Sci. 2: 356–359.

Komulainen P., Brown G.R., Mikkonen M., Karhu A., Garcia M.R., O'Malley D. et al. Comparing EST based genetic maps between Scots pine and Loblolly pine. (submitted).

Kriebel H.B. 1985. DNA sequence components in the *Pinus strobus* nuclear genome. Can. J. For. Res. 15: 1–4.

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E. et al. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174–181.

Myers R.M., Maniatis T. and Lerman L.S. 1987. Detection and localization of single base changes by denaturing gradient gel electrophoresis. Methods Enzymol. 155: 501–527.

Neale D.B., Sewell M.M. and Brown G.R. 2002. Molecular dissection of the quantitative inheritance of wood property traits in loblolly pine. Ann. For. Sci. 59: 595–605.

Orita M., Iwahana H., Kanazawa H., Hayashi K. and Sekiya T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand cinformation polymorphisms. Proc. Natl. Acad. Sci. USA 86: 2766–2770.

Paterson A.H. 1995. Molecular dissection of quantitative traits: progress and prospects. Genome Research 5: 321–333.

Plomion C., Durel C.-E. and O'Malley D.M. 1996. Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. Theor. Appl. Genet. 93: 849–858.

Plomion C., Costa P. and Bahrman N. 1997. Genetic analysis of needle protein in Maritime pine. 1. Mapping dominant and codominant protein markers assayed on diploid tissue, in a haploid-based genetic map. Silvae Genetica 46: 161–165.

Plomion C., Hurme P., Frigerio J.M., Ridolfi M., Pot D., Pionneau C. et al. 1999. Developing SSCP markers in two *Pinus* species. Molecular Breeding 5: 21–31.

Plomion C., Le Provost G. and Stokes A. 2001. Wood formation in trees. Plant Physiol. 127: 1513–1523.

Plomion C., Cooke J., Richardson T., McKay J. and Tuskan G. 2003. Conference Report on the Forest Trees Workshop at the Plant and Animal Genome Conference. January 12th 2003; San Diego, CA, USA. Comp Funct Genom. (in press).

Perry D.J. and Bousquet J. 1998. Sequence-tagged-site (STS) markers of arbitrary genes: development, characterization and analysis of linkage in black spruce. Genetics 149: 1089–1098.

Prager E.M., Fowler D.P. and Wilson A.C. 1976. Rates of evolution in conifers. Evolution 30: 637–649.

Price R.A., Liston A. and Strauss S.H. 1998. Phylogeny and systematics of *Pinus*. In: Richardson D.M. (ed.), Ecology and Biogeography of *Pinus*. Cambridge Univ. Press, Cambridge, pp. 49–68.

Saylor L.C. 1971. Karyotype analysis of the genus *Pinus* – subgenus *Pinus*. Silvae Genetica 21: 5.

Schiex T., Chabrier P., Bouchez M. and Milan D. 2001. Boosting EM for Radiation Hybrid and Genetic Mapping. In: The proceedings of WABI'2001 (First Workshop on Algorithms in Bioinformatics), LNCS 2149. Website: http://www.inra.fr/bia/T/CarthaGene/.

Sewalt V.J.H., Ni W., Blount J.W., Jung H.G., Massoud S.A., Howles P.A. et al. 1997. Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of L-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase. Plant Physiol. 115: 41–50.

Sewell M.M., Bassoni D.L., Megraw R.A., Wheeler N.C. and Neale D.B. 2000. Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. Theor. Appl. Genet. 101: 1273–1281.

Sewell M.M., Davis M.F., Tuskan G.A., Wheeler N.C., Elam C.C., Bassoni D.L. et al. 2002. Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). II. Chemical wood properties. Theor. Appl. Genet. 104: 214–222.

Temesgen B., Brown G.R., Harry D.E., Kinlaw C.S., Sewell M.M. and Neale D.B. 2001. Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.). Theor. Appl. Genet. 102: 664–675.

Van Ooijen J.W. and Voorrips R.E. 2001. Joinmap 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands, Website: http://www.joinmap.nl.

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 QTLs in interspecific hybrid progeny, stability of QTL expression across different ages. Theor. Appl. Genet. 95: 597–608.

Voorrips R.E. 2001. MapChart version 2.0: Windows software for the graphical presentation of linkage maps and QTLs. Plant Research International, Wageningen, The Netherlands.

Wakayima I., Newton R., Johnston J.S. and Price H.J. 1993. Genome size and environmental factors in the genus *Pinus*. Am. J. Bot. 80: 1235–1241.

Whetten R., Sun Y.H., Zhang Y. and Sederoff R. 2001. Functional genomics and cell wall biosynthesis in loblolly pine. Plant Mol. Biol. 47: 275–291.

ANNEXE IV

Cross species transferability and mapping of genomic and cDNA SSRs in pines

Development of SSR loci					Transfer of SSR loci		
Species	SSR origin	Number of primer pairs tested	Number of polymorphic single copy SSR loci	Reference	Number of markers tested in other conifers	Number of successfully transferred markers	Reference
Pinus radiata $^{H\text{-}am}$	EGL	2	2	Smith and Devey 1994	2	2^{H} 0^{S} 0^{P}	Echt et al. 1999
					2	$1^{H\text{-}me}$	Karhu et al. 2000
					2	$0^{H\text{-}me}$	Mariette et al. 2001
					2	$0^{H\text{-}am}$	Shepherd et al. 2002
Pinus radiata $^{H\text{-}am}$	TGL }	43*	2	Fisher et al. 1998			
	EGL }		11		7	7^{H} 3^{S} 1^{P}	Fisher et al. 1998
					4	$3^{H\text{-}me}$ $3^{H\text{-}am}$ 0^{S} 0^{P}	Echt et al. 1999
					2	$0^{H\text{-}me}$	Mariette et al. 2001
					7	$4^{H\text{-}am}$	Shepherd et al. 2002
					20	$11^{H\text{-}am}$	Devey at al. 1999
Pinus radiata $^{H\text{-}am}$	EGL	50	10	Devey et al. 2002			
Pinus taeda $^{H\text{-}am}$	ELCL	18	16	Elsik et al. 2000b	7	$7^{H\text{-}am}$ $5^{H\text{-}me}$	Kutil and Williams 2001
					25**	$13^{H\text{-}am}$	Shepherd et al. 2002
					19**	$10^{H\text{-}me}$	Gonzalez-Martinez et al. 2003
Pinus taeda $^{H\text{-}am}$	ELCL	29	15	Elsik and Williams 2001			
	EGL	37	8				
Pinus taeda $^{H\text{-}am}$	ELCL	8	8	Kutil and Williams 2001	8	$8^{H\text{-}am}$ $2^{H\text{-}me}$	Kutil and Williams 2001
Pinus taeda $^{H\text{-}am}$	EUML	36	19	Zhou et al. 2002			
Pinus contorta $^{H\text{-}am}$	EGL	5	5	Hicks et al. 1998			
Pinus sylvestris $^{H\text{-}me}$	TGL	2	0	Kostia et al. 1995			
Pinus sylvestris $^{H\text{-}me}$	EGL	37	7	Soranzo et al. 1998	3	$3^{H\text{-}me}$	Gonzalez-Martinez et al. 2003
Pinus halepensis $^{H\text{-}me}$	EGL	25	8	Keys et al. 2000	8	$7^{H\text{-}me}$	Keys et al. 2000
					8	$1^{H\text{-}me}$	Mariette et al. 2001
Pinus pinaster $^{H\text{-}me}$	EGL	29	2	Mariette et al. 2001			
Pinus densiflora $^{H\text{-}as}$	EGL	14	6	Lian et al. 2000	6	$6^{H\text{-}as}$ $5^{H\text{-}am}$ 0^{S}	Lian et al. 2000
Pinus strobus ^{S}	EGL	77	19	Echt et al. 1996	15	12^{S} 0^{H} 0^{P}	Echt et al. 1999
					28	3^{H}	Karhu et al. 2000
					4	0^{H}	Mariette et al. 2001
					5	0^{H}	Shepherd et a. 2002
Pinus strobus ^{S}	EGL	4	0	Echt et al. 1999	4	0^{S}	Echt et al. 1999
Picea sitchensis ^{SP}	EGL	7	4	van de Ven and Mac Nicol 1996			
Picea abies ^{SP}	EGL	36	7	Pfeiffer et al. 1997			
Picea abies ^{SP}	EGL	96	34	Paglia et al. 1998			
Picea abies ^{SP}	ECDL	6	6	Scotti et al. 2000			
Picea abies ^{SP}	EGL	55	16	Scotti et al. 2002a (tri)			
Picea abies ^{SP}	EGL	53	16	Scotti et al. 2002b			
Picea glauca ^{SP}	EGL	13	13	Hodgetts et al. 2001	13	12^{SP}	Hodgetts et al. 2001
Picea glauca ^{SP}	EGL	16	6	Rajora et al. 2001	6	6^{SP}	Rajora et al. 2001
Pseudotsuga menziesii ^{P}	EGL	102	48	Amarasinghe and Carlson 2002	50	31^{P}	Amarasinghe and Carlson 2002
Tsuga heterophylla ^{P}	EGL	16	11	Amarasinghe et al. 2003			
Cryptomeria japonica ^{P}	EST	3	2	Moriguchi et al. 2003			
	EGL }	67*	31				
	TGL }		1				
Total	-	776	333 (43%)		213	108 (50%)	

Table 1: Literature review on SSR development and transferability in the Pinaceae.
TGL: total genomic DNA library; EGL: enriched genomic DNA library; ELCL: enriched low-copy DNA library; ECDL: enriched cDNA library; EUML: enriched undermethylated DNA library; EST: developed from EST sequences.
*: primer pairs designed from both EGL and TGL (not differentiated in the publication);
**: primer pairs combined with Elsik and Williams (2001) primers (not differentiated in the publication).
Amplification in other species: H: hard pine subgenus *Pinus*; S: soft pine subgenus *Strobus*; SP: spruce genus *Picea*; P: Pinaceae species other than *Pinus*; am: American; as: Asian; me: Mediterranean pines.

Cross species transferability and mapping of genomic and cDNA SSRs in pines.

Chagné David[1], Chaumeil Philippe[1], Ramboer Agnès[1], Collada Carmen[2], Guevara Angeles[3], Cervera Maria Teresa[3], Vendramin Giovanni Giuseppe[4], Garcia Virginie[5], Frigerio Jean-Marc[1], Echt Craig[6], Richardson Thomas[6], Plomion Christophe[1#]

1: UMR BIOGECO – INRA, Equipe de génétique – 69 route d'Arcachon, 33610 Cestas Cédex, France
2: Departamento de Biotecnología, Escuela Técnica Superior de Ingenieros de Montes de Madrid, UPM, Ciudad Universitaria sn, 28040 Madrid, Spain
3: Departamento de Genética Forestal, CIFOR-INIA, Carretera de la Coruna Km 7, 28040 Madrid, Spain
4: Istituto di Genetica Vegetale, Sezione di Firenze, Consiglio Nazionale delle Ricerche, Via Madonna del Piano, 50019 Sesto Fiorentino, Firenze, Italy
5: UMR Physiologie et Biotechnologie Végétale, INRA Bordeaux, 71 avenue E. Bourleaux, 33883 Villenave d'Ornon Cédex, France
6: New Zealand Forest Research Institute Ltd, Private Bag 3020, Rotorua, New Zealand

Author for correspondence: plomion@pierroton.inra.fr
Phone: 33 5 57 12 28 38
Fax: 33 5 57 12 28 81

Abstract

Two unigene datasets of *Pinus taeda* and *Pinus pinaster* were screened to detect di-, tri- and tetranucleotide repeated motifs using the *SSRIT* script. A total of 419 SSRs were identified, from which only 12.8% overlapped between the two sets. The position of the SSRs within coding sequence were predicted using *FrameD*. Trinucleotide was the most abundant repeated motif (63% and 51% in *P. taeda* and *P. pinaster*, respectively) and tend to be found within translated regions (76% in both species), whereas dinucleotide repeats were preferentially found within the 5'- and 3'-untranslated regions (75% and 65%, respectively). Fifty three primer pairs amplifying a single PCR fragment in the source species (mainly *P. taeda*), were tested for amplification in six other pine species. Amplification rate was high and agreed with the phylogenetic distance between species, varying from 64.6% in *P. canariensis* to 94.2% in *P. radiata*. Comparatively, genomic SSRs were found to be less transferable: 58 of the 107 primer pairs (i.e. 54%) derived from *P. radiata* amplifying a single fragment in *P. pinaster*. The level of polymorphisms of cDNA- and genomic-SSRs was compared in *P. pinaster*. No differences in terms of mapped loci, heterozygosity and number of alleles were found. This study suggests that cDNA-SSR developed from pine ESTs are valuable genetic markers for the genus Pinus and can be used for a variety of purposes, in particular as orthologous loci for comparative genome mapping in conifers.

Introduction

Microsatellites, also known as Simple Sequence Repeats (SSRs), are short motifs (1-6 bp) that are tandemly repeated, like (AC)n, (CCG)n or (AAAG)n. Due to their genetic features (codominant mode of inheritance and highly polymorphic), they are commonly used for genetic mapping, QTL detection, population genetic studies and genetic resource management in plant and animal species (Goldstein and Schlotterer 1999). Compared to other plants, and despite a number of attempts, only few polymorphic and single copy nuclear microsatellite markers have been developed in the Pinaceae (reviewed in table 1). The genome structure of these species, characterised by a large physical

size (22 pg/C, Leitch et al. 2001) with a high amount of repeated sequences (Kriebel 1985, Kamm et al. 1996, Kossack and Kinlaw 1999, Elsik and Williams 2000a) has been the main obstacle to the development of useful SSR markers. In addition, the ancient divergence time between coniferous species (Price et al. 1998) and the high complexity of their genome means that transferability of single-copy SSRs between genera and even between *Pinus* species (the most studied genus), is generally poor, resulting in a large proportion of amplification failure, non-specific amplification, multibanding patterns or lack of polymorphism (Echt et al. 1999, Mariette et al. 2001). Given the high cost of developing useful SSR markers, cross-species transferability is a valuable attribute.

In an attempt to circumvent these genome related problems, Elsik and Williams (2001) removed most of the repetitive portion of the genome using a DNA reassociation kinetics-based method and Zhou et al. (2002) targeted the low-copy portion of the genome using undermethylated regions enrichment method. Both approaches yielded remarkable enrichment for useful SSR markers in *Pinus taeda*. Scotti et al. (2002), used an alternative strategy based on the pre-screening of single-copy microsatellite containing clones, using dot blot hybridisation analysis, and also obtained a high amount of single-copy polymorphic SSR markers in *Picea abies*. The *Pinus taeda* SSRs developed by Elsik and Williams (2001) and Zhou et al. (2002), transferred quite well between American hard pines (Shepherd et al. 2002), but were shown to be less transferable in the phylogenetically more distant Mediterranean hard pines (Gonzalez-Martinez et al. 2003). Interestingly, perfect trinucleotide SSRs transferred better than other motifs from American to Mediterranean pines (Kutil and Williams 2001).

SSRs have been proven to be present in all genomic regions, including coding regions (Toth et al. 2000). By developing a cDNA library enriched in SSRs, Scotti et al. (2000) showed the presence of microsatellites within the coding regions of Norway spruce (*Picea abies*), a species belonging to the Pinaceae. The availability of Expressed Sequence Tags (ESTs) resulting from large sequencing projects is potentially a valuable source of SSRs, that can be evaluated without intensive laboratory development. Recently, cDNA-SSRs were obtained from EST databases developed in several plant species such as grape (Scott et al. 2000), cereals (Temnykh et al. 2000, 2001, Cho et al. 2000, Cordeiro et al. 2001, Kantety et al. 2002, Eujayl et al. 2002, Varshney et al. 2002, Gao et al. 2003) and Arabidopsis (Cardle et al. 2000, Morgante et al. 2002). These EST-derived markers showed good transferability between phylogenetically related species (Eujayl et al. 2003, Gupta et al. 2003).

The objectives of this study were three fold: i) quantify the amount and identify the types of SSRs present in the coding regions of two pine genome sequences, ii) compare the polymorphism information content of SSRs derived from cDNA and genomic sources, and iii) evaluate the transferability of cDNA versus genomic sequence derived SSRs across several pine species. To address the first question, 9 cDNA and 10 genomic SSR loci were genotyped in a sample of *Pinus pinaster* individuals. Since only three genomic SSRs were available for Pinus pinaster (Mariette et al. 2001), we transferred a number of unpublished genomic SSRs from *Pinus radiata* to *Pinus pinaster*. Both experiments contributed to evaluate the transferability rate of SSRs among pines.

Material and methods

In silico SSR detection in pine ESTs

The *P. pinaster* and *P. taeda* contigs and singletons in publicly available EST

Table 2: Sequences used for the construction of the pine interpolated Markov model.

Species	Accession	Coding sequence length (bp)
Pinus pinaster	AB084493	1575
Pinus pinaster	AF448201	2780
Pinus pinaster	AJ315675	926
Pinus pinaster	AJ490522	1087
Pinus pinaster	AY168850	1209
Pinus pinaster	AY168851	1209
Pinus pinaster	AY321087	1209
Pinus pinaster	AY321088	2222
Pinus pinaster	AY321089	2222
Pinus radiata	AF023615	682
Pinus radiata	AF036095	794
Pinus radiata	AF039566	303
Pinus radiata	AF049065	700
Pinus radiata	AF049066	648
Pinus radiata	AF049067	1691
Pinus radiata	AF049068	898
Pinus radiata	AF049069	666
Pinus radiata	AF060491	1093
Pinus radiata	AF109149	1258
Pinus radiata	AF110333	977
Pinus radiata	AF119225	1170
Pinus radiata	AF120097	676
Pinus radiata	U42399	752
Pinus radiata	U42400	743
Pinus radiata	U70873	1170
Pinus radiata	U76725	1560
Pinus radiata	U76726	801
Pinus radiata	U76756	1575
Pinus radiata	U76757	1237
Pinus radiata	U90341	1194
Pinus radiata	U90342	392
Pinus radiata	U90343	310
Pinus radiata	U90344	679
Pinus radiata	U90345	657
Pinus radiata	U90346	685
Pinus radiata	U90347	657
Pinus radiata	U90348	502
Pinus radiata	U90349	611
Pinus radiata	U90350	285
Pinus radiata	U92008	1258
Pinus radiata	AF001136	1340
Pinus taeda	AF013802	511
Pinus taeda	AF013803	428
Pinus taeda	AF013804	700
Pinus taeda	AF013805	630
Pinus taeda	AF081678	944
Pinus taeda	AF085330	776
Pinus taeda	AF096998	1548
Pinus taeda	AF101785	724
Pinus taeda	AF101786	267
Pinus taeda	AF101787	404
Pinus taeda	AF101788	624
Pinus taeda	AF101789	428
Pinus taeda	AF101790	371
Pinus taeda	AF101791	395
Pinus taeda	AF103808	1685
Pinus taeda	AF130440	1045
Pinus taeda	AF132119	1792
Pinus taeda	AF132120	1761
Pinus taeda	AF132121	1755
Pinus taeda	AF132122	1743
Pinus taeda	AF132123	1807
Pinus taeda	AF132124	1767
Pinus taeda	AF132125	1697
Pinus taeda	AF132126	1764
Total	65	67969

databases gave us access to the non-redundant coding portion of the genome (unigene). ESTs were independently assembled for each species. A total of 18,498 maritime pine ESTs were assembled using *StackPack* (Christoffels et al. 2001) providing 2,893 contigs and 5,001 singletons available at the following URL: http://cbi.labri.u-bordeaux.fr/outils/SPAM/. For loblolly pine, the 8,070 contigs and 12,307 singletons resulting from 75,047 ESTs were available at the following URL: http://web.ahc.umn.edu/biodata/nsfpine/contig_dir16/.

