Association genetics in maritime pine (Pinus pinaster Ait.) for growth and wood quality traits

Camille Lepoittevin

N° d'ordre : 3923

THÈSE

PRÉSENTÉE A

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE SCIENCES ET ENVIRONNEMENTS

Par Camille LEPOITTEVIN

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : Ecologie évolutive, fonctionnelle et des communautés

Génétique d'association chez le pin maritime (*Pinus pinaster* Ait.) pour la croissance et les composantes de la qualité du bois

Association genetics in maritime pine (*Pinus pinaster* Ait.) for growth and wood quality traits

Soutenue le : 10 décembre 2009

Devant la commission d'examen formée de :

M. Santiago GONZALEZ-MARTINEZ	Chargé de Recherche, INIA Madrid	Rapporteur
Mme Joëlle RONFORT	Directeur de Recherche, INRA Montpellier	Rapporteur
M. Patrick BABIN	Professeur, Université de Bordeaux I	Examinateur
Mme Catherine BASTIEN	Directeur de Recherche, INRA Orléans	Examinateur
Mme Pauline GARNIER-GERE	Chargée de Recherche, INRA Bordeaux	Co-Directeur de thèse
M. Luc HARVENGT	Directeur Laboratoire de Biotechnologies, FCBA	Co-Directeur de thèse
M. Christophe PLOMION	Directeur de Recherche, INRA Bordeaux	Directeur de thèse

Université Bordeaux 1
Les Sciences et les Technologies au service de l'Homme et de l'environnement

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

~ Sir Ronald Aylmer Fisher

La fin est dans les moyens comme l'arbre est dans la semence.

~Gandhi

Remerciements

Tout d'abord je tiens à remercier Antoine Kremer pour m'avoir reçue au sein de BIOGECO, ces quatre années ont été formidables, et ça va continuer !

Ensuite, merci à Christophe Plomion de m'avoir proposé ce sujet de thèse, le seul sujet qui pouvait me donner envie de poursuivre mes études. Merci pour la confiance et la liberté que tu m'as accordées, pour les nombreux colloques auxquels j'ai pu participer, et pour ton enthousiasme communicatif pour les nouvelles techniques de génotypage ou de phénotypage.

Merci à Luc Harvengt pour avoir financé cette thèse CIFRE, m'avoir accueillie au sein de l'AFOCEL (maintenant FCBA) et pour les bilans réguliers qui ont permis de tenir le cap durant ces quatre ans (encore désolée pour la quatrième année !).

Cette thèse n'aurait pas été ce qu'elle est sans Pauline Garnier-Géré : je te remercie mille fois pour ta disponibilité et ta pédagogie, les heures passées à m'expliquer la théorie de la coalescence, la sélection naturelle, la génétique d'association, pour nos longues discussions sur les modèles démographiques… J'ai beaucoup appris grâce à toi. Nous avons encore beaucoup de choses à écrire toutes les deux !

Merci à mes deux rapporteurs pour avoir accepté de relire ce manuscrit, Joëlle Ronfort et Santiago Gonzalez-Martinez (c'est grâce à toi que je me suis lancée dans les modèles démographiques, merci !), et à Patrick Babin et Catherine Bastien pour avoir accepté de faire partie du jury.

Je tiens aussi à remercier vivement toutes les personnes qui ont contribué à ce travail, de près ou de loin, votre aide a été précieuse :

- L'UE de Pierroton pour l'échantillonnage et la collecte d'aiguilles, en particulier Christophe, Fred, Nico, Laurent, Henri, Hervé et Bernard ;
- L'équipe FCBA de la Station Sud-Ouest pour leur aide à la collecte et au broyage des échantillons de bois, en particulier Jean-Pierre, Thomas, François, mais aussi Pierre Alazard pour la collecte d'aiguilles et ses réponses rapides à mes questions concernant les essais de terrain ;
- Denilson et Audrey de l'équipe IntechFibres du FCBA, pour leur accueil chaleureux à Grenoble, l'initiation à la spectrométrie proche infrarouge, mais aussi pour toutes les données chimiques qu'ils ont produites. Merci aussi Denilson pour ta participation à mes comité de thèse et pour la relecture du chapitre III ;

- Fredo, François, Delphine, Maëlys, les deux Guillaume(s) et Pierre pour leur participation aux manips de broyage et de spectrométrie. Mentions spéciales pour François, qui m'a aussi aidée pour le séquençage des gènes candidats, et Fredo, qui a brillamment géré le labo QB pendant que je terminais ma thèse ;
- Jorge Paiva (Obligada pour ta bonne humeur envahissante !), John MacKay et Frank Bedon pour le choix des gènes candidats ;
- Jean-Marc Frigerio pour le support bioinformatique ;
- Frank Salin pour la manip' de génotypage sur la plateforme de Toulouse ;
- Rémy Petit et Valérie Le Corre pour la relecture de mon premier article ;
- Un grand merci à mon comité de thèse pour ses conseils pertinents, Alain Charcosset, Valérie Le Corre, Brigitte Mangin, Philippe Rozenberg et Leopoldo Sanchez.

Merci aussi à tous les Pierrotonais qui ont fait de cette thèse quatre années inoubliables, notamment Grégoire, pour me supporter tous les jours au bureau (on est p'têt ben un peu Normands tous les deux, sûrement pour ça qu'on s'entend bien… Tes dons en déchiffrage de runes plomionesques m'impressionneront toujours !) ; Céline, Valérie et Patrick pour les pauses de midi qui me changent bien les idées ; Philou, François, Erwan, Fred, Loic, Corine et bien d'autres pour parler d'autre chose que de génétique ; le club de mots-fléchés Pierrotonais pour établir de nouveaux records de vitesse chaque midi ; et bien sûr tous les membres de BIOGECO, visiteurs ou permanents, je suis très heureuse de pouvoir rester parmi vous !

Et parce qu'il n'y a pas que le travail dans la vie, je souhaite aussi remercier le Club de Judo de Cestas (promis maintenant je ne louperai plus un entraînement, Hajime !) et l'Ecole de Musique du Val de Leyre. Et oui, pas besoin de génétique pour exécuter un Tai Otoshi ou jouer du Bach !

Le plus important pour la fin : merci à ma famille et mes amis pour leur soutien et leur affection, Joëlle pour les escapades Québécoises ou shopping, les Forains pour la course annuelle au trophée de Theil Rabier (Hancore !), mes grands-parents et mes parents qui m'ont toujours encouragée à poursuivre mes études, et Chloé pour ses coups de fil quasi-quotidiens (Haha ! Tu ne t'y attendais pas à celle-là !).
Enfin, MERCI à Thomas : tu as vaillamment supporté mes états d'âme statistiques et moléculaires, maintenant apprécions ensemble quelques vacances bien méritées et pensons à nous !

Contents

Introduction

Wood is an endlessly renewable material which uses solar energy for its biosynthesis. It is one of our most important natural resources, and has been exploited for hundreds years as fuel, building material, and source of paper. In the next 20 years, the world's population is projected to increase by more than 30%, followed by an equivalent increase of wood consumption (FAO 2009). In the meantime, we expect a decline in harvesting from natural forests, which is greatly encouraged by environmental policies. This will likely contribute to the emergence of planted forests as the major source of wood supply. The FAO recently modeled the global wood production from planted forests by 2030 (FAO 2006). The outlook was calculated based on predicted changes in planted forest area as well as opportunities for increased productivity from more efficient management practices and genetic improvement. All models indicated an increase of planted forests, with the highest increase for pine forests (FAO 2006).

Pine species play an especially important role in modern plantation forestry worldwide and form a large part of both the annual wood harvest and the immature plantation forests that will provide wood in the future (BURDON 2002; FAO 2006). Indeed, the large number of pine species allows choice for widely varying environments, some species presenting high yields even under unfavorable conditions. Pines are also well suited for reforestation and monocultures, and their wood is easily processed and utilized for a wide variety of end uses such as lumber, pulp and paper or particleboard (PLOMION *et al.* 2007).

Maritime pine (*Pinus pinaster* Ait.) is a Mediterranean species which naturally occurs in southwestern Europe and northwestern Africa, across a fragmented surface of about 4 millions hectares (Figure 1). It is the main tree species used for reforestation in southwestern Europe, and it has been introduced in several countries out of its natural range, namely Chile, South Africa (where it is now considered as an invasive species), Greece, Western Australia (~50,000 ha, and more than 150,000 ha of planed plantations, T. Butcher, personal communication) and Turkey (~57,000 ha, N. Bilir, personal communication). It is the most common softwood in France in terms of both surface area (~1.4 million hectares) and production, with ~8 million cubic meters of timber each year (ALAZARD *et al.* 2005). Aquitaine is the largest maritime pine producing region in France: the Landes forest, which covers about one million hectares, is among the largest and more productive planted conifer forest of Europe (ALAZARD *et al.* 2005). Maritime pine was first planted in this region at the end of the 18th century, to stabilize the coastal dunes and drain the marshes, and was mainly

used for resin production. Since then, the Landes forest area has increased by more than 4 times. When the resin market collapsed in the fifties due to the emergence of oil by-products, Aquitaine foresters chose to reorient their objectives toward timber and pulp production.

Figure 1: Distribution map of maritime pine (from ALIA and MARTIN 2003)

In the sixties, the genetic value of the reproductive material used in plantations was recognized as an important issue for sustainable productivity and quality. Trials comparing different provenances revealed the growth superiority of the Landes provenance and its good adaptation to Aquitaine sandy podzol moorlands ranging from very dry (the dunes along the coast), dry (with *Caluna vulgaris* and *Erica cinerea*), semi-humid (with *Pteridium aquilinum*) and humid (with *Molinia caerulea*) lands. The local resources of the Aquitaine forest thus provided the best genetic basis to build up a breeding population. Two different breeding programs started with the aim to improve both growth and stem straightness, at INRA[*] and AFOCEL[†]. The two institutions joined forces in 1995, in the context of the GIS PMF[‡] which also included the CPFA[§] and the ONF[**]. Today, 100% of plantations are made with improved seedlings from the second generation of seed orchards, and the genetic gain of improved

[*] Institut National de la Recherche Agronomique / National Institute for Agricultural Research

[†] Association Forêt-Cellulose / Association for Forests and Cellulose. AFOCEL has been replaced by the FCBA in June 2007 (http://www.fcba.fr).

[‡] Groupement d'Intérêt Scientifique Pin Maritime du Futur / Maritime Pine Group for the Future.

[§] Centre de Productivité et d'Action Forestière Aquitaine / Centre for Forest Action and Productivity for Aquitaine

[**] Office National des Forêts / National Forestry Office

varieties is about 30% for both volume and straightness (ALAZARD *et al.* 2005). The 3rd generation of seed orchards was planted in 2002 and should start producing improved seeds within the next two years, with expected genetic gains of 40-45% for growth and stem straightness compared to the original sampled material (Annie Raffin, personal communication). The recurrent selection scheme proved successful in improving these two traits, but the introduction of wood quality selection criteria is now an important need for the timber wood market and pulp/paper industries. Several recent observations corroborate this demand: first the gain in productivity has been followed by a reduction in the harvest age (*e.g.* from 60 to 40 years), which has led to a greater proportion of harvested juvenile wood causing a fall of overall intrinsic wood quality (POT *et al.* 2002; BOUFFIER *et al.* 2009). Secondly, preliminary investigations have reported negative correlations between productivity and quality (POT *et al.* 2002).

New breeding objectives are difficult to define, mainly because the relationships between selected traits and end uses are not straightforward. POT *et al.* (2004) defined wood quality as a multi-feature concept that can be studied at different levels, either chemical (*e.g.* composition and content of lignins, cellulose and hemicellulose contents), anatomical (*e.g.* fibre morphology), physical (*e.g.* modulus of elasticity) or technological (*e.g.* wood density, pulp yield). Traditional quantitative genetic studies have shown that wood quality traits are variable and heritable and can therefore be selected and provide significant genetic gains (ZOBEL and VAN BUIJTENEN 1989; NYAKUENGAMA *et al.* 1999; POT *et al.* 2002; RAYMOND 2002; DA SILVA PEREZ *et al.* 2007). The inclusion of wood properties in most breeding programs is however still limited by the time and cost required to screen numerous wood quality traits in a large number of individuals, and the necessity to wait until trees are nearly mature to assess wood properties. Optimisation of genetic gains per unit of time could be achieved using indirect and early selection criteria for target traits measured at adult stage. In this context, molecular markers open the possibility for early selection as it is becoming cheaper and feasible to perform selection prior to trait expression (LANDE and THOMPSON 1990; RIBAUT and HOISINGTON 1998; MOREAU *et al.* 2000; BLANC *et al.* 2008).

From the molecular point of view, variability of wood quality related traits can be defined as the result of variation either at the DNA, transcriptome, proteome or metabolome levels, which can all affect the differentiation of cambial cells also subject to environmental and developmental factors (PAIVA 2006). Indeed, numerous candidate genes for wood properties have been highlighted by substantial studies of maritime pine proteome (GION *et al.* 2005), metabolome (PAIVA 2006) and transcriptome (LE PROVOST *et al.* 2003; PAIVA *et al.* 2008a;

PAIVA *et al.* 2008b) in the past few years. These genes need however to be validated, either by reverse genetic approaches (transgenesis), or by association mapping studies giving further evidence of their involvement in trait variation. This was the goal of the ANR* funded project (GenoQB, Genomics of wood formation and molecular tools for breeding wood quality in maritime pine, GNP05013C) which took place from 2006 to 2009 as a tight partnership between INRA and FCBA (formerly AFOCEL). I was recruited by FCBA as a CIFRE† PhD student in January 2006, and took responsibility for the GenoQB workpackage 2. The main objectives were twofold: i/ study the landscape of nucleotide diversity in previously identified candidate genes for wood formation, and ii/ relate naturally occurring nucleotide polymorphisms with the variability of wood quality related traits measured in experimental designs of the maritime pine breeding program. The strategy developed to achieve these goals is summarized in Figure 2 and discussed below. The samples and markers used in the different chapters are described in Figure 3.

In **Chapter I** – "Disentangling diversity patterns due to demography or natural selection in *Pinus pinaster* transcription factors involved in wood formation" (submitted to Genetics) –, I describe the nucleotide diversity patterns of nine transcription factors putatively involved in wood formation. The rationale behind a focus on transcription factors was that, if substantial sequencing has been done on candidate genes of known molecular and physiological mechanisms (*e.g.* genes of the lignification or cellulose pathways) (POT *et al.* 2005; EVENO *et al.* 2008), little is still known about transcription factors diversity in conifers. Nevertheless, many of these genes have been identified as strong candidates for wood formation mainly from expressional studies (reviewed in DEMURA and FUKUDA 2007). In this chapter, I use coalescent simulations to detect possible departures of transcription factors' nucleotide patterns from neutral molecular evolution, which can be due to either natural selection effects on genes playing an important adaptive role, or past demographic changes in the population considered. The demographic history of Aquitaine maritime pine population is discussed in the light of the results obtained, and a coalescent demographic model is proposed for neutrality testing in this population.

* Agence Nationale de la Recherche / National Research Agency

† Convention Industrielle de Formation par la Recherche

Chapters II and **III** are technical prerequisites to the association mapping study described in Chapter IV. In **Chapter II** – "Developing a SNP genotyping array for *Pinus pinaster*: comparison between *in vitro* and *in silico* detected SNPs" –, the Aquitaine breeding population is genotyped for 384 SNPs* selected either from resequenced candidate genes for wood formation (including the transcription factors examined in Chapter I) and drought stress tolerance, or from contigs of assembled Expressed Sequence Tags (ESTs). Genotyping success rates of both categories are compared, and the interest of *in silico* resources is discussed. The genotyping error rate is also estimated based on Mendelian inconsistencies in a pedigree, and recommendations are made for the use of high-throughput genotyping methods in large and complex genomes of non-model species. In **Chapter III** -"Genetic parameters of growth and wood chemical-properties in *Pinus pinaster*"- phenotypic variation in the Aquitaine breeding population for growth and wood chemical properties is assessed in a progeny and a clonal trial. Among the traits studied, near infrared spectrometry combined with a non-destructive wood sampling method is also used to rapidly assess lignin content in hundreds of trees. Heritabilities and genetic correlations between traits are estimated, and Best Linear Unbiased Predictors (BLUPs) are calculated for each tree as inputs for Chapter IV.

Finally in **Chapter IV** –"Association mapping for growth and wood chemical-properties in the *Pinus pinaster* Aquitaine population"-, an association study is performed using both the molecular markers developed in Chapter I and II, and the phenotypes and BLUPs obtained in Chapter III. First we search for a putative hidden stratification in the breeding population, since population structure is an important bias producing false-positive associations. Then a one-stage and a two-stage approaches accounting for familial relatedness are used to test for significant genotype-phenotype associations. The power of experiments is discussed, and a strategy for upcoming association studies in maritime pine breeding population is proposed.

* Single Nucleotide Polymorphisms

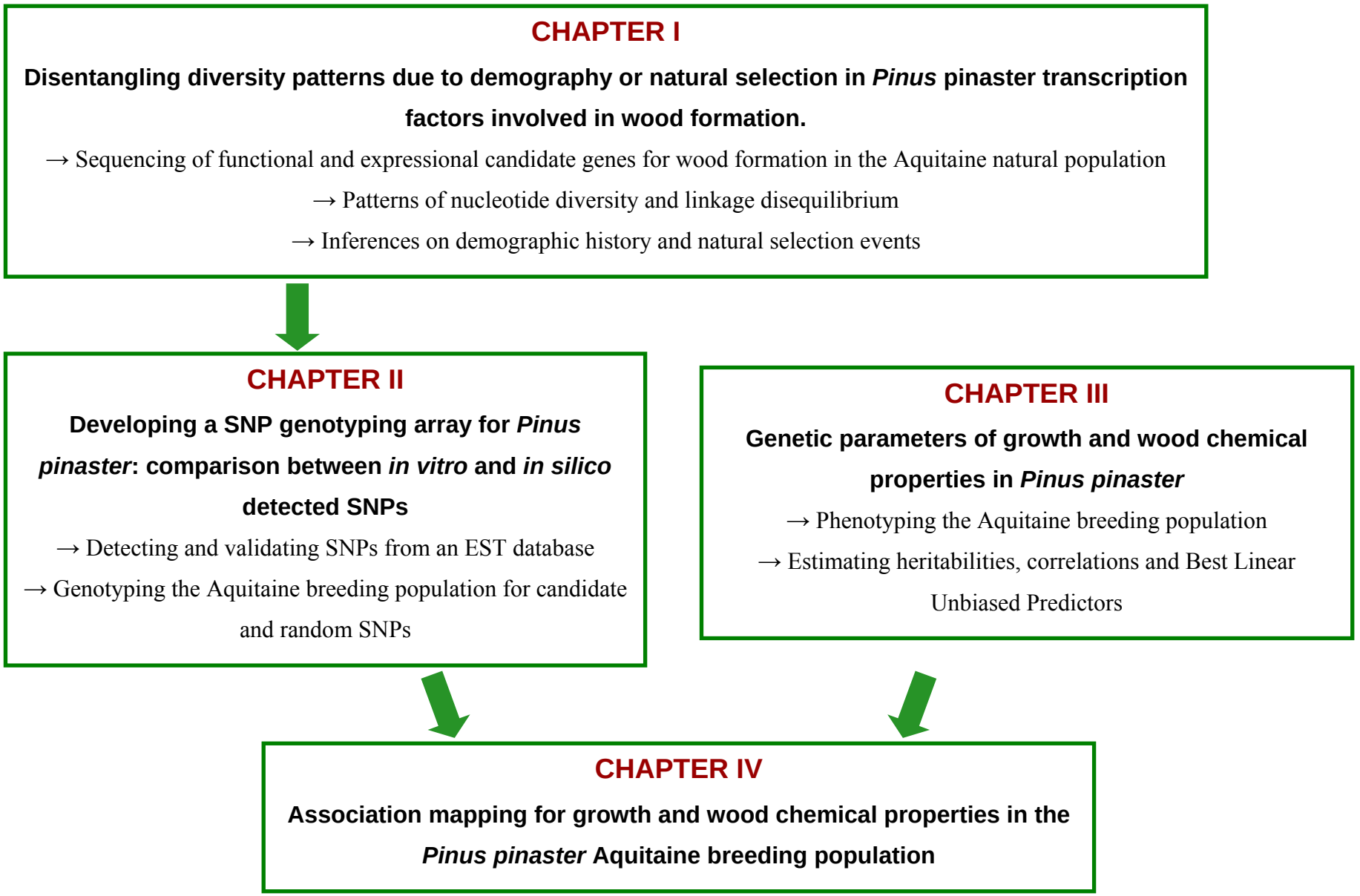


Figure 2: Strategy of the GenoQB project workpackage 2.

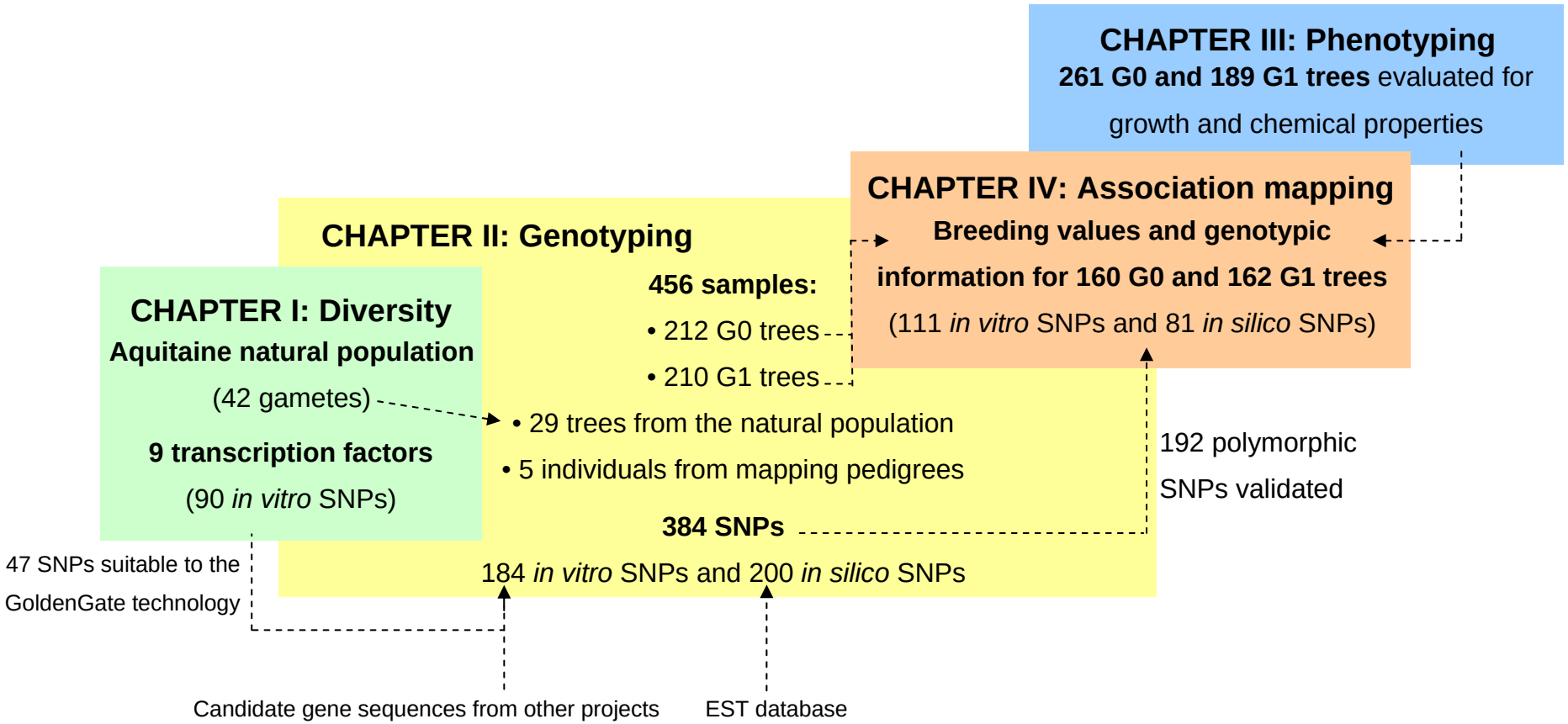


Figure 3: Samples and markers used in the different chapters of the thesis manuscript.

References

ALAZARD, P., D. CANTELOUP, L. CRÉMIÈRE, A. DAUBET, T. LESGOURGUES *et al.*, 2005 Genetic breeding of the maritime pine in Aquitaine: an exemplary success story. By Groupe Pin Maritime du Futur, PG_EDITION, Cestas.