We searched the *P. pinaster* and *P. taeda* unigene sets for tandemly repeated motifs of 2, 3 and 4 bp using the *SSRIT* SSR search tool (Temnykh et al. 2001; http://www.gramene.org/db/searches/ssrtool), with 14, 15 and 20 as minimum repeat length, respectively. We associated the *SSRIT Perl* script with the *FrameD* gene prediction software (Schiex et al. 2003) to determine if the detected repeated motifs were located in the 5' or 3' untranslated regions (UTRs) or in the open reading frame (ORF). *FrameD* was developed to predict the position of the translated regions in EST sequences. Because *FrameD* uses Interpolated Markov Models (IMM; Salzberg et al. 1998) to build probabilistic models of coding sequences, a pine-specific IMM was constructed to enhance the prediction in *P. taeda* and *P. pinaster* sequences. We used 67 kb from 65 pine full-length coding sequences to build the *Pinus* IMM (table 2). Finally, the sequences containing microsatellites in maritime and loblolly pines were compared in order to check the redundancy of the sequences containing SSRs in both species.

PCR primer design and amplification

We designed 56 PCR primer pairs flanking the microsatellites identified with our in situ analysis using *Primer v3.0* software (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) with default parameters, except that we used a range of 40-55% for the *Primer GC%*, *GC clamps* of 2 bases and a *Max Tm difference* of 10. We kept the expected amplified fragment length below 500 bp to avoid the risk of the presence of intron, which may induce PCR failure. Fifty-three out of 56 PCR primers were designed based on *P. taeda* sequences and 3 were developed from *P. pinaster* sequences. The PCR primers were chosen to represent the broadest range of SSR possible considering the repeat type (di-, tri- or tetranucleotide), the motif (e.g. AG, AT), the length (5 to 26 repeats) and the position (UTR or ORF). In addition to these new SSRs, we also included in our study a set of cDNA-SSRs previously developed from *P. taeda* sequences (C. Echt, unpublished) and available at the following URL: http://dendrome.ucdavis.edu/Gen_res.htm. This set resulted from a SSR search using a preliminary sequence dataset of about 10,000 *P. taeda* ESTs.

A third set of 107 PCR primers was developed from *P. radiata* genomic SSRs and screened for amplification success in pine species (C. Echt and T. Richardson, unpublished).

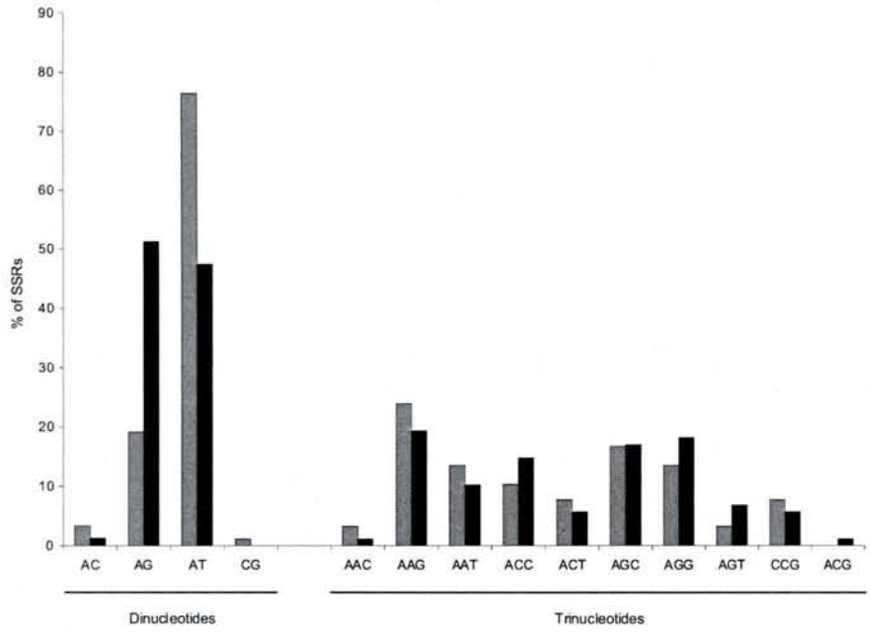
DNA was isolated using the protocol described by Doyle and Doyle (1990). PCR reactions were performed with 15 ng of genomic DNA in a total reaction volume of 10 µl, with 1X reaction buffer (Gibco BRL), 2 mM MgCl2, 1 µM of each primer, 0.2 mM of dNTP and 0.5 U of Taq polymerase (Gibco BRL) on a Stratagene Robocycler□ Gradient 96 (Stratagene, La Jolla, CA, USA) using the following cycles: preliminary denaturing (94°C, 5 min) followed by 30 cycles of denaturing (94°C, 30s), annealing (locus-specific temperature, 30s), and extension (72°C, 1 min), and a final extension (72°C, 10 min). An additional touchdown was performed for some loci (10 cycles with annealing temperature decreasing for 1°C every cycle).

Amplification success was checked on 1.5% agarose gels. We checked that the

Table 4: Di-, tri- and tetranucleotide SSR detection in *Pinus pinaster* and *Pinus taeda* unigenes using *SSRIT* software.

	Pinus taeda			*Pinus pinaster*		
	Contigs	Singletons	Total	Contigs	Singletons	Total
Number of sequences analysed	8,070	12,307	20,377	2,893	5,001	7,894
Number of dinucleotide SSRs (n>7)	41	48	89	41	35	76
Number of trinucleotide SSRs (n>5)	103	56	159	43	44	87
Number of tetranucleotide SSRs (n>5)	1	2	3	3	2	5
Total number of SSRs	145	106	251	87	81	168
Enrichment (%)			1.2%			2.1%

Figure 1: Distribution of the different classes of di- and trinucleotide SSRs in *Pinus taeda* (grey box) and *Pinus pinaster* (black box) unigenes.



amplification showed a single band pattern with a size corresponding to the expected length. Amplification resulting in multiple bands were discarded from further analysis since they could result from non-specific amplification or paralogous loci.
The useful loci were then run on a LICOR automated sequencer with the same conditions as described by Mariette et al. (2001) to precisely determine the length of each amplification product (i.e., allele).

Sequencing

Amplified fragments in maritime pine were cloned and sequenced as described by Dubos et al. (2003) in order to check the orthology of same markers based on sequence identity.

Plant material

Polymorphism and reliable co-dominant inheritance was tested in three maritime pine mapping pedigrees (INRA-F2 pedigree, Costa et al. 2000; INRA-G2 pedigree Chagné et al. 2002; AFOCEL-F1 pedigree, Ritter et al. 2002) for which saturated genetic maps are available, and a fourth (INIA-F1) which is under construction (M.T. Cervera, unpublished). Loci that were polymorphic in at least one mapping pedigree were also tested on twenty-six unrelated *P. pinaster* elite trees from the Aquitaine region (south-western France). These trees are the first generation of the maritime pine breeding programme and were used to estimate the level of diversity (heterozigozity and number of alleles) of the SSRs.
Samples from 7 species belonging to the genus *Pinus* (subgenus *Pinus*): *P. canariensis*, *P. halepensis*, *P. pinaster*, *P. pinea*, *P. radiata*, *P. sylvestris*, and *P. taeda* were used to test the amplification rate of the cDNA-SSR markers.

Data analysis

The markers segregating in the INRA-G2 and INRA-F2 mapping pedigree were visually scored and assigned two allele genotypes. We used *Joinmap* v3.0 (VanOijen et al. 2001) using a minimum LOD of 6 for genetic map construction. The *Arlequin* software (Schneider et al. 2000) was used to estimate genetic diversity parameters based on the genotypes of the 26 unrelated P. pinaster individuals.

Results

SSR detection in pine ESTs and sequence annotation

A total of 251 and 168 SSRs were found in *P. taeda* and *P. pinaster* unigene sets (table 3). This corresponds to enrichment rates of 1.2% and 2.1%, respectively (table 4). The most common repeat types were trinucleotides (63% in *P. taeda* and 51% in *P. pinaster*), followed by dinucleotides (36% in *P. taeda* and 45% in *P. pinaster*). Tetranucleotide repeats were almost absent (1% in *P. taeda* and 3% in *P. pinaster*). These results were obtained for a minimum repeat number of 7, 5 and 5 for di-, tri- and tetranucleotide motifs, respectively. These thresholds are comparable to those used by Cardle et al. (2000) and Scott et al. (2000), and correspond to perfect motifs only. If we used less stringent detection criteria (e.g. minimum of 5 repeats for dinucleotides, as in Morgante et al. 2002) and allowed the detection of compound motifs we have estimated that the SSR enrichment would increase by two fold.
Regarding the types of repeated motif (figure 1), the AT and AG motifs were the most represented among the dinucleotides (76% and 19% in *P. taeda*, and 47% and 51% in *P. pinaster*, respectively), whereas the AC and CG types were rare (less than 3% in both species). Regarding trinucleotides, the AAG motif was the most common repeat type (23.9% and

Table 5: Cross-specific amplification of 53 cDNA-SSR markers: locus ID and amplification in 7 hard pine species.

Locus nomenclature follows the recommendations of the Treegenes database (http://dendrome.ucdavis.edu/Tree_Page.htm) for pine STS, also described by Brown et al. (2001). Accession: **Pinus taeda* unigene contig numbering (http://web.ahc.umn.edu/biodata/nsfpine/contig_dir16/), **Genbank accession or ***Pinus pinaster* unigene contig numbering (http://cbi.labri.u-bordeaux.fr/outils/SPAM/). Position in the gene: UTR: untranslated region; ORF: open reading frame; NP: no protein. PCR conditions: annealing temperature (°C) or touchdown temperature range. Amplification: *Pp*: *Pinus pinaster* (subsection Sylvestres); *Pt*: *Pinus taeda* (subsection Australes); *Pr*: *Pinus radiata* (subsection Oocarpae); *Ps*: *Pinus sylvestris* (subsection Sylvestres); *Ph*: *Pinus halepensis* (subsection Sylvestres); *Ppi*: *Pinus pinea* (subsection Pineae); *Pc*: *Pinus canariensis* (subsection Canarienses); +: single locus amplification; -: no amplification; na: no data.

Loci information									Amplification						
Locus name	Identification	Repeated motif	Number of repeat	Position in gene	Forward primer	Reverse primer	Annealing temperature	Expected length (bp)	*Pp*	*Pt*	*Pr*	*Ps*	*Ph*	*Ppi*	*Pc*
RPtest1	Contig4518*	AAT	7	5' UTR	gatcgttattcctcctgcca	ttcgatatcctccctgcttg	50	125	+	+	+	+	+	+	+
RPtest5	Contig6309*	AAC	6	ORF	acaacaataataacgggggc	acgctttagatcctcctgca	55	197	+	+	+	+	+	+	+
RPtest6	Contig3845*	TGC	5	ORF	aggattccaacagcatcacc	ctgaacatgaagcgcagtgt	55	147	+	+	+	+	+	+	+
RPtest8	Contig8048*	CCG	6	ORF	ggtgcgagattgaaattcgt	tttgcagtctgttgcctttg	60-50	196	-	+	+	na	na	na	na
RPtest9	Contig1667*	AGC	10	ORF	ccagacaacccaaatgaagg	gcctgctatcgaatccagaa	51	289	+	+	+	+	+	+	+
RPtest11	Contig3631*	ATC	7	3' UTR	aggatgcctatgatatgcgc	aaccataacaaaagcggtcg	56	213	+	+	+	+	-	+	+
RPtest13	AA739656**	CTG	5	ORF	gatttttcaggaagaccccc	tgtaaggcacaagccctctt	51	277	+	+	+	+	-	-	+
RPtest15	Contig8064*	ACC	6	ORF	gaacgtggttatggcggtag	ccagggacagttaccagcat	56	246	+	+	+	+	+	+	+
RPtest16	AA739818**	AGT	5	ORF	cagaaatggcgtccaaattc	accccacttatatccccagc	56	132	+	+	+	+	-	+	+
RPtest20	Contig6393*	AGC	5	ORF	gttcccactcaagggttgaa	acatcatttgttgccgcata	56	259	+	+	+	-	-	-	-
RPtgbLP5	AF013805**	AAT	6	5' UTR	agaggttccaaacgagaggt	tcgacttctgatttctttacatga	60-50	176	-	+	-	na	na	na	na
ssrPp_cn524	Contig524***	AG	14	5' UTR	cgattgtttttgccttttaagc	aaatatggcggggtgtgc	50	156	+	+	+	+	+	+	-
ssrPt_AA739797	AA739797**	AT	11	3' UTR	actttgcggtgaatcagacc	aaagtaaggctgcttgcatga	51	281	+	+	+	-	-	-	-
ssrPt_AW010960	AW010960**	AT	9	ORF	atcgactaggcatcaggtgg	tcctcgtagcccagcttttta	49	225	+	+	+	+	+	+	+
ssrPt_AW225917	AW225917**	AT	9	3' UTR	tgcattgaaaaatacagcgg	attatgtacgaggccccaca	49	198	+	+	+	+	+	+	+
ssrPt_AW981642	AW981642**	AAG	7	ORF	gtggcacagggttttctgat	caaaccttcggtagcctcac	60-50	245	-	+	+	na	na	na	na
ssrPt_AW981772	AW981772**	CCT	4	ORF	gatcctgttcctcctcctcc	cctggacagaaacagcaaca	49	266	+	+	+	+	+	+	+
ssrPt_BF049767	BF049767**	AG	22	ORF	ttttgggtcgtaggaacctg	taaaacgggtgtctcttcgg	51	227	+	+	+	+	+	+	-
ssrPt_BF778306	BF778306**	AG	7	NP	gaagatggagacgaagcagg	tttgcagtctgttgcctttg	60-50	172	-	+	+	na	na	na	na
ssrPt_ctg1376	Contig1376*	AT	20	NP	cgatattatggattttgcttgtga	aaatgcatgccaaacttaaatac	60-50	145	+	+	+	+	-	-	-
ssrPt_ctg1525	Contig1525*	AGG	7	ORF	ttgaaaccatataagcaatgcc	aggacctgggtaaggaggc	60-50	173	+	+	+	+	+	+	+
ssrPt_ctg16480	Contig16480*	AAAT	13	NP	ctaaaacatcggtcggaagc	atttagtccaggccatgtcg	60-50	151	+	+	na	na	na	na	na
ssrPt_ctg16811	Contig16811*	AT	11	5' UTR	tccatgatgttgcagattgg	gtgttccccaatggtctgtc	56	199	+	+	+	-	-	-	+
ssrPt_ctg17607	Contig17607*	AAG	9	ORF	cgccattaatatgcctaccg	atctctgcgctgcttgaagt	54	225	+	+	+	+	+	+	+
ssrPt_ctg18103	Contig18103*	AT	10	NP	cctggattcatttgtggctaa	catgccaacttcttgcattg	60	184	+	+	-	+	+	+	+
ssrPt_ctg2300	Contig2300*	CCG	6	ORF	cactttgcgagagactgcac	acgctgaaggaaatcgagaa	49	173	+	+	+	+	+	+	+
ssrPt_ctg275	Contig275*	AT	16	3' UTR	acggagatatattgctggcg	aaagaataacgtgaaacaaaccc	60-50	137	+	+	+	+	-	-	-
ssrPt_ctg3021	Contig3021*	AGC	14	ORF	ctcagattcctccaaatgcg	catgcaacatatgcaaaccg	60-50	234	+	+	+	+	+	+	+
ssrPt_ctg3089	Contig3089*	AT	17	NP	ctttcttcacgttggacttctt	ttagccatggagagtgcaga	45	482	-	+	+	+	+	+	+
ssrPt_ctg3754	Contig3754*	AGC	6	5' UTR	tctttgggtttctggagtgg	gctgttgctgttgttcttgg	60-50	421	+	+	+	+	+	+	+
ssrPt_ctg4363	Contig4363*	AT	10	3' UTR	taataattcaagccaccccg	agcaggctaataacaacacgc	60-50	100	+	+	+	+	+	+	+
ssrPt_ctg4487a	Contig4487*	CCG	5	ORF	tctgctgtgtggacaaacct	ttcttggctcaaaatctcgg	60-50	155	+	+	+	+	+	-	-
ssrPt_ctg4487b	Contig4487*	CCG	10	3' UTR	atgacgcattatcaggggaa	ttgcacagaaagcaggtttg	45	254	+	+	+	+	+	-	-
ssrPt_ctg4698	Contig4698*	ATC	10	ORF	cgaaaaggtggttctgatgg	ttttccgctggatttaccac	49	246	+	+	+	+	+	+	+
ssrPt_ctg5167	Contig5167*	AAC	7	ORF	tgcagagagattcgatggg	attttggtttgtttgctggc	60-50	293	+	+	+	+	+	+	+
ssrPt_ctg5333	Contig5333*	AGC	7	ORF	gaaggagtcggcgataacag	gggaattcgacctgtgaaga	49	163	+	+	+	+	-	-	-
ssrPt_ctg6390	Contig6390*	AAG	8	5' UTR	atccacgacttgtcgacgc	atcaaccaacttaggcagcg	45	440	-	+	+	-	+	+	+
ssrPt_ctg64	Contig64*	CCG	7	ORF	ggaagctgttacaagtgcgg	atcgagaagagaggaagggc	60-50	284	+	+	+	+	+	+	+
ssrPt_ctg7024	Contig7024*	AAG	7	ORF	gggaattctgaaagacaaggg	aacttaccatcgagagcccc	60-50	277	+	+	+	+	+	+	-
ssrPt_ctg7081	Contig7081*	AAG	7	ORF	gtcatccacgttcattggc	tcacaactgaccaaactgcc	60-50	442	+	+	+	+	+	+	-
ssrPt_ctg7141	Contig7141*	CGG	8	ORF	gaatgacgcattatcagggg	tcacctttctcacctctgcc	45	381	-	+	+	+	+	+	+
ssrPt_ctg7170	Contig7170*	AGC	5	ORF	ggtttttcgatttctgaggc	aacaggtgtgcaaatagccc	60-50	385	+	+	+	+	+	+	-
ssrPt_ctg7425	Contig7425*	AAG	6	ORF	aataagaccccagaggagcc	gacgtctttcaccaaatcgc	60-50	384	+	+	+	+	+	-	-
ssrPt_ctg7444	Contig7444*	AT	10	5' UTR	tcttcaccatcggtttctcc	tggatctgtcacctcctcatc	58	285	+	+	+	+	+	+	+
ssrPt_ctg7731	Contig7731*	AT	12	5' UTR	agtggtgaagggtccatctg	gcataacacaaaagccagca	51	217	+	+	+	+	+	+	+
ssrPt_ctg7824	Contig7824*	AT	12	3' UTR	tgacctgtcttgtgagacgc	ttttgaaacagattgcagcc	60-50	501	+	+	+	+	+	-	-
ssrPt_ctg7867	Contig7867*	CCG	6	5' UTR	ggtcgtggaggaggtaggg	actgataacagctgccccc	45	154	+	+	+	+	-	+	+
ssrPt_ctg8064	Contig8064*	ACC	6	ORF	gaacgtggttatggcggtag	tcgtggcaactatctcctcc	50	147	+	+	+	+	+	+	+
ssrPt_ctg865	Contig865*	AT	15	3' UTR	tttcagaagctcccgatttg	cttgtggacatggttaatgaag	45	232	+	+	+	+	+	+	+
ssrPt_ctg8767	Contig8767*	AGC	8	ORF	tggggaaaaatggcatacat	ggagcagacacccatggact	55	180	+	+	+	-	-	-	-
ssrPt_ctg9249	Contig9249*	AAG	7	5' UTR	ctgctccctcagctcttcc	agacgtcactgccattaccc	55	156	+	+	-	+	+	+	+
ssrPt_ctg946	Contig946*	AGG	9	3' UTR	tatcaggtataggcctccgc	aaataggagcccttctggga	53	287	+	+	+	-	-	-	-
ssrPt_ctg988	Contig988*	AT	7	3' UTR	taataattcaagccaccccg	aacattttgcacgatagccc	51	319	+	+	+	-	-	-	-
							Amplification rate (%)		86.8	100	94.2	85.4	72.9	70.8	64.6

19.3% in *P. taeda* and *P. pinaster*, respectively), followed by AGC and AGG motifs.

Figure 2 shows the position of the detected SSRs in the gene sequences of both species based on the results obtained with *FrameD* (Schiex et al. 2003). Significant differences between di- and trinucleotide SSRs were observed. Dinucleotides were found mostly in the UTRs (75% and 65% in P. taeda and *P. pinaster*, respectively), whereas trinucleotides were more frequent in the ORFs (76% in both species). For both type of repeats, SSRs were less abundant in the 5' UTR than in the 3' UTR.

By assembling the *P. taeda* and *P. pinaster* contigs and singletons that contained SSRs using *StackPACK* (Christoffels et al. 2001), we found that only 22 of the 171 (12.8%) *P. pinaster* sequences matched in the *P. taeda* unigene, providing a catalogue of 397 non-redundant putative SSR markers for pines.

Transferability of cDNA and genomic SSRs in pines

As a representative sample, 72 primer pairs were designed from EST sequences. Fifty-two out of the 69 P. taeda and 1 out of the 3 *P. pinaster* cDNA-SSRs amplified a single band of the expected size in the source species. The multi-banding pattern observed for 5 loci could be attributed to non-specific amplifications or the presence of multigene families, that are frequent in pines (Kinlaw and Neale 1997). The lack of amplification obtained for 14 loci, could be explained by the quality of the primer pairs and/or the presence of introns. Table 5 summarises the amplification success for these 53 cDNA-SSR markers in 7 pine species. Overall, the amplification rates in non-source species ranged between 64.6% in *P. canariensis* to 94.2% in *P. radiata*. This transferability rate was comparable to the result obtained with EST-derived markers in pines (Brown et al. 2001, Chagné et al. 2003, Komulainen et al. 2003).

Fifty-eight out of 107 (54%) *P. radiata* SSR markers amplified a single band in *P. pinaster*. This transferability rate was higher to that of Gonzalez-Martinez et al. (2003) in *P. pinaster* using *P. taeda*-derived SSRs (36%), and to that of Shepherd et al. (2002) in P. elliottii and P. caribaea using *P. radiata*-derived SSRs (44%). Overall, the interspecific transferability of cDNA-SSR markers was higher than genomic SSRs.

Polymorphism and genetic mapping of cDNA and genomic SSRs in Pinus pinaster

Among the 46 single copy cDNA and 58 genomic SSR loci that amplified in *Pinus pinaster*, 9 (19.5%) and 7 (12%) were found to be polymorphic in at least one of the 4 mapping pedigrees, respectively. Their location in the INRA-G2 and INRA-F2 genetic maps (Chagné et al. 2002, Costa et al. 2000) and polymorphism in two other maritime pine pedigrees is presented in table 6. All the loci were linked with a minimum LOD of 6, except for locus ssrPt_ctg275 that was not linked to any linkage group in both maps. The 3 SSR markers previously developed by Mariette et al. (2001) were also mapped in both pedigrees. Overall, these SSRs made it possible to align 8 of the 12 linkage groups between the two maps. Linkage group homology was also confirmed using a set of ESTPs mapped in the INRA-G2 (Chagné et al. 2003) and INRA-F2 pedigrees (D. Chagné and P. Semat, unpublished).

If we consider the repeat type and position of the cDNA-SSRs, it should be noted that 5 out of 17 dinucleotide cDNA-SSRs (29%) were polymorphic in at least one maritime pine mapping pedigree whereas 4 out of 35 (11%) for the trinucleotides. In the same line, 6 out of 18 (33%) of the SSRs belonging to UTRs were polymorphic, compared to 3 out of 30 (10%) in the ORF. Moreover, 55% (5 out of 9 loci) of the cDNA-SSRs that corresponded to dinucleotide motifs

Figure 2: Distribution of the di- and trinucleotide SSRs within the Open Reading Frame (ORF; in white) or in the 5' untranslated regions (UTR; dark grey) and 3' UTR (light grey) in *Pinus taeda* and *Pinus pinaster* contigs. Sequences for which no ORF could be detected were not considered.

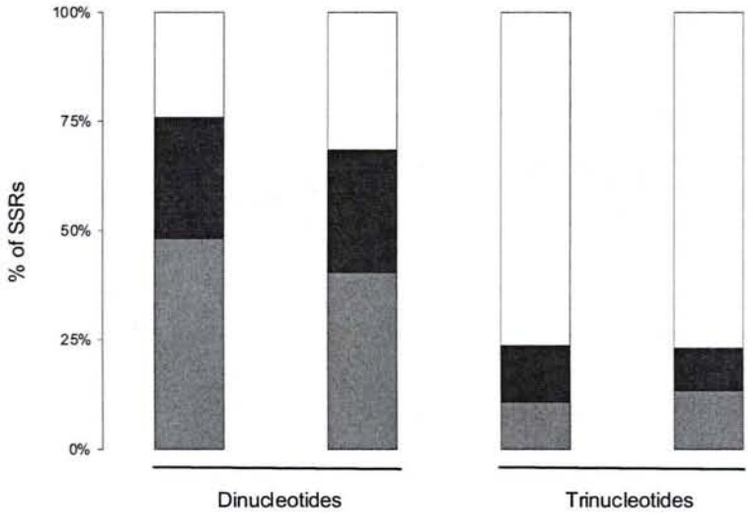


Table 6: Mapping information and genetic diversity parameters of the 3 classes of microsatellites genotyped on 26 trees. The mapping location in the *INRA-G2* (Chagné et al. 2002) and *INRA-F2* maps (Costa et al. 2000) are indicated following the linkage group numbers. M: monomorphic; P: polymorphic; UL: unlinked; *H*: heterozygosity; *A*: number of alleles; *: these values were not taken into account for the comparison of diversity parameters between cDNA and genomic SSRs.

		Mapping pedigree				Genetic diversity	
Marker type	Locus ID	*INRA-G2*	*INRA-F2*	*AFOCEL-F1*	*INIA-F1*	*H*	*A*
cDNA-SSR	ssrPp_cn524	6	1	P	M	0.81	5
	ssrPt_ctg275	P/UL	P/UL	P	P	0.74	8
	ssrPt_ctg4363	M	12	P	M	0.68	4
	ssrPt_ctg7824	10	M	M	M	0.35	2
	ssrPt_ctg988	11	M	P	M	0.55	3
	RPtEST11	5	2	P	M	0.74	4
	RPtEST13	10	M	M	M	0.66	3
	ssrPt_ctg1525	M	11	M	M	0.16	2
	ssrPt_ctg64	3	3	M	P	0.68	4
P. radiata genomic SSR	NZPR1078	2	7	P	M	0.68	4
	NZPR114	M	5	M	P	0.68	5
	NZPR1702_b	11	6	P	M	0.38*	2*
	NZPR413	4	8	P	P	0.58	4
	NZPR472	1	M	P	P	0.67	4
	NZPR544	M	3	M	P	0.41	4
	NZPR823_a	5	M	P	P	0.67	3
P. pinaster and *P. halepensis* genomic SSR	FRPp91	1	9	P	P	0.85	9
	FRPp94	10	5	P	P	0.80	8
	ITPh4516	3	3	P	P	0.84	8

belonging to untranslated regions were polymorphic. This result suggests that a pre-annotation of the sequences containing SSRs is a prerequesite for an efficient development of cDNA-SSR markers.
We verified the orthology for the 7 polymorphic SSR loci originated from *P. radiata* genomic library by sequencing the PCR products obtained from amplifying *P. pinaster* DNA. High sequence identity were found for 6 of the loci (table 7) to a level comparable to that found between orthologous ESTs in pines (Brown et al. 2001, Chagné et al. 2003, Komulainen et al. 2003). Interestingly, one locus (NZPR1702_b) did not show any sequence homology and did not contain an SSR motif. Electrophoresis on acrylamide gel showed that this locus presented two distinct bands, 30 bp apart (i.e. two alleles corresponding to an insertion-deletion polymorphism). Moreover, this locus presented the lowest genetic diversity (H=0.38). This locus was subsequently discarded for the comparison between genomic and cDNA SSRs (see next section).

Level of diversity of cDNA and genomic SSRs in Pinus pinaster

The 9 polymorphic cDNA-SSRs were genotyped in 26 maritime pine trees. Their heterozygosity (*H*) and number of alleles (*A*) are shown in table 6 and were compared with those of the 6 genomic SSRs transferred from *P. radiata* and the 3 genomic SSRs described by Mariette et al. (2001). Within the cDNA-SSRs, there was no significant difference between the heterozygosity obtained in the ORF and the UTRs, or between tri- and dinucleotide SSRs. Within the genomic SSRs, a significant difference of the diversity parameters was found between the loci transferred from *P. radiata* and those developed from *P. pinaster* and *P. halepensis* by Mariette et al. (2001). This difference suggests that genomic SSRs tend to be less polymorphic when transferred from phylogenetically distant species: *P. radiata* belongs to the Oocarpeae subsection, whereas *P. pinaster* and *P. halepensis* belongs to the Sylvestres subsection of the pine genus (Mirov 1967). Finally, the level of diversity was not different between the transferred *P. radiata* genomic SSRs and the cDNA-SSRs.

Discussion

Composition and distribution of SSRs in the expressed genome of pine

The SSR composition of the coding region of the pine genome was first compared to the results published in other plant species. In dicotyledonous species where cDNA-SSR have been searched: i.e. *Vitis vinifera* (Scott et al. 2000) and *Arabidopsis thaliana* (Cardle et al. 2000, Morgante et al. 2002), the most represented repeat types, i.e. AG, AT, AAG, AGG and AGC, were also found to be the most frequent in pines (figure 1). Conversely, the most common repeated motif in monocotyledonous species (Varshney et al. 2002), CCG, was quite rare in pines (5.2% and 7.2% in *P. pinaster* and *P. taeda*, respectively). This result suggests that the SSR composition of gymnosperms genes is more similar to that of dicots than monocots. However, given the few number of species analysed, this interpretation remains to be confirmed.
The presence of a majority of trinucleotides in the ORFs (figure 2) was also in agreement to what has been described in other plants. Morgante et al. (2002) showed a strong positive selection for trinucleotides in the translated regions of *A. thaliana*. Metzgar et al. (2000) explained the excess of triplet repeat microsatellites in the coding regions by the effect of important mutation pressures. Indeed, a mutation in a mono-, di-, tetra- or pentanucleotide SSR in the ORFs would result in a frameshift that could change the translated protein structure and function.

Table 7: *Pinus radiata* SSRs for the 7 informative loci: repeated motif, PCR primers, amplification conditions, product length and sequence homology between *Pinus pinaster* and *Pinus radiata*.

Locus name	Repeated motif	Forward primer	Reverse primer	Annealing temperature (°C)	Expected length (bp)	Sequence homology
NZPR1078	AC_{10}	tggtgatcaagccttttcc	gttgatgagtgatggcatgg	53	342	91.5%
NZPR114	CA_{15} .. CA_{13} TA_{22}	aagatgacccacatgaagtttgg	ggagctttataacatatctcgatgc	56	193	88.2%
NZPR1702_b	AC_{15} CA_{13}... AT_5	tatgattggaccattggggt	ccaaaccctcctccacatatc	53	187	no homology
NZPR413	TG_{23} GT_6	tgaacctcgatggaatagcc	cccgccttgcatcaatta	53	253	89.1%
NZPR472	AC13	gagaaaattcaaccaccgga	ggttgtagggcagtgaatcc	53	309	89.4%
NZPR544	CA_5 AC_{12} TA_5	gcgatgtgcaacccttgata	tgctattccgtcaaaaaccc	56	286	86.1%
NZPR823	AC_{57}	tatcgggagcaagttatgcc	tgcactctttttcgtctcca	53	296	92.5%

Morgante et al. (2002) detected much higher levels of SSRs in the 5' UTRs, especially AG/CT repeats. The rather small amount of SSRs detected in the 5' UTRs of pine genes (17.4%, table 3) contrasted with their results could reflect a true feature of pine genes or it could simply be that the low coverage of the 5'-end in the pine ESTs has provided an anomaly. Some support for the latter view comes from ESTs obtained from the sequencing of the 5'-ends of 3' anchored cDNAs (Frigerio et al. 2003, Kirst et al. 2003). Therefore, the 5' UTRs were probably under-represented in the two pine EST collections analyzed.

Transferability of cDNA and genomic SSRs in pines

From 64.6% to 94.2% of the pine cDNA-SSRs transferred to one or more of the 7 pine species tested (table 5). It has been clearly shown that the transferability of molecular markers (including SSRs) depends on the phylogenetic distance between species. Most of the developed markers in this study originated from P. taeda, an American pine which belongs to the *Pinus* section of the subgenus *Pinus* (Mirov 1967). Then, it is not surprising that the highest transfer rate was observed for *P. radiata* (94.2%), another American pine belonging to the same section. In the same line, the transfer rate decreased for Mediterranean pines of the same section (*P. pinaster*, 86.8%; *P. sylvestris*, 85.4%; *P. halepensis*, 72.9%), and was even lower with Mediterranean pines of the more distant section Pinea (*P. pinea*, 70.8%; *P. canariensis*, 64.6%). Nevertheless, this study was only carried out on hard pines (subgenus *Pinus*). We also anticipate a lower transferability of cDNA-SSRs in the subgenus *Strobus*, or even within other genera of the Pinaceae family. However, the transferability rates in these more distant species should be higher for cDNA-SSRs compared to genomic SSR (Echt et al. 1999).

Similar rates of cross-species transferability has been reported using EST-derived SSRs in the genus *Medicago* (Eujayl et al. 2003, 89%) and within the Poaceae (Gupta et al. 2003, 55%). Comparatively, genomic SSRs have shown to be less transferable in pine (54% between *P. radiata* and *P. pinaster*, this study; 29% between P. strobus and P. radiata, Echt et al. 1999; and 36% between *P. taeda* and *P. pinaster*, Gonzalez-Martinez et al. 2003). This rate is low compared to other plant genera (e.g. up to 85% between *Glycine spp.*, Peakall et al. 1998). These results suggest that the data mining of pine cDNA libraries is valuable approach to develop transferable SSRs. Furthermore, it should be noted that the cDNA-SSRs were obtained without library screening. Clearly the development of pine sequence databases and the in silico approach described here provides a cost effect approach to SSR marker development.

In rice and wheat, EST-derived SSRs have been reported to have lower rate of polymorphisms compared to SSRs derived from genomic libraries (Cho et al. 2000, Eujayl et al. 2002). However, such difference was not found in *Medicago* (Eujayl et al. 2003) and *Picea* (Scotti et al. 2000) two highly polymorphic genera compared to the highly domesticated cereal crops. Our findings in maritime pine revealed that non-source species genomic SSRs and cDNA-SSRs have similar levels of diversity and thus cDNA-SSRs are not less polymorphic.

The markers developed in this study were mapped in the maritime pine genetic map that was aligned with the loblolly pine map using comparative genome mapping (Chagné et al. 2003). Therefore, they can be used as orthologous markers in other conifer species. At the intraspecific level, these markers have been mapped in the different genetic maps of maritime pine, which will make it possible to construct a consensus map of this species.

Nevertheless, more markers will be needed to reach the saturation levels desired.

Conclusion

We have shown in this study that database-sourced cDNA-SSRs can be efficiently developed for, and transferred across pine species. Pine SSR markers developed in this way are less expensive to produce and are as informative as SSR markers derived from other (genomic-based) methods. However, since these markers correspond to transcribed regions one should further study if they behave as neutral marker or not, if they are to be used in genetic diversity analysis and in association studies.

Acknowledgements

DC was funded by the French Ministry of Research. This research was supported by grants from France (Ministère de l'Agriculture et de la Pêche-DERF No 61.45.80.15/02) and the European Union (TREESNIPS project: QLK3-CT-2002-01973). The maritime pine ESTs were produced with the support of the Aquitaine Région (n°2002 0307002A) and INRA (Lignome) as well as the European Union (GEMINI: QLK5-CT-1999-00942). The work at New Zealand Forest Research was funded by New Zealand's Foundation for Research, Science and Technology (CO4X005).

References

Amarasinghe V, Brown GR, Mank JE, Carlson JE (2002) Microsatellite DNA loci for Western Hemlock [*Tsuga heterophylla* (Raf.) Sarg]. Mol Ecol 2: 236-238.

Amarasinghe V, Carlson JE (2002) The development of microsatellite DNA markers for genetic analysis in Douglas-fir. Can J For Res 32: 1904-1915.

Brown GR, Kadel EE III, Bassoni DL, Kiehne KL, Temesgen B, Van Buijtenen JP, Sewell MM, Marshall KA, Neale DB (2001) Anchored reference loci in Loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics 159: 799-809.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847-854.

Chagné D, Lalanne C, Madur D, Kumar S, Frigerio JM, Krier C, Decroocq S, Savoure A, Bou-Dagher KM, Bertocchi E, Brach J, Plomion C (2002) A high density genetic map of maritime pine based on AFLPs. Ann For Sci 59: 627-636.

Chagné D, Brown G, Lalanne C, Madur D, Pot D, Neale D, Plomion C (2003) Comparative genome and QTL mapping between maritime and loblolly pines. Mol Breed 12: 185-195.

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCough SR, Park WD, Ayer N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and Genbank sequences in rice (*Oriza sativa*). Theor Appl Genet 100: 713-722.

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nuc Ac Res 29(1): 238-238.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160: 1115-1123.

Costa P, Pot D, Dubos C, Frigerio J-M, Pionneau C, Bodénès C, Bertocchi E, Cervera MT, Remington DL, Plomion C (2000) A genetic map of maritime pine based on AFLP, RAPD and protein markers. Theor Appl Genet 100: 39-48.

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

Dubos C, Plomion C (2003) Identification of water-deficit responsive genes in maritime pine (*Pinus pinaster* Ait.) roots. Plant Mol Biol 51: 249-262.

Echt CS, May-Marquardt P, Hseih M, Zahorchak R (1996) Characterization of microsatellite markers in eastern white pine. Genome 39: 1002-1008.

Echt CS, May-Marquardt P (1997) Survey of microsatellite in pine. Genome 40: 9-17.

Echt CS, Vendramin GG, Nelson CD, May-Marquardt P (1999) Microsatellite DNA as shared

genetic markers among conifer species. Can J For Res 29: 365-371.

Elsik CG, Williams CG (2000) Retroelements contribute to the excess low-copy number DNA in pine. Mol Gen Genet 264: 47-55.

Elsik CG, Minihan VT, Hall SE, Scarpa AM, Williams CG (2000) Low-copy microsatellite markers for *Pinus taeda* L. Genome 43: 550-555.

Elsik CG, Williams CG (2001) Low-copy microsatellite recovery from a conifer genome. Theor Appl Genet 103: 1189-1195.

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104: 399-407.

Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, Zwonitzer JC, Mian MAR (2003) Medicago trunculata EST-SSRs reveal cross-species genetic markers for *Medicago* spp. Theor Appl Genet, in press.