ALIA, R., and S. MARTIN, 2003 EUFORGEN Technical guidelines for genetic conservation and use of Maritime pine (*Pinus pinaster*). By International Genetic Resources Institute, Rome.

BLANC, G., A. CHARCOSSET, J.-B. VEYRIERAS, A. GALLAIS and L. MOREAU, 2008 Marker-assisted selection efficiency in multiple connected populations: a simulation study based on the results of a QTL detection experiment in maize. Euphytica **161:** 71-84.

BOUFFIER, L., A. RAFFIN, P. ROZENBERG, C. MEREDIEU and A. KREMER, 2009 What are the consequences of growth selection on wood density in the French maritime pine breeding programme? Tree Genetics & Genomes **5:** 11-25.

BURDON, R. D., 2002 An introduction to pines, pp. x-xxi in *Pines of silvicultural importance*. CAB International, Wallingford, UK.

DA SILVA PEREZ, D., A. GUILLEMAIN, P. ALAZARD, C. PLOMION, P. ROZENBERG *et al.*, 2007 Improvement of *Pinus pinaster* Ait elite trees selection by combining near infrared spectroscopy and genetic tools. Holzforschung **61:** 611-622.

DEMURA, T., and H. FUKUDA, 2007 Transcriptional regulation in wood formation. Trends in Plant Science **12:** 64-70.

EVENO, E., C. COLLADA, M. A. GUEVARA, V. LEGER, A. SOTO *et al.*, 2008 Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Molecular Biology and Evolution **25:** 417-437.

FAO, 2006 Global planted forests thematic study: results and analysis. By FAO, A. DEL LUNGO, J. BALL and J. CARLE, Rome.

FAO, 2009 *State of the world's forests 2009*. Food and Agriculture Organization of the United Nations, Rome.

GION, J. M., C. LALANNE, G. LE PROVOST, H. FERRY-DUMAZET, J. PAIVA *et al.*, 2005 The proteome of maritime pine wood forming tissue. Proteomics **5:** 3731-3751.

LANDE, R., and R. THOMPSON, 1990 Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics **124:** 743-756.

LE PROVOST, G., J. PAIVA, D. POT, J. BRACH and C. PLOMION, 2003 Seasonal variation in transcript accumulation in wood-forming tissues of maritime pine (*Pinus pinaster* Ait.) with emphasis on a cell wall glycine-rich protein. Planta **217:** 820-830.

MOREAU, L., S. LEMARIE, A. CHARCOSSET and A. GALLAIS, 2000 Economic Efficiency of One Cycle of Marker-Assisted Selection. Crop Science **40:** 329-337.

NYAKUENGAMA, J. G., R. EVANS, C. MATHESON, D. SPENCER and P. VINDEN, 1999 Wood quality and quantitative genetics of *Pinus radiata* D. Don: fibre traits and wood density. Appita Journal **52:** 348-350.

PAIVA, J., 2006 Phenotypic and molecular plasticity of wood forming tissues in Maritime pine (*Pinus pinaster* Ait.), PhD Thesis, University of Bordeaux 1 (France).

PAIVA, J. A. P., M. GARCES, A. ALVES, P. GARNIER-GERE, J. C. RODRIGUES *et al.*, 2008a Molecular and phenotypic profiling from the base to the crown in maritime pine wood-forming tissue. New Phytologist **178:** 283-301.

PAIVA, J. A. P., P. H. GARNIER-GERE, J. C. RODRIGUES, A. ALVES, S. SANTOS *et al.*, 2008b Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. New Phytologist **179:** 1180-1194.

PLOMION, C., D. CHAGNE, D. POT, D. KUMAR, P. L. WILCOX *et al.*, 2007 Pines, pp. 29-92 in *Genome Mapping and Molecular Breeding in Plants: Volume 7 Forest Trees*. Springer Verlag, Berlin.

POT, D., 2004 Déterminisme génétique de la qualité du bois chez le pin maritime : du phénotype aux gènes, PhD Thesis, ENSA de Rennes.

POT, D., G. CHANTRE, P. ROZENBERG, J. C. RODRIGUES, G. L. JONES *et al.*, 2002 Genetic control of pulp and timber properties in maritime pine (*Pinus pinaster* Ait.). Annals of Forest Science **59:** 563-575.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist **167:** 101-112.

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

RIBAUT, J.-M., and D. HOISINGTON, 1998 Marker-assisted selection: new tools and strategies. Trends in Plant Science **3:** 236-239.

ZOBEL, B. J., and J. P. VAN BUIJTENEN, 1989 *Wood variation: its causes and control*. Springer Verlag, Berlin.

Disentangling diversity patterns due to demography or natural selection in *Pinus pinaster* transcription factors involved in wood formation.

Camille Lepoittevin[*†], Pauline Garnier-Géré[*], François Hubert[*], Luc Harvengt[†], Christophe Plomion[*]

[*] INRA, UMR1202 Biodiversité Gènes & Communautés, F-33610 Cestas, FRANCE

[†] FCBA, Laboratoire de Biotechnologies, F-77370 Nangis, FRANCE

Submitted to Genetics[1]

[1] Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. EU482890-EU482903.

Introduction

Unraveling the role of gene regulation on phenotypic adaptation and evolutionary diversification has triggered much research and debates in the scientific community since the mid 70s, when KING and WILSON (1975) showed that the level of protein divergence between humans and chimpanzees was too small to account for their phenotypic differences. They suggested that minor changes in gene regulation could lead to major phenotypic changes. The increasing knowledge on genomes' organization has since confirmed the importance of both regulatory regions and proteins on variation in gene expression, thus reinforcing the belief that these changes can generate most phenotypic evolution in contrast to variation at structural genes (GALANT and CARROLL 2002; CARROLL 2008; WAGNER and LYNCH 2008). Regulatory proteins are coded by transcription factors (TFs) that usually stimulate or inhibit gene expression when binding to their targets, *ie* short, non-coding DNA regulatory regions also called *cis*-regulatory elements. In the genome, the potentially very large number of *trans*-acting TFs and *cis*-regulatory elements combinations finely tune the level, location and chronology of gene activity (WRAY *et al.* 2003). Increasing experimental evidence is showing that this complex network can significantly affect among and within species adaptive phenotypic variation (OLSON and VARKI 2003; WRAY *et al.* 2003; GOMPEL *et al.* 2005; DOEBLEY *et al.* 2006; JEONG *et al.* 2008). Post-transcriptional regulation of gene expression involving microRNAs or small interfering RNAs is also a recognized mechanism in eukaryotes (JONES-RHOADES *et al.* 2006; FILIPOWICZ *et al.* 2008), but the extent to which its variation affects phenotypic changes is still unclear (CHEN and RAJEWSKY 2006), although sequence variants have been recently detected in Arabidopsis (EHRENREICH and PURUGGANAN 2008).

Most TFs can be grouped into families according to their DNA binding domains, which are highly conserved over large evolutionary distances (RIECHMANN *et al.* 2000; WRAY *et al.* 2003). These families vary considerably in size among organisms, with expansion rates higher in plants than in animals, and also higher compared to other gene families in plants (SEOIGHE and GEHRING 2004; SHIU *et al.* 2005). As gene duplication can lead to functional innovations, one of the hypotheses proposed for explaining the parallel expansion of orthologous TFs in Arabidopsis and rice, or TF duplications in *Lotus japonicus*, is an adaptive response to selection pressures caused by environmental stresses that higher plants commonly experience (SHIU *et al.* 2005; YOSHIDA *et al.* 2008, respectively). In the last decade, the debate has evolved from discerning the relative contributions of coding (TFs) and non-coding (*cis*-

elements) regulatory mutations to phenotypic evolution (HOEKSTRA and COYNE 2007), to asking more specific questions about the mechanisms involved (WRAY 2007; WAGNER and LYNCH 2008). The "*cis*-regulatory evolution" theory puts forward that the main driving force behind adaptation would be caused by changes within *cis*-regulatory regions rather than within TFs. The conservation of functions in TFs across organisms would be due to the action of strong purifying selection because of the large number of downstream target genes with which they can interact, and thus possible deleterious pleiotropic effects in case of mutations. In contrast, *cis*-regulatory regions would be preserved from consequences of such negative effects due to their modular architecture (PURUGGANAN 2000; STERN 2000; CARROLL 2005). The gain and loss of TF binding sites by mutational changes would then have a higher chance of being adaptive and trigger phenotypic evolution (WAGNER and LYNCH 2008). The main alternative to the "*cis*-regulatory evolution" theory is that not only *cis*-elements but also TFs can directly contribute to adaptation, with the less conserved TFs domains being as likely to promote morphological evolution than *cis*-elements (HOEKSTRA and COYNE 2007; WAGNER and LYNCH 2008). In a recent genome scan for high diversity regions in Arabidopsis, CORK & PURUGGANAN (2005) found that TFs could harbor significant within-species variation as a consequence of balancing selection. Selection has also been inferred on plants' TFs, for example at the *CAULIFLOWER* gene in *Brassica oleracea*, where TF variation was associated with the differences in inflorescence structure observed among subspecies (PURUGGANAN *et al.* 2000), or at the *teosinte glume architecture* gene, which has triggered a major morphological change during maize domestication (WANG *et al.* 2005).

In the past few years, TFs from different families have been shown to affect wood formation, through their role in lignin metabolism (ROGERS and CAMPBELL 2004; GROOVER and ROBISCHON 2006; BOMAL *et al.* 2008), cell differentiation (OHASHI-ITO *et al.* 2005) or cell fate (ROGERS and CAMPBELL 2004; PAUX *et al.* 2005; GROOVER and ROBISCHON 2006). Wood formation is a fundamental process in terrestrial plants evolutionary history. It has been associated to a number of advantageous features such as tallness and life-long increase in storage capacity and fecundity (PETIT and HAMPE 2006). Wood plays a key role in sap conduction, mechanical support or biotic and abiotic stress resistance (VANCE *et al.* 1980; PLOMION *et al.* 2001). In this work, we focused on nine maritime pine (*Pinus pinaster* Ait.) TFs putatively involved in wood formation. Our objectives were first to assess nucleotide diversity patterns in those TFs, and second to test whether these patterns can be explained by neutral models (integrating demographic history or not), or if they show any signature of selection.

P. pinaster, like other mid-latitude conifers in Europe, is likely to have experienced several population size changes with successive expansions and contractions due to large temperature oscillations during the Quaternary period (BENITO GARZON *et al.* 2007; PETIT *et al.* 2008). These changes probably affected its current diversity patterns at neutral markers (GONZALEZ-MARTINEZ *et al.* 2004; BUCCI *et al.* 2007), but also at adaptive traits (DANJON 1994; CORREIA *et al.* 2008) and at their underlying candidate genes (EVENO *et al.* 2008). Demographic events affect all the genes in a genome, while natural selection will affect only particular loci or genomic regions. Revealing imprints of selection at the molecular level without genome-wide information is thus a difficult task. Moreover, demographic events such as a moderate bottleneck can affect gene genealogies in a way that is very similar to that of a selective sweep due to positive selection (DEPAULIS *et al.* 2003) or to balancing selection at this locus (TAJIMA 1989a). Patterns of nucleotide differentiation in *P. pinaster* populations have been recently associated with the potential action of both directional and balancing selection in candidate loci for wood formation and drought stress (POT *et al.* 2005; EVENO *et al.* 2008). However, detailed analyses of site frequency spectrum accounting for possible effects of both selective and demographic effects have not been performed so far in this species. The F_{st}-based outlier methods (from BEAUMONT and NICHOLS 1996; BEAUMONT and BALDING 2004) that were used in EVENO *et al.* (2008), although recognized as being robust to variations in demographic histories, still do not explicitly model past demographic events. In two other European conifer species, HEUERTZ *et al.* (2006) and PYHAJARVI *et al.* (2007) modeled different bottleneck *scenarii* accounting for the species demographic histories, and showed that ancient bottlenecks followed by expansion could explain the patterns observed at 22 loci in *Picea abies* and 16 loci in *Pinus sylvestris*. *P. pinaster* has a more southern and scattered geographical distribution, which may have led to a smaller effective population size. To meet our second objective, we thus simulated a large range of demographic models likely to encompass maritime pine natural features and history, since they could possibly explain departures from neutral expectations in the targeted genes. Different levels of recombination were also considered, and the different *scenarii* were used as null hypotheses against which to test for potential selection events. We show that the best demographic models for the studied population is a bottleneck that might have occurred during the last ice age (~120,000 to 15,000 years ago), and discuss if this finding is consistent with knowledge on past climatic fluctuations and on *P. pinaster* recolonization history.

Materials and Methods

Population sampling and DNA extraction

Seeds were collected from six *P. pinaster* natural populations distributed along the Atlantic coast of France: namely Petrock, Mimizan, Hourtin, Le Verdon, Olonne-sur-Mer and St-Jean de Monts (see Supplementary Table 1). Megagametophytes (endosperm), the haploid maternal tissues surrounding the embryo in conifer seeds, were harvested from germinated seedlings. We used 42 gametes (seven per population, each from a different individual) of *P. pinaster* for nucleotide diversity analyses, and one genotype (two gametes) of *Pinus taeda* as a reference species to determine ancestral allelic states. Genomic DNA was extracted from megagametophytes using the method described by PLOMION *et al.* (1995), and from the needles of *P. taeda* using Invisorb® Spin Plant Mini Kit (Invitek, Berlin, Germany).

Candidate gene selection and sequencing

Nine transcription factors were chosen for their known or putative implication in conifer wood formation, with emphasis on the R2R3-*MYB* gene family (Table 1): *PtMYB1* and *PtMYB8* from loblolly pine are potential regulators of phenylpropanoid metabolism (PATZLAFF *et al.* 2003b; BOMAL *et al.* 2008), and *PtMYB4* is involved in the control of monolignol biosynthesis (PATZLAFF *et al.* 2003a; NEWMAN *et al.* 2004; BEDON *et al.* 2007); *PtMYB14* has been shown to be associated with the activation of defense mechanisms such as isoprenoid and terpene biosynthesis, but also proved to play a role in the vascular system organization (BEDON 2007); *PgMYB2* is preferentially expressed in secondary xylem and in compression wood-forming tissues in spruce stems (BEDON *et al.* 2007); *PpMYB5* is preferentially expressed in late wood compared to early wood, and *PpSCL1* from the *SCARECROW-LIKE* family has been hypothesized to be implicated in the radial organization of vascular tissue in *P. pinaster* (PAIVA 2006). Finally, spruce *PgHDZip31* from the homeodomain leucine-zipper family and *PgLIM2* from the *LIM*-domain family are expressional candidates for wood formation (J. MacKay, personal communication).

Since the contig available in the *P. pinaster* EST dabatase for *SCL1* did not represent the whole transcript, we resequenced the clone of one of its 3'-end ESTs (cDNA clone PP102H03 from *P. pinaster* differentiating xylem library), and increased its sequence length by 1034 base pairs (bp). The primers for *SCL1* were designed from this sequence. All the other primers were designed either from tentative contigs obtained from the Pinus gene index database (http://compbio.dfci.harvard.edu/tgi/plant.html), or directly from full length gene

sequences using Primer3 (ROZEN and SKALETSKY 2000). The 39 primer pairs are described in Supplementary Table 2.

For PCR reactions, we used 10 ng of genomic DNA, 0.8 U of Taq DNA polymerase (New England Biolabs®, Inc., Ipswich, MA, USA), 1.5 mM of $MgCl_2$ and 0.2 µM of each primer in a total volume of 20 µL. Then, we performed a touchdown 64-55° program (DON *et al.* 1991) in a GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA, USA). The size of each PCR product was determined on a 1% agarose gel using GeneRuler™ 1kb DNA Ladder (Fermentas, Burlington, Ontario, Canada) as a control. All amplification products were purified on MultiScreen® PCR_{96} Filter Plates (Millipore™, Billerica, MA, USA) before being quantified on a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, LLC, Wilmington, DEL, USA). Sequencing reactions were realized with 20 ng of PCR product in a total volume of 10 µl using the BigDye v1.1 terminator cycle sequencing kit (Applied Biosystems). Sequencing reaction products were purified by an ethanol/EDTA/sodium acetate precipitation before loading onto an ABI 3730 automatic sequencer (Applied Biosystems).

For each fragment, sequencing was first performed on a discovery panel of eight to ten gametes randomly chosen from the six sampled populations. This allowed a 0.25 frequency polymorphism to be detected with a probability of at least 90%. The sequencing of the complete panel was not performed when there was no more than one polymorphism per kilobase.

Sequence processing and polymorphic sites detection

Base calling was performed from the raw chromatograms using the KB algorithm of the Sequencing Analysis software v.5.2 (Applied Biosystems). Sequence alignment and nucleotide polymorphism detection (single nucleotide polymorphisms (SNPs), insertion-deletions (INDELs)) were carried out using CodonCode Aligner v.2.0.4 (Codon Code Corporation, Dedham, MA, USA). Nucleotides with phred scores below 20 were considered as missing data. Since all studied TFs belong to multigenic families, we carefully checked that chromatograms did not present any double peaks, which would have been due to paralog amplification. The use of haploid tissue allowed the direct determination of haplotypes (*i.e.* multilocus combinations of polymorphisms) without cloning PCR products.

Diversity, molecular differentiation and recombination rate estimates

Two estimates of nucleotide diversity were computed: Watterson's θ_W (WATTERSON 1975), based on the number of segregating sites, and $\theta\pi$ (NEI 1987), based on the average number of

pairwise differences. Under the Wright-Fisher standard neutral model (WRIGHT 1931), θ_W and $\theta\pi$ estimate the population mutation parameter $\theta=4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per gene per generation. Both nucleotide diversity estimates, and also the haplotype diversity (H_d according to NEI 1987), were computed with the software DnaSP 4.2 (ROZAS *et al.* 2003), considering both SNPs and INDELs (INDELs were only present in non-coding regions, and were not excluded from the mutations considered in the analyses).

We estimated differentiation among the six sampled populations using analyses of molecular variance (AMOVA, EXCOFFIER *et al.* 1992) for each gene on the pairwise difference matrix between haplotypes (*p*-distance), and for each SNP. We also estimated F_{st} parameters between all pairs of populations at the haplotype level, and tested the hypothesis of random distribution of the individuals in pairs by performing exact tests of population differentiation for each gene (RAYMOND and ROUSSET 1995). AMOVA, F_{st} estimates and exact tests were performed using Arlequin v. 3.01 (EXCOFFIER *et al.* 2005).

Recombination allows linked sites to have different coalescent histories, and therefore should be estimated and accounted for when performing neutrality tests using coalescent simulations (STUMPF and MCVEAN 2003; ZENG *et al.* 2007a). DnaSP v. 4.2 was used to estimate the minimum number of recombination events (R_m) using the four-gamete test (HUDSON and KAPLAN 1985). The population recombination parameter ($\rho=4N_er$, where N_e is the effective population size and r is the per gene per generation recombination rate) was also estimated with LDhat v. 2.1 (http://www.stats.ox.ac.uk/~mcvean/LDhat/index.html) using a composite-likelihood estimator (HUDSON 2001), and assuming a constant recombination rate along genes. θ_W was used as input for coalescent likelihood estimations in LDhat. Approximate 95% confidence intervals defined by the 2.5% and 97.5% upper and lower bounds for ρ were set using values observed for ± 2 on each side of the maximum likelihood value (HUDSON 2001).

Extent of linkage disequilibrium

The linkage disequilibrium (LD) between pairs of informative polymorphic sites (minimum allelic frequency > 5% for 40 gametes) within and between genes was computed as the squared allele frequency correlation r^2 (HILL and ROBERTSON 1968) using the Tassel software (BRADBURY *et al.* 2007). This statistic takes into account both recombination and mutation histories and is less sensitive to sample size than other common LD statistics (FLINT-GARCIA *et al.* 2003). Statistical significance of r^2 was computed with a one-tailed Fisher's exact test, applying Bonferroni corrections for multiple testing for each gene. The overall LD decay with

distance between informative sites was estimated by non-linear regression of r^2 within each gene, fitting our data to estimate an "experimental" recombination rate C (C=4Nc, where N is the effective population size, and c is the recombination fraction between sites) with the non linear regression function (nls) in the R software v. 2.6.2 (R Development Core Team 2008). We used the HILL and WEIR (1988) expectation of r^2 for a low level of mutation, adjusted for a sample of size n:

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

Neutrality tests under the standard neutral model

To assess any potential departure of the observed nucleotide diversity patterns from expectations under the standard neutral model (SNM), we first tested observed $\theta\pi$ and H_d values against those expected under neutrality. We then performed four categories of tests: 1) site-frequency-spectrum-based tests applying Tajima's D statistic (TAJIMA 1989b), which represents a standardized difference between $\theta\pi$ and $\theta_{W,}$ and Fay and Wu's H statistic (FAY and WU 2000), which allows to detect an excess or a lack of high-frequency-derived variants, 2) haplotype-based tests, computing the Fu's *Fs* statistic (FU 1997), which reflects the expected number of haplotypes given the observed nucleotide diversity, then the Ewens-Watterson (*EW*) test, which considers the haplotype homozygosity conditional on the number of haplotypes (WATTERSON 1978), and finally the haplotype number K-test (DEPAULIS and VEUILLE 1998) which compares the expected and observed haplotype numbers conditional on the number of segregating sites, 3) LD tests based on associations between sites, performing the Kelly *ZnS* test (KELLY 1997) which averages the squared correlations between alleles at different sites across all pairwise comparisons, 4) the recently described *DH*, *DHEW* and *HEW* compound tests combining site-frequency and/or haplotype-frequency tests, which have been shown in particular cases to be relatively less sensitive than their component tests (Tajima's *D*, Fay & Wu's *H* and Ewens-Watterson *EW*) to recombination, background selection and demography (ZENG *et al.* 2007b).

Tajima's *D*, Fu's *Fs*, Fay & Wu's *H*, Depaulis & Veuille *K*-test and Kelly's *ZnS* statistics were computed in DNAsp v. 4.2. Their significance was tested by comparing their observed value to distributions obtained from 10,000 coalescent simulations using the DNAsp software coalescent tool. Coalescent simulations were conducted successively given $\theta\pi$ and given the number of segregating sites S, and with recombination rates spanning the previously estimated confidence intervals for ρ. *EW* and compound tests were performed thanks to Dr.

Kai Zeng's program (personal communication), which uses θ_W as input for coalescent simulations under the assumption of no recombination. *P. taeda* sequences (Table 1) were used as outgroups for deriving the allelic ancestral states required for Fay & Wu's *H* and associated compound tests (*DH*, *DHEW* and *HEW*).

Simulations of demographic scenarii

We simulated a range of simple bottleneck models with the mlcoalsim v1.42 software (RAMOS-ONSINS and MITCHELL-OLDS 2007), which consisted in (thinking forward) an instantaneous decrease of population size from N_a (ancestral population size) to N_e (present population size) at a time T_a (in units of $4N_e$ generations until present). We did not simulate any expansion after the population contraction (as explained below), so the time at which the bottleneck occurred (T_a) coincides with its duration. Only two parameters are then necessary to define a bottleneck, its intensity N_a/N_e and its timing T_a. These bottlenecks seemed more suitable to the demographic history of maritime pine than classical bottlenecks where the contraction is followed by a population expansion, since in Europe, the last expansion of tree species from their Mediterranean refugia took place after the last glacial maximum, 25,000 to 17,000 years ago (PETIT *et al.* 2003). This means that the number of generations (340 to 1250 taking a generation time of 20-50 years) is probably too short on the evolutionary scale for the expansion to leave traces on diversity patterns for average mutation rates previously estimated for conifer species ($0.70\text{-}1.31 \times 10^{-9}$ per site and per year for silent mutation rate estimates in *Pinus* nuclear genomes, see ANN *et al.* 2007). In the case of the maize domestication bottleneck, TENAILLON *et al.* (2004) showed in their simulations that the size of the population after the final expansion had little effect on the outcome of simulations because of the small number of generations since the domestication event (about 7,000 generations for an effective population size ranging between 500,000 and 5,000,000). In maritime pine, both N_e and the number of generations since the expansion are probably lower. Moreover, preliminary simulation work showed that for the observed levels of diversity on the studied genes, positive values of Tajima's *D* and Fu's *Fs*, and negative values of Fay & Wu's *H* (see results section), a population contraction is a more conservative hypothesis than a contraction followed by an expansion (see Supplementary Methods 1).

The severity of a bottleneck is approximately proportional to the product of its intensity and duration, which corresponds to the time that would lead to the same amount of coalescent events if the size of the population was not reduced (providing that no or few mutations occurred during the bottleneck period) (FAY and WU 1999; GALTIER *et al.* 2000; DEPAULIS *et*

al. 2003). FAY and WU (1999) showed that bottlenecks of different intensities and durations but of similar severity could lead to very similar Tajima's *D* values. We therefore determined a grid of intensities and timings covering a large range of severities: N_a/N_e varying from 1 (the SNM) to 100, and the time at which the bottleneck occurred (T_a) varying between 10^{-2} and 5, which amounted to 80 different *scenarii* tested. The range for T_a values was optimized after a first round of simulations (see Supplementary Methods 2).