Fisher PJ, Richardson TE, Gardner RC (1998) Characteristics of single- and multi-copy microsatellites from *Pinus radiata*. Theor Appl Genet 96: 969-979.

Frigerio JM, Dubos C, Chaumeil P, Salin F, Garcia V, Barré A, Plomion C (2003) Using transcriptome analysis to identify osmotic stress candidate genes in maritime pine (Pinus pinaster Ait.). In: Sustainable Forestry, Wood products & Biotechnology, BIOFOR proceeding, in press.

Gao L, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed 12: 245-261.

Goldstein DB, Schlotterer C (1999) Microsatellites. Evolution and applications. Goldstein and Schlotterer eds, Oxford University Press.

Gonzalez-Martinez SC, Robledo-Arnuncio JJ, Collada C, Diaz A, Williams CG, Alia R, Cervera MT (2003) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theor Appl Genet, in press.

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Gen, in press.

Hicks M, Adams D, O'Keefe S, MacDonald E, Hodgetts R (1998) The development of RAPD and microsatellite markers in lodgepole pine (*Pinus contorta* var. *latifolia*). Genome 41: 797-805.

Hodgetts RB, Aleksiuk MA, Brown A, Clarke C, MacDonald E, Nadeem S, Khasa (2001) Develoment of microsatellite markers for white spruce (*Picea glauca*) and related species. Theor Appl Genet 102: 1252-1258.

Kamm A, Doudrick RL, Heslop-Harrison JS, Schmidt T (1996) The genomic and physical organization of Ty1-copia-like sequences as component of large genomes in *Pinus elliottii* var. *elliottii* and other gymnosperms. Proc Natl Acad Sci 93: 2708-2713.

Kantety RV, LaRota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum wheat. Plant Mol Biol 48: 501-510.

Karhu A, Dieterich JH, Savolainen O (2000) Rapid expansion of microsatellite sequences in pines. Mol Bio Evol 17: 259-265.

Keys RN, Autino A, Edwards KJ, Fady B, Pichot C, Vendramin GG (2000) Characterization of nuclear microsatellites in *Pinus halepensis* Mill. and their inheritance in *P. halepensis* and *Pinus brutia* Ten. Mol Ecol 9: 2155-2234.

Kinlaw CG, Neale DB (1997) Complex gene families in pine genomes. Trends Plant Sci 2: 356-359.

Kirst M, Johnston AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with Arabidopsis thaliana. Proc Natl Acad Sci USA 100: 7383-7388.

Komulainen P, Brown GR, Mikkonen M, Karhu A, Garcia-Gil MR, O'Malley D, Lee B, Neale DB, Savolainen O (2003) Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*. Theor Appl Genet 107: 667-678.

Kossack DS, Kinlaw CS (1999) IFG, a *gypsy*-like retrotransposon in *Pinus* (Pinaceae), has an extensive hystory in pines. Plant Mol Biol 39: 417-426.

Kostia S, Varvio SL, Vakkari P, Pulkkinen P (1995) Microsatellite sequences in a conifer, *Pinus sylvestris*. Genome 38: 244-1248.

Kriebel HB (1985) DNA sequences components in the *Pinus strobus* nuclear genome. Can J For Res 15: 1-4.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

Kutil BL and Williams CG (2001) Triplet-repeat microsatellites shared among hard and soft pines. J Hered 92: 327-332.

Leitch IJ, Hanson L, Winfield M, Parker J, Bennett MD (2001) Nuclear DNA C-values complete familial representation in gymnosperms. Annals of Botany 88: 843-849.

Lian C, Miwa M, Hogetsu T (2000) Isolation and characterization of microsatellite loci from the Japanese red pine, *Pinus densiflora*. Mol Ecol 9: 1171-1193.

Mariette S, Chagné D, Decroocq S, Vendramin GG, Lalanne C, Madur D, Plomion C (2001) Microsatellite markers for *Pinus pinaster* Ait. Ann For Sci 58: 203-206.

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10: 72-80.

Mirov NT (1967) The Genus Pinus. New York: Ronald press.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30: 194-200.

Moriguchi Y, Iwata H, Ujino-Ihara T, Yoshimura K, Taira H, Tsumura Y (2003) Development and characterization of microsatellite markers for *Cryptomeria japonica* D. Don. Theor Appl Genet 106: 751-758.

Paglia GP, Olivieri AM, Morgante M (1998) Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol Gen Genet 258: 466-478.

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross-species amplification of soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15: 1275-1287.

Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome 40: 411-419.

Price RA, Liston A, Strauss SH (1998) Phylogeny and systematics of *Pinus*. In Richardson D.M. (ed) Ecology and biogeography of *Pinus*. Cambridge University Press, pp 49-68.

Rajora OP, Rahman MH, Dayanandan S, Mosseler A (2001) Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. Mol Gen Genet 264: 871-882.

Ritter E, Aragones A, Markussen T, Achere V, Espinel S, Fladung M, Wrobel S, Faivre-Rampant P, Jeandroz S, Favre J-M (2002) Construction and exploitation of a multifunctional and saturated genetic map for coniferous species. Ann For Sci 59: 637-643.

Salzberg SL, Delcher AL, Kasif F, White O (1998) Microbial gene identification using Markov interpolated models. Nuc Ac Res 26: 544-548.

Schiex T, Gouzy J, Moisan A, de Oliveira Y (2003) FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nuc Ac Res 31: 3738-3741.

Schneider S, Roessli D, Excoffier L (2000) ARLEQUIN v.2000 (Genetics and Biometry Laboratory, University of Geneva, Switzerland).

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100: 723-726.

Scotti I, Magni F, Fink R, Powell W, Binelli G, Hedley PE (2000) Microsatellite repeats are not randomly distributed within Norway spruce (*Picea abies* K.) expressed sequences. Genome 43: 41-46.

Scotti I, Magni F, Paglia GP, Morgante M (2002a) Trinucleotide microsatellites in Norway spruce (*Picea abies*): their features and the development of molecular markers. Theor Appl Genet 106: 40-50.

Scotti I, Paglia GP, Magni F, Morgante M (2002b) Efficient Development of Dinucleotide microsatellite markers in Norway spruce (*Picea Abies* Karst.) through dot-blot selection. Theor Appl Genet 104: 1035-1041.

Shepherd M, Cross M, Maguire TL, Dieters MJ, Williams CG, Henry RJ (2002) Transpecific microsatellites for hard pines. Theor Appl Genet 104: 819-827.

Smith DN and Devey ME (1994) Occurrence and inheritance of microsatellites in *Pinus radiata*. Genome 37: 977-983.

Soranzo N, Provan J, Powell W (1998) Characterization of microsatellite loci in *Pinus sylvestris* L. Mol Ecol 7: 1247-1263.

Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCough SR (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100: 697-712.

Temnykh S., DeClerck G., Lukashova A., Lipovich L., Cartinhour S., McCouch S. (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Gen Res 11: 1441-1452.

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Gen Res 10: 967-981.

Van Ooijen JW and Voorrips RE (2001) Joinmap 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands. Website: http://www.joinmap.nl

Van de ven WTG and Mac Nicol RJ (1996) Microsatellites as DNA markers in Sitka spruce. Theor Appl Genet 93: 613-617.

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol 7:537-546.

Zhou Y, Bui T, Auckland LD, Williams CG (2002) Undermethylated DNA as a source of microsatellite from a conifer genome. Genome 45: 91-99.

ANNEXE V

Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences

Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences

Loïck Le Dantec[1,3]*, David Chagné[2]*, David Pot[2], Olivier Cantin[1,3], Pauline Garnier-Géré[2], Frank Bedon[2], Jean-Marc Frigerio[2], Philippe Chaumeil[2], Patrick Léger[2], Virginie Garcia[4], Frédéric Laigret[1], Antoine de Daruvar[3], Christophe Plomion[2]

1/ Unité de Recherche sur les Espèces Fruitières et la Vigne, INRA, 71 avenue Edouard Bourlaux, BP 81, 33883 Villenave d'Ornon cedex, France
2/ UMR 1202 BIOGECO, INRA, Equipe de Génétique, 69 route d'Arcachon, 33612 Cestas Cédex, France
3/ Centre de Bioinformatique de Bordeaux, Université V. Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cédex, France
4/ UMR 619 Physiologie et Biotechnologie Végétales, INRA, IBVM, 71 avenue Edouard Bourlaux, BP 81, 33883 Villenave d'Ornon cedex, France

* these authors have contributed equally to this work
for correspondence: Christophe Plomion: e mail: plomion@pierroton.inra.fr

Keywords
SNP - *in silico* detection- EST– maritime pine

Abstract

We developed an automated pipeline for the detection of Single Nucleotide Polymorphisms (SNPs) in Expressed Sequence Tag (EST) data sets, by combining three DNA sequence analysis programs: *Phred*, *Phrap* and *PolyBayes*. This application requires access to the individual electrophoregram traces. First, a reference set of 65 SNPs was obtained from the sequencing of 30 gametes in 13 maritime pine (*Pinus pinaster* Ait.) gene fragments (6,671 bp), resulting in a frequency of 1 SNP every 102.6 bp. Second, parameters of the three programs were optimized in order to retrieve as many true SNPs, while keeping the rate of false positive as low as possible. Overall, the efficiency of detection of true SNPs was 83.1%. However, this rate varied largely as a function of the rare SNP allele frequency: down to 41% for rare SNP alleles (frequency < 10%), up to 98% for allele frequencies above 10%. Third, the detection method was applied to the 18,498 assembled maritime pine (*Pinus pinaster* Ait.) ESTs, allowing to identify a total of 1,400 candidate SNPs, in contigs containing between 4 and 20 sequence reads. These genetic resources, described for the first time in a forest tree species, were made available at http://www.pierroton.inra/genetics/Pinesnps. We also derived an analytical expression for the SNP detection probability as a function of the SNP allele frequency, the number of haploid genomes used to generate the EST sequence database, and the sample size of the contigs considered for SNP detection. The frequency of the SNP allele was shown to be the main factor influencing the probability of SNP detection.

Introduction

Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA variation in the human genome (Sachidanandam et al. 2001). They have recently become the marker of choice in genetic analyses due to their abundance and stability compared to tandem repeat sequences (microsatellites) (Gray et al. 2000). SNPs have been mainly developed and used in human genetics for complex diseases mapping and in association studies (Collins et al. 1999, Emahazion et al. 2001, Kruglyak 1997),

pharmacogenomics (Riley et al. 2000), forensic studies (Wilson et al. 1995, Andreasson et al. 2002) and inference of population history (Brumfield et al. 2003). In plants, information on the frequency, nature, and distribution of SNPs is still in its infancy. Only in recent years, SNPs have been developed for a handful of species such as maize (Ching et al. 2002, Thornsberry et al. 2001), Arabidopsis (Cho et al. 1999, Nordborg et al. 2002), sugarcane (Grivet et al. 2003) and wheat (Somers et al. 2003).

The most classical way to identify SNP polymorphism is direct sequencing of amplicons of candidate genes from a set of individuals that represent the diversity in the population of interest. The drawback of this approach is the time and money investments. An alternative method takes advantage of the redundancy of gene sequences generated by Expressed Sequence Tag (EST) sequencing programs (Picoult-Newberg et al. 1999, Useche et al. 2001, Batley et al. 2003, Kim et al. 2003). The development of ESTs has been initiated for many agronomically important plant species (Rounsley et al. 1998) and has generated a huge amount of data. It is now possible to use these data to search for DNA variation, providing enough individuals are included in the cDNA libraries. Mining SNPs from EST sequences generally involves the following steps: treatment of raw EST sequences (trace files), clustering, assembly, and detection of SNPs in aligned sequence data. Several bioinformatic tools have been developed for *in silico* SNP detection. They are based on sequence trace file analysis to filter out sequence errors. Among them, *PolyBayes* (Marth et al. 1999) is a software that applies a Bayesian inference to calculate the probability of a given site being polymorphic within aligned EST sequences, and *PolyPhred* (Nickerson et al. 1997) is a program that automatically detects the presence of SNP in heterozygotes by fluorescence-based sequencing of PCR products.

Other tools for SNP detection are based on a conservative approach, and use redundancy of SNPs in EST sequences combined with the co-segregation of haplotypes to differentiate between "true" SNPs and sequencing errors (Barker 2003 et al. 2003; Kota et al. 2003).

Here, we present a bioinformatic approach to the automated detection of SNPs in EST sequences using the combination of *Phred*, *Phrap* and *PolyBayes* softwares. Rates of true SNP recovery and false positive sites were estimated using a reference set of true SNPs obtained from the sequencing of 13 maritime pine gene fragments representing a total of 6.6 kb. SNPs were then searched in the maritime pine ESTs, providing a valuable resource of DNA markers for this economically and ecologically important forest tree species in south-western Europe.

Material and methods

SNP discovery pipeline

An automated system for SNP discovery was developed using a combination of three publicly available programs: *Phred* (Ewing et al. 1998, Ewing and Green 1998), *Phrap* (www.phrap.org) and *PolyBayes* (Marth et al. 1999). *Phred* makes the base identification and assigns a quality score to each position in a sequence. *Phrap* assembles the reads into contigs. *PolyBayes* is an engine using Bayesian inference methods, which determines, for each site within a multiple alignment, the probability of the site being polymorphic. This process takes into account the depth of coverage, the base quality values of the sequences, and the a priori expected rate of polymorphic sites in the considered species. Prior to its use for SNP mining in an EST dataset, the accuracy of the automatic detection method was tested with a reference set of true SNPs.

Table 1: Result of SNP detection in the 13 gene fragments of the reference set using the visual inspection (A) and the automated method (B): gene identification, sequenced fragment size, comparison between the two methods, efficiency of the automated method and number of false positives (SNP detected by the method B only). Gene ID are as follows: C4H: Trans-cinnamate-4-monooxygenase; CAD: Cinnamyl alcohol dehydrogenase (two fragments F1 and F2); PAL: Phenylalanine ammonia lyase; Korrigan: endo-1,4-β-D-glucanase; GRP: Glycine-rich protein; MYB-like TF: MYB-like transcription factor; ACC oxidase: amino-cyclopropane-carboxylic acid oxidase; CesAn: Cellulose synthase; CCoAOMT: Caffeoyl CoA O-methyltransferase; AGP: Arabinogalactan protein.

Gene identification	Fragment size (bp)	True SNPs detected by method A	dbSNP accession	SNPs detected by automated method B	Shared SNPs between A and B	Number of true SNPs not detected by method B	Number of false positive
C4H	539	10	ss16208982;8985-8987;8991;8994-8998;9000	10	10	0	0
CAD-F1	537	4	ss16209001;9005;9008;9010	3	3	1	0
CAD-F2	523	4	ss16209011;9012;9015;9016	4	3	1	1
PAL	547	9	ss16209020-9025;9028;9029;9032	7	7	2	0
Korrigan	937	5	ss12709589-ss12709593	7	5	0	2
GRP	479	9	ss12709575-ss12709585	8	8	1	0
MYB-like TF	494	2	ss12709586-ss12709587	3	2	0	1
ACC oxydase	270	1	ss12709588	1	1	0	0
CesA7	553	2	ss12709573-ss12709574	2	2	0	0
CesA4	489	1	ss12709572	0	0	1	0
CesA3	490	7	ss12709565-ss12709571	6	6	1	0
CCoAOMT	492	7	ss16209076-ss16209082	3	3	4	0
AGP	321	4	ss16209070;9071;9073;9074	4	4	0	0
Total	6671	65		58	54	11	4

Development of a reference set of true SNPs

Sixty-five SNPs (Table 1) were used as a reference set of true SNPs. These SNPs were deduced from the expert visual inspection of maritime pine (*Pinus pinaster* Ait.) nucleotide sequences, resulting from the amplification of 13 DNA fragments in 30 megagametophytes (haploid tissue surrounding the embryo) representing the whole natural range of the species. Details about plant material, DNA extraction, primer design, PCR amplification, DNA sequencing and electrophoregram analysis are presented elsewhere (Le Provost et al. 2003). The use of megagametophytes lowered the risk of confusing polymorphism at a unique locus with differences between paralogous loci. With haploid tissue, amplification of two or even more members of a gene family would have been easily detected by the observations of double peaks in the sequences. The risk of confusing true single nucleotide polymorphisms with sequencing errors was also minimized by a systematic visual inspection (referred to as method A in Table 1) of the electrophoregrams using *Sequencher* (GeneCode, Ann Arbor, MI, USA). This reference set was used as a positive control to test the accuracy of the automated method (referred to as method B in the next sections).

Recovery of true SNPs using the automated method

For each gene fragment, a file containing all sequence traces was used for *Phred-Phrap* analysis through a web interface generated using the *PISE* package (Letondal 2001). Output files were accessible from a dynamically generated web page. *Polybayes* was then run on the *ACE* file generated by *Phrap*. *PolyBayes* generated two output files: a result file in a text format and a modified *ACE* file with polymorphism "tags" that were viewed by *Consed* (Gordon et al. 1998, http://www.phrap.org/consed/consed.html) . *Consed* displays the assembly and the associated tags along with the multiple alignment of the chromatograms. *Xml* files used by *PISE* as well as *Perl* scripts to create the directory structures required for the functioning of *Phred/Phrap/PolyBayes/Consed* are available upon request to the first author.
The efficiency of the pipeline to detect true SNPs was tested using the reference set described above. Several parameters of Phrap (i.e. retain exact duplicate reads, relaxed assembly stringency set to 10) and *Polybayes* (i.e. no paralog filtering) were optimised by comparing the automated pipeline output of the reference set to the known true sets of SNPs. This combination of parameters allowed a maximum of true SNPs to be detected while keeping the rate of false positive sites as low as possible.

SNP detection in the maritime pine expressed sequence tag dataset

For SNP discovery in maritime pine ESTs with method B, a total of 28,128 plasmid clones were available. These were derived from the following five cDNA libraries: one cDNA library constructed using differentiating xylem collected on four genotypes of the Corsican provenance (Canton et al. 2003), and four cDNA libraries constructed with root and needle tissues of 240 drought stressed and 120 well watered seedlings of the Aquitaine provenance (Frigerio et al. 2003). EST data were assembled into contigs using a semi-automatic pipeline for EST processing and annotation (V. Garcia, unpublished), mainly based on *StackPACK* (Christoffels et al. 2001). For each contig, a directory structure (directory with the name of the contig and the following subdirectories: *edit_dir*, *chromat_dir*, *phd_dir*, *poly_dir*) required for the functioning of *Phred-Phrap* and *PolyBayes* was created. Chromatograms for all ESTs in each contig were first copied from the chromatogram

repository into the associated chromat_dir directory.

The *ACE* files resulting from *Phred-Phrap* analysis were then passed to the SNP detection program *PolyBayes*. Based on the reference SNP dataset, one SNP every 100 bp was used as input value. *PolyBayes* output files were parsed to collect the most relevant SNP information. Functional annotation of the consensus sequence for each contig analysed was extracted from the EST database. Finally, the data were displayed on a dynamically generated *HTML* page with links to the *Phred-Phrap* and *PolyBayes* output files for further examination and SNP detection results for each contig (i.e. functional annotation, nucleotide variation, SNP probability P_{SNP}, SNP position on the consensus sequence, number of SNPs with $P_{SNP}>0.99$ and contig length). A brief summary displayed the total number of SNP with $P_{SNP}>0.99$, the number of contigs with no SNP detected and the total length of analysed contigs. Several home made PHP (*PHP Hypertext Processor*) scripts (available upon request from the first author) were used to combine *Phred-Phrap* and *Polybayes* and also to make interfaces for user query, EST database access, output files parsing and generate *HTML* result pages.

SNP detection probability in EST data

To estimate the probability of SNP detection in an already assembled EST data set, two main parameters have to be taken into account: (i) the number of haploid genomes "g" segregating in the EST database (e.g. g=10 for 5 diploid genotypes used to construct the cDNA library), and (ii) the number of ESTs "e" in the contigs considered for SNP detection. Let X_A be the random variable corresponding to the number of allele A in the "e" ESTs, Y_A the random variable corresponding to the number of allele A in the "g" genome, and P_A the frequency of one the SNP allele in the population in which these genotypes are sampled (SNPs are considered as diallelic, Brookes 1999). The probability of SNP detection (i.e. the two alleles detected in a given contig) is equal to $1-[\ P(X_A=0, or, X_A=e)\ ] = 1-P(X_A=0)-P(X_A=e)$, with $P(X_A=e)$ the probability that all ESTs carries the A allele, and $P(X_A=0)$ the probability that all ESTs carries the alternative allele. To compute these two probabilities we used the Bayse theorem as follows:

$$P(X_A=0)=\sum_{y_A} P(Y_A=y_A)\times P(X_A=0/Y_A=y_A)$$

and

$$P(X_A=e)=\sum_{y_A=1}^{g} P(Y_A=y_A)\times P(X_A=e/Y_A=y_A)$$

with

$$P(Y_A=y_A)=\beta(y_A,g,P_A)=\frac{g!P_A^{\ y_A}(1-P_A)^{g-y_A}}{(g-y_A)!y_A!}$$

$$P(X_A=0/Y_A=y_A)=\left[(1-y_A)/g\right]^e$$

and $P(X_A=e/Y_A=y_A)=\left[(y_A)/g\right]^e$

The probability of SNP detection in a contig was finally computed for different values of the three parameters P_A , g and e. It should be noted that the probability of SNP detection as estimated above depends on several assumptions regarding: (i) the quality of the sequences, (ii) the constitution of the contigs (mixture of paralogs, sequence length homogeneity), and (iii) the efficiency of the SNP detection algorithm. All were assumed optimal, i.e. high sequence quality, absence of paralog within each contig, DNA fragment with identical length, polymorphism detection algorithm independent of the frequency of the SNP alleles in the EST dataset.

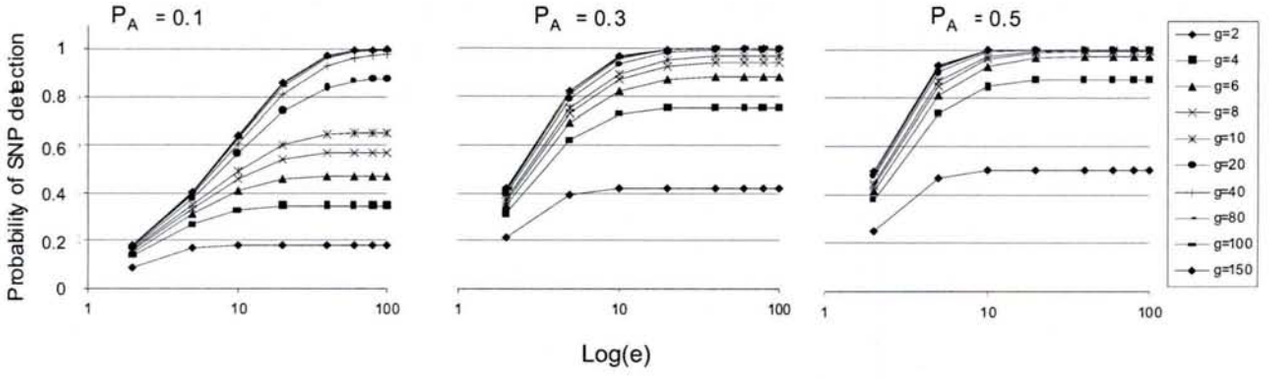
Table 2: Efficiency of the automated method for the detection of true SNPs in relation with the SNP allele frequency (P_A). A and B referred to the automated and visual method for SNP detection, respectively.

Rare allele frequency	True SNPs (A)	Shared SNPs (A and B)	% recovery of true SNPs
p<10%	17	7	41%
p>10%	48	47	98%
Total	65	54	83.1%

Table 3: Multiple regression on SNP detection probability. P_A, g and e were used as independant variables. P_A: SNP allele frequency; g: number of haploid genomes in the database; e: number of ESTs in a contig. r^2 is the determination coefficient.

source of variation	df	SS	F value	P-value	r^2 (%)
p	1	3.711962	82.66597	0	42.62
g	1	2.121951	47.25617	5.51E-11	24.36
e	1	2.876731	64.06526	5.44E-14	33.02

Figure 1: SNP detection probability in relation to (i) the number of alleles in the database (2<g<150), (ii) the number of sequences in a contig (2<e<100) and (iii) the rare SNP allele frequency in the population (0.1< PA <0.5)



Results and discussion

Validation of the automated method for SNP detection

Sixty-five SNPs could be detected by visual inspection of 6,671 bp in 13 gene fragments, resulting in a frequency of one SNP every 102.6 bp, which is comparable to the results obtained in other plant species, such as maize where one SNP per 100 bp were detected in contigs comprising at least 20 ESTs (Batley et al. 2003). In the reference sequence dataset 58 SNPs were detected using the automated system (method B, Table 1). These 58 SNPs were checked by visualizing the sequence chromatograms using *Consed*, and compared to the 65 true SNPs previously detected by visual inspection of aligned sequences (method A, Table 1). While 54 SNPs corresponded to true SNPs, indicating an efficiency of discovery of true SNPs of 83.1%, 4 SNPs among 58 (6.9%) corresponded to false positive sites. Visual inspection clearly showed that they were due to weaker sequence quality inducing wrong base calling. Because this occurred at the 3' and 5' ends of the sequences it is anticipated that most of these artifacts could be circumvented using trimmed EST sequences based on optimized *Phred* scores.

More critical was the inability of the pipeline to detect 11 true SNPs. Interestingly, most of them had a low allele frequency (p<10%) in the reference dataset. Table 2 shows the results obtained with both the manual and the automated methods, for SNP allele frequencies above or below 10%: 17 among the 65 true SNPs had rare alleles (p<10%), from which only 7 were found using the automated method, indicating a 41% efficiency rate for rare allele SNP detection, whereas 98% of the SNPs with allele frequency greater than 10% were found by the automated method. This result clearly shows the limits of the automated method for detecting rare SNP alleles, and this should be taken into account for the use of the pipeline in EST database.

SNP detection probability

Figure 1 shows the probability of SNP detection according to (i) the number of ESTs "e" in a contig considered for SNP detection (e=2 to 100), (ii) the frequency "p" of one SNP allele in the population (p=0.1, 0.3 and 0.5), and (iii) the number of haploid genomes "g" segregating in the EST database (g=2 to 150). A 95 % SNP detection probability was achieved with the following combination of parameters: p=0.1, g>80, e>40, p=0.3, g>20 and e>20, p=0.5, g>8, e>10, again showing the difficulty to detect low frequency SNP alleles.

These three parameters were taken as independent variables in a multiple linear regression analysis in order to test their effect on SNP detection probabilities. All parameters had a strong effect on the ability to recover SNPs (overall determination coefficient of 90%; Table 3). The factor explaining the highest proportion of variation (43%) in the probability of SNP detection was the SNP allele frequency; the number of ESTs in the contig still explaining 33%, and finally the number of haploid genomes segregating in the database was the less limiting factor.

Recovery of true SNPs from a reference data set using ESTs

For three genes of the reference dataset, *AGP*, *CCoAOMT* and *GRP*, 19, 19 and 21 maritime pine ESTs were available, respectively. This made it possible to validate the efficiency of the pipeline to detect true SNPs in EST data. The four true SNPs detected in *AGP* were found using the automated method. For *CCoAOMT* and *GRP*, one SNP out of three, and six SNPs out of nine were detected in the EST sequences, respectively. Interestingly, the missing SNPs corresponded to rare alleles

Table 4: Cluster profile of maritime pine ESTs resulting from the *StackPACK* analysis.

Number of ESTs per contig	Number of contigs	Total length (bp)
4	286	251,108
5	166	152,166
6	113	112,073
7	71	74,002
8	49	54,314
9	41	43,579
10	31	30,482
11	29	31,958
12	23	27,779
13	17	17,766
14	14	14,102
15	9	11,915
16	15	17,252
17	7	7,306
18	6	5,751
19	9	11,387
20	12	15,539
21	5	6,021
22	4	5,872
23	6	5,449
24	6	8,178
25	5	5,262
26	2	1,490
27	1	1,583
28	3	4,130
29	5	5,602
31	1	524
32	3	2,754
33	3	2,196
37	1	1,331
39	1	1,597
40	1	1,759
41	1	3,264
44	1	1,763
49	1	1,586
50	1	1,353
51	1	2,460
53	1	2,535
55	1	607
59	1	863
60	1	2,300
65	1	2,155
78	1	728
80	1	1,392
154	1	740
179	1	765

Figure 2: Proportion of contigs containing at least one SNP, in relation with the number of ESTs per contig e=2 to 100.

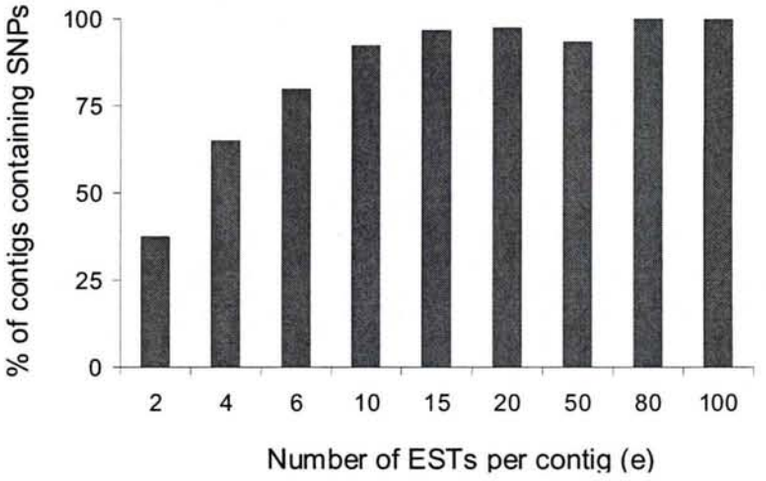


Figure 3: Comparison between the quantity of SNP detected by the automated method and the quantity expected considering the SNP frequency in the reference set (1 SNP / 90 bp) and the total length of the sequences used.

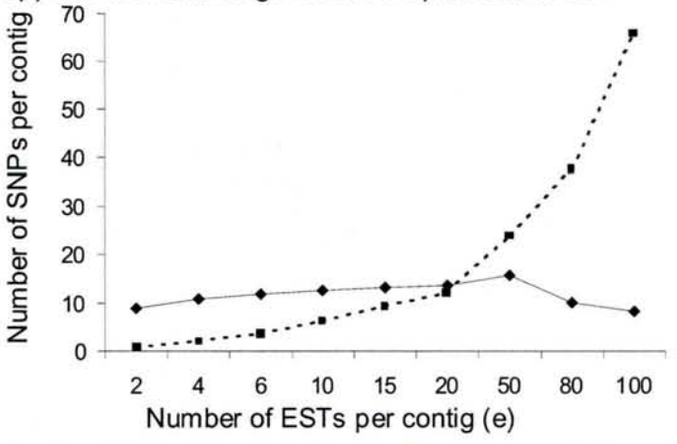


Table 5: SNP detection in maritime pine ESTs with a minimum and a maximum number of 4 and 20 ESTs per contig: variant type (transition *vs.* transversion) and SNP frequency.

Number of contigs 4<e<20		940
Total length of the sequences		924,216 bp
Number of contig with no SNP		568
SNP frequency		1 SNP / 660 bp
Number of SNP detected (P_{SNP}>0.99)		1400
More than 3 alleles		0
Transition	ag	374
	ct	444
	total	818
Transversion	ac	148
	at	167
	cg	122
	gt	145
	total	582

in the Aquitaine provenance that was the main source of mRNA for the construction of cDNA libraries. Again, this result demonstrates the limitation of detecting rare SNP alleles in EST datasets, even in contig size containing about 20 reads.

Automated SNP detection in maritime pine ESTs

Untreatable chromatograms and stringent quality checking led us to eliminate 9,630 sequences from the whole set of ESTs, resulting in a total of 18,498 ESTs used in the clustering and assembly steps. These ESTs had an average length of 480 bp and included 89% of bases with a quality score greater or equal to 20 (99% per base accuracy according to *Phred* definition). As a result of clustering and assembly, 13,447 sequences were members of one of the 2,697 multi-sequence clusters found, whereas 5,051 ESTs remained as singletons (Table 4). Thus a gene index of 7,748 different sequences was created, assuming that singletons corresponded to unique transcripts.

Considering that "g" was not a limiting factor in our EST dataset (see the Material and methods section), the SNP detection probability, P_{SNP}, corresponding to a contig size of e=20 ESTs was thus 0.86, 1 and 1 for a rare SNP allele frequency of p=0.1, 0.3 and 0.5, respectively. If 10 reads were considered, P_{SNP} dropped to 0.65 for p=0.1. In the whole EST database, the number of contigs containing at least one SNP was recorded for e values increasing from 2 to 100 (Figure 2). SNP abundance increased with increasing contig sample size and almost all contigs were found to contain at least one SNP for a number of sequence reads e>20. More interesting was the mean number of SNPs per contig in relation to the contig size. The abundance of SNPs per contig remained low for e<20 and then increased exponentially with e (Figure 3, dotted line), reaching a value of one SNP every 11 bp in the two contigs containing more than 100 reads. Such disproportionate number of SNPs was compared with the expected number of SNPs obtained using the SNP frequency observed in the reference set (i.e. one SNP per 100 bp) and the total length in bp used for the detection (Figure 3, plain line). To explain the striking difference between the observed and expected abundance of SNPs for contig size above 20 ESTs, it could be first hypothesized that rare SNP alleles were only detected in contigs containing a high number of sequence reads. However, it is more likely that such difference resulted from either accumulation of sequencing errors or the presence of paralogous sequences within large contigs, as reported by Batley et al. (2003) in maize. As a conclusion, it was decided to ignore those contigs with more than 20 sequence reads for the automatic detection of SNPs in the maritime pine EST dataset. On the other hand, a minimum of 4 reads was considered as a minimum for SNP mining. It should be noted that rare SNP alleles were probably overlooked for contigs represented by a small number of sequences (Figure 1).

The results of the automated SNP search are summarized in Table 5 and were obtained within 54 minutes on a 3.06 Ghz Intel Pentium IV PC with 512 MB RAM running Mandrake Linux 9.1). A total of 1400 SNPs (*Polybayes* P_{SNP}>0.99) were found in a set of 940 contigs containing between 4 and 20 reads and representing 942,216 bp in total length. Five hundred and sixty-eight contigs (60%) did not have any SNP. All the detected SNPs were biallelic, and more transitions (818) than transversions (582) were obtained, which agrees with other results obtained in previous SNP discovery programs (Picoult-Newberg et al. 1999, Batley et al. 2003). Overall, SNP abundance was one SNP per 660 bp, ranging from SNP frequencies of one per 1457 bp of aligned sequences for 4-read contigs, to one per 137 bp of aligned sequences for 20-read contigs.

Conclusion

This study demonstrates the usefulness of bioinformatics-based sequence analysis tools to extend the scope of EST data in the field of marker development. A bioinformatic pipeline was developed and allowed for the first time the detection of SNPs in a EST dataset of a forest tree species. By comparing automated and visual methods for SNP mining on a reference set of true SNPs, we showed the intrinsic power of the developed engine: i.e. 83.1% of all the SNPs and 98% of the "non-rare" SNP alleles could be detected. From the calculation of SNP detection probabilities, it was clearly shown that the efficiency to uncover true SNPs dropped drastically for rare SNP alleles ($p<0.1$). It was also shown that considering more than 20 reads per contig considerably increased the number of false SNPs, likely resulting from random errors accumulating during the sequencing step and/or the mixture of paralogous loci.

Finally, most of the SNPs detected in this study, for contig size ranging from 4 to 20, were found in known function genes (data not shown). Such "functional SNPs" will provide a valuable source of molecular markers for genetic mapping (Chagné et al. 2002), comparative mapping in pines (Chagné et al. 2003), nucleotide diversity analysis, and association studies, four topics that are currently explored in the frame of the maritime pine genomics program.

Acknowledgements

LLD was funded by the Conseil Régional d'Aquitaine, DC was funded by the French Ministry for Research. Maritime pine ESTs were produced with the support of the Aquitaine Région, INRA (Lignome project) and European Union (GEMINI project, QLK5-CT1999-00942) . We thank Drs Gabor T. Marth, Mark D. Yandell, Ian F. Korf and Warren R. Gish who gave us access to the Polybayes software.

References

Andreasson H., Asp A., Alderborn A., Gyllensten U., Allen M. 2002. Mitochondrial sequence analysis for forensic identification using pyrosequencing technology. Biotechniques 32: 124-6.

Barker G., Batley J., O' Sullivan H., Edwards K.J., Edwards D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19: 421-2.

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

Brookes A.J. 1999. The essence of SNPs. Gene 234: 177-186.

Brumfield R.T., Beerli P., Nickerson D.A., Edwards S.V. 2003. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology & Evolution 18: 249-256.

Cantón F.R., Le Provost G., Garcia V., Barré A., Frigério J-M., Paiva J., Fevereiro P., Ávila C., Mouret J-F., de Daruvar A., Cánovas F.M., Plomion C. 2003. Transcriptome analysis of wood formation in maritime pine. In : Sustainable Forestry, Wood products & Biotechnology, BIOFOR proceeding (in press).

Chagné D., Lalanne C., Madur D., Kumar S., Frigerio J-M., Krier C., Decroocq S., Savoure A., Bou-Dagher K-M., Bertocchi E., Brach J., Plomion C. 2002. A high density genetic map of maritime pine based on AFLPs. Ann For Sci 59: 627-636.

Chagné D., Brown G., Lalanne C., Madur D., Pot D., Neale D., Plomion C. 2003. Comparative genome and QTL mapping between maritime and loblolly pines. Mol Breed 12: 185-195.

Ching A., Caldwell K.S., Jung M., Dolan M., Smith O.S., Tingey S., Morgante M., Rafalski A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 1-19.

Cho R.J., Mindrinos M., Richards D.R., Sapolsky R.J., Anderson M., Drenkard E., Dewdney J, Reuber T.L., Stammers M., Federspiel N., Theologis A., Yang W.H., Hubbell E., Au M., Chung E.Y., Lashkari D., Lemieux B., Dean C., Lipshutz R.J., Ausubel F.M., Davis R.W., Oefner P.J. 1999. Genome-wide mapping with biallelic

markers in Arabidopsis thaliana. Nat genet 23: 203-207.

Christoffels A., van Gelder A., Greyling G., Miller R., Hide T., Hide W. 2001. STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nuc Ac Res 29(1): 238-238.

Collins F.S., Guyer M.S., Charkravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. Science 278: 1580-1.

Collins A., Lonjou C., Morton N.E. 1999. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96: 15173-15177.

Emahazion T., Feuk L., Jobs M., Sawyer S.L., Fredman D., St Clair D., Prince J.A., Brookes A.J. 2001. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet 17: 407-413.

Ewing B., Green P. 1998. Base calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 8: 186-194.

Ewing B., Hiller L.D., Wendl M.C., Green P. 1998. Base calling of automated sequencer traces using Phred. II. Accuracy assessment. Genome Res 8: 175-185.