For each model, 10,000 replicates of data samples for six loci (*HDZ31*, *LIM2*, *MYB1*, *MYB2*, *MYB14*, and *SCL1*) were simulated using their respective observed numbers of sites, sample sizes and estimated recombination rates. An inherent problem to current estimates, recombination rates in particular, is that they might not be representative of those occurring before and during the *scenarii* departing from the SNM. To account for a possible larger variation in effective recombination rate (ρ) during *scenarii* of population size changes and their potential effects on computed test statistics, we simulated for each gene and each bottleneck model a range of ρ values (per nucleotide) that was successively sampled from the following uniform distributions with their respective lower and upper bounds: U(0,0.0015), U(0.0015,0.0038), U(0.0038,0.0077) and U(0.0077,0.015). These were considered as different levels of recombination tested, from very low, low, medium to high, respectively.

Simulations were conducted conditional on θ, which was adjusted for each gene so that the mean $\theta\pi$ of output simulated data was similar to that observed. Indeed, θ is used to place a number of mutations on the coalescent tree according to a Poisson distribution. Thus, simulating a bottleneck results in a backward increase of θ nearly proportional to the increase of the effective population size (N_e) if the demographic event is not too old ($< 4N_e$ generations in the mean). To obtain a simulated θ (θ_{sim}) close to that observed for our data (θ_{obs}), two rounds of simulations were performed for each model. The first round was conducted conditional on θ_{obs}, and the mean θ of simulated data was used to compute a correction factor $Corr = \theta_{obs} / \theta_{sim}$. The second round was conducted conditional on $\theta_{corr} = Corr \times \theta_{obs}$. We checked that for this second run, the θ_{sim} value obtained was not significantly different from θ_{obs}. For each bottleneck scenario we monitored three summary statistics, Tajima's *D*, Fu's *Fs* and Fay & Wu's *H*. Their *P*-values were given by mlcoalsim outputs by comparing the observed values to the distributions obtained from 10,000 replicates.

Results

Nucleotide diversity

A total of 15,244 bp were sequenced in *P. pinaster*, covering ~9,000 bp of coding regions at nine candidate loci (Table 2) and representing between 74% to 100% of their respective full-length coding sequences. Alignments with the *P. taeda* reference sequences gave amino-acid identities ranging from 96% to 99.4% (Table 1), suggesting that they corresponded to orthologous pairs among *P. taeda* and *P. pinaster*. Three out of the nine genes studied showed no or very low polymorphism in the discovery panel of about 10 megagametophytes, and were not sequenced further (*MYB4* and *MYB8* with one SNP, and *MYB5* monomorphic across a >1300 bp region). Good quality sequence data could not be obtained with *MYB14* for 14 out of the 42 megagametophytes because of non-specific PCR amplification, probably due to coamplification of other family members. Diversity results for this gene could thus be biased and will be taken with caution.

In total, we identified 90 segregating sites (including 12 singletons), which corresponded to one polymorphic site every 169 bp on average. Among these, seven were short INDELs located in non-coding regions while 83 were bi-allelic SNPs; 41 were located in coding regions (comprising 16 non-synonymous mutations), and 49 in non-coding regions. The number of polymorphisms observed per gene varied from 0 to 27, with mean nucleotide diversity ($\theta\pi_{total}$) ranging from 0 (for *MYB5*) to 0.00777 (for *MYB14*), with an average of 0.002 (Table 2). Average $\theta\pi_s$ (silent) was around five times higher than $\theta\pi_{ns}$ (nonsynonymous) (respective ranges [0; 0.01073] and [0; 0.00392]). The averaged estimates of θ_W and $\theta\pi$ were very close whether for all polymorphic sites or for both silent and non-synonymous sites (Table 2). The number of haplotypes per locus was low, varying from one (*MYB5*) to eight (*MYB2*), with an average of 3.9. The mean haplotype diversity H_d was 0.49, ranging from 0 to 0.82 (Table 2).

Population differentiation

To assess the subdivision in the studied population, AMOVA were performed on the six most polymorphic genes (*MYB1*, *MYB2*, *MYB14*, *HDZ31*, *LIM2* and *SCL1*). At the haplotype level, none indicated a significant subdivision between the six sampled subpopulations at a 5% type I error level. On a *per* SNP basis, only five out of 88 sites showed significant differentiation between populations in four genes (*MYB2*, *MYB14*, *HDZ31* and *SCL1*), half of them being low-frequency variants (data not shown).

Table 1. Candidate genes list with corresponding GenBank Accessions.

Gene name[a]	Origin of reference sequence	GenBank reference sequence accession	TIGR tentative contig[b]	*P. pinaster* / *P. taeda* sequences[c] (GenBank Accession)	cds coverage[d]	cds Identity[e]	Amino acid similarity[f]
MYB1	*P. taeda*	AY356372	TC84357	EU482890 / EU482902	100%	98.8%	99.1%
MYB2	*P. taeda*	DQ399060	TC105208	EU482893 / EU482900	87%	98.2%	97.4%
MYB4	*P. taeda*	AY356371	TC95526	EU482894 / none	90%	97.8%	97.9%
MYB5	*P. pinaster*	BX250447	TC84446	EU482896 / none	90%	-	-
MYB8	*P. taeda*	DQ399057	TC106834	EU482895 / none	76%	98.3%	97.8%
MYB14	*P. taeda*	DQ399056	TC81452	EU482897 / EU482901	86%	95.6%	96%
HDZ31	*P. taeda*	DQ657210	TC101825	EU482891 / EU482899	74%	98.9%	99.4%
LIM2	*P. taeda*	BF777520	TC99625	EU482892 / EU482903	94%[g]	98.9%[g]	98.9%[g]
SCL1	*P. pinaster*	BX254698	TC99461	EU482898 / none	unknown[h]	-	-

[a] Gene nomenclature followed BEDON *et al.* 2007. *HDZ31* stands for *class III HDZip 31*, *SCL1* stands for *SCARECROW-LIKE 1*;

[b] http://compbio.dfci.harvard.edu/tgi/plant.html, Pine assembly release 7.0 (July 23, 2008);

[c] Sequences obtained within this study;

[d] Coding sequence coverage; "cds" stands for "coding sequence";

[e] Pairwise nucleic acid coding sequences identity calculated between the *P. taeda* reference sequence and the *P. pinaster* orthologous sequence;

[f] Pairwise amino acid sequences similarity calculated between the *P. taeda* reference sequence and the *P. pinaster* orthologous sequence;

[g] For *LIM2*, the full *P.taeda* cds was not available compared to other genes. The cds coverage, identity and amino acid similarity were based on the TIGR tentative contig expected cds;

[h] The complete cds for this gene is not known.

Table 2. Nucleotide diversity and recombination rates estimates in nine *P. pinaster* transcription factors.

			Nucleotide diversity									Haplotype diversity		Recombination rates	
			Total			Non synonymous sites			Silent sites						
Gene name	Sample size[a]	Length (bp) (coding + non coding)	S[b] (singl.)	θπ[c]	θw[d]	S[b]	θπ[c]	θw[d]	S[b]	θπ[c]	θw[d]	k[e]	Hd[f]	ρ[g]	CI[h]
MYB1	40	1304 (941+363)	8	3.19	1.48	2	1.45	0.67	6	5.65	2.62	2	0.51	0	[0 ; 0.2]
MYB2	38	1819 (1051+768)	13	1.46	1.7	1	0.18	0.29	12	2.52	2.87	8	0.82	0	[0 ; 1.17]
MYB4	10	1315 (803+512)	1 (1)	0.15	0.27	0	0	0	1	0.29	0.51	2	0.2	-	-
MYB5	8	1383 (689+694)	0	0	0	0	0	0	0	0	0	1	0	-	-
MYB8	10	1345 (1212+133)	1	0.26	0.26	0	0	0	1	0.87	0.87	2	0.36	-	-
MYB14	28	847 (481+366)	27 (6)	7.77	8.36	7	3.92	4.91	20	10.73	10.74	5	0.69	0.52	[0.2 ; 1.4]
HDZ31	39	3255 (1773+1482)	20 (3)	2.45	1.45	2	0.11	0.36	18	4.06	2.21	6	0.62	0	[0 ; 0.12]
LIM2	39	2130 (374+1756)	9 (1)	1.59	1.01	0	0	0	9	1.85	1.17	4	0.69	1.54	[0.2 ; 3.72]
SCL1	40	1846 (1846+0)	11 (1)	1.15	1.4	4	0.57	0.66	7	3.18	3.98	5	0.49	0	[0 ; 0.36]
TOTAL		**15244 (9170+6074)**	**90 (12)**			**16**			**74**						
AVERAGE				**2.00**	**1.77**		**0.69**	**0.77**		**3.24**	**2.77**	**3.89**	**0.49**		

[a] Total number of haploid sequences analysed;

[b] Number of segregating sites (number of singletons);

[c] Nucleotide diversity $\theta\pi$ (NEI 1987) per site ($\times 10^{-3}$);

[d] Nucleotide diversity θ_W (WATTERSON 1975) per site ($\times 10^{-3}$);

[e] Number of haplotypes;

[f] Haplotype diversity (NEI 1987);

[g] Per gene population recombination rate estimate;

[h] Population recombination rate confidence interval;

Using pairwise exact tests of population differentiation (RAYMOND and ROUSSET 1995), we only observed significant differentiation for *MYB2* between Petrock and St-Jean de Monts populations (*P*-value < 2%), which are located at both extremes of the sampled area. Therefore a low level of population structure, although undetected for five other genes, may exist. For the following analyses, an homogeneous subset of four populations (Petrock, Mimizan, Hourtin and Olonne-sur-Mer) was then created based on a dendrogram using Ward's clustering method (WARD 1963) on populations pairwise F_{st} (see EVENO *et al.* 2008).

Recombination and LD

The minimum number of recombination events (R_m) for the six polymorphic genes was null. The population recombination rate estimates ρ were either null or very low (Table 2). This is consistent with the high levels of LD detected between informative sites for these genes: 235 among the 525 pairwise LD exact tests were significant after Bonferroni correction. In addition, four out of six TFs showed significant Kelly's *ZnS* values (*MYB1*, *HDZ31*, *LIM2* and *SCL1*). The lowest average r^2 was observed for the *MYB2* gene (r^2=0.24), and the highest for *MYB1* where all sites were in complete LD (r^2=1) (Figure 1); the mean r^2 estimated across all loci was 0.58. Within gene LD was not decaying significantly with distance except for *MYB2*, with r^2 values dropping from 0.47 to 0.17 within 1000 bp. These analyses reveal strong haplotype structures for all genes studied except *MYB2*. We did not detect any significant LD between genes (Supplementary Figure 1).

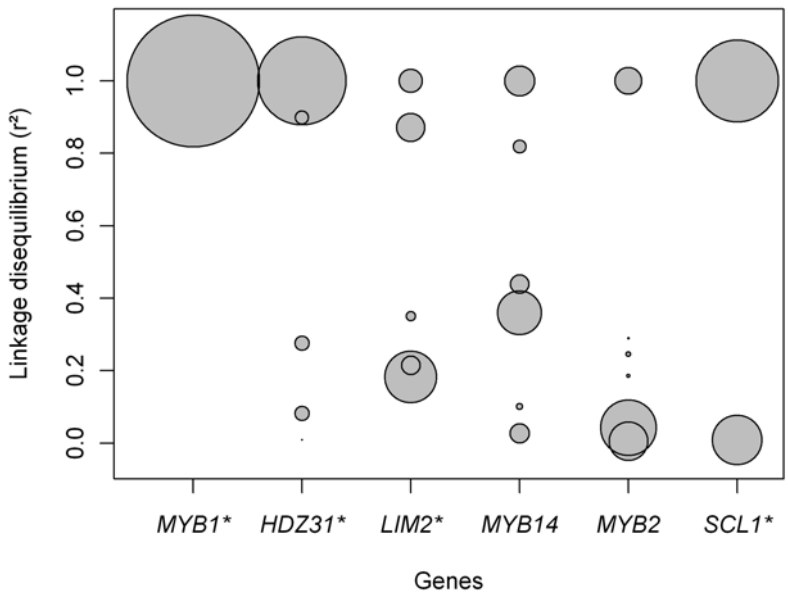


Figure 1. Pairwise linkage disequilibrium values (r^2) between informative sites within six transcription factors. For each gene, the area of each bubble is proportional to the percentage of site pairs showing the r^2 value indicated by the center of the bubble. * Genes showing a significant Kelly's *ZnS* statistic under the SNM (*P*-value ≤ 5%).

Neutrality testing

Eight different statistical tests and three compound tests were performed on the six most polymorphic TFs to detect departures from the SNM. For all tests, significance levels accounting for recombination did not depend on the range of recombination rates tested in ρ confidence interval. In the same way, coalescent simulations performed with $\theta\pi$ and S as conditional parameters gave similar results. Neutrality tests for the *MYB2* gene were computed successively on the total data set and on the homogeneous subset, and gave very similar outputs (Table 3). The only significant test for this gene was Fay & Wu's *H*, with a negative value suggesting an excess of high-frequency derived variants. *SCL1* departed from the SNM only for Kelly's *ZnS* statistics, which indicated higher than expected pairwise correlations between polymorphic sites (*P*-value < 5%). *MYB14* was the most polymorphic of the nine TFs with an average of one mutation every 31 bp ($\theta\pi_{total}$ = 0.0077), but within the range of neutral expectations, as for the *D*, *EW*, *ZnS* and the compound statistics (Table 3). *MYB14* haplotype diversity (H_d) and number of haplotypes were significantly lower than expected (H_d *P*-value < 5%, *K-test P*-value < 1%, Fu's *Fs P*-value < 1% Table 3), and a significant excess of high-frequency derived variants (Fay & Wu's $H < 0$ with a *P*-value < 5%) was detected for this gene. This pattern would need however to be confirmed to exclude any bias due to incomplete sequencing of all gametes.

The *MYB1*, *HDZ31* and *LIM2* loci showed a common pattern of: i/ higher than expected (according to S) $\theta\pi$ values, ii/ highly significant positive values for Tajima's *D*, implying an excess of intermediate-frequency variants, iii/ lower than expected haplotype numbers *K*, consistently with a highly significant positive value for Fu's *Fs*, and iv/ LD patterns incompatible with the neutral expectations according to Kelly's *ZnS* statistics (Table 3). Ewens-Watterson's test was only significant for *MYB1* among all TFs tested. *DH* and *DHEW* compound tests were significant for *MYB1* and *HDZ31*, whereas the *HEW* compound test was never significant. Overall, *MYB1*, *HDZ31* and *LIM2* showed the strongest departures from the SNM, with *MYB1* in particular presenting only 2 haplotypes in similar frequencies (45% and 55%), and thus high significance levels for most tests. In summary, the patterns observed didn't fit well the SNM, and even if significant statistics differed among genes, the common pattern is that of an excess of intermediate frequency variants, or a lack of haplotypes and/or high levels of LD.

Table 3. Neutrality tests in six polymorphic *P. pinaster* transcription factors.

Gene name	Sample size[a]	S[b]	$\theta\pi$[b]	D[c]	H[d]	Hd[b]	EW[e]	Fs[f]	K[g]	ZnS[h]	*DH* *P*-value[i]	*HEW* *P*-value[i]	*DHEW* *P*-value[i]
MYB1	40	8	$3.19^{***(+)}$	$3.31^{***(+)}$	-0.17^{ns}	$0.51^{*(-)}$	$0.50^{*(-)}$	$11.86^{**(+)}$	$2^{***(-)}$	$1^{***(+)}$	0.9996^{***}	0.9388^{ns}	0.9988^{**}
HDZ31	39	20	$2.45^{**(+)}$	$2.27^{**(+)}$	0.40^{ns}	$0.62^{**(-)}$	0.40^{ns}	$8.8^{**(+)}$	$6^{**(-)}$	$0.59^{**(+)}$	0.9791^{*}	0.1744^{ns}	0.9653^{*}
LIM2	39	9	$1.59^{*(+)}$	$1.71^{*(+)}$	-0.49^{ns}	0.69^{ns}	0.33^{ns}	$5.49^{**(+)}$	$4^{**(-)}$	$0.39^{*(+)}$	0.8922^{ns}	0.8405^{ns}	0.8574^{ns}
MYB14	28	27	7.77^{ns}	-0.19^{ns}	$-8.36^{*(-)}$	$0.69^{*(-)}$	0.34^{ns}	$7.08^{**(+)}$	$5^{**(-)}$	0.36^{ns}	0.2555^{ns}	0.3919^{ns}	0.2681^{ns}
MYB2	38	13	1.46^{ns}	-0.45^{ns}	$-5.02^{*(-)}$	0.82^{ns}	0.20^{ns}	0.11^{ns}	8^{ns}	0.24^{ns}	0.1740^{ns}	0.6077^{ns}	0.4974^{ns}
MYB2 (subset)[j]	25	10	1.15^{ns}	-0.69^{ns}	$-5.98^{**(-)}$	0.72^{ns}	0.31^{ns}	0.24^{ns}	6^{ns}	0.46^{ns}	0.1096^{ns}	0.2338^{ns}	0.1593^{ns}
SCL1	40	11	1.15^{ns}	-0.54^{ns}	-1.47^{ns}	0.49^{ns}	0.55^{ns}	1.99^{ns}	5^{ns}	0.51^{*}	0.1765^{ns}	0.1412^{ns}	0.1290^{ns}

[a] Total number of haploid sequences analyzed;

[b] S, $\theta\pi$ and Hd as in Table 1;

[c] Tajima's D (TAJIMA 1989);

[d] Fay & Wu's H (FAY & WU 2000);

[e] Ewens-Watterson statistics (WATTERSON 1978);

[f] Fu's Fs (FU 1997);

[g] Number of haplotypes (K-test, DEPAULIS & VEUILLE 1998);

[h] Kelly's ZnS (KELLY 1997);

[i] *P*-values for *DH*, *HEW* and *DHEW* compound tests (ZENG 2007);

[j] For this gene, statistics computed for a subset of four populations, see Materials and Methods;

Significance of the tests under the SNM: ns *P*-value > 5%, * *P*-value ≤ 5%, ** *P*-value ≤ 1%, *** *P*-value ≤ 0.1%.

(-) and (+): the observed statistic is lower or higher, respectively, than expected under the SNM

Assessment of alternative demographic models

After preliminary trials that helped to decide on the best parameters' range (Supplementary Methods 1 and 2), we simulated 80 different bottleneck *scenarii* for the six TFs, combined with five different recombination rates (the genes' estimates plus four different levels, from very low to high), which allowed a total of 2,400 *scenarii* to be explored. Coalescent simulations involve many stochastic processes, which lead to slight variations among outcomes of a particular scenario even after 10,000 iterations: for example under the SNM (grid squares corresponding to intensity of $N_a/N_e = 1$ for an absence of bottleneck in Figures 2 and 3), the significance level of Fu's *Fs* fluctuates between 1% and 5% depending on the runs (see Figure 2, A and C). For a range of T_a values comprised between 0.01 and 0.1 to 0.3 (depending on the genes), as expected, the significance levels of Tajima's *D*, Fu's *Fs* and Fay & Wu's *H* decrease as the bottleneck intensity increase. This trend is less clear or not observed for *scenarii* with higher T_a values (T_a above 2) for which a major part of the coalescent events already took place during the bottlenecks and thus increasing the intensity does not affect its detection (Figures 2 and 3, see also Supplementary Methods 2 for the range of duration and intensity which allow the highest departures from statistics expected neutral values).

When using either the genes' own estimates or very low recombination rates, a small number of *scenarii* could explain the patterns observed at the six loci: at a 5% threshold, Tajima's *D*, Fu's *Fs* and Fay & Wu's *H* were not significant for bottlenecks of intensity (N_a/N_e) above 10 and of length (T_a) between 0.15 and 2 (Figures 2 and 3). At a 10% threshold, only bottlenecks of intensity above 30 and of length between 0.3 and 1 could explain the statistics observed values. This result is mainly due to *MYB1* Tajima's *D* which was not significant only for this narrow range of timings and intensities (Figure 3 A). For the other genes, the three statistics could be explained by bottlenecks of intensity above 4 and timing between 0.04 and 2 in most cases (Figure 2 and 3, and Supplementary Figure 2). As observed in the SNM case, the patterns differed among genes for the statistics tested: for example, Tajima's *D* was never significant across models for *MYB2*, *MYB14* and *SCL1*, while it was significant for *HDZ31*, *LIM2* and *MYB1* for a range of *scenarii* (Figure 2, Supplementary Figure 2).

When examining the *scenarii* with increased recombination rates, significance levels decreased for many statistics, across genes and models. For medium or high recombination levels, no bottleneck scenario could explain the diversity patterns observed at all genes: *MYB1* Tajima's *D* was significant in all models tested for medium recombination rates, and so did *MYB1* and *HDZ31* Fu's *Fs* for high recombination rates (Figure 3 B and Supplementary

Figure 3). It should be noticed that a large number of tests became significant for some of the bottleneck models even at low recombination rates (which corresponded to a slight increase compared to observed values), for example Fu's *Fs* test for *SCL1* (Figure 3C) or Fay & Wu's *H* for *MYB2* (Supplemental Figure 6). Complete outputs of the 2,400 simulations across different genes are given in Supplementary Figures 2-8.

Discussion

This work reports nucleotide diversity patterns for nine TFs putatively involved in wood formation in a large unstructured *P. pinaster* population of the French Atlantic coast. Three of these genes presented very low levels of nucleotide diversity, which may be a sign of strong purifying selection. The others were more polymorphic, although their nucleotide diversity was generally lower than that observed for genes encoding structural proteins in the same species. Strong departures from the standard neutral model (SNM) were observed for three polymorphic TFs (*HDZ31*, *LIM2* and *MYB1*), based on several statistics, with common patterns of excess of intermediate frequency variants, lack of haplotypes and/or high levels of LD. Such patterns usually result either from a bottleneck affecting the whole genome, from balancing selection affecting specific loci, or from both. We modeled a large range of bottleneck *scenarii* likely to encompass *P. pinaster* natural features and history, in order to explore their impact on the significance of test statistics. For the very low recombination rates estimated under the SNM, we showed that a small number of these *scenarii* could explain the diversity patterns observed for all studied genes. However, no bottleneck was sufficient to explain the Tajima's *D* value observed for one of the genes (*MYB1*) when allowing for even a medium level of recombination rate, suggesting that this gene could have been affected by balancing selection. Below, we first argue in favor of the action of purifying selection in TFs. We then look at the relevance of the bottleneck *scenarii* that best fit our data given the current knowledge of past climatic events and *P. pinaster* history and demography, and compare our findings to results from other conifer species in Europe.

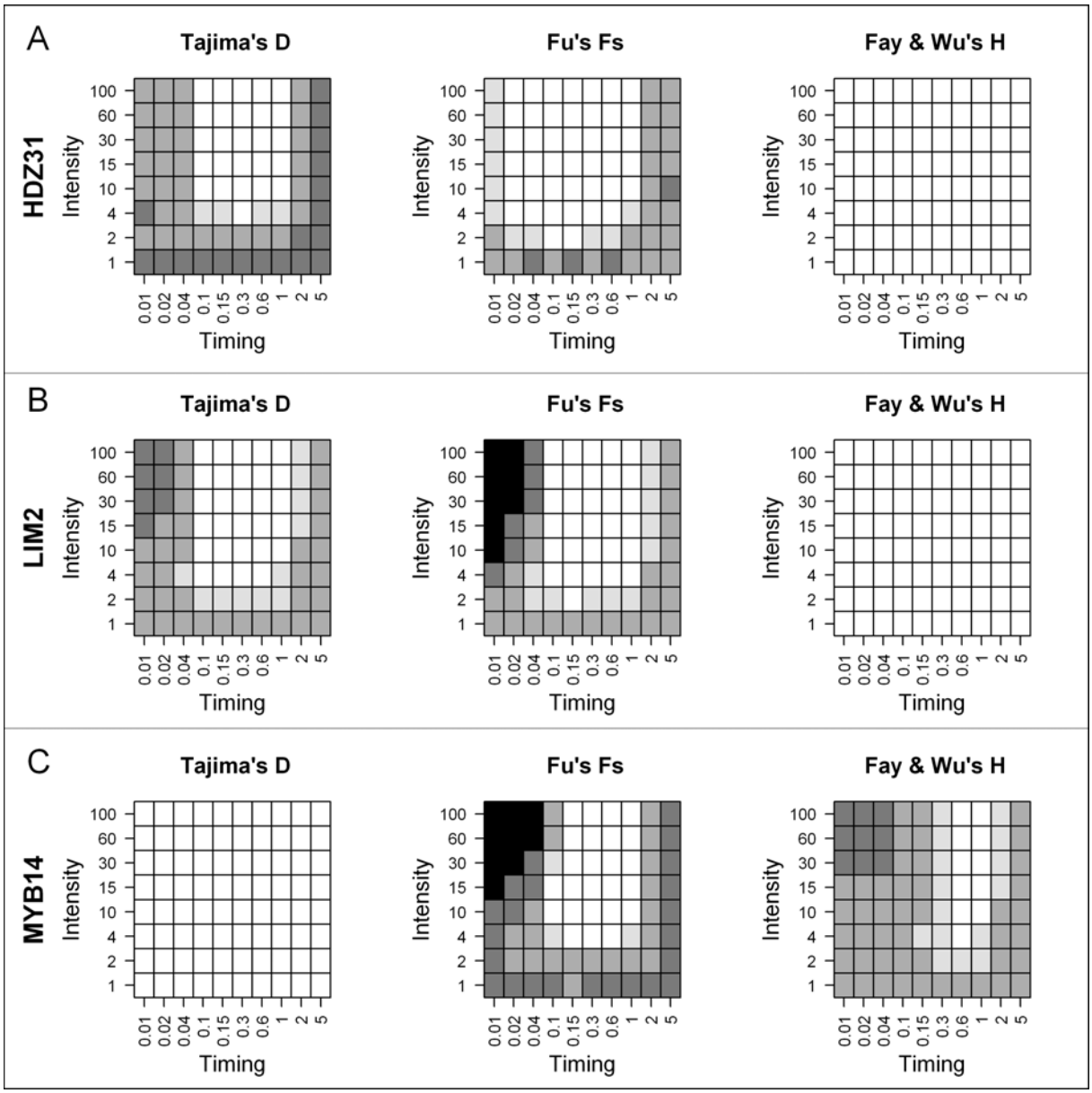


Figure 2. Effects of demography on the significance levels of neutrality tests statistics.
Each grid represents the significance levels of a neutrality test (Tajima's *D*, Fu's *Fs* or Fay & Wu's *H*) under different bottleneck models for A) *HDZ31* B) *LIM2* and C) *MYB14*. From lighter to darker square shading: *P*-value > 0.1; 0.05 < *P*-value ≤ 0.1; 0.01 < *P*-value ≤ 0.05; 0.001 < *P*-value ≤ 0.01; *P*-value ≤ 0.001. Recombination rate estimated for each gene were used in the coalescent simulations. Intensity is the *Na/Ne* ratio, *Na* being the ancestral population size and *Ne* being the present population size. Timing is expressed in units of 4.*Ne* generations until present.

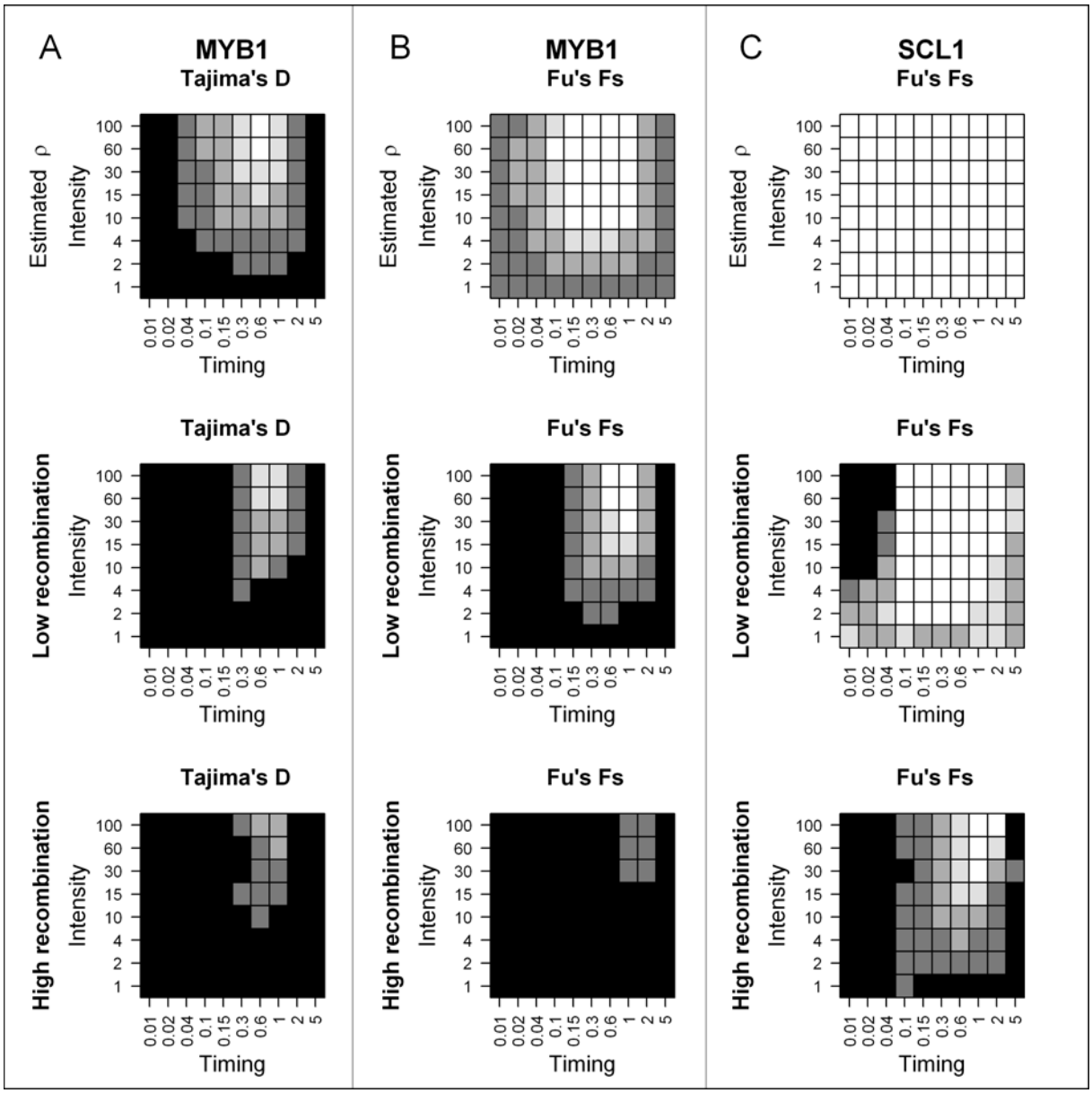


Figure 3. Effects of demography and recombination on the significance levels of neutrality tests statistics.
Each grid represents the significance levels of A) Tajima's *D* for *MYB1*, B) Fu's *Fs* for *MYB1* or C) Fu's *Fs* for *SCL1* under different bottleneck models, and for different recombination rates. The coalescent simulations have been conducted successively with: ρ estimated for each gene, a low level of recombination with ρ sampled in the uniform distribution U(0.0015,0.0038), and a high level of recombination with ρ sampled in the uniform distribution U(0.0077,0.015). Intensity, timing and square shading are as in Figure 2.

Low levels of nucleotide diversity in transcription factors

Apart from the three genes with no or very low observed diversity (*MYB4, MYB5* and *MYB8*), the average $\theta\pi_s$ across the six more polymorphic genes (*HDZ31, LIM2, MYB1, MYB2, MYB14* and *SCL1*) was around 0.006, which is lower although of the same order of magnitude than the mean silent diversity estimate of 0.0075 across a set 25 candidate genes in the same population (compiling data on drought stress resistance genes in the same material after EVENO *et al.* 2008, and our own unpublished data on other genes involved in wood formation, Supplementary Figure 9). Excluding *MYB14* for which diversity could be over-estimated due to incomplete data, the highest diversity value was around 0.0057, with an average of 0.0034 (for the five polymorphic genes left), which corresponds to the lower range of values found to date in *P. pinaster* (see Supplementary Figure 9). The very low diversity observed in *MYB4*, *MYB5* and *MYB8* could be due to a strong effect of purifying selection, which commonly acts against deleterious mutations in regions of functional importance (KIMURA and OHTA 1974). Given that TFs could interact with the promoters of tens to hundreds of genes (WRAY *et al.* 2003), a low diversity at TFs would therefore be caused by the negative impact of their pleiotropic action, consistently with the *cis*-regulatory evolution theory (CARROLL 2005). Additional insights into the functional importance of genes belonging to *MYB* and *HDZip III* gene families come from their ubiquity in all plant lineages and the observation that their basic features in conifers is similar to that observed in angiosperms, suggesting that their expansion predated the gymnosperm – angiosperm split about 300 million years ago (PRIGGE and CLARK 2006; BEDON *et al.* 2007). The action of strong purifying selection has also been proposed as an important process in the evolution of duplicated genes or multigenic families, following a brief period of relaxed selection (LYNCH and CONERY 2000; VAN DE PEER *et al.* 2001; LYNCH and CONERY 2003). These processes were for example inferred from a phylogeny of *Knox-I* TFs across four different conifer species (GUILLET-CLAUDE *et al.* 2004).
Most analyses of plant TFs so far have focused on building phylogenies and searching for homologies among family members across species, but very few have investigated their intra-specific nucleotide diversity. In one study in *Arabidopsis thaliana*, distinct patterns of polymorphisms across gene families were identified (CLARK *et al.* 2007): families involved in transcriptional regulation such as *MYB* and basic helix-loop-helix TFs showed less changes with major effects (premature stop codons, alteration of methionine residues initiation, non-functional splicing sites, deletions or other strong sequences shifts) than other gene families, consistently with the constant action of purifying selection. However, purifying selection may act with more or less strength, as illustrated by the range of values for the ratio of

nonsynonymous and synonymous substitution rates, *Ka/Ks*, across the more polymorphic genes in *P. pinaster*. The mean *Ka/Ks* value is around 0.25 (ranging from 0.17 to 0.59, except for *LIM2* having a *Ka* of 0, using *P. taeda* as the reference species). This is higher than mean values for both control (0.18) and candidate (0.08) loci in *P. abies* (HEUERTZ *et al.* 2006), and than the median value of 0.16 for a set of 77 fragments in *P. tremula* (INGVARSSON 2008). This range of values also corresponds to the high *Ka/Ks* ratio class of genes defined by PALMÉ *et al.* (2008) in *P. sylvestris*, which could have been more affected by positive selection than those with lower *Ka/Ks* ratios. Thus, apart from the obvious action of purifying selection on the least polymorphic genes, the relatively high *Ka/Ks* ratios for the other genes would suggest the potential impact of other processes on their diversity patterns.

Power of neutrality tests

We have shown that the six more polymorphic genes also presented significant departures from neutrality and demographic equilibrium. With no *a priori* knowledge on these forces, different types of neutrality tests were chosen (either based on frequency spectrum or LD and haplotype structure), since they are not sensitive to the same processes, and can thus complement each other (FU 1997; RAMOS-ONSINS and ROZAS 2002; DEPAULIS *et al.* 2003; RAMIREZ-SORIANO *et al.* 2008). For example, Fu's *Fs* and Kelly *ZnS* are among the most powerful statistics to detect population expansions, but they are strongly affected by recombination, which should be carefully examined when making inferences. Expansion signals associated to negative Tajima's *D* are in contrast more robust to recombination rate variation (RAMIREZ-SORIANO *et al.* 2008).

The strong departures from the SNM observed were detected principally using simple statistics like Tajima's *D*, Fu's *Fs* and Kelly's *Zns*, while the compound statistics in our case were not as powerful. First, this could be due to the difference they show in terms of robustness to various processes when considering either ends of their distributions. Indeed, compound statistics have been shown to be relatively robust against the presence of recombination, background selection and demography when trying to detect positive selection (*i.e.* for the left tails of Tajima's *D* and Fay and Wu's *H* null distributions, and for the right tail of Ewens-Watterson's *EW* distribution) (ZENG *et al.* 2007b). However, they may be more sensitive to those processes when trying to detect balancing selection. For example, the *EW* statistic is minimal when all haplotypes are equally represented in a sample (WATTERSON 1978). Since recombination is likely to create new haplotypes at low frequencies and thus increase the statistic value, its lower tail will be sensitive to recombination, whereas the upper tail has been shown to be less affected (ZENG *et al.* 2007a). Second, depending on the pattern

tested (in our case, a general deficit of haplotypes but an excess for those with intermediate frequencies), compound tests power can also vary between both tails of their distribution, which in turn depend on the tails of their component tests. This is the case for *MYB1*, which showed an extreme but only moderately significant value for the *EW* statistic (0.505, whereas the statistic can vary between 0.5 and 0.95 in the case of two haplotypes observed in a sample of 40 sequences). WATTERSON (1978) already noticed that the *EW* test could lack power in such cases. In our sample, the Fay and Wu's *H* test was also not significant for the genes showing the highest departures from the SNM according to other statistics, which probably reduced the power of the compound tests. Actually, Fay and Wu's *H* can only detect incomplete hitchhiking, *i.e.* when recombination allowed few variants to escape the hitchhiking effect (FAY and WU 2000). Our candidates may be either very close to selected variants (since in case of complete hitchhiking, Fay and Wu's *H* is not powerful) or not submitted to selection.

Impact of demographic history on diversity patterns in the Atlantic maritime pine population

We saw that for low recombination rates, the best fitting bottlenecks (*i.e.* those that could explain the patterns observed on the polymorphic genes for different statistics at a *P*-value < 5%) were those reducing the population size by at least 90% (intensity $N_a/N_e > 10$), and with a T_a value (indicating both the start and duration of the bottleneck in our case) between 0.15 and 2, the mode of the range being at 0.6, in units of $4N_e$ generations. Excluding *MYB1*, the lower bound for T_a drops to 0.04 for *scenarii* with intensities above 4 (Figures 2 and 3). If we allow for more variation in recombination rates while still excluding very high recombination rates, the range of *scenarii* that remains compatible with most genes (except *MYB1*) is the same ($0.15 < Ta < 2$ in $4Ne$ generations).

In order to convert those results into absolute time, we can either deduce *Ne* from substitution rate estimates in order to try and date the bottlenecks assuming a range of generation times, or we can propose a dating based on past climatic changes and *P. pinaster* demographic history, and see to what plausible range of *Ne* it corresponds. The consistency of deduced N_e values can be discussed in both cases. Conifer species are generally expected to show large N_e values because of their highly outcrossing mating system, extensive gene flow and wide geographic distributions (NEALE and SAVOLAINEN 2004; BOUILLE and BOUSQUET 2005; SAVOLAINEN and PYHAJARVI 2007). N_e is usually calculated from $\theta=4N_e\mu$ estimates, having first estimated the mutation rate μ by the rate of synonymous substitution, which should be equal under the SNM, and which can be expressed as *Ks/2T*, *Ks* being the divergence rate between two

species, and *T* their divergence time (GRAUR and LI 2000). For the Aquitaine population, with a mean *Ks* (with *P. taeda*) of 0.051 (compiling the same data than above on 25 genes), and the more recent estimate of divergence time between the pine sub-genera of *P. taeda* and *P. pinaster* of 15 millions years (Figure 2 in ANN *et al.* 2007), we get a substitution rate of around 1.7×10^{-9} per site and per year. This estimate is close to the range of estimates given by ANN *et al.* (2007) across *Pinus* species ($0.70\text{-}1.31\times10^{-9}$). Then taking the mean $\theta\pi_s$ given above of 0.0075 per site, and a generation time ranging from 20 to 50 years, we obtain an approximate range of values for N_e of 22,000-55,000 individuals. The derived N_e estimates are in the low range of that published for conifers (Supplementary Table 3), as could be expected in only one population of a species with a fairly scattered distribution. Using such estimates, the onset of the bottleneck should be dated back to around 2.6 million years ago (MYA), for a period between 0.7 MYA and as old as 9 MYA. However, fairly large amplitude temperature shifts (~10°C) have occurred recurrently since the late and middle Pleistocene era (~0.7 – 0.4 MYA), and more recently in the last glacial-interglacial transition in Europe (PETIT *et al.* 1999; PETIT *et al.* 2008). Also, the population size of conifer species did fluctuate during the Würm glaciation (~120,000 to 15,000 years ago), according to phylogeographic studies, palynological records, or climate modelling (CHEDDADI *et al.* 2006; BENITO GARZON *et al.* 2007). Given those historical data, the deduced range for the bottleneck age (9-0.7 MYA) does not seem realistic, as it is not possible that the population studied would have kept a constant size since its onset. The problem may come from the approximate calculation we used for N_e due to successive errors (large confidence intervals for θ and *Ks* when too few loci are available, fossil calibrations impacting on divergence time, generation time), but also to assumptions of neutrality and demographic equilibrium after the divergence of both species, which are clearly unrealistic. A large variance has thus to be considered around this N_e estimate, which could be largely over-estimated, thus giving a range of far too ancient bottlenecks which are not compatible with the assumption of constant N_e after they occurred.

Several elements agree with a smaller N_e in the studied population. Phylogeographic studies employing both organellar and nuclear DNA markers (VENDRAMIN *et al.* 1998; DERORY *et al.* 2002; BURBAN and PETIT 2003; BUCCI *et al.* 2007) indicated that *P. pinaster* is subdivided into at least three gene pools reflecting its survival in different refugia. The Aquitaine population is part of the "Atlantic gene pool", originating from the largest center of re-expansion after the last glaciation period, from south-eastern and central Spain. According to GONZALEZ-MARTINEZ *et al.* (2002), a low adaptive differentiation between central and

northern Spain indicated a rapid spread of the species along this post-glacial migration pathway. This suggests a picture of a population which established quite rapidly, with likely founder effects, although very strong ones would have been avoided in such a tree species with long generations, recurrent gene flow (AUSTERLITZ *et al.* 2000), and repeated introductions. A lower N_e also seems consistent with bioclimatic modeling in Spain (BENITO GARZON *et al.* 2007), the original gene pool from which the Aquitaine population expanded, which suggests for *P. pinaster* a reduction in population size during the last glacial maximum larger than 90% (corresponding to an intensity above 10). This is coherent with the possible range of best fitting *scenarii* (Figures 2 and 3). Given the patterns observed (maintenance of haplotypes), we can see that not all genealogies have coalesced at the start of the bottleneck, which can be qualified as intermediate, in contrast with strong ones where only one lineage survives (DEPAULIS *et al.* 2003). FAY and WU (1999) also defined bottlenecks of severity comprised between 0.25 and 4 as intermediate bottlenecks. The lower severity estimate we observe (using the product S of intensity and duration) in the best fitting *scenarii* is at least 9 (Supplementary Figure 10), with values increasing very fast for older or more intense bottlenecks. Even if the effect of mutation for older events is not accounted for in S, these very high values (> 30 for most bottlenecks) are difficult to interpret. Previous studies modeling bottlenecks in species known to have been submitted to strong decrease in N_e (> 90%) report severities within the range of 4-10 (see TENAILLON *et al.* 2004 for maize; VOIGHT *et al.* 2005 for humans; ASPI *et al.* 2006 for wolfes). All this suggests that, among the *scenarii* that best fit our data (see Supplementary Figure 10), the ones with the lowest severities are already strong bottlenecks, and might be more realistic than older or more intense ones.

A more credible period for a strong bottleneck, *i.e.* since the start of the last glacial period, ~120,000 years ago, would then correspond to 2,400-6,000 maritime pine generations, when taking a larger range for generation time of 20-50 years around our assumed mean of 35 years. If the length of this bottleneck has to be lower than 6,000 generations, it means that N_e for the Aquitaine population would have to be smaller than 10,000 individuals (for T_a=0.15) with a lower bound of 750 individuals (for T_a =2). Assuming a N_e value of a few thousands is however plausible for one single population of a species in which the differentiation among populations at neutral markers can be quite large (up to 30%, EVENO *et al.* 2008). For these reasons, and given the arguments developed above, it seems reasonable to conclude in the occurrence of a bottleneck after the start of the last glacial period, rather than a much older one.

When comparing our results and interpretation with previous studies exploring past demographic *scenarii* in European tree species (HEUERTZ *et al.* 2006; PYHAJARVI *et al.* 2007; INGVARSSON 2008), we noted a few striking differences. First, the *scenarii* that fit their data best include an expansion time that is 2 to 16 times longer than the actual duration of the size reduction, which was quite short in their simulations (fixed at 0.0015, 0.006 and 0.015 × 4*Ne* in HEUERTZ *et al*. 2006, PYHARJARVI *et al.* 2007 and INGVARSSON 2008, respectively). The observed patterns in those three studies are in fact more characteristic of expansion signals following population size reductions, with negative mean Tajima's *D* (-0.88 to -0.16, depending on the region and species, and averaged across many loci). In contrast, we only modelled a population contraction without expansion, since the Holocene expansion seemed too recent to be detected using nuclear SNPs (based on maritime pine history, its observed nucleotide diversity and on preliminary simulations, Supplementary Methods 1). Besides, those long expansion times (starting at the end of population contraction) dated back to periods going from ~388 thousands years ago (INGVARSSON 2008) to 2 MYA (PYHAJARVI *et al.* 2007). This would imply a constant N_e (or *quasi* constant, given the long exponential growth and low rate used, see HEURTZ *et al.* 2006) for a very large number of generations, given their N_e estimates (see Supplementary Table 3), which does not seem consistent with the assumed climate fluctuations since the middle Pleistocene.

Finally, if a bottleneck that took place ~120,000 years ago best accounts for observed genetic patterns, we need to reconcile our results with signals of population growth detected across the whole geographical range of maritime pine (BUCCI *et al.* 2007). Using a method adapted by NAVASCUÉS *et al.* (2006) for microsatellite data, BUCCI *et al.* (2007) found uni-modal mismatch distributions for most maritime pine populations (including the Atlantic one), which are characteristic of past expansion (ROGERS and HARPENDING 1992). Assuming a higher bound for a silent substitution rate for *P. pinaster* of around 10^{-8} for 1 kb (see above), and noting that the mutation rate employed by BUCCI *et al.* (2007) by binary recoding the data of five cpSSRs is around 5×10^{-5} (PROVAN *et al.* 1999) for a similar sequence length, we can observe that the latter is at least 5000 times higher than that used in our simulations. Therefore, the evolutionary scales of both studies are very different but not incompatible. Imprints of expansions in the last 300 generations (since ~10,000 years ago, after the end of the last glacial age) can thus probably be detected with a much higher mutation rate than that characterizing the genes we studied.

Detection of selection signals in transcription factors?

We have seen earlier that purifying selection could be acting on TFs, and that the strong departures from neutrality observed in polymorphic genes could be explained by a credible range of bottleneck *scenarii*. The very low recombination rates based on observed estimates that were used in the simulations may be under-estimated, as both demography and selection can affect them considerably (STUMPF and MCVEAN 2003). Low recombination rates also represent a conservative assumption for many neutrality tests (RAMIREZ-SORIANO *et al.* 2008). When allowing for higher recombination rate values in the different *scenarii*, observed patterns at *MYB1* could no longer be explained by demographic models. The action of balancing selection might therefore be inferred at this locus in addition to that of purifying selection. Indeed, *MYB1* harbors eight mutations but only two haplotypes in similar frequencies, constituting a strong case of balancing selection. Recent functional genomic studies suggest that this gene is a strong candidate for wood formation in conifers: it is most abundantly expressed in *P. taeda* differentiating xylem (PATZLAFF *et al.* 2003b), and its over-expression in *P. glauca* causes a reduced root growth, an enhanced lignin deposition, and the up-regulation of many genes encoding phenylpropanoid enzymes involved in lignin monomer synthesis (BOMAL *et al.* 2008). The genes involved in lignification pathways and wood formation are likely to be important for adaptation, based on evidence of ecotype differences for biotic and abiotic stress responses (NAGY *et al.* 2000; YANG *et al.* 2005). The maintenance of alleles at intermediate frequencies at this locus could therefore result from the heterogeneity of environmental conditions across years, as previously hypothesized for three genes involved in drought stress resistance in the same species (EVENO *et al.* 2008). Selection does not appear necessary to explain the patterns observed for *HDZ31* and *LIM2*, since much higher recombination rates are required for them to show significant tests across all *scenarii*.

In two previous studies looking at sequence diversity patterns in *P. pinaster* (POT *et al.* 2005; EVENO *et al.* 2008), the strongest evidence for possible selection came from analyses based on F_{st} scans along genes. In the POT et al. (2005) study, a strong differentiation between two geographic regions was observed for one gene involved in the cellulose pathway, which could be due to diversifying selection among those regions. Using F_{st}-based outlier methods, EVENO *et al.* (2008) detected both positive (higher F_{st} than expected under neutrality), and negative (lower F_{st}) outlier genes for five out of 11 drought stress candidate genes, suggesting that some of them could not be explained by demography alone. It would be interesting to see whether these two positive outliers show signals of directional selection. These would be opposite, and thus not consistent with those induced by bottleneck *scenarii*.