Frigerio J-M., Dubos C., Chaumeil P., Salin F., Garcia V., Barré A., Plomion C. 2003. Using transcriptome analysis to identify osmotic stress candidate genes in maritime pine (*Pinus pinaster* Ait.). In : Sustainable Forestry, Wood products & Biotechnology, BIOFOR proceeding (in press).

Gordon D., Abajian C., Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8: 195-202.

Gray I.C., Campbell D.A., Spurr N.K. 2000. Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9: 2403-2408.

Kota R., Rudd S., Facius A., Kolesov G., Thiel T., Zhang H., Stein N., Mayer K., Graner A. 2003. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Mol Gen Genom (in press).

Kruglyak L. 1997. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17: 21-4.

Le Provost G., Paiva J., Pot D., Brach J., Plomion C. 2003. Seasonal variation in transcript accumulation in wood forming tissues of maritime pine (Pinus pinaster Ait.) with emphasis on a Cell wall Glycin Rich Protein. Planta. 217: 820-830.

Letondal C. 2001. A Web interface generator for molecular biology programs in Unix. Bioinformatics 17: 73-82.

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P., Gish W.R. 1999. A general approach to single-nucleotide polymorphism discovery. Nat Genet 23: 452-456.

Nickerson D.A., Tobe V.O., Taylor S.L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nuc Ac Res 25: 2745-51.

Nordborg M., Borevitz J.O., Bergelson J., Berry C.C., Chory J., Hagenblad J., Kreitman M., Maloof J.N., Noyes T., Oefner P.J., Stahl E.A., Weigel D. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30: 190-3.

Grivet L., Glaszmann J-C., Vincentz M., da Silva F., Arruda P. 2003. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. Theor Appl Genet 106:190-197.

Somers D.L., Kirkpatrick R., Moniwa M., Walsh A. 2003. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 49: 431-437.

Picoult-Newberg L., Ideker T.E., Pohl M.G., Taylor S.L., Donaldson M.A., Nickerson D.A., Boyce-Jacino M. 1999. Mining SNPs from EST databases. Genome Res. 9: 167-74.

Rounsley S., Xiaoying L., Ketchum K.A. 1998. Large-scale sequencing of plant genomes. Cur Op Plant Biol 1:136-141.

Riley J.H., Allan C.J., Lai E., Roses A. 2000. The use of single nucleotide polymorphisms in the isolation of common disease genes. Pharmacogenomics 1: 39-47.

Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., Fulton L., Hillier L., Waterston R.H., McPherson J.D., Gilman B., Schaffner S., Van Etten W.J., Reich D., Higgins J., Daly M.J., Blumenstiel B., Baldwin J., Stange-Thomann N., Zody M.C., Linton L., Lander E.S., Atshuler D. 2001. A map of human genome sequence variation

containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

Thornsberry J.M., Goodman M.M., Doebley J., Kresovich S., Nielsen D., Buckler E.S. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28: 286-289.

Useche F.J., Gao G., Harafey M., Rafalski A. 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. Genome Inform Ser Workshop Genome Inform. 12: 194-203.

Wilson M.R., Di Zinno J.A., Polanskey D., Replogle J., Budowle B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med. 108: 68-74.

ANNEXE VI

Analysis of the distribution of marker classes in a genetic linkage map of Norway spruce (*Picea abies* Karst.).

Analysis of the distribution of marker classes ina genetic linkage map of Norway spruce (*P. abies* Karst.)

Ivan Scotti[1]*, Andrea Burelli[1], David Chagné[2], John Fuller[3], Peter E. Hedley[3], Gunnar Jansson[4], David Neale[5], Christophe Plomion[2], Wayne Powell[3], Michela Troggio[5], Michele Morgante[1]

1 : Dipartimento di Produzione Vegetale e Tecnologie Agrarie, University of Udine - Via delle Scienze 208 - 33100 Udine (Italy)
2 : INRA - Laboratoire de Genetique et Amelioration des Arbres Forestiers - 69 route d'Arcachon - F-33612 Cestas (France)
3: Scottish Crop Research Institute - Invergowrie - Dundee DD2 5DA (United Kingdom)
4: Skogforsk - Uppsala Science Park, SE 75183, Uppsala (Sweden)
5: Institute of Forest Genetics - Pacific Southwest Research Station, USDA Forest Service - Department of Environmental Horticulture, University of California - Davis, CA 95616 (USA)
*Current address: Department of Biology, Indiana University, Jordan Hall A135 – Bloomington, IN 47405- (USA)

Corresponding author:
Michele Morgante
Dipartimento di Produzione Vegetale e Tecnologie Agrarie, University of Udine - Via delle Scienze 208 - 33100 Udine (Italy)
Phone 0039 0342558606 - Fax 00390432558603 - E-mail: michele.morgante@uniud.it

Abstract

A genetic linkage map of Norway spruce (*Picea abies* Karst.) was constructed using a total of 691 markers comprising six marker types, including dominant and codominant loci, and loci belonging to the low and high copy-number fractions of the genome (AFLPs, S-SAPs, SSRs, ESTPs, IRAPs, and SCARs), using a pseudo-testcross scheme. Upon generation of the female and male parent maps, thirteen linkage groups were obtained, including at least one linkage group from each of the two parental maps, bridged by at least one shared marker, using a selective LOD threshold. Upon relaxation of this threshold, further linkage groups found a counterpart, resulting in a map covering a total of 2316 cM and 1668 cM, respectively in the female and male parent, representing fifteen shared linkage groups. The analysis of the distribution of marker types over linkage groups and relative to each other shows that marker classes are not randomly represented in the map, with some classes (e.g. ESTs) contributing more than others, in relative terms, to the set of linked markers. IRAP and AFLP markers show clustering and S-SAPs and IRAPs are separated from ESTs and SSRs, respectively. In general, low copy- and high copy-sequence derived markers displayed a tendency to occupy separate regions of the linkage groups. Despite the variety of marker types used in this study, the number of linkage groups in each parental map was not significantly lower than that obtained in previous studies in the *Picea* genus, when the same linkage analysis criteria are used.

Keywords: *Picea abies* - conifers - linkage map - genome structure - molecular markers

Introduction

Genetic linkage maps are an essential tool in genetics. They provide information on the distribution of markers and polymorphisms across the genome, which represents a base for the study of genome structure and evolution. They also allow to localise genetic loci underlying ecologically and economically important traits. In that way genetic linkage maps are useful for several applications, such as marker-assisted selection. In conifers, the need for saturated linkage maps is made even more pressing by the long generation time and by the strictly outbreeding mating system, which hampers the design of

suitable controlled cross schemes and the application of selection over a large number of generations.

Mapping in conifers has some advantages, represented by the possibility to analyse large numbers of megagametophytes, genetically equivalent to a population of gametes. Single-tree maps have been built with this method for several species (Tulsieram et al.,1992; Binelli and Bucci, 1994; Paglia et al., 1998; Gosselin et al., 2002; Remington et al., 1999; Echt and Nelson, 199t; Costa et al., 2000), providing a large set of data (http://www.pierroton.inra.fr/genetics/labo/mapreview.html) that allow the comparison of genetic features of the different taxa. On the other hand, this strategy has the basic disadvantage of producing experimental "dead-ends", since the addition of markers to the map is limited by the amount of DNA that can be extracted from the chosen set of endosperms. Other maps, based on the living offspring of a controlled cross (e.g. Devey et al., 1994; Chagné et al., in press), can be extended indefinitely, because trees can be grafted or rooted and therefore potentially immortalised. Grattapaglia and Sederoff (1994) introduced pseudo-testcross mapping in tree genetics. This scheme is based on the use of the offspring of two heterozygote parents as the mapping population, and results in two maps, one for each parent tree. Codominant markers segregating in both parents, and dominant inter-cross markers (segregating in both parents), can be used to identify homologous linkage groups between the two maps. Therefore, the combination of large numbers of markers in different segregation configurations, together with the application of the pseudo-testcross scheme, allows to identify linkage groups that result from the juxtaposition of male and female linkage groups. If more than one marker is shared between the pairs of linkage groups, then the biological properties of male and female meiotic processes (such as recombination frequency) can also be compared, the groups can be aligned ("multiple parallel linkage"), and a consensus map can be built, if a sufficiently large set of markers is shared between the two maps.

Maps of species belonging to the genus *Picea* (Binelli and Bucci 1994; Paglia et al., 1998; Gosselin et al., 2002) typically do not show the expected number of linkage groups, corresponding to the haploid chromosome number (n = 12), despite the fact that some (Paglia et al., 1998) included a number of markers comparable with those mapped in Pinus species saturated maps (Echt and Nelson, 1997; Remington et al., 1999; Costa et al., 2000), for which 12 linkage groups were identified. One possible explanation for this observation is the selection of one or few types of markers for the construction of the linkage maps, that may not evenly cover the chromosomes (Gosselin et al. 2002). We have decided here to use different types of markers, corresponding to different classes of sequence families. Amplified fragment length polymorphisms (AFLPs; Vos et al., 1995) were used to produce a large amount of markers located into potentially expressed regions of the genome using a hypo-methylated DNA-digesting restriction enzyme (PstI); Expressed Sequence Tag Polymorphisms (ESTPs) are by definition expressed regions and usually represent the low copy-number fraction of the genome; Simple Sequence Repeats (SSRs or microsatellites) are a class of repeated DNA which shows a rather even distribution in Norway spruce genome relative to other sequence families (Pfeiffer et al., 1997; Scotti et al., 2002a and 2002b). However, SSRs have recently been shown to be associated with non-repetitive DNA in plants (Morgante et al., 2002), and to cluster to some extent in conifers (Elsik and Williams, 2001). Sequence-Specific Amplified Polymorphisms (S-SAP; Waugh et al., 1997) are based on the amplification of DNA regions flanking retrotransposons,

as do the Inter-Retrotransposon Amplified Polymorphisms (IRAP; Kalendar et al., 1999), and therefore target one of the most highly repeated classes of sequences in conifers, representing a large fraction of the genome (Cattonaro et al., in prep.). Therefore, our data set includes markers acting as probes for DNA fragments with a variety of sequence, function, and distribution properties. This has allowed us to study the distribution of different marker types along and between linkage groups, exploiting the differences among them in sequence class content. Moreover, we were able to describe a large-scale "geography" of the genome of Norway spruce, by the application of statistical methods, borrowed from geographical sciences (autocorrelation; Cliff and Ord, 1973), that allowed the testing of the clustering of markers of the same or different types. This approach, applied for the first time here to the analysis of the distribution of markers in a linkage map, produced an overview of the architecture of the genome of Norway spruce, a step toward a better understanding of the structure of the genome of conifers.

Materials and Methods

Mapping population

Eighty-five plants were selected among 135 offsprings of the cross between mother tree N2022 and father tree E2006. This family belongs to a partial diallelic cross grown at four different locations in Sweden (Osby, 56°26'N, 14°20'E, 150 m a.s.l.; Hörby, 55°54'N, 13°46'E, 155 m a.s.l.; Tönnersjö, 56°40'N, 13°05'E, 90 m a.s.l.; Vetlanda, 57°25'N, 15°09'E, 230 m a.s.l.). These plants are presently mature, and data were collected on them for several quantitative characters, which will not be discussed here. DNA was extracted from needles of each plant either using the Qiagen Plant DNeasy kit (Qiagen, Hilden, Germany) (for AFLP, ESTP, and S-SAP markers) or with the protocol described in Scotti et al., 2002c (for SSRs and inter-LTR markers).

Markers

Nomenclature. Markers were named, as far as possible, following the nomenclature used in the Treegenes database for reference markers in forest tree genetics (http://dendrome.ucdavis.edu/Treegenes/locusname.html). Details on the meaning of the names are given along with the description of protocols below; details on the identity of bands are available from the institutions that have produced the markers (as identified in each marker name) on request.

AFLP reactions were performed as described in Paglia and Morgante (1998), but with half the amount of isotope, digesting genomic DNA with the enzymes *Pst*I and *Mse*I. Marker names indicate the pair of markers used for the selective amplification plus a progressive number identifying the band in the gel pattern it was scored from. Two groups of markers with different names, run in two separate instances, are present. In AFLP names including the letters "pm", the following letter indicates the *Pst*I selective primer (with letters corresponding to the primers described in Paglia and Morgante (1998), and "d" corresponding to a primer with sequence: 5'-GACTGCGTACATGCAGCAC-3'), the first two digits indicate the MseI selective primer (in the Keygene code; see http://grain.jouy.inra.fr/ggpages for a reference list), the last two digits indicate the progressive number of the band in the gel pattern it was derived from. In AFLP names including the letters "pst", the first pair of digits represents the PstI selective primer, the second pair of digits indicates the MseI selective primer (both in Keygene number code), the remaining digit(s) indicate the progressive number of the band in the gel pattern it was derived from.

S-SAP reactions were performed according to Waugh et al. (1997), for markers indicated by the letters "pmp", using the same template as the AFLPs, and using a primer designed on the LTR of the

Table 1. (a) List of primers designed on *alisei* and *copia* LTR sequences. F/R refers to the orientation, and digits to the starting position, of the primer sequence relative to Norway spruce Gypsy LTR sequence (Cattonaro unpublished) (b) primer pairs used for the inter-LTR markers, with the corresponding marker names. Isotope-labelled primer of each pair in rows, cold primer in columns.

(a)

Primer name	**GenBank accession**	**coordinates**	**Primer Sequence**
LTR-F15	AF180435		ATCTGAAATTATCTCTTTTCAAGG
LTR-F67	AF180428		ATGCTATATTGATGTTTGGTTCTT
LTR-F214	AF180429		TTATCTCTTTTTCAAGGCGTT
LTR-R35	AF180936		ATTAAGGGGTAATGAGCCAAT
LTR-R195	AF180429		AAATGAAAATGTCTCCTTGGT
*Copia*LTR-F105	AF104492		TGGTGGTATGTAGAACGTCAC
*Copia*LTR-R111	AF104495		ACCACCAGAAATGTCATGAC

(b)

	LTR-F15	LTR-F67	LTR-F214	LTR-R35	LTR-R195	*Copia*LTR-F105
LTR-F15						LTR106-LTR110
LTR-F214				LTR044-LTR056	LTR088-LTR094	LTR030-LTR043
LTR-R35		LTR001-LTR009				
*Copia*LTR-F105		LTR057-LTR067		LTR017-LTR029		
*Copia*LTR-R111	LTR102-LTR105	LTR068-LTR087	LTR010-LTR016		LTR095-LTR101	

copia element of Pinus contorta (5'-GTACTATATGTTTACGACATG -3'; kindly provided by Jim Provan, Scottish Crop Research Institute) in the selective PCR. In these markers' names, the first two digits indicate the *Mse*I primer that was used in combination with the copia primer in the selective amplification, and the following digits indicate the progressive number of the band in the gel pattern it was derived from. For S-SAP markers indicated by the letters "al", DNA was digested with the enzymes *Eco*RI and *Mse*I, ligated, and pre-amplified according to the AFLP protocol described in Paglia and Morgante (1998). Following pre-amplification, a linear PCR was performed using the primer LTR3'F15 (table 1a) designed on the LTR of the Alisei *gypsy* retroelement of Norway spruce (Cattonaro et al., unpublished; see Table 1 for the position, orientation and original sequence of all *Alisei* primers used here), using the following protocol: 0.3 µM primer, 200 µM each dNTP, 0.5 U of AmpliTaq Gold *Taq* Polymerase (Applied Biosystems, Foster CA), 5 µL of a 1:10 dilution of pre-amplified DNA, in 1X AmpliTaq Buffer, in a total volume of 20 µL, and with the following PCR cycle: 5' at 94 °C followed by 25 cycles (30" at 94 °C, 30" at 55 °C, 1' at 72 °C). Following the linear PCR, the selective PCR was performed with the primer LTR3'F67 (table 1a), labelled as described in Paglia and Morgante (1998), with the following protocol: 0.25 µM labelled primer, 0.30 µM *Eco*RI primer, 200 µM each dNTP, 0.5 U AmpliTaq Gold DNA Polymerase, 5 µL of linear-amplified DNA, 1:5 diluted, in 1X AmpliTaq Buffer in a total volume of 20 µL, and with the cycle described in Paglia and Morgante (1998). In these marker names, the first two digits represent the *Eco*RI primer used in the selective amplification (Keygene code) and the following digits indicate band size.

IRAP markers (Kalendar et al. 1999). A further set of primers were designed on *Alisei* LTRs and on Norway spruce *copia* LTRs (Zuccolo et al., unpublished) (Table 1a) and used, single or in pairs (Table 1b), to amplify genomic DNA of Norway spruce, following a protocol identical to the one used for SSR amplification in Scotti et al. (2002a, 2002b). For all the primers annealing temperature was set to 56 °C; PCR cycle and detection were as described in Scotti et al. (2002a). Correspondence between IRAP marker names and primer pairs are displayed in table 1b. Only a subset of pairs, yielding a substantial amount of polymorphic bands, was amplified on the whole mapping population.

SSRs are amplified according to Pfeiffer et al. (1997) and to Scotti et al. (2002a, 2002b). As described in the cited papers, some primer pairs produced multi-band patterns. When more than one locus could be scored, they were given names corresponding to the primer pair name, plus an extension in digits and/or letters.

ESTP markers are in part described in Chagné et al. (2003); those that are not described in the cited paper will be described in a paper companion to the present one (Troggio et al. in prepapration).

In addition to these, a handful of SSCP markers (with the conditions described in Plomion et al., 1999) derived from RAPD, SCAR, and SSR markers (Binelli and Bucci, 1994; Scotti et al., 1998, 1999) were included into the data set.

Construction of the genetic linkage map

To construct the linkage map, a step-by-step procedure was followed, combining the features of *MAPMAKER/EXP* (Lander et al. 1987) and *JOINMAP* (Stam 1993), and adding markers progressively according to their informativeness and their quality. At the first step, all markers informative in each of the parents (segregating in 1:1 or in 1:1:1:1 configuration), and without segregation distortion, were used to build the female and the male map separately using *MAPMAKER*, using the group command with LOD>4 and $\Theta<0.25$ and then

ordering the markers with multi-point LOD>5. Here, data sets were doubled by adding to each one the set of re-coded data, in order to allow *MAPMAKER* to analyse pairs of markers in repulsion phase. A "framework" map was therefore obtained, and its linkage groups were used for map length, map coverage estimation, and autocorrelation analysis. In the second step, all unlinked markers in the *MAPMAKER* data set were checked for linkage to the groups obtained above at LOD>2 and $\Theta<0.40$, and assigned as accessory markers when they displayed linkage to a mapped marker. In the third step, distorted markers and markers segregating in 3:1 configuration were checked for linkage in the two maps using JOINMAP, and added as accessory markers (to one or both maps) when they displayed linkage to a mapped or accessory marker with thresholds LOD>2, $\Theta<0.40$. In the fourth step, the male and female "framework" maps were searched for common markers, and the groups were aligned according to the markers (either framework or accessory) they shared. Linkage groups that found no match in the opposite map were not used in the subsequent analyses. An "inclusive" map was obtained this way, and used for the analysis of the distribution of markers mapped in the different linkage groups.

Map coverage and estimation of genome length.

Estimates of genome length were obtained on the "framework" map using the methods of Hulbert et al. (1988). Confidence intervals with $\alpha = 0.05$ were obtained as in Gerber and Rodolphe (1994).