Conclusion and perspectives

We have shown that strong departures from the standard neutral model in several transcription factors could be explained by bottleneck *scenarii* in one largely sampled population of *P. pinaster*. One exception is the *MYB1* gene, which is thus a good candidate for further functional and association studies. A larger number of loci would be useful to better estimate demographic parameters, for example with methods based on approximate Bayesian computation (BEAUMONT *et al.* 2002), using informative statistics summarizing the data. This is likely to become possible in the near future by combining climate models, fossil records, and the growing genomic resources in this species. It would also be interesting to compare intensity and timing parameters of the best *scenarii* among different populations, and to integrate explicitly the bottleneck *scenarii* in a larger setting of structured populations representative of the *P. pinaster* geographic distribution. This could help in detecting outliers in genome or multilocus scans, as further evidence for the potential role of particular candidates might come from data in populations contrasted for the targeted adaptive traits. Differentiation among genes submitted to selection might also be more easily detectable, since directional selection signals could have been lost within populations but persist among populations. In natural accessions of *A. thaliana*, ZHEN and UNGERER (2008) found that purifying selection on TFs involved in freezing tolerance could be relaxed in southern environments compared to northern ones, with higher levels of non-synonymous nucleotide polymorphisms associated to adaptive phenotypic variation.

Given good candidate genes, association mapping should then be performed to see whether significant associations can be detected at specific alleles or combinations of alleles with wood formation related traits. Identifying the specific target(s) of selection would require to further examine the levels and patterns of nucleotide variation at the surrounding loci (CORK and PURUGGANAN 2005), since we found surprisingly high levels of LD for all the TFs except *MYB2*. This result differs from those obtained in previously published studies in conifer species, which showed a rapid decay of LD (NEALE and SAVOLAINEN 2004; HEUERTZ *et al.* 2006). This difference might be due to the choice of this specific set of genes, or to the use of one larger within-population sample which may have resulted in more accurate LD estimates in comparison with the mixture of gametes from larger geographic areas used in past studies. However, no LD was detected here between TFs and there was no differentiation among sub-populations, considering either haplotypes or SNPs. This agrees with previous findings on the lack of stratification in this environmentally homogenous and flat area, both at the molecular (MARIETTE *et al.* 2001; DERORY *et al.* 2002; RIBEIRO *et al.* 2002; EVENO *et al.* 2008) and

phenotypic (DANJON 1994; DANJON 1995) levels of variation. This suggests the potential efficiency of future candidate gene mapping association studies in this population, as in other conifer species.

Acknowledgements

We thank John MacKay from Université Laval (Québec, CA) for providing data about the implication of conifer transcription factors in wood formation, Kai Zeng for providing his program for compound neutrality tests, and Rémy Petit and Valérie Le Corre for their helpful comments on the manuscript. This research was supported by grants from ANR Genoplante (GenoQB, GNP05013C), from ANR PFTV (BOOST-SNP, 07PFTV002), and from the Aquitaine Region. Sequencing was performed at the Genome and Transcriptome Facility of Bordeaux, FR (http://www.pierroton.inra.fr/biogeco/site_pole_agro/genoseq.html). C. Lepoittevin was supported by CIFRE contract between FCBA and INRA. F. Hubert was funded by the EVOLTREE Network of Excellence (http://www.evoltree.org).

Author's contributions:

CP and LH organized the funding of the study. CL and CP selected the candidate genes; CL and FH performed the sequencing; CL performed the coalescent simulations; CL and PGG analyzed the data and wrote the paper.

References

ANN, W., J. SYRING, D. S. GERNANDT, A. LISTON and R. CRONN, 2007 Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Molecular Biology and Evolution **24:** 90-101.

ASPI, J., E. ROININEN, M. RUOKONEN, I. KOJOLA and C. VILA, 2006 Genetic diversity, population structure, effective population size and demographic history of the Finnish wolf population. Molecular Ecology **15:** 1561-1576.

AUSTERLITZ, F., S. MARIETTE, N. MACHON, P.-H. GOUYON and B. GODELLE, 2000 Effects of colonization processes on genetic diversity: differences between annual plants and tree species. Genetics **154:** 1309-1321.

BEAUMONT, M. A., and D. J. BALDING, 2004 Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology **13:** 969-980.

BEAUMONT, M. A., and R. A. NICHOLS, 1996 Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B-Biological Sciences **263:** 1619-1626.

BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. Genetics **162:** 2025-2035.

BEDON, F., 2007 Structure génique et caractérisation fonctionnelle de facteurs de transcriptions MYB-R2R3 impliqués dans la formation du xylème chez les conifères, PhD Thesis, Université Laval (Québec, Canada).

BEDON, F., J. GRIMA-PETTENATI and J. MACKAY, 2007 Conifer R2R3-MYB transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). BMC Plant Biology **7:** 17.

BENITO GARZON, M., R. SANCHEZ DE DIOS and H. SAINZ OLLERO, 2007 Predictive modelling of tree species distributions on the Iberian Peninsula during the Last Glacial Maximum and Mid-Holocene. Ecography **30:** 120-134.

BOMAL, C., F. BEDON, S. CARON, S. D. MANSFIELD, C. LEVASSEUR *et al.*, 2008 Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. Journal of Experimental Botany **59:** 3925-3939.

BOUILLE, M., and J. BOUSQUET, 2005 Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. American Journal of Botany **92:** 63-73.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BUCCI, G., S. C. GONZALEZ-MARTINEZ, G. LE PROVOST, C. PLOMION, M. M. RIBEIRO *et al.*, 2007 Range-wide phylogeography and gene zones in *Pinus pinaster* Ait. revealed by chloroplast microsatellite markers. Molecular Ecology **16:** 2137-2153.

BURBAN, C., and R. J. PETIT, 2003 Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. Molecular Ecology **12:** 1487-1495.

CARROLL, S. B., 2005 Evolution at two levels: On genes and form. PLoS Biology **3:** 1159-1166.

CARROLL, S. B., 2008 Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134:** 25-36.

CHEDDADI, R., G. G. VENDRAMIN, T. LITT, L. FRANCOIS, M. KAGEYAMA *et al.*, 2006 Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. Global Ecology and Biogeography **15:** 271-282.

CHEN, K., and N. RAJEWSKY, 2006 Natural selection on human microRNA binding sites inferred from SNP data. Nature Genetics **38:** 1452-1456.

CLARK, R. M., G. SCHWEIKERT, C. TOOMAJIAN, S. OSSOWSKI, G. ZELLER *et al.*, 2007 Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science **317:** 338-342.

CORK, J. M., and M. D. PURUGGANAN, 2005 High-diversity genes in the Arabidopsis genome. Genetics **170:** 1897-1911.

CORREIA, I., M. H. ALMEIDA, A. AGUIAR, R. ALIA, T. S. DAVID *et al.*, 2008 Variations in growth, survival and carbon isotope composition (delta C-13) among *Pinus pinaster* populations of different geographic origins. Tree Physiology **28:** 1545-1552.

DANJON, F., 1994 Stand features and height growth in a 36-year-old maritime pine (*Pinus-Pinaster* Ait) provenance test. Silvae Genetica **43:** 52-62.

DANJON, F., 1995 Observed selection effects on height growth, diameter and stem form in maritime pine. Silvae Genetica **44:** 10-19.

DEPAULIS, F., S. MOUSSET and M. VEUILLE, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. Journal of Molecular Evolution **57:** S190-S200.

DEPAULIS, F., and M. VEUILLE, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution **15:** 1788-1790.

DERORY, J., S. MARIETTE, S. C. GONZALEZ-MARTINEZ, D. CHAGNE, D. MADUR *et al.*, 2002 What can nuclear microsatellites tell us about maritime pine genetic resources conservation and provenance certification strategies? Annals of Forest Science **59:** 699-708.

DOEBLEY, J. F., B. S. GAUT and B. D. SMITH, 2006 The molecular genetics of crop domestication. Cell **127:** 1309-1321.

DON, R. H., P. T. COX, B. J. WAINWRIGHT, K. BAKER and J. S. MATTICK, 1991 'Touchdown'PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research **19:** 4008.

EHRENREICH, I. M., and M. D. PURUGGANAN, 2008 Sequence variation of microRNAs and their binding sites in Arabidopsis. Plant Physiology **146:** 1974-1982.

EVENO, E., C. COLLADA, M. A. GUEVARA, V. LEGER, A. SOTO *et al.*, 2008 Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Molecular Biology and Evolution **25:** 417-437.

EXCOFFIER, L., G. LAVAL and S. SCHNEIDER, 2005 Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online **1:** 47-50.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data. Genetics **131:** 479-491.

FAY, J. C., and C. I. WU, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Molecular biology and evolution **16:** 1003.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FILIPOWICZ, W., S. N. BHATTACHARYYA and N. SONENBERG, 2008 Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews Genetics **9:** 102-114.

FLINT-GARCIA, S. A., J. M. THORNSBERRY and E. S. BUCKLER, 2003 Structure of linkage disequilibrium in plants. Annual Review of Plant Biology **54:** 357-374.

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

GALANT, R., and S. B. CARROLL, 2002 Evolution of a transcriptional repression domain in an insect Hox protein. Nature **415:** 910-913.

GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981-987.

GOMPEL, N., B. PRUD'HOMME, P. J. WITTKOPP, V. A. KASSNER and S. B. CARROLL, 2005 Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature **433:** 481-487.

GONZALEZ-MARTINEZ, S. C., R. ALIA and L. GIL, 2002 Population genetic structure in a Mediterranean pine (*Pinus pinaster* Ait.): a comparison of allozyme markers and quantitative traits. Heredity **89:** 199-206.

GONZALEZ-MARTINEZ, S. C., S. MARIETTE, M. M. RIBEIRO, C. BURBAN, A. RAFFIN *et al.*, 2004 Genetic resources in maritime pine (*Pinus pinaster* Aiton): molecular and quantitative measures of genetic variation and differentiation among maternal lineages. Forest Ecology and Management **197:** 103-115.

GRAUR, D., and W. H. LI, 2000 *Fundamentals of molecular evolution*. Sinauer Associates Sunderland, Mass.

GROOVER, A., and M. ROBISCHON, 2006 Developmental mechanisms regulating secondary growth in woody plants. Current Opinion in Plant Biology **9:** 55-58.

GUILLET-CLAUDE, C., N. ISABEL, B. PELGAS and J. BOUSQUET, 2004 The evolutionary implications of knox-I gene duplications in conifers: Correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. Molecular Biology and Evolution **21:** 2232-2245.

HEUERTZ, M., E. DE PAOLI, T. KALLMAN, H. LARSSON, I. JURMAN *et al.*, 2006 Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics **174:** 2095-2105.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics **38:** 226-231.

HILL, W. G., and B. S. WEIR, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theoretical Population Biology **33:** 54.

HOEKSTRA, H. E., and J. A. COYNE, 2007 The locus of evolution: Evo devo and the genetics of adaptation. Evolution **61:** 995-1016.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147-164.

INGVARSSON, P. K., 2008 Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics **180:** 329-340.

JEONG, S., M. REBEIZ, P. ANDOLFATTO, T. WERNER, J. TRUE *et al.*, 2008 The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell **132:** 783-793.

JONES-RHOADES, M. W., D. P. BARTEL and B. BARTEL, 2006 MicroRNAs and their regulatory roles in plants. Annual Review of Plant Biology **57:** 19-53.

KELLY, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197-1206.

KIMURA, M., and T. OHTA, 1974 On some principles governing molecular evolution. Proceedings of the National Academy of Sciences of the United States of America **71:** 2848-2852.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

LYNCH, M., and J. S. CONERY, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

LYNCH, M., and J. S. CONERY, 2003 The evolutionary demography of duplicate genes. Journal of Structural and Functional Genomics **3:** 35:44.

MARIETTE, S., D. CHAGNE, C. LEZIER, P. PASTUSZKA, A. BAFFIN *et al.*, 2001 Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. Heredity **86:** 469-479.

NAGY, N. E., V. R. FRANCESCHI, H. SOLHEIM, T. KREKLING and E. CHRISTIANSEN, 2000 Wound-induced traumatic resin duct development in stems of Norway spruce (*Pinaceae*): anatomy and cytochemical traits. American Journal of Botany **87:** 302-313.

NAVASCUÉS, M., Z. VAXEVANIDOU, S. C. GONZALEZ-MARTINEZ, J. CLIMENT, L. GIL *et al.*, 2006 Chloroplast microsatellites reveal colonization and metapopulation dynamics in the Canary Island pine. Molecular Ecology **15:** 2691-2698.

NEALE, D. B., and O. SAVOLAINEN, 2004 Association genetics of complex traits in conifers. Trends in Plant Science **9:** 325-330.

NEI, M., 1987 *Molecular evolutionary genetics*. Columbia University Press.

NEWMAN, L. J., D. E. PERAZZA, L. JUDA and M. M. CAMPBELL, 2004 Involvement of the R2R3-MYB, AtMYB61, in the ectopic lignification and dark-photomorphogenic components of the det3 mutant phenotype. The Plant Journal **37:** 239-250.

OHASHI-ITO, K., M. KUBO, T. DEMURA and H. FUKUDA, 2005 Class III homeodomain leucine-zipper proteins regulate xylem cell differentiation. Plant and Cell Physiology **46:** 1646-1656.

OLSON, M. V., and A. VARKI, 2003 Sequencing the chimpanzee genome: Insights into human evolution and disease. Nature Reviews Genetics **4:** 20-28.

PAIVA, J., 2006 Phenotypic and molecular plasticity of wood forming tissues in Maritime pine (*Pinus pinaster* Ait.), PhD Thesis, University of Bordeaux 1 (France).

PALME, A. E., M. WRIGHT and O. SAVOLAINEN, 2008 Patterns of Divergence among Conifer ESTs and Polymorphism in *Pinus sylvestris* Identify Putative Selective Sweeps. Molecular Biology and Evolution **25:** 2567-2577.

PATZLAFF, A., S. MCINNIS, A. COURTENAY, C. SURMAN, L. J. NEWMAN *et al.*, 2003a Characterisation of a pine MYB that regulates lignification. The Plant Journal **36:** 743-754.

PATZLAFF, A., L. J. NEWMAN, C. DUBOS, R. WHETTEN, C. SMITH *et al.*, 2003b Characterisation of PtMYB1, an R2R3-MYB from pine xylem. Plant Molecular Biology **53:** 597-608.

PAUX, E., V. CAROCHA, C. MARQUES, A. M. DE SOUSA, N. BORRALHO *et al.*, 2005 Transcript profiling of Eucalyptus xylem genes during tension wood formation. New Phytologist **167:** 89-100.

PETIT, J. R., J. JOUZEL, D. RAYNAUD, N. I. BARKOV, J. M. BARNOLA *et al.*, 1999 Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature **399:** 429-436.

PETIT, R. J., I. AGUINAGALDE, J. L. DE BEAULIEU, C. BITTKAU, S. BREWER *et al.*, 2003 Glacial refugia: Hotspots but not melting pots of genetic diversity. Science **300:** 1563-1565.

PETIT, R. J., and A. HAMPE, 2006 Some evolutionary consequences of being a tree. Annual Review of Ecology, Evolution, and Systematics **37:** 187-214.

PETIT, R. J., F. S. HU and C. W. DICK, 2008 Forests of the past: A window to future changes. Science **320:** 1450-1452.

PLOMION, C., G. LEPROVOST and A. STOKES, 2001 Wood formation in trees. Plant Physiology **127:** 1513-1523.

PLOMION, C., D. M. O'MALLEY and C. E. DUREL, 1995 Genomic analysis in maritime pine (*Pinus pinaster*). Comparison of two RAPD maps using selfed and open-pollinated seeds of the same individual. Theoretical and Applied Genetics **90:** 1028-1034.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist **167:** 101-112.

PRIGGE, M. J., and S. E. CLARK, 2006 Evolution of the class III HD-Zip gene family in land plants. Evolution & Development **8:** 350-361.

PROVAN, J., N. SORANZO, N. J. WILSON, D. B. GOLDSTEIN and W. POWELL, 1999 A low mutation rate for chloroplast microsatellites. Genetics **153:** 943-947.

PURUGGANAN, M. D., 2000 The molecular population genetics of regulatory genes. Molecular Ecology **9:** 1451-1461.

PURUGGANAN, M. D., A. L. BOYLES and J. I. SUDDITH, 2000 Variation and selection at the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. Genetics **155:** 855-862.

PYHAJARVI, T., M. R. GARCIA-GIL, T. KNURR, M. MIKKONEN, W. WACHOWIAK *et al.*, 2007 Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics **177:** 1713-1724.

RAMIREZ-SORIANO, A., S. E. RAMOS-ONSINS, J. ROZAS, F. CALAFELL and A. NAVARRO, 2008 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179:** 555-567.

RAMOS-ONSINS, S. E., and T. MITCHELL-OLDS, 2007 Mlcoalsim: Multilocus coalescent simulations. Evolutionary Bioinformatics **3:** 41:44.

RAMOS-ONSINS, S. E., and J. ROZAS, 2002 Statistical properties of new neutrality tests against population growth. Molecular Biology and Evolution **19:** 2092-2100.

RAYMOND, M., and F. ROUSSET, 1995 An exact test for population differentiation. Evolution **49:** 1280-1283.

RIBEIRO, M. M., S. MARIETTE, G. G. VENDRAMIN, A. E. SZMIDT, C. PLOMION *et al.*, 2002 Comparison of genetic diversity estimates within and among populations of maritime pine using chloroplast simple-sequence repeat and amplified fragment length polymorphism data. Molecular Ecology **11:** 869-877.

RIECHMANN, J. L., J. HEARD, G. MARTIN, L. REUBER, C. Z. JIANG *et al.*, 2000 Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. Science **290:** 2105-2110.

ROGERS, A. R., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Molecular biology and evolution **9:** 552.

ROGERS, L. A., and M. M. CAMPBELL, 2004 The genetic control of lignin deposition during plant growth and development. New Phytologist **164:** 17-30.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology **132:** 365-386.

SAVOLAINEN, O., and T. PYHAJARVI, 2007 Genomic diversity in forest trees. Current Opinion in Plant Biology **10:** 162-167.

SEOIGHE, C., and C. GEHRING, 2004 Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends in Genetics **20:** 461-464.

SHIU, S. H., M. C. SHIH and W. H. LI, 2005 Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiology **139:** 18-26.

STERN, D. L., 2000 Perspective: Evolutionary developmental biology and the problem of variation. Evolution **54:** 1079-1091.

STUMPF, M. P. H., and G. A. T. MCVEAN, 2003 Estimating recombination rates from population-genetic data. Nature Reviews Genetics **4:** 959-968.

TAJIMA, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597-601.

TAJIMA, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TENAILLON, M. I., J. U'REN, O. TENAILLON and B. S. GAUT, 2004 Selection versus demography: A multilocus investigation of the domestication process in maize. Molecular Biology and Evolution **21:** 1214-1225.

VAN DE PEER, Y., J. S. TAYLOR, I. BRAASCH and A. MEYER, 2001 The ghost of selection past: Rates of evolution and functional divergence of anciently duplicated genes. Journal of Molecular Evolution **53:** 436-446.

VANCE, C. P., T. K. KIRK and R. T. SHERWOOD, 1980 Lignification as a mechanism of disease resistance. Annual Review of Phytopathology **18:** 259-288.

VENDRAMIN, G. G., M. ANZIDEI, A. MADAGHIELE and G. BUCCI, 1998 Distribution of genetic diversity in *Pinus pinaster* Ait. as revealed by chloroplast microsatellites. Theoretical and Applied Genetics **97:** 456-463.

VOIGHT, B. F., A. M. ADAMS, L. A. FRISSE, Y. D. QIAN, R. R. HUDSON *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences of the United States of America **102:** 18508-18513.

WAGNER, G. P., and V. J. LYNCH, 2008 The gene regulatory logic of transcription factor evolution. Trends in Ecology & Evolution **23:** 377-385.

WANG, H., T. NUSSBAUM-WAGLER, B. L. LI, Q. ZHAO, Y. VIGOUROUX *et al.*, 2005 The origin of the naked grains of maize. Nature **436:** 714-719.

WARD, J. H., 1963 Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58:** 236-244.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7:** 256-276.

WATTERSON, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405-417.

WRAY, G. A., 2007 The evolutionary significance of *cis*-regulatory mutations. Nature Reviews Genetics **8:** 206-216.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution **20:** 1377-1419.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97-159.

YANG, S.-H., H. WANG, P. SATHYAN, C. STASOLLA and C. A. LOOPSTRA, 2005 Real-time RT-PCR analysis of loblolly pine (*Pinus taeda*) arabinogalactan-protein and arabinogalactan-protein-like genes. Physiologia Plantarum **124:** 91-106.

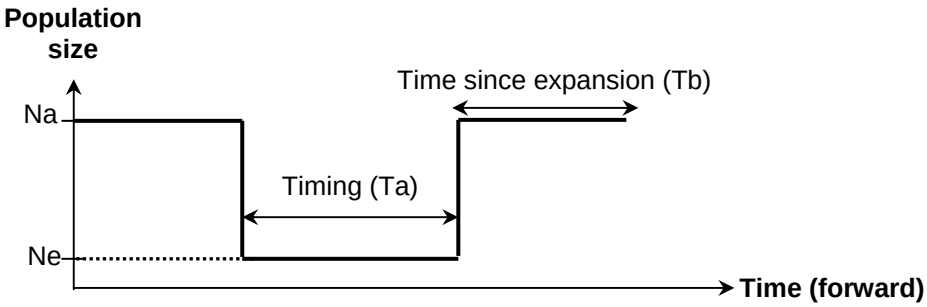
YOSHIDA, K., R. IWASAKA, T. KANEKO, S. SATO, S. TABATA *et al.*, 2008 Functional differentiation of *Lotus japonicus* TT2s, R2R3-MYB transcription factors comprising a multigene family. Plant and Cell Physiology **49:** 157-169.

ZENG, K., S. H. MANO, S. H. SHI and C. I. WU, 2007a Comparisons of site- and haplotype-frequency methods for detecting positive selection. Molecular Biology and Evolution **24:** 1562-1574.

ZENG, K., S. SHI and C. I. WUT, 2007b Compound tests for the detection of hitchhiking under positive selection. Molecular Biology and Evolution **24:** 1898-1908.

ZHEN, Y., and M. C. UNGERER, 2008 Relaxed selection on the CBF/DREB1 regulatory genes and reduced freezing tolerance in the southern range of *Arabidopsis thaliana*. Molecular Biology and Evolution **25:** 2547.

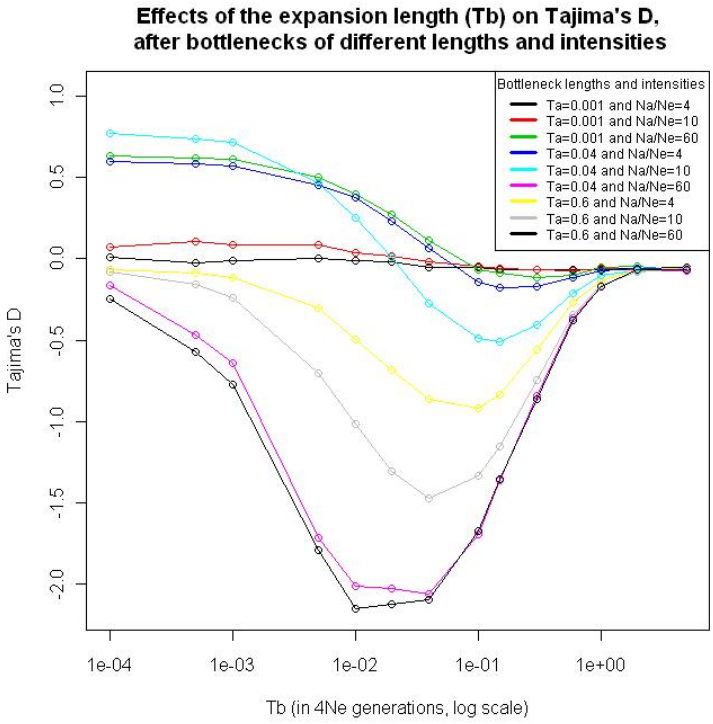
Supplementary Methods 1. Effects of a population expansion after bottlenecks of different intensities and durations on Tajima's *D*, Fu's *Fs* and Fay & Wu's *H*.

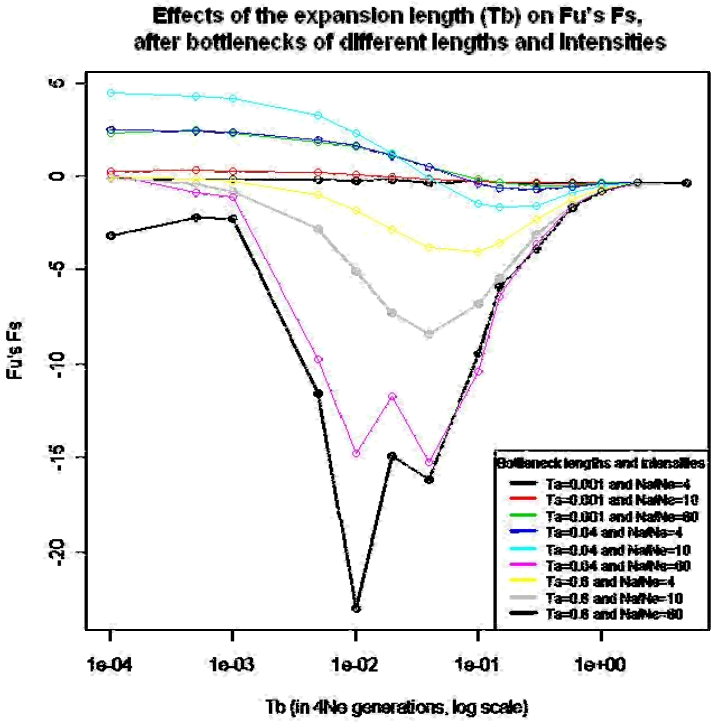


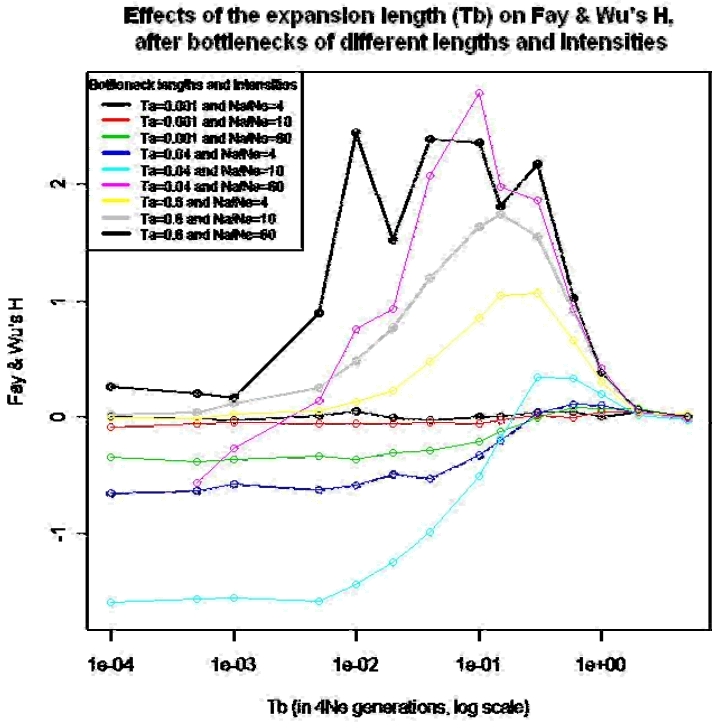
We tested the effects of an expansion of duration *Tb* (in 4.*Ne* generations) after bottlenecks of different intensities (*Na/Ne*) and durations (*Ta*, in 4.*Ne* generations) on Tajima's *D*, Fu's *Fs* and Fay & Wu's *H.*

Each point of the curves represents the mean of a statistic (Tajima's *D,* Fu's *Fs* or Fay & Wu's *H*) over 10,000 simulated samples of 35 sequences of length 1 kb. The recombination rate was sampled in a uniform distribution U(0,0.002). $\theta\pi$ was adjusted as explained in the materials and methods section to be equal at 3 in the simulated samples.

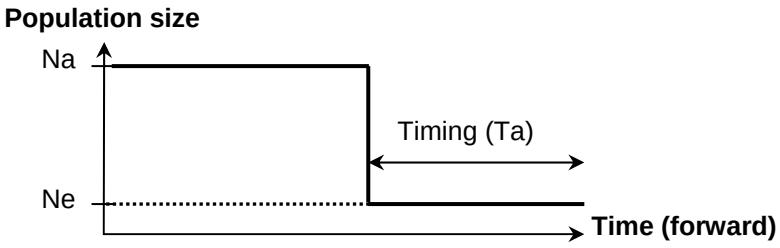
One can observe that, for expansions of duration above 0.0001 x 4.*Ne* generations, Tajima's *D* and Fu's *Fs* become more negative, and Fay & Wu's *H* becomes more positive. A model without expansion is then more conservative than models with expansion for positive values of Tajima's *D* and Fu's *Fs* and for negative values of Fay & Wu's *H*.



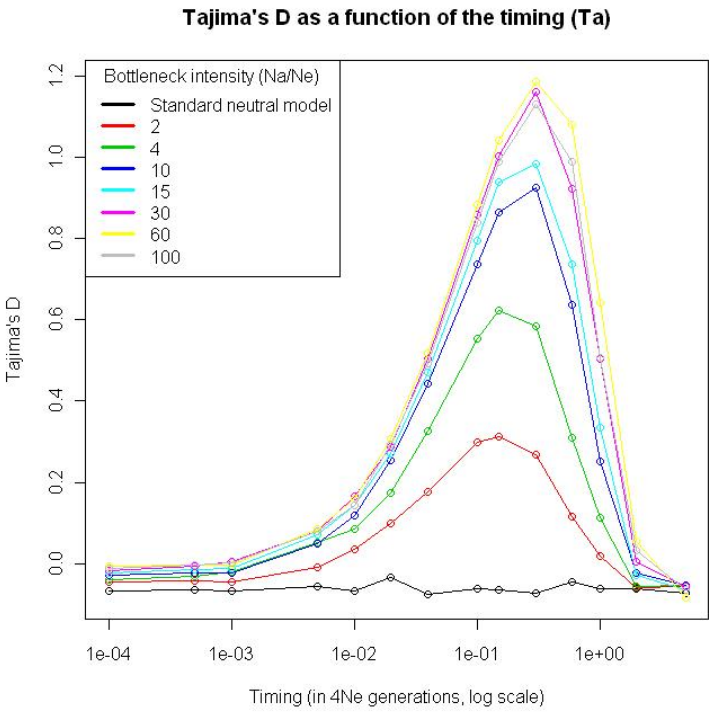


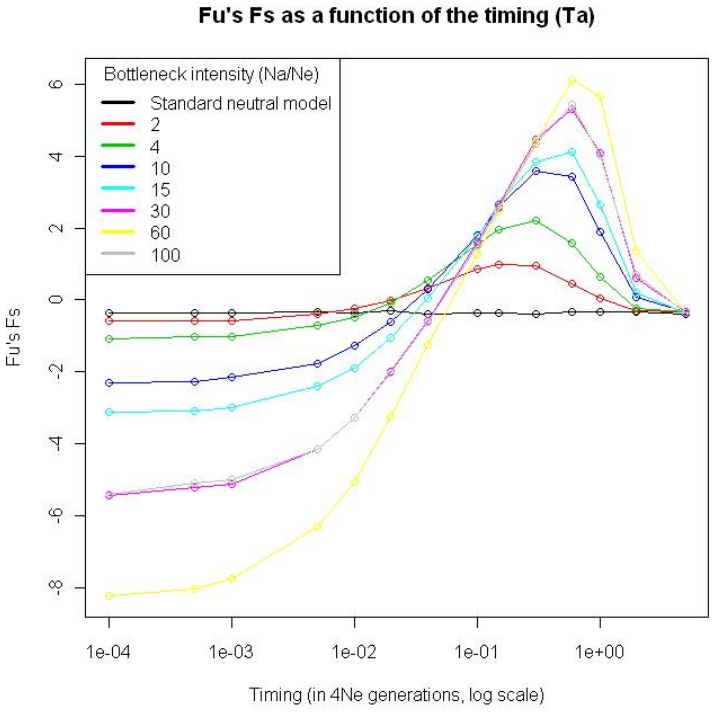


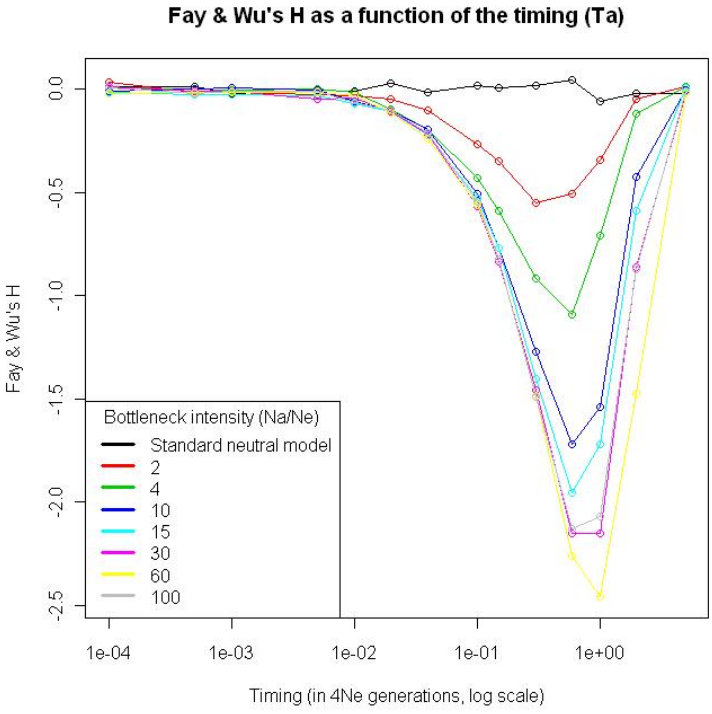
Supplementary Methods 2. Effects of the bottleneck length (*Ta*) and intensity (*Na/Ne*) on Tajima's *D*, Fu's *Fs* and Fay & Wu's *H*.



We tested the effects of the bottleneck duration Ta (in 4.*Ne* generations) and intensity (*Na/Ne*) on Tajima's *D*, Fu's *Fs* and Fay & Wu's *H.*
Each point of the curves represents the mean of a statistic (Tajima's *D,* Fu's *Fs* or Fay & Wu's *H*) over 10,000 simulated samples of 35 sequences of length 1 kb. The recombination rate was sampled in a uniform distribution U(0,0.002). *θπ* was adjusted as explained in the materials and methods section to be equal at 3 in the simulated samples.
This preliminary study helped us choosing a range of bottleneck durations for which the statistics reacted to the population size changes.







Supplementary Table 1. Geographic coordinates of the sampled populations.

Population	Latitude	Longitude
Petrock	44.05	-1.32
Mimizan	44.13	-1.3
Le Verdon	45.57	-1.22
Hourtin	45.17	-1.13
Olonne-sur-Mer	46.57	-1.83
St-Jean de Monts	46.77	-2.02

Supplementary Table 2. Gene names, primer pairs and sequence characteristics.

Gene name	Genbank Accession	Primer ID	Primer sequence 5' 3'	Fragment length (bp)	Length sequenced (bp)	Start position on the sequence	End position on the sequence	Sequencing[a]
MYB1	EU482890	F7_Myb1 R3_Myb1	TCGACGATTCTGAAGAGCTTATACT CTCCTCCTCCATATGTACTTGGTTA	1517	1426	1	1426	F / R
MYB2	EU482893	F1_Myb2 R1_Myb2	AACCAGCAGAAAGTGAAACGTGG ATCTGTGCGAATTAGAAGGCTATCG	1412	1233	1	1233	F / R
		F3_Myb2 R3_Myb2	ATAATCTGCCCCAGTTCTTGTTCG GACGATGGTGAAGATGATGAGAATC	932	801	1162	1962	F / R
MYB4	EU482894	F1_Myb4 R1_Myb4	CTCCAAACCCAAGCTAAGGAAAG CTGACTCTGTACTTGCTCATGTC	1071	937	1	937	F / R
		F2_Myb4 R3_Myb4	GATCATACACCTGCATTCCATTCT CGTAGAAGTTGGTGTTGTACTGTTG	1342	1174	523	1696	F / R
MYB5	EU482896	F2_Myb_rel R1_Myb_rel	CTCCCTGCTGTGAAAAAGCTCATACA TCATTATTTACACGGGCATGTGAC	1511	1383	1	1383	F / R
MYB8	EU482895	F1_Myb8 R2_Myb8	GATCAAGAACCTCTGGAACTCGT TCAGACAACGGAGTACTGAGCAT	1254	1118	1	1118	F / R
		F2_Myb8 R3_Myb8	GCTACAAACAGAAACTCAGGAAGG CGTGTTGTTTTCTTTCAGTCTGCG	1817	749	597	1345	R
MYB14	EU482897	F1_Myb14 R2_Myb14	CTGCTGTGAGAAGGCTCATACTAA CATCCCTGTGAGACACCAAAGTC	1037	971	1	971	F / R
HDZ31	EU482891	F7_HDZ R9_HDZ	AGTCTGTGGTAACTAGTGGTCAG ACTACATTGCAGCTTCTAGCCATC	1173	1040	1	1040	F / R
		F12_HDZ R10_HDZ	GTGATAACTGAACAATGACACATGC GATCCTGGAACACTATAAGGACTTGA	1206	665	855	1519	F
		F8_HDZ R3_HDZ	AAGCCACAGGTGAAGTAGTTTTTG GAATAACCCTGAACCCAGAAGGA	1159	1030	1439	2468	F / R
		F16_HDZ R12_HDZ	CCTTACTCAGGAGGAAGCTGTC GTAGCTCTGACAAATCCAACGG	755	651	2277	2927	F / R
		F9_HDZ R4_HDZ	AATGCTTCTAATGGGCTAACAACTC ACTTTTGCGACCATTTTCATCCAGA	1793	754	2612	3365	R

		F6_HDZ	TGCATTCCAGTTTACTTATGAGAGC	637	535	2918	3452	F
		R4_HDZ	ACTTTTGCGACCATTTTCATCCAGA					
LIM2	EU482892	F1_LIM2	CAAGATTATCAAGCGGGACTGTAG	1465	1341	1	1341	F / R
		R5_LIM2	CGAAGAATGAGATCGGCTAGAATC					
		F3_LIM2	AACTATTCGTCATTTGAGGGAGTTC	1231	1132	1277	2408	F / R
		R1_LIM2	ATCACTTGTCGTCTGTGTCTGAGTT					
SCL1	EU482898	F7_SCL	TAGAGATCTTTCAGGAAGCAATGTC	695	625	1	625	R
		R1_SCL	CCAACAGTGCTGTTTCTATTTCCT					
		F8_SCL	GTGGAAGTCCTTTGTCTTCACAG	655	579	594	1172	F
		R2_SCL	AATATGTCCTCCTGAGAAATGCAAC					
		F4_SCL	CTATGCCTATGGGGATCATACAC	1255	1072	806	1877	F / R
		R8_SCL	TTAACATGTGCACTCAGCGGGTA					

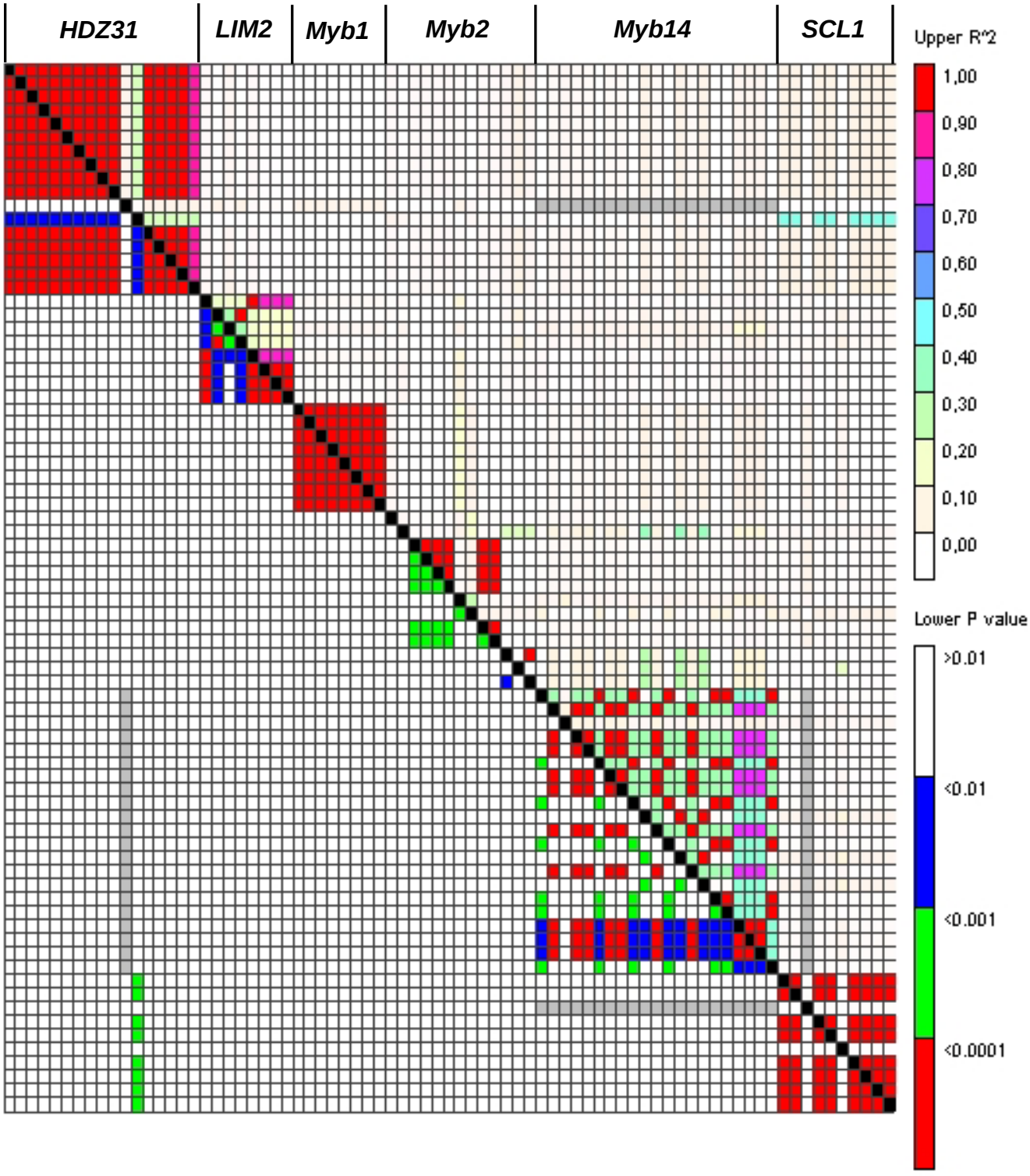
[a] F and R stand for forward and reverse sequencing, respectively.

Supplementary Table 3. Published values for estimated effective population size in different conifer tree species and in *Populus tremula*

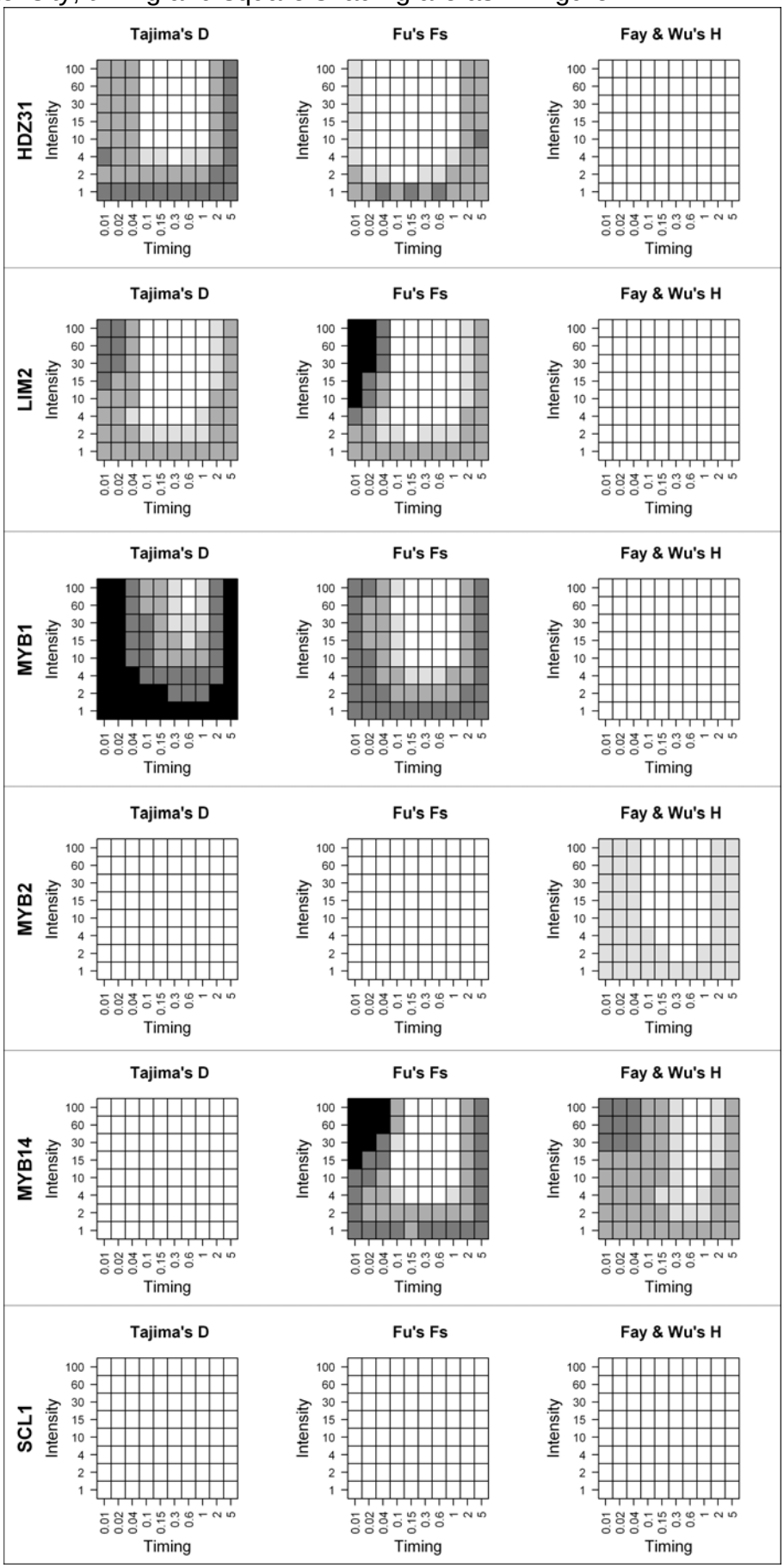
Species	Reference	Estimated mutation rate per site and per year	Generation time used (years)	θ_{silent} (per site)	Number of loci used to estimate θ	Number of gametes used to estimate θ	*Ne* estimate
Pinus taeda (whole range)	Brown *et al.* 2004	1.17×10^{-10}	25	0.00658	19	32	560,000
Pinus taeda (whole range)	Ann *et al.* 2006	7×10^{-10}	25	0.00658	19	32	94,000
Pinus sylvestris (North population)	Pyhajarvi *et al.* 2007	6.05×10^{-10}	20	0.0056*	16	40	46,281*
Picea abies (whole range)	Bouille *et al.* 2005	2.23×10^{-10} - 3.42×10^{-10}	50	0.0066	3	20	96,491-148,649
Picea glauca (whole range)	Bouille *et al.* 2005	2.23×10^{-10} - 3.42×10^{-10}	50	0.0073	3	20	106,725-164,414
Picea mariana (whole range)	Bouille *et al.* 2005	2.23×10^{-10} - 3.42×10^{-10}	50	0.0081	3	20	118,421-182,432
Populus tremula (Sweden,France and Austria)	Ingvarsson 2008	2.5×10^{-9}	15	0.0177	77	24-38	118,000
Pinus pinaster (Aquitaine)	This study	1.7×10^{-9}	20-50	0.0075	25	24	22,000-55,000

*To obtain this estimate of *Ne* , we used a nucleotide silent diversity of 0.0056 as estimated by the authors for the North population.