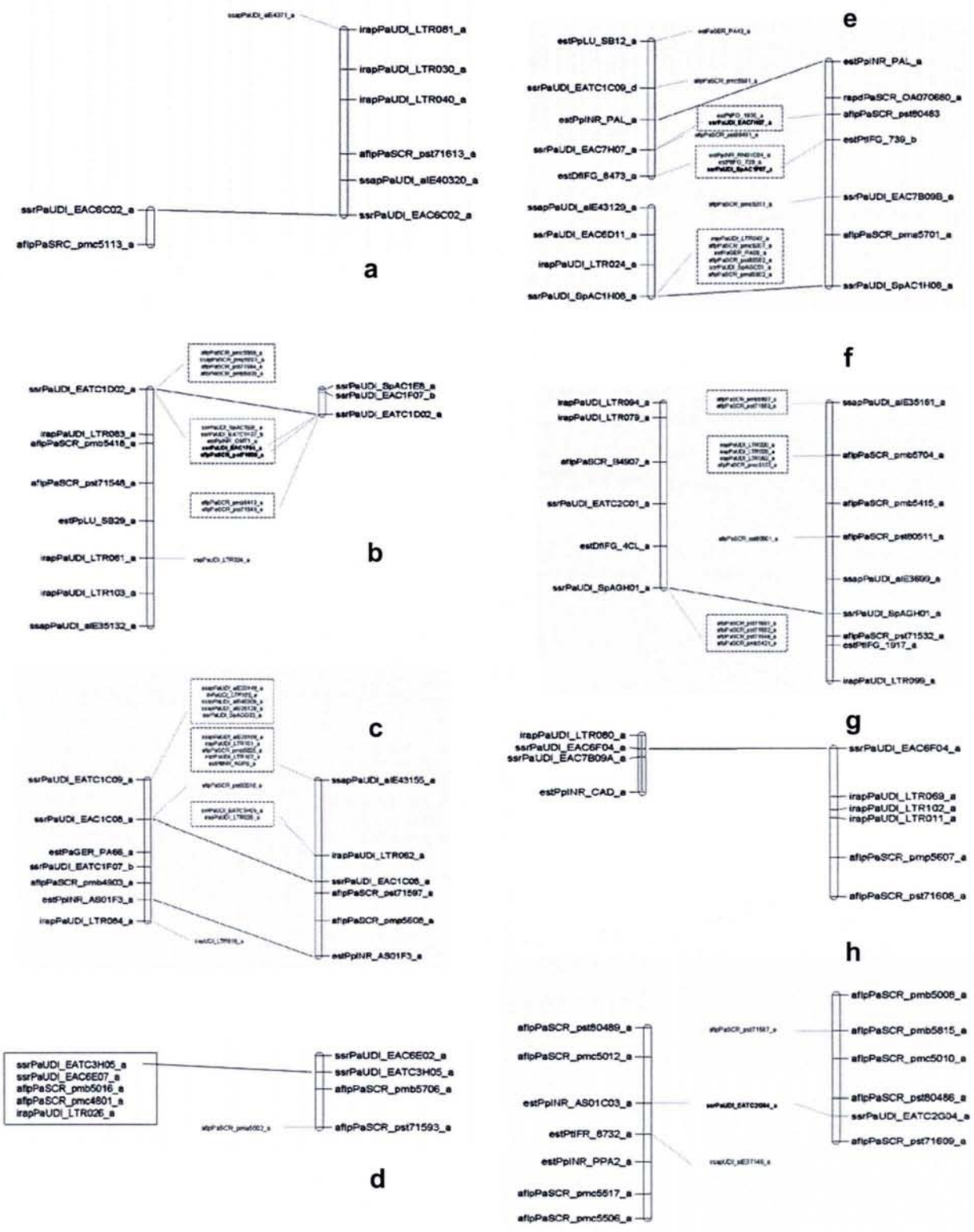
Markers distribution analysis.

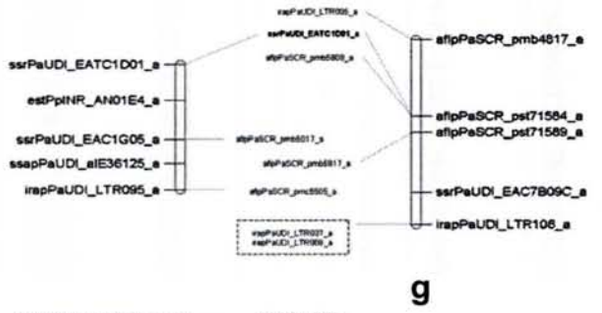
Marker distribution over linkage groups was analysed in the "inclusive" map with standard statistical methods ($\chi 2$ test, G test). The observed frequency of markers of each type in the data set was used to derive the expected number of linked markers for each class; the observed number of linked markers for each class was used to derive the expected number of markers of each type that should be found in each linkage group (joining together all the male and female linkage groups belonging to the same group). Heterogeneity of recombination between the two parents was tested according to the methods described by Liu (1998).

The distribution of marker types relative to each other was tested in the "framework" map using spatial autocorrelation (Cliff and Ord, 1973). Standard normal deviate (s.n.d.; Sokal and Oden, 1978) was used as the statistic to test autocorrelation. This method allows to test whether the number of pairs ("joins") of objects (loci in this case) of the same type (same marker type, in our case), or of different types, observed within a range of distances (here, map distances) meets the expectation based on the number of pairs of objects (loci) observed in that range and on the frequency of objects of each type (marker type) observed in the sample (marker data set). Positive significant values of s.n.d. (s.n.d. > 1.96 for $\alpha = 0.05$) indicate clustering, negative significant values of s.n.d. (s.n.d. < -1.96 for $\alpha = 0.05$) indicate separation. Since this analysis can be applied to markers of the same type or of different types, this allows to identify both the tendency of markers of one type to cluster together (or to spread) and the tendency of one marker type to be found close (or distant) from another.

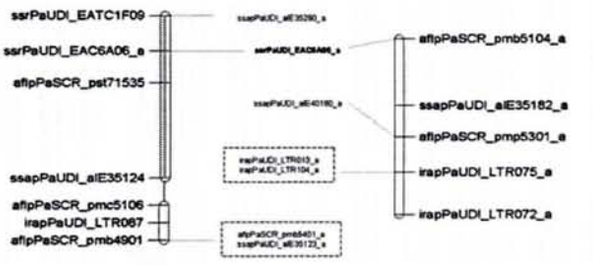
In order to include marker pairs lying at all possible genetic distances, the "first order" command of *MAPMAKER* v2.0 was used to order all markers previously grouped at LOD>4.0 in the "framework" map. Although this order is not supported statistically, and is not presented here, it allowed to include all markers in the analysis. Inspection of the result of the "*ripple*" analysis, performed using *MAPMAKER* on this order, shows that

Figure 1: Parallel groups from the "framework" linkage map of Norway spruce. For each linkage group, female groups are displayed on the left, male groups on the right. Multilocus order supported at LOD>5.0.

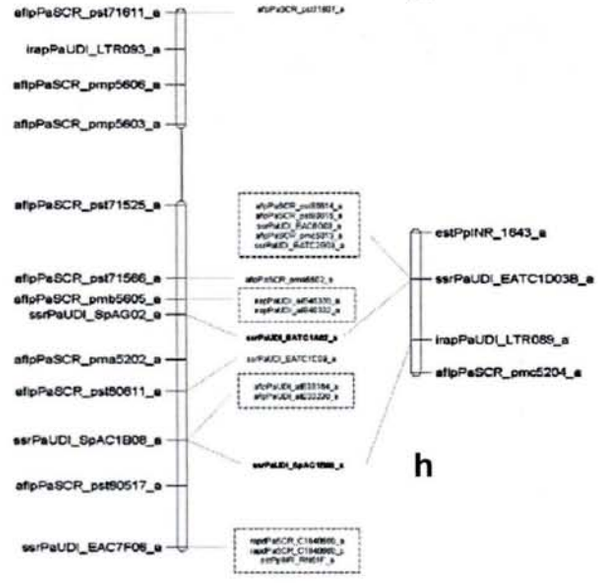



ssrPaUDI_EATC1D01_a
estPpINR_AN01E4_a
ssrPaUDI_EAC1G05_a
ssapPaUDI_aIE36125_a
irapPaUDI_LTR095_a
aflpPaSCR_pmb4617_a
aflpPaSCR_pst71584_a
aflpPaSCR_pst71589_a
ssrPaUDI_EAC7B09C_a
irapPaUDI_LTR106_a

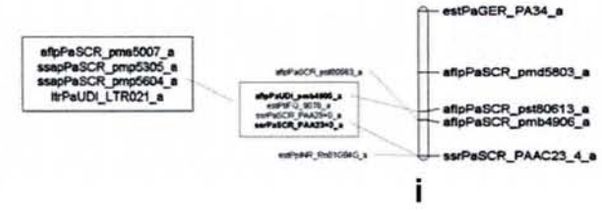
g


ssrPaUDI_EATC1F09
ssrPaUDI_EAC6A06_a
aflpPaSCR_pst71535
ssapPaUDI_aIE35124
aflpPaSCR_pmc5106
irapPaUDI_LTR067
aflpPaSCR_pmb4901
aflpPaSCR_pmb5104_a
ssapPaUDI_aIE35182_a
aflpPaSCR_pmp5301_a
irapPaUDI_LTR075_a
irapPaUDI_LTR072_a

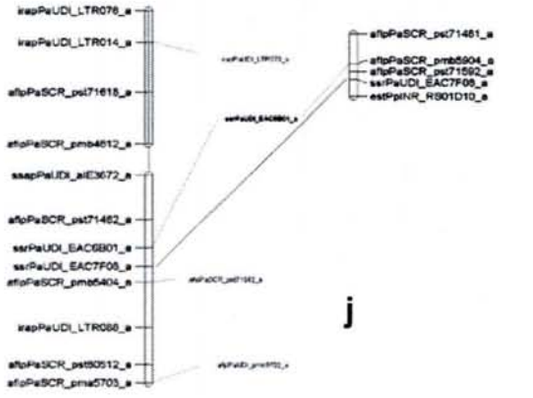
k


aflpPaSCR_pst71611_a
irapPaUDI_LTR093_a
aflpPaSCR_pmp5606_a
aflpPaSCR_pmp5603_a
aflpPaSCR_pst71525_a
aflpPaSCR_pst71566_a
aflpPaSCR_pmb5605_a
ssrPaUDI_SpAG02_a
aflpPaSCR_pma5202_a
aflpPaSCR_pst50611_a
ssrPaUDI_SpAC1B08_a
aflpPaSCR_pst50517_a
ssrPaUDI_EAC7F06_a
estPpINR_1643_a
ssrPaUDI_EATC1D03B_a
irapPaUDI_LTR089_a
aflpPaSCR_pmc5204_a

h


aflpPaSCR_pma5007_a
ssapPaSCR_pmp5305_a
ssapPaSCR_pmp5604_a
ltrPaUDI_LTR021_a
estPaGER_PA34_a
aflpPaSCR_pmd5803_a
aflpPaSCR_pst80613_a
aflpPaSCR_pmb4906_a
ssrPaSCR_PAAC23_4_a

i


irapPaUDI_LTR076_a
irapPaUDI_LTR014_a
aflpPaSCR_pst71618_a
aflpPaSCR_pmb4612_a
ssapPaUDI_aIE3672_a
aflpPaSCR_pst71462_a
ssrPaUDI_EAC6B01_a
ssrPaUDI_EAC7F06_a
aflpPaSCR_pmb5404_a
irapPaUDI_LTR088_a
aflpPaSCR_pst50512_a
aflpPaSCR_pma5703_a
aflpPaSCR_pst71461_a
aflpPaSCR_pmb5904_a
aflpPaSCR_pst71592_a
ssrPaUDI_EAC7F06_a
estPpINR_R901D10_a

j

local uncertainties in marker order do not seriously affect distances between markers. Therefore, this map order was used as the starting point for calculating the numbers of pairs in each distance class (see results). All calculations were made by hand based on the formulas given in Sokal and Oden (1978).

Results

Molecular markers

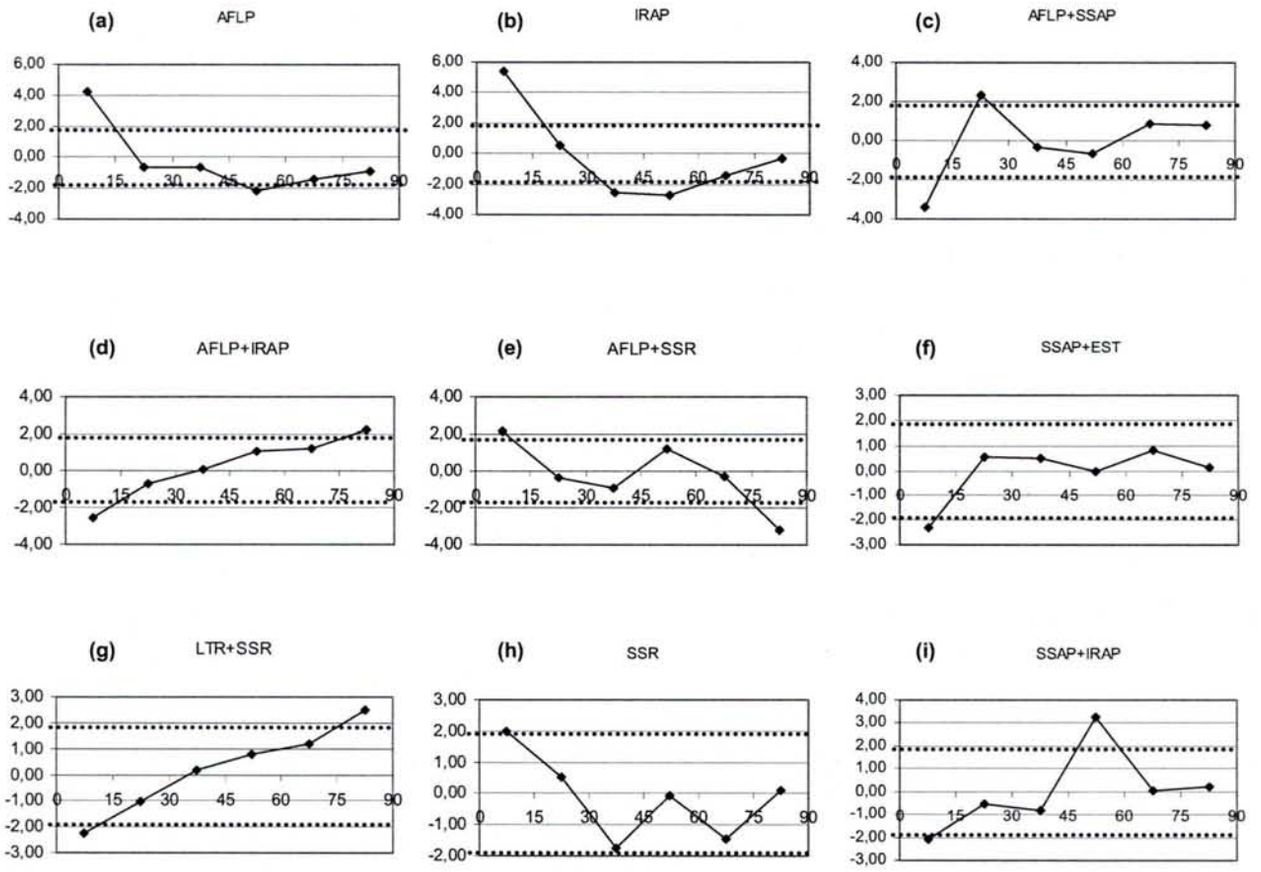
On average, 7.6 segregating bands per primer combination were obtained from 39 primer combinations for AFLP markers; 117 markers segregated in the female, 101 in the male, and 78 in 3:1 configuration. For S-SAP markers, 118 markers were obtained from 9 primer combinations, with an average of 13.1 polymorphic bands per gel, 35 markers segregating in the female, 31 in the male and 52 in 3:1 configuration. For IRAP markers, the average number of segregating bands per gel was 11.0, from 10 primer combinations, with 40 markers segregating in the female, 48 in the male, and 22 in 3:1 configuration. SSR and ESTP markers produced a subset of fully informative markers (as expected for codominant markers). The total number of loci segregating for SSRs was 76, with 48 informative markers in the female, 44 in the male, and three segregating in 3:1 configuration. The ESTP markers were 69, with 42 informative loci in the female, 34 in the male, six segregating in 3:1 configuration. The total number of loci in the two data sets was 469 and 434, for the female and the male respectively (691 globally).

In order to check consistency of data sets produced in different labs, all sample sets were checked a posteriori for identity at the University of Udine using two fully-informative SSR markers; no inconsistency was found.

Construction of the linkage map

The total number of linked markers at a two-point LOD score of 4.0 or more was 258 for the female (55.0%) and 218 for the male (50.2%), respectively. The "framework" linkage groups that would eventually collapse into the final fifteen shared linkage groups (21 for the female, 17 for the male) of the "inclusive" map described here include 203 (9.7 markers per group on average) for the female, and 152 markers (8.9 markers per linkage group on average) for the male. Smaller "framework" linkage groups, non included in the "inclusive" map, are not further discussed here. The average distance between adjacent markers was 12.8 cM for the female and 12.5 cM for the male, and the "framework" maps covered a total distance of 2316.4 cM and 1668.6 cM in the female and the male respectively (Haldane function). The total number of markers in the "inclusive" map was 422 (or 61% of the 691 markers that composed the global data set). Markers shared between the two parent maps, both framework and accessory, were used to combine linkage groups of the two maps into sets of homologous linkage groups. A total of 13 "parallel linkage groups" were obtained at the "framework" map stage (displayed in Figure 1), while the "inclusive" map (displayed as accessory material at the webpage http://www.pierroton.inra.fr/genetics/Picea) presented 15 "parallele linkage groups". Map stability was further checked by selecting a subset of the most trustable markers (50% of all dominant markers, selected based on their scorability, plus all co-dominant markers) and re-running map construction analyses. Overall map order was confirmed for all linkage groups, although the subset of markers that contributed to the framework map may slightly change. Heterogeneity of recombination frequency between the two parents was tested for all the couples of framework markers present in both parents, which showed linkage in the two-

Figure 2. Autocorrelograms for the eight s.n.d. plots that showed at least one significant value, and for SSRs. In each plot, on the x-axis: distance classes; on the y-axis: s.n.d. Note that each plot is scaled differently for ease of reading. Horizontal dashed lines indicate limits of significance (+1.96 and -1.96).



point analysis. Out of eleven couples, none showed heterogeneity in a $\chi 2$ test with 1 d.f., $\alpha = 0.05$. Heterogeneity of average distances between adjacent markers in linkage groups of the two parent maps was also tested both assuming normal distribution of the average distances (Student's t test) and assuming no normal distribution (Mann-Whitney U test). In both cases, no heterogeneity was found ($P = 0.63$ and $P = 0.57$, respectively). Regression of linkage group length against number of markers appearing in each linkage group shows a linear relationship both for the female ($R^2 = 0.83$, slope $b = 13.9$) and for the male ($R^2 = 0.88$, $b = 15.9$), with no overall marker density heterogeneity between linkage groups and between parents. Genome length estimates were obtained from the two-point analyses of the female and male framework maps separately. Two estimates were obtained from each data set, one with threshold LOD = 4, and one with threshold LOD = 6. For the female data set, $G_{LOD=4} = 3454 \pm 42$ cM, $G_{LOD=6} = 3540 \pm 37$ cM. For the male data set, $G_{LOD=4} = 2218 \pm 39$ cM, $G_{LOD=6} = 2074 \pm 34$ cM (Haldane function). Averaging over the four estimates, we obtain an estimated genome length of 2821 cM.