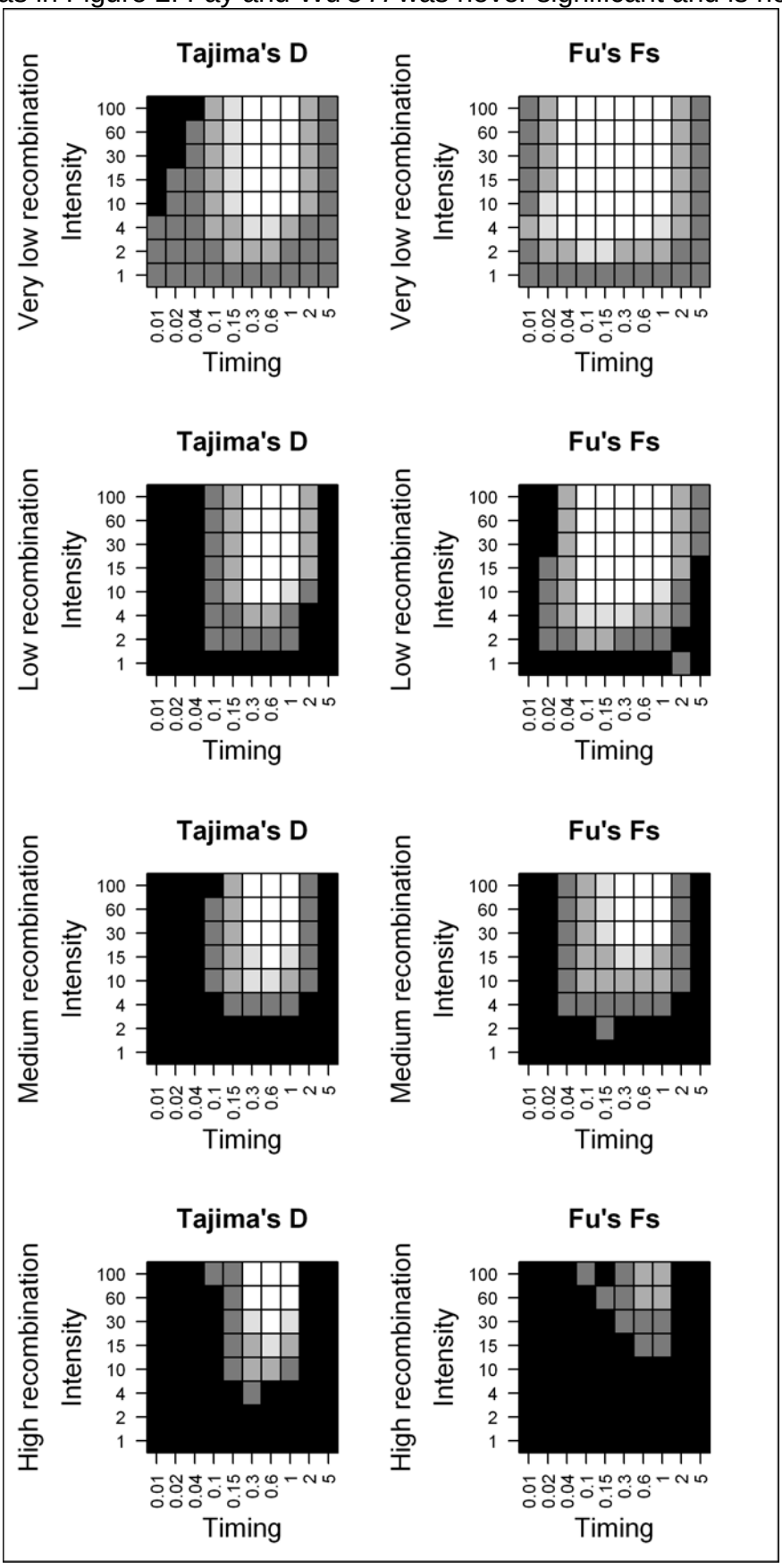
Supplementary Figure 1.
Linkage disequilibrium between polymorphic SNPs (MAF > 5%) of six TFs, considering gametes from all populations together. Square correlations between allelic frequencies (r^2) are represented above the diagonal, and probabilities for exact tests of independence are indicated below the diagonal.



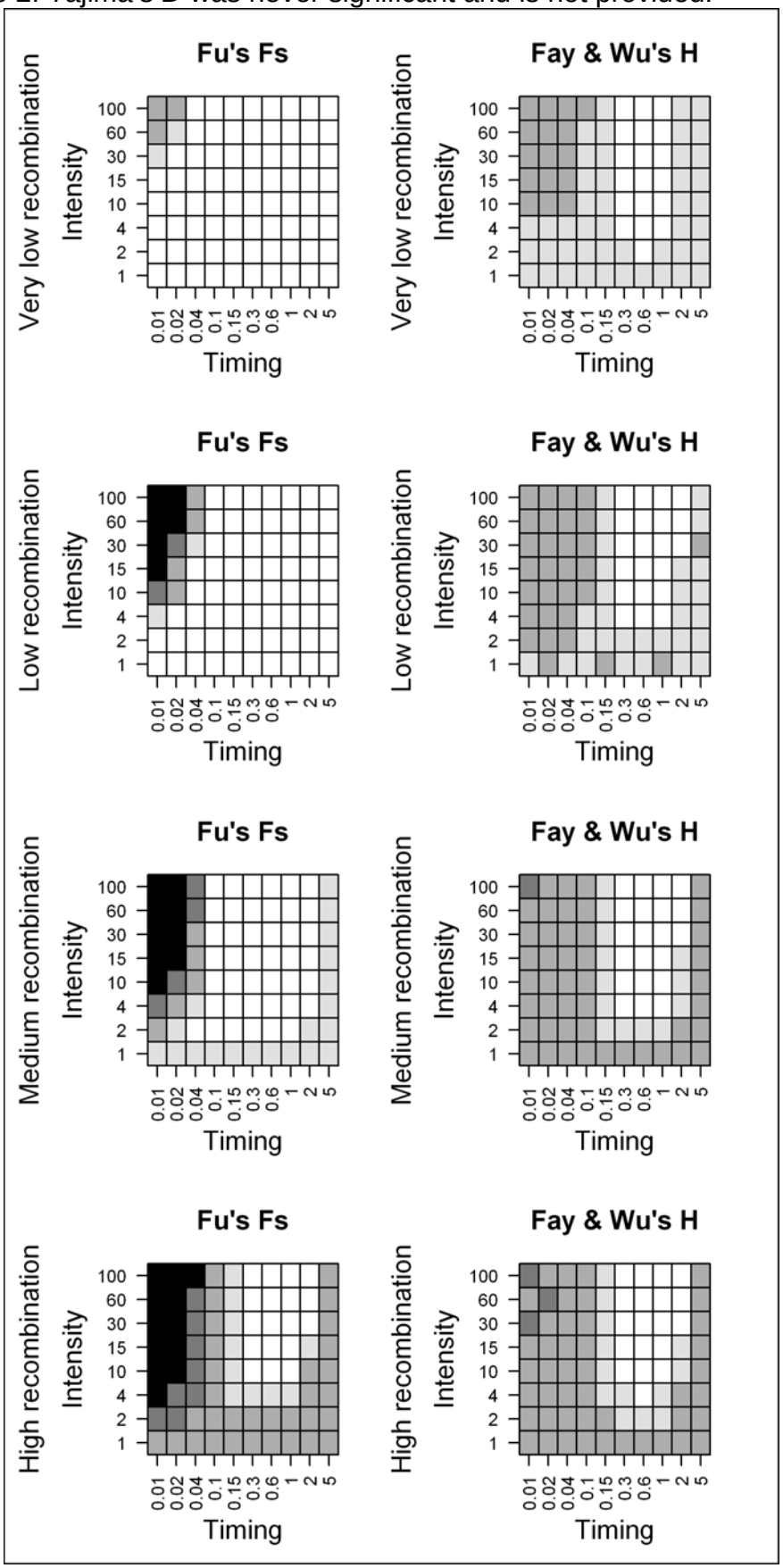
Supplementary Figure 2. Effects of bottlenecks with estimated levels of recombination on the neutrality tests statistics significance for six transcription factors. Intensity, timing and square shading are as in Figure 2.



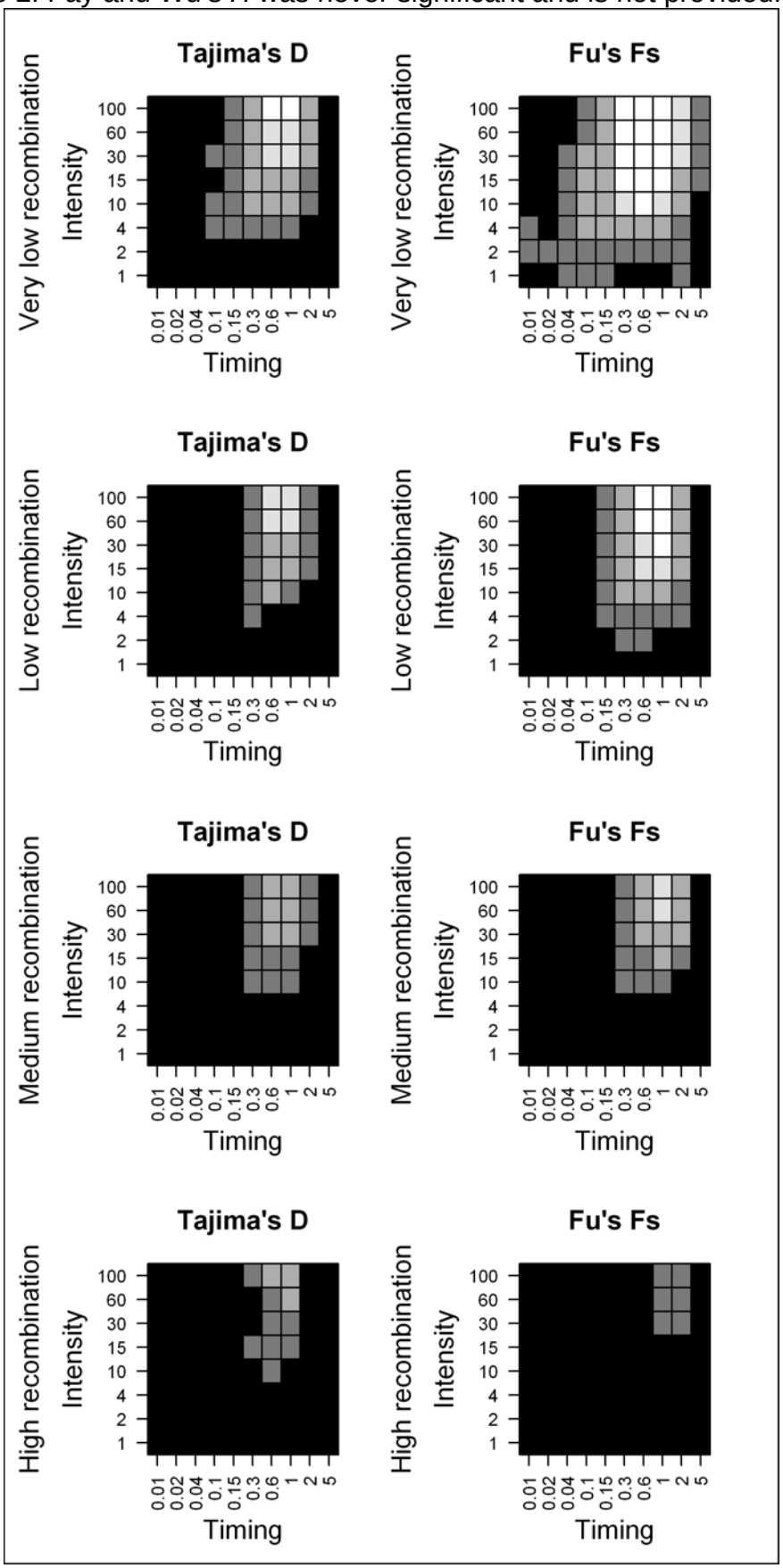
Supplementary Figure 3. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *HDZ31.* Intensity, timing and square shading are as in Figure 2. Fay and Wu's *H* was never significant and is not provided.



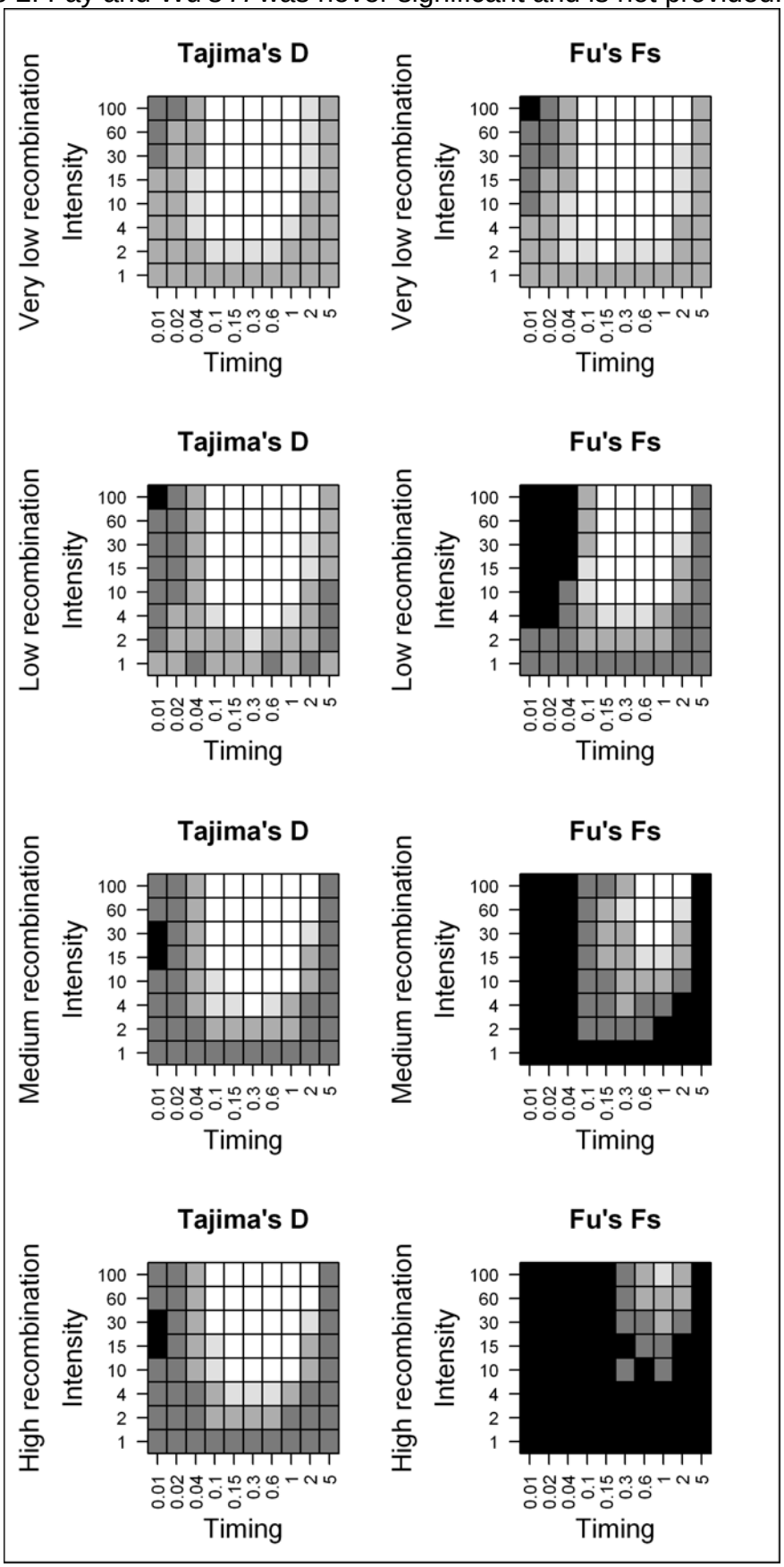
Supplementary Figure 4. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *MYB2.* Intensity, timing and square shading are as in Figure 2. Tajima's *D* was never significant and is not provided.



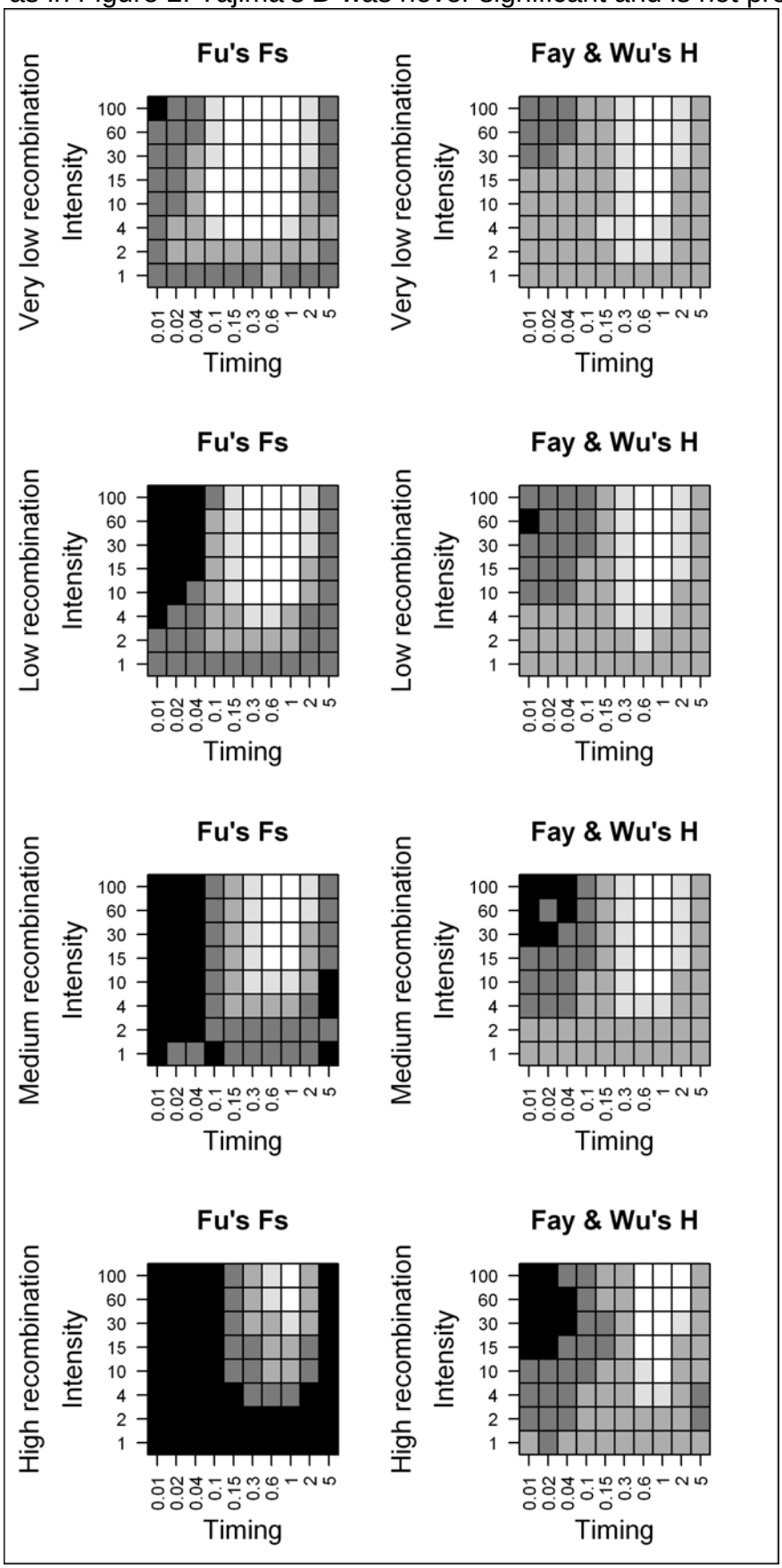
Supplementary Figure 5. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *MYB1.* Intensity, timing and square shading are as in Figure 2. Fay and Wu's *H* was never significant and is not provided.



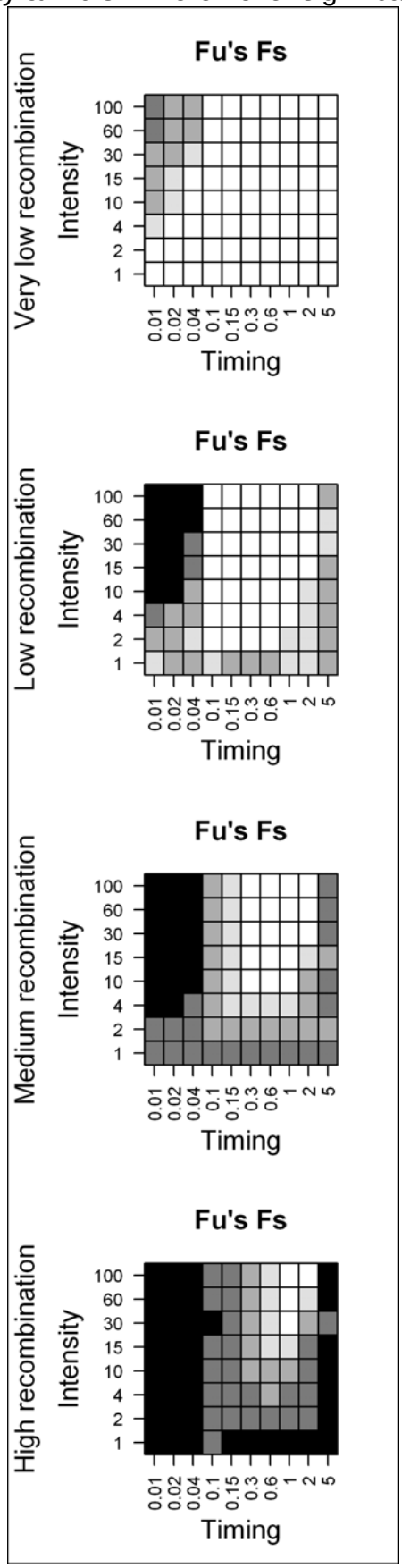
Supplementary Figure 6. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *LIM2.* Intensity, timing and square shading are as in Figure 2. Fay and Wu's *H* was never significant and is not provided.



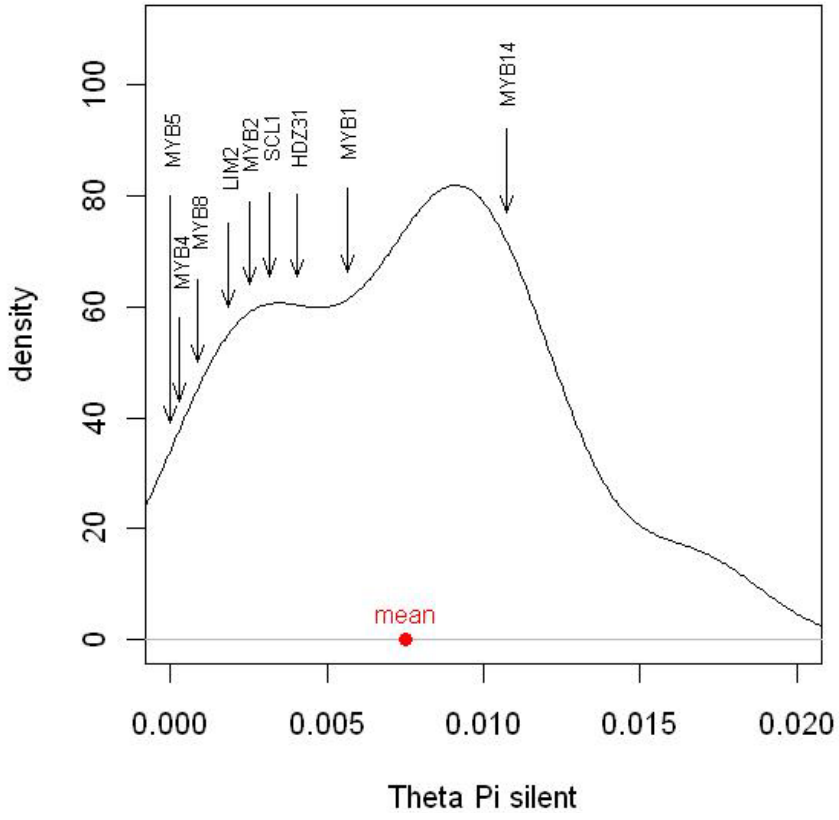
Supplementary Figure 7. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *MYB14.* Intensity, timing and square shading are as in Figure 2. Tajima's *D* was never significant and is not provided.



Supplementary Figure 8. Effects of bottlenecks with varying levels of recombination on the neutrality tests statistics significance for *SCL1.* Intensity, timing and square shading are as in Figure 2. Tajima's *D* and Fay & Wu's *H* were never significant and are not provided.



Supplementary Figure 9. Distribution of silent diversity estimates on 25 candidate genes for the Aquitaine population of *P. pinaster*, including drought stress resistance genes (EVENO *et al.* 2008) and wood formation candidate genes (*PAL-1*, *CesA02, CesA04, CesA05*, *GRP1*, *C4H2*, *THE1*, *BOTERO*, *CHS*, *COBRA*, *4CL1*, *CAD*, *LACCASE4*, *SUSY2*, *AGP1*, *KORRIGAN*).



Supplementary Figure 10. Severities of the different bottleneck *scenarii*.

		Severity of the bottleneck (*Ta* x *Na/Ne*)									
Intensity (*Na/Ne*)	**100**	4	8	16	40	**60**	**120**	**240**	**400**	**800**	2000
	60	2.4	4.8	9.6	24	**36**	**72**	**144**	**240**	**480**	1200
	30	1.2	2.4	4.8	12	**18**	**36**	**72**	**120**	**240**	600
	15	0.6	1.2	2.4	6	**9**	**18**	**36**	**60**	**120**	300
	10	0.4	0.8	1.6	4	6	12	24	40	80	200
	4	0.16	0.32	0.64	1.6	2.4	4.8	9.6	16	32	80
	2	0.08	0.16	0.32	0.8	1.2	2.4	4.8	8	16	40
	1	SNM	SNM	SNM	SNM	SNM	SNM	SNM	SNM	SNM	SNM
		0.01	**0.02**	**0.04**	**0.1**	**0.15**	**0.3**	**0.6**	**1**	**2**	**5**
		Timing (*Ta*, in 4*Ne* generations)									

Best fitting *scenarii* for the six TFs, or excluding *MYB1* for medium recombination rates.

Developing a SNP genotyping array for *Pinus pinaster*: comparison between *in vitro* and *in silico* detected SNPs

Camille Lepoittevin[*†], Jean-Marc Frigerio[*], Pauline Garnier-Géré[*], Franck Salin[*], Frank Bedon[*], Emmanuelle Eveno[*], Maria-Teresa Cervera[‡], Luis Cancino[*§], François Hubert[*], Barbara Vornam[**], Luc Harvengt[†], Christophe Plomion[*]

[*] INRA, UMR1202 Biodiversité Gènes & Communautés, F-33610 Cestas, FRANCE

[†] FCBA, Laboratoire de Biotechnologies, F-77370 Nangis, FRANCE

[‡] INIA, Departamento de Mejora Genética y Biotecnologia, 28040 Madrid, SPAIN

[§] Instituto de Biologia Vegetal y Biotechnologia, Talca, CHILE

[**] University of Goettingen, 37077 Goettingen, GERMANY

Introduction

In the last few years, the development of high-throughput methods for the detection and genotyping of single nucleotide polymorphisms (SNPs) has led to a revolution in their use as molecular markers (HENRY 2008). Their abundance in animal and plant genomes, the reduction in cost and the increased throughput of SNP assays have made these markers attractive for high-resolution genetic mapping, fine mapping of QTLs, linkage-disequilibrium based association mapping, genetic diversity analyses, genotype identification, marker-assisted selection and characterization of genetic resources (GUPTA *et al.* 2001; RAFALSKI 2002a; RAFALSKI 2002b; HENIKOFF and COMAI 2003; GIBBS and SINGLETON 2006; SLATE *et al.* 2009).

In non-model species, large scale SNP genotyping involves two main steps: first the discovery of polymorphisms, and second the genotyping of a set of specimens. SNP identification can proceed either from *in vitro* or *in silico* approaches. *In vitro* methods, such as the re-sequencing of targeted amplicons, are generally more appropriate when sequence data is limited or when one is interested in polymorphisms in specific genotypes or candidate genes. This approach is generally costly and time consuming, but has been proved successful to detect SNPs in many organisms (reviewed by EDWARDS *et al.* 2007). In contrast, *in silico* discovery is the most obvious method for *de novo* SNP identification. This approach offers a low cost source of abundant SNPs, providing that sufficient redundancy is present in the pre-existing sequence databases and that the sequences have been generated from DNA or cDNA extracted from a diverse set of individuals (PICOULT-NEWBERG *et al.* 1999; RAFALSKI 2002a; GANAL *et al.* 2009). Such SNPs have been validated by large scale genotyping for a number of plant species including *Arabidopsis* (SCHMID *et al.* 2003), maize (BATLEY *et al.* 2003), grapevine (PINDO *et al.* 2008), melon (DELEU *et al.* 2009), tomato (LABATE and BALDO 2005), spruce (PAVY *et al.* 2006) or pine (LE DANTEC *et al.* 2004).

There is no one ideal method for SNP genotyping and the selection of an appropriate technique largely depends on many factors including cost, accuracy, multiplexing capacity and throughput, equipment and difficulty of assay development (SOBRINO *et al.* 2005; SYVÄNEN 2005). A range of high-throughput methods are currently developed for model species such as humans, but their use in non-model species with large genome size, high level of ploidy or redundancy is often a challenge (CHAGNÉ *et al.* 2007). Recently, PAVY *et al.* (2008) and ECKERT *et al.* (2009) achieved the multiplexed genotyping of hundreds of SNPs in conifers, a group of plants that is characterized by a large genome size (MURRAY 1998). They

used the Illumina bead array platform combined with GoldenGate assay (OLIPHANT *et al.* 2002; FAN *et al.* 2003). This genotyping platform was also successfully used for genomes containing a high number of paralogous genes such as barley (ROSTOKS *et al.* 2006), soybean (HYTEN *et al.* 2008) or tetraploid and hexaploid wheat (AKHUNOV *et al.* 2009).

Maritime pine (*Pinus pinaster Ait.*) genome is extremely large (up to 23.8 Gb/C, which is 150 times larger than that of *Arabidopsis thaliana*) (MURRAY 1998). Despite the economical and ecological importance of this species in south-western Europe, where it covers over 4M ha, it will be many years before its full genome sequence is available. However, about 30,000 *P. pinaster* expressed sequence tags (ESTs) were produced in the past decade, followed by the re-sequencing of more than 40 wood-quality and drought-stress related candidate genes. We report here the valorization of these resources to the first highly multiplexed SNP genotyping array in *P. pinaster*. Our objectives were three-fold: i/ validate a number of SNPs for future linkage mapping and candidate-gene-based association studies, and ii/ compare the conversion rate of SNPs derived from *in vitro versus in silico* datasets, as to our knowledge no other study in conifers has attempted to genotype a large number of *in silico* SNPs without preliminary re-sequencing, and iii/ estimate the genotyping error rate of the GoldenGate technology for a conifer genome, which has not been reported so far. The SNPs validated in this study have been made available through the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/index.html) and GnpSNP (http://urgi.versailles.inra.fr/GnpSNP/snp/welcome) databases.

Methods

Plant material

Plant material consisted of 456 individuals, including: 212 unrelated trees resulting from mass selection in the natural forest of South West of France in the Aquitaine region (first-generation breeding population, referred as the "G0" population), 210 offspring resulting from open-pollinated or controlled crosses among the G0 trees (second-generation breeding population, referred as "G1"), 29 trees randomly sampled in the same geographical area as the G0 trees, and 5 trees involved in two- and three-generation outbreed pedigrees, used for linkage and QTL mapping. DNA was extracted from needles using Invisorb® Spin Plant Mini Kit (Invitek, Berlin, Germany), and quantified on a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, LLC, Wilmington, DEL, USA).

SNP discovery

For SNP discovery two sets of sequences were considered. The first dataset comprised maritime pine sequences for 41 different genes involved in plant cell wall formation (candidate genes for wood quality) or drought stress resistance (Supplemental Table 1). For each fragment, an average of 50 megagametophytes (haploid tissue surrounding the embryo) from different populations were sequenced. The chromatograms were visually checked (nucleotides with phred scores below 30 were considered as missing data) and the SNPs were considered as true. Indeed, the use of megagametophytes lowered the risk of confusing polymorphism at a unique locus with differences between paralogous loci, as amplification of two or even more members of a gene family would have been easily detected by the visualization of double peaks in the chromatograms. This first set of SNPs will be referred as *in vitro* SNPs. The second sequence dataset consisted in a collection of 3,995 non-singleton contigs of the maritime pine unigene (G. Le Provost, unpublished, http://cbi.labri.fr/outils/SAM2/COMPLETE/ version pinaster_14_02_2007). We used the Polybayes software (MARTH *et al.* 1999) to detect SNPs with a high probability with the parameters described for maritime pine in LE DANTEC *et al.* (2004). This second set of SNPs will be referred as *in silico* SNPs.

SNP selection for array construction

We developed a Perl script, *snp2illumina*, for automatically extracting SNPs from multifasta sequence files and output them as a SequenceList file compatible with the Illumina Assay Design Tool software (ADT; http://www.illumina.com/downloads/Illumina_Assay_Design_Tool.pdf). This file contains the SNP names and surrounding sequences with polymorphic loci indicated by IUPAC codes for degenerated bases. The Perl script *snp2illumina* can work in batch mode and is available upon request from the corresponding author.

The functionality score provided by ADT is similar to a predicted probability of genotyping success, taking into account the sequence conformation around the SNP, the lack of repetitive elements in the surrounding sequence, and in the case of model species the sequence redundancy against the available sequence database (SHEN *et al.* 2005). In the case of maritime pine, no sequence database was available to test for sequence redundancy. All the SNPs presenting a functionality score below 0.4, which is considered as a lower limit for genotyping success by the manufacturer, were discarded.

Two contrasted strategies "depth vs. breath of SNP coverage" were adopted to select informative SNPs. In respect to *in vitro* SNPs, our objective was to include as many polymorphisms as possible for each gene fragment, therefore depth of coverage was preferred. For *in silico* SNPs, our goal was to include a low number of markers per unigene in a large number of unigenes, thus giving more emphasis to breath of coverage. The main technical constraint for selecting *in vitro* SNPs was that the selected polymorphisms should not be less than 60 nucleotides away from each other. When several SNPs stood within this limit it was decided to filter out lowest frequency variants and polymorphisms showing high level of linkage disequilibrium with other selected SNPs of the same fragment. Rare variants (minor allele frequency < 5%) were discarded. To select *in silico* SNPs we used the log-file of the *snp2illumina* script that records for each SNP the number of ESTs considered for the detection, the minor allele frequency and the PolyBayes score. To minimize the number of false positives we included in the assay only SNPs with a PolyBayes score above 99%, with either a minor allele appearing at least twice within four to ten ESTs, or a minor allele frequency above 20% when more than ten ESTs were available. Indeed, it is highly unlikely that sequencing errors of two independently sequenced ESTs occur at the same base location. We also excluded SNPs that were surrounded by other polymorphisms in the immediate 60 bases to avoid technical problems due to neighboring polymorphisms. In both cases, chromatograms were visually checked to ensure the quality of the flanking sequences, and we used BLASTn analysis (ALTSCHUL *et al.* 1990) to ensure that *in vitro* and *in silico* SNPs belonged to different genes.

SNP genotyping array

The Illumina GoldenGate technology (Illumina Inc., San Diego, CA, USA) was used to carry out the genotyping reactions in accordance with the manufacturer's protocol (LIN *et al.* 2009). To assess the reproducibility of the genotyping assay, 19 DNA samples were duplicated across the different plates. Negative controls were also added to each 96-well plate. Highly multiplexed extension reactions were conducted using 250ng of template DNA per sample. The clustering was realized with the BeadStudio software (Illumina Inc.), and a quality score for each genotype was generated. A GenCall score cutoff of 0.25 was used to determine valid genotypes at each SNP and the SNPs retained had to get a minimum GenTrain score of 0.25, which represents a stringent criterion used in human genetic studies (www.illumina.com / FAN *et al.* 2003). GenCall and GenTrain scores measure the reliability of SNP detection based on the distribution of genotypic classes (AA, AB and BB). Clusters were visually inspected to

ensure high quality data (Figure 1). When we observed cluster compression (*i.e.* when the homozygous clusters normalized theta values were not in the [0, 0.1] or [0.9, 1] ranges, as illustrated in Figure 1 B, C and D), we considered that the genotyping failed, as this is likely due to genome redundancy (HYTEN *et al.* 2008). Indeed, the compression of the BB homozygous cluster towards the AA cluster could result from a paralog gene matching the A allele, increasing the signal for the A dye for both BB and AB genotypes. We also considered as genotyping failures monomorphic SNPs for which clusters could be divided in two or more subgroups like in Figure 1E.

Measuring the error rate using pedigree data

We used the breeding population pedigree information (relationships between first and second generation) to detect possible Mendelian Inconsistencies (MIs) between parents and offspring using the PedCheck software (O'CONNELL and WEEKS 1998). Then, we used the method described in SAUNDERS *et al.* (2007) to estimate the genotyping error rate Π from MIs. Not all genotyping errors (GEs) are detectable as MIs, but there is a linear relationship between the GE and the MI counts has shown by HAO *et al.* (2004). The expected number of MIs at a marker ($\Pi.P_{MI}$) in a family in which one or both parents and m children have been genotyped can be derived from the marker allele frequency p, m and Π as follows (SAUNDERS *et al.* 2007): if only one parent has been genotyped

$$\Pi P_{MI} = \Pi p(p-1)\left(2-\left(1-\frac{1}{2}(p-1)\right)^{m}-\left(1-\frac{1}{2}p\right)^{m}+\frac{1}{2}m\right) \qquad (1)$$

and if both parents have been genotyped

$$\Pi P_{MI} = (m+2)\Pi - 2\Pi\left\{p^2+(1-p)^2+\left(\frac{1}{2}\right)^{m-1}p(1-p)+4\left[\left(\frac{3}{4}\right)^{m}-\left(\frac{1}{2}\right)^{m}\right]p^2(1-p)^2\right\}$$
$$-m\Pi p(1\_p)\{3p^2+4p(1\_p)+3(1-p)^2\} \qquad (2)$$

These relationships can be easily generalized to large non-inbred pedigrees and many SNPs, by summation of $\Pi.P_{MI}$ over all families and averaging Π over all SNPs. This procedure allows to estimate a *per* SNP and a global genotyping error rate (SAUNDERS *et al.* 2007). We performed this analysis on 17 unrelated families from the breeding population, using for each marker the allele frequency (p) estimated on the Aquitaine G0 genotyping dataset (212 samples).

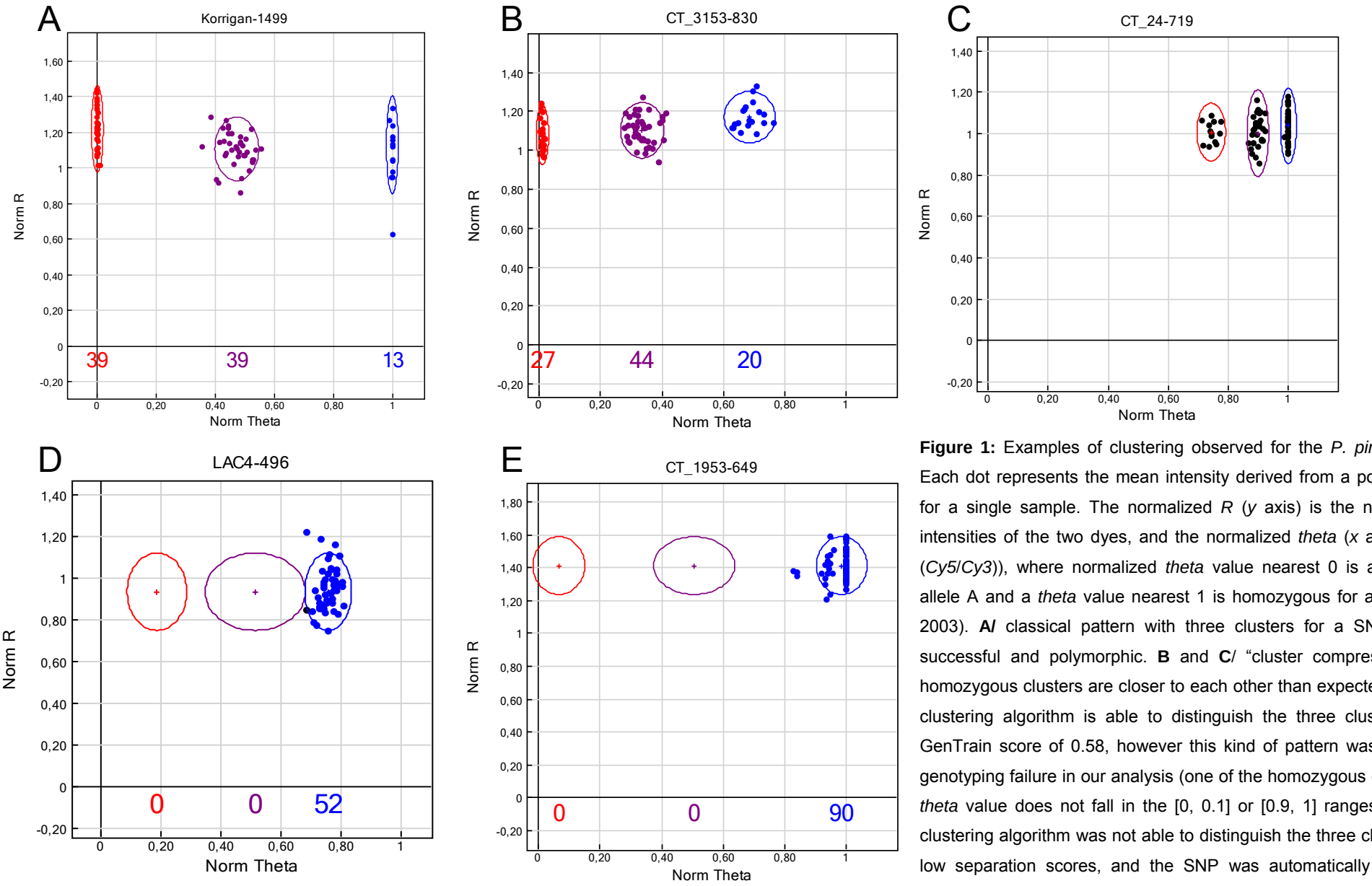


Figure 1: Examples of clustering observed for the *P. pinaster* SNP array. Each dot represents the mean intensity derived from a population of beads for a single sample. The normalized *R* (*y* axis) is the normalized sum of intensities of the two dyes, and the normalized *theta* (*x* axis) is $((2/\Pi)\mathrm{Tan}^{-1}(Cy5/Cy3))$, where normalized *theta* value nearest 0 is a homozygous for allele A and a *theta* value nearest 1 is homozygous for allele B (FAN *et al.* 2003). **A**/ classical pattern with three clusters for a SNP considered as successful and polymorphic. **B** and **C**/ "cluster compression" when both homozygous clusters are closer to each other than expected. In panel **B**, the clustering algorithm is able to distinguish the three clusters and gives a GenTrain score of 0.58, however this kind of pattern was considered as a genotyping failure in our analysis (one of the homozygous cluster normalized *theta* value does not fall in the [0, 0.1] or [0.9, 1] ranges). In panel **C** the clustering algorithm was not able to distinguish the three clusters because of low separation scores, and the SNP was automatically considered as a genotyping failure because of its low GenTrain score. SNPs in panels **D** and **E** were also interpreted as genotyping failures either because of abnormal *theta* values (**D**) or because of the presence of subgroups in a cluster (**E**).

Results

SNP detection and construction of the SNP array

A total of 448 *in vitro*-SNPs were detected in the dataset of re-sequenced fragments. Overall 155, 81 and 28 SNPs were discarded because of low functionality scores, neighboring polymorphisms, or because they corresponded to rare variants, respectively. The 184 remaining SNPs included in the assay represented 40 different fragments (Supplemental Table 2).

Similarly, 9,364 *in silico*-SNPs were detected in the unigene set, and we selected 200 of them satisfying our very stringent criteria, *i.e.* PolyBayes and functionality scores, polymorphism proximity, minimum number of ESTs for the detection, minor allele frequency and visual validation of the chromatograms. They represented 146 different unigene elements. Figure 2 shows the number of ESTs considered for the detection of the 200 *in silico*-SNPs.

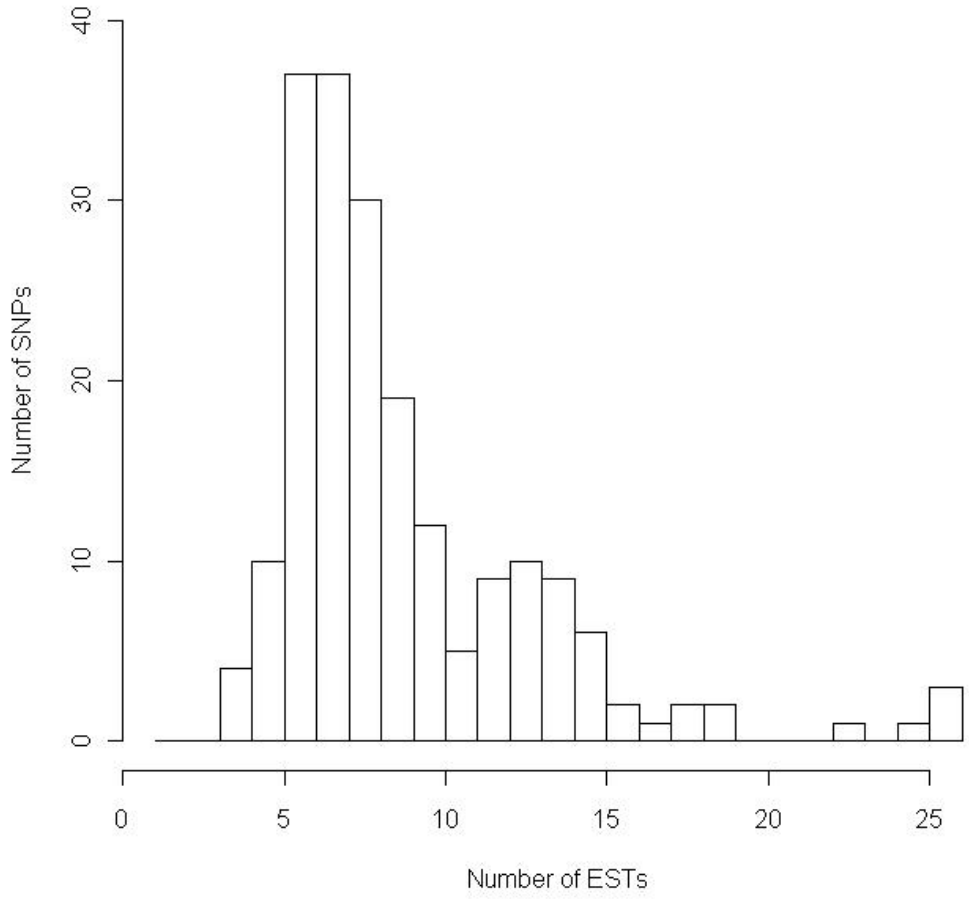


Figure 2: Number of ESTs considered for the detection of 200 *in silico*-SNPs.

Reproducibility and overall success rate of the SNP assay

No discordance was detected between the 19 replicated samples, *i.e.* the same genotype was observed over the replicates, yielding a reproducibility rate of 100%. For nine polymorphic

SNPs we observed cluster compression (as in Figure 1B), and for nine monomorphic SNPs we found either unexpected normalized theta values, or subgroups in a homozygous cluster (as in Figure 1D and 1E, respectively). In those cases we considered that the genotyping failed despite acceptable GenTrain scores.

To measure the global success of the genotyping assay we first estimated the success-rate, which corresponds to the number of SNPs that are successfully genotyped (considering both monomorphic and polymorphic SNPs) divided by the total number of SNPs in the assay, and second the conversion rate, which is the number of polymorphic SNPs divided by the total number of SNPs in the assay, as defined in FAN *et al.* (2003). Among the 384 SNPs analyzed, 257 were successfully genotyped (Table 1), leading to a global success-rate of 66.9%. The minimum GenTrain score observed for these SNPs was 0.53. A total of 60 SNPs were found to be monomorphic in the tested samples, yielding to a conversion rate of 51% (Table 1).

Table 1: Success rate of SNP markers.

Category	Number of SNPs (*in vitro* / *in silico*)	Percentage of SNPs (*in vitro* / *in silico*)
Failed SNPs (GenTrain score < 0.25)	127 (50 / 77)	33% (27% / 38.5%)
Successful SNPs but monomorph	60 (22 / 38)	16% (12% / 19%)
Successful SNPs and polymorph	197 (112 / 85)	51% (61% / 42.5%)
Total	384 (184 / 200)	100% (100% / 100%)

The mean call rate across successful SNPs, which is 1 minus the rate of missing data, exceeded 98% for four of the five plates analyzed, but dropped to 77.5% for the fifth plate where we noticed evaporation problems during the genotyping reactions. We found significant differences depending on the origin of the markers: *in vitro*-SNPs generally showed significantly higher genotyping-success and conversion rates compared to *in silico*-SNPs (+11.5% and +18.5% with *P*-values of 0.0115 and 3.10^{-4}, respectively, estimated using 10,000 permutations among *in vitro*-SNP and *in silico*-SNP classes).

The distribution of allelic frequencies for *in vitro*- and *in silico*-SNPs is shown in Figure 3. Among successfully genotyped SNPs, monomorphic loci were twice more abundant for *in silico