Marker distribution analysis

The thirteen "parallel linkage groups", each one considered as a single unit, were tested for heterogeneity of the distribution of markers belonging to different classes. A G-test was applied to the number of markers of each class that appeared in each linkage group. Expected numbers of markers of each class in each linkage group were obtained as described above, and the G-test shows no departure from an even distribution of marker types over linkage groups; the different chromosomes appear therefore to be compositionally similar, as far as marker class representation can tell. On the contrary, a $\chi 2$ test applied to the overall number of markers of each class that fall into the pool of linkage groups, compared to the numbers expected based on the frequency of markers in the data set, turned out to be significant at $\alpha=0.001$, showing that some marker classes (e.g. ESTPs, Inter-LTRs, and SSRs) contributed to the mapped marker set more than expected, while others (e.g. AFLPs) contributed less.

Spatial autocorrelation analysis was applied to the five classes of markers: AFLPs, S-SAPs, IRAPs, SSRs, and ESTPs, and to all possible pairs of classes; a total of fifteen autocorrelograms were obtained. The subset of eight correlograms, showing at least one s.n.d. value with significant departure from random distribution, and the autocorrelogram for SSRs, which displays one value very close to significance, are displayed in Figure 2. Distance classes were chosen in order to find a compromise between distance class width and number of pairs of loci falling into each distance class. Six distance windows of equal width were chosen: 0-15 cM, 15-30 cM, 30-45 cM, 45-60 cM, 60-75 cM, 75-90 cM. The last possible class, with distances larger than 90 cM, was not considered here, since it would include too few pairs. For each distance class, the number of pairs, of pairs for each of the fifteen possible kinds, and the number of markers contributing to the pairs, were counted on the framework map linkage groups (each single female and male group taken separately). Expected numbers of pairs of each type in each distance class were computed directly form the frequency of markers in the framework map and from the counts of pairs in each distance class. The results obtained show that AFLP (Figure 2a) and IRAP markers (Figure 2b) tend to cluster (positive significant autocorrelation at the shortest distance class), and that AFLP and IRAP markers tend not to lie far from each other (negative significant autocorrelation at the longest distance class; figures 2a and 2b). When the pairs of two different marker types are analysed, the trend for AFLP

markers is not to be close to retro-transposon based markers (Figures 3c and 3d), and to be not found at large distances from SSRs (figure 2e), while S-SAPs are not found close to ESTPs (Figure 2f) and IRAPs show a clear tendency to separation from SSRs and AFLPs (Figures 2d and 2g). Finally, SSRs (Figure 2h) show a value of s.n.d. just below the significance threshold (s.n.d = 1.94) at the shortest distance class, a hint of the tendency to some degree of clustering.

Discussion

A variety of markers have been applied here for the construction of a genetic linkage map of Norway spruce. AFLPs were the most represented set (43%), S-SAPs covered 18% of the markers, Inter-LTR markers amounted to 16% of the data set, SSRs and ESPTs corresponded to 11% and 10% of the loci respectively. In this map, therefore, each class of markers we have taken into account is well represented and gave us indications on their features and distribution. AFLP markers performed worse than reported in Paglia et al. (1998), with 7.6 polymorphic bands per primer combination; this low average figure is due to a large number of primer combinations from which only few bands were scored, while usually in AFLP analyses primer combinations yielding low numbers of polymorphisms are discarded at the parent screening stage. S-SAP markers showed an average of 13.1 bands per primer combination, a figure slightly higher than reported for barley in Waugh et al. (1997). However, since our method was modified by the introduction of a linear PCR, direct comparisons are only of limited value. Porceddu et al (2002), on the other hand, detected a remarkably higher number of polymorphisms in Medicago sativa, using the method of Waugh et al. (1997). In this case, however, the number of polymorphisms was proportionally higher for AFLP markers, and therefore the differences with our work should be ascribed to species-specific differences or to differences in the selectivity in the choice of scorable markers. Inter-LTRs, with 11.0 bands per primer combination on average, displayed intermediate multiplex ratio (include literature reference to the concept of multiplex ratio, Powell et al. 1996). Interestingly, all primer combinations involving primers with the same orientation relative to the LTR sequence (see Table 1), including those with only one primer, produced poor banding patterns, and were discarded. This is presumably and indication of the rarity, in the genome, of pairs of retrotransposons lying at short distance in "head to head" or "tail to tail" orientation. On the other hand, the amplification of fragments in the combinations we have used cannot be taken directly as an indication of pairs of retrotransposons in the same orientation at short distance, since we do not know what fraction of the bands is due to amplification of single, internally deleted retrotransposons. Perhaps the only strong indication of association between LTR retrotransposons is the amplification of bands from copia and *Alisei* primers together. Indeed, most of the primer pairs that worked well are a combination of two primers from the two elements (Table 1). Microsatellites displayed in some cases a large number of (sometimes dominant) segregating bands, indicating that they often amplify multiple loci, in spite of the fact that they belong to the low copy-number fraction of the genome (Scotti et al., 2002a, 2002b). ESTP markers were successfully transferred from species of the *Pinus* genus (*P. pinaster* and *P. taeda*), with a level of polymorphism comparable to SSRs, providing the base for the alignment of Norway spruce linkage groups with other maps in comparative mapping analyses (Temesgen et al., 2001; Brown et al., 2001; Troggio et al., in prep.). Interestingly, a primer derived from a pine Copia element also worked in Norway spruce, with levels of polymorphism comparable to AFLPs. This

sequence, however, does not match perfectly with any clone in our rather large spruce repeated-sequence DNA database and therefore may not belong to this subset of the genome.

In our map, a relatively large amount of markers remained unlinked (39%) compared to other maps of Norway spruce (Paglia et al., 1998) and of white spruce (Gosselin et al., 2002). The framework maps covered 2316.4 and 1668.6 cM for the female and the male respectively, not far from the figure of the map of Paglia et al. (1998), but with much less framework markers; on the other hand, the other map of Norway spruce (Binelli and Bucci, 1994) covered a much larger map distance (3584 cM) with a number of markers similar to ours. However, the latter map does not appear to have been built with the same criteria, since the authors mention no difference between framework and accessory markers; comparisons between our result and theirs may consequently be misleading. Genome length estimates given here differ widely between the two parental maps, while they are rather insensitive to the LOD threshold used. On average we obtained an estimate of 2821 cM, similar to what reported in Paglia et al.; nevertheless, comparing the three estimates, there seems to be a rough proportionality between map length and genome length estimate, so that each map covers approximately the same proportion of the (estimated) genome length. Estimates obtained by Gosselin et al. (2002) for white spruce do not differ greatly from these. In general, genome length estimates for spruces tend to be higher than for pines (Echt and Nelson 1997). Thirteen linkage groups were obtained when the two parental maps were merged together taking advantage of the markers shared between them, thus approaching the expected number of chromosomes n = 12; otherwise, the number of linkage groups obtained in the two maps separately lies far from the expected, as it occurred in all spruce genetic linkage maps published so far. Since a subset of SSR markers is shared between the present map and the one by Paglia et al. (1998), it was possible to check the two maps for consistency. Three pairs of markers appear here and in the previous map, so that LG3 (this map) corresponds to LG F (Paglia et al., 1998), LG10 to LG D, LG6 to LG S. Other pairs occurring in the previous map appear to be separated in our map, and in one case two loci, mapping on separate groups in Paglia et al. 1998, appear to be tightly linked here. One possible explanation for this feature could be that paralogous, instead of orthologous, loci have been mapped in the two maps, if two different loci amplified by the same primer pair segregate in the two progenies. This possibility cannot be ruled out, considering the complexity of the amplification patterns of some SSR loci and the level of duplication of spruce genome.

Marker distribution showed some peculiarity. Some marker classes (especially ESTPs) were more represented than expected in the set of linked markers, while others (mainly AFLPs) were less. This may mean that ESTPs are found in regions more marker-rich than AFLPs, thus increasing the probability of being linked; or, on the other hand, there may be a relationship between sequence composition and recombination frequency (Kong et al., 2002), so that some sequence classes are preferentially found in regions were genetic distances are shorter (perhaps due to reduced recombination), and therefore linkage has a stronger statistical support. Akhunov et al. (2003) indeed show that regions with low copy-number sequences also show low recombination in wheat; however, this contrasts to some extent with the features of *Arabidopsis thaliana* genome, where regions of low recombination are associated with repeated elements such as retrotransposons. In our case, however, the proportion of linked retroelement-based markers did not deviate from the expectation, and therefore we

cannot say that they map to regions where recombination rate is different from average. However, differences in overall data-scoring methods between marker classes, that may affect their (apparent) linkage, cannot be completely ruled out, in spite of all efforts made to cross-check data produced in different conditions.

Autocorrelation was used to test spatial distribution of marker classes in the map. If we consider that haploid genome size of spruce is 15000 Mbp, and that genome length is approximately 2800 cM, the distance classes chosen here correspond to blocks of approximately 83 Mbp on average (not counting for heterogeneity in recombination rate per unit physical size, on which no report is currently available for *Picea abies*), and therefore the information we get on the architecture of spruce genome is at a relatively coarse scale. Nevertheless, the set of significant autocorrelograms clearly show that retrotransposon-based markers tend to be excluded from the regions where low copy-number markers are found (including AFLPs, which by their design will belong to the hypo-methylated fraction of the genome) and that conversely different types of single-copy markers (AFLP and SSR) tend to be found close to one another. This agrees with what was found for maize (cit.) and A. thaliana (The Arabidopsis Genome Initiative, 2000), although comparisons with the latter species may be misleading due to large differences in genome size and presumably in evolutionary history. These results fit with the view of a genome where large expanses of repeated DNA separate regions that carry low-copy number regions, including genes and SSRs (Morgante et al., 2002). This would also be in agreement with the observation of gene-rich and gene-poor regions in other plant species (Feuillet and Keller, 2002). Clustering of some marker classes was also found, in accordance with the observation of the sub-telomeric distribution of retrotransposons in Norway spruce genome (Cattonaro unpublished). Also microsatellites showed a trend to clustering, although non significant, which agrees with the observations by Elsik and Williams (2001) of physical clustering of SSRs in conifer genome.

Recapitulation of the data presented here suggests that the genome of Norway spruce is not compositionally uniform, and perhaps it has not evolved uniformly, in accordance to what has been observed for angiosperms (Feuillet and Keller, 2002). Observations such as those presented here, and in other papers (e.g. Elsik and Williams, 2001; Hizume et al, 2002), indicate that the characterisation at a finer scale of the genome of conifers may allow the identification of traces of their genome's past evolution.

Acknowledgements.

This project was supported by the EU grant no. BIO4-CT972125 "Anacongen". The authors wish to thank Federica Cattonaro for helping with the retrotransposon sequences and Nicoletta Felice for technical support.

References

Akhunov, E. D., Goodyear, A. W., Geng, S., Qi, L.-L., Echalier, B., Gill, B. S., Miftahudin, Gustafson, J. P., Lazo, G., Chao, S., Anderson, O. D., Linkiewicz, A. M., Dubcovsky, J., La Rota, M., Sorrells, M. E., Zhang, D., Nguyen, H. T., Kalavacharla, V., Hossain, K., Kianian, S. F., Peng, J., Lapitan, N. L.V., Gonzalez-Hernandez, J. L., Anderson, J. A., Choi, D.-W., Close, T. J., Dilbirligi, M., Gill, K. S., Walker-Simmons, M. K., Steber, C., McGuire, P. E., Qualset, C. O., Dvorak, J. (2003) The Organization and Rate of Evolution of Wheat Genomes Are Correlated With Recombination Rates Along Chromosome Arms. Genome Research 13, 1-11.

Binelli, G., Bucci, G. (1994) A genetic linkage map of *Picea abies* Karst., based on RAPD markers, as a tool in population genetics. Theor. Appl. Genet. 88, 283-288.

Brown, G.R., Kadel, E.K. III, Bassoni, D.L., Kiehne, K.L., Temesgen, B., van Buijtenen, J.P., Sewell, M.M., Marshall, K.A., Neale, D.B. (2001) Anchored reference loci in Loblolly pine (*Pinus*

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

taeda L.) for integrating pine genomics. Genetics 159, 799-809.

Chagné, D., Brown, G., Lalanne, C., Madur, D., Pot, D., Neale, D., Plomion, C. (2003) Comparative genome and QTL mapping between maritime and loblolly pines. Molecular Breeding, in press.

Cliff, A. D. ,Ord, J. K. (1973) Spatial autocorrelation. Pion, London.

Costa, P., Pot, D., Dubos, C., Frigerio, J.M., Pionneau, C., Bodenes, C., Bertocchi, E., Cervera, M.-T., Remington, D.L., Plomion, C. (2000) A genetic map of Maritime pine based on AFLP, RAPD and protein markers. Theor. Appl. Genet. 100, 39-48.

Echt, C.S., Nelson, C.D. (1997) Linkage mapping and genome length in eastern white pine (*Pinus strobus* L.). Theor. Appl. Genet. 94, 1031-1037.

Elsik CG, Williams CG (2001) Families of clustered microsatellites in a conifer genome. Mol. Genet. Genomics 265, 535-542.

Feuillet, C., Keller, B. (2002). Comparative genomics in the grass family: molecular characterisation of grass genome structure and evolution. Ann. Botany 89, 3-10.

Gerber, S, Rodolphe, F. (1994) An estimation of the genome length of maritime pine (*Pinus pinaster* Ait.). Theor. Appl. Genet. 88, 289-292.

Gosselin, I., Zhou, Y., Bousquet, J., Isabel, N. (2002) Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD, SCAR, and ESTP markers. Theor. Appl. Genet. 104, 987-997.

Grattapaglia, D., 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.

Hizume, M., Shibata, F., Matsumoto, A., Maruyama, Y., Hayashi, E., Kondo, T., Kondo, K., Zhang, S., Hong, D. (2002) Tandem repeat DNA localising on the proximal DAPI bands of chromosomes in *Larix*, pinaceae. Genome 45, 777-783.

Hulbert, S.H., Ilott, T.W., Legg, E.J., Lincoln, S.E., Lander, E.S., Michelmore, R.W. (1988) Genetic analysis of the fungus, *Bremia lactucae*, using restriction fragment length polymorphisms. Genetics 120, 947-958.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. , Stefansson, K. (2002) A high-resolution recombination map of the human genome. Nature genet. 31, 241-247.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., Newburg, L. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1, 174-181.

Liu, B.H. (1998) Statistical genomics. CRC Press, Boca Raton - New York.

Morgante, M., Hanafey, M., Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature genet. 30, 194-200.

Paglia, G.P., Morgante, M. (1998) PCR-based multiplex DNA fingerprinting techniques for the analysis of conifer genomes. Mol. Breed., 4, 173-177.

Paglia, G.P., Olivieri, A.M., Morgante, M. (1998) Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol. Gen. Genet. 258, 466-478.

Pfeiffer, A., Olivieri, A.M., Morgante, M. (1997) Identification and characterisation of microsatellites in Norway spruce (*Picea abies* K.) Genome 40, 411-419.

Porceddu, A., Albertini, E., Barcaccia, G., Marconi, G., Bertoli, F.B., Veronesi F. (2002) Development of S-SAP markers based on an LTR-like sequence from *Medicago sativa* L. Mol. Genet. Genomics 267, 107-114.

Remington, D.L., Whetten, R.W., Liu, B.-H., O'Malley, D.M. (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor. Appl. Genet. 98, 1279-1292.

Scotti, I., Paglia, G.P., Magni, F., Morgante, M. (2002a) Efficient Development of Dinucleotide microsatellite markers in Norway spruce (*Picea Abies* Karst.) through dot-blot selection. Theor. Appl. Genet. 104, 1035-1041.

Scotti, I., Magni, F., Paglia, G.P., Morgante, M. (2002b) Trinucleotide microsatellites in Norway spruce (*Picea abies*): their features and the development of molecular markers. Theor. Appl. Genet. in press.

Scotti I., Mariani A., Verona V., Candolini A., Cenci C.A., Olivieri A.M. (2002c) AFLP markers and cytotaxonomic analysis reveal hybridisation in the genus *Schœnus* (Cyperaceae). Genome 45, 222-228.

Sokal, R., Oden, N.L. (1978) Spatial autocorrelation methods in biology. 1. Methodology. Biol. J. Linn. Soc. 10, 199-228.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: JOINMAP. Plant J. 3, 793-744.

Temesgen, B., Brown, G.R., Harry, D.E., Kinlaw, C.S., Sewell, M.M., Neale, D.B. (2001). Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.). Theor. Appl. Genet. 102, 664-675.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796-815.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. Nucl. Ac. Res. 21, 4407-4414.

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Thomas, B.B.T., Powell, W. (1997) Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol. Gen. Genet. 253, 687-694.

S.C.D. - U.H.P. NANCY 1
BIBLIOTHÈQUE DES SCIENCES
Rue du Jardin Botanique
54600 VILLERS-LES-NANCY

université HENRI POINCARÉ NANCY 1

Monsieur **CHAGNE David**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY 1

en BIOLOGIE FORESTIERE

VU, APPROUVÉ ET PERMIS D'IMPRIMER N°933

Nancy, le 26 avril 2004

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ, NANCY I
LE PRÉSIDENT

J.P. FINANCE

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél : 03 83 68 20 00 - Fax : 03 83 68 21 00

Développement de marqueurs moléculaires chez le pin maritime (*Pinus pinaster* Ait.) et cartographie génétique comparée des conifères

Résumé

Le pin maritime (*Pinus pinaster* Ait.) est une espèce de première importance écologique et économique pour le Sud-Ouest de l'Europe. En France, il fait l'objet d'un programme d'amélioration génétique depuis les années 1960 et bénéficie depuis une dizaine d'années des apports de la génomique. Afin de construire une carte génétique saturée et de la comparer aux cartes d'autres espèces de conifères, différents types de marqueurs moléculaires ont été développés en utilisant à la fois des techniques de biologie moléculaire et des outils bioinformatiques. Au total, 766 marqueurs AFLP, 54 EST, 9 SNP et 30 SSR polymorphes ont été cartographiés. Les marqueurs AFLP ont permis de saturer rapidement le génome, les marqueurs orthologues (EST et SSR) ont alors permis de comparer les cartes de *Pinus pinaster* et de *Pinus taeda* et d'identifier des groupes de liaison homologues entre ces deux espèces. Cette étude a été complétée par une analyse bibliographique et a permis de montrer que le génome des Pinaceae n'avait pas subi de profond bouleversement depuis la divergence des genres et des espèces il y a 150 à 100 millions d'années. D'un point de vue appliqué, cette étude a permis de vérifier la position de QTL et de gènes candidats liés à la qualité du bois chez les pins.

Mots clefs : *Pinus pinaster* - cartographie génétique comparée - marqueurs moléculaires - EST - SSR - AFLP – SNP.

Molecular markers development in maritime pine (*Pinus pinaster* Ait.) and comparative genome mapping in conifers.

Abstract

Maritime pine (*Pinus pinaster*) is an economically and ecologically important forest tree species in southwestern Europe. A breeding program was started by INRA in the 60's and was recently enhanced by the toolkits of genomics. The objectives of the present work were twofold: 1/ construct a saturated genetic linkage map for maritime pine and 2/ compare this map to that of other conifers. Molecular and *in silico* approaches were used to develop molecular markers. A total of 766 AFLPs, 54 ESTs, 30 SSRs and 9 SNPs were mapped. While AFLP markers made it possible to quickly saturate the linkage map, ESTs and SSRs were used as orthologous markers to align the *P. taeda* and *P. pinaster* maps. Homologous linkage groups between the two species were identified. Together with literature data our results indicate that the genome structure of conifer species did not profoundly change since their ancient divergence time. As an application to this study, QTL and candidate genes for wood quality in pines were verified.

Keywords: *Pinus pinaster* - comparative genome mapping - molecular markers - EST - SSR - AFLP – SNP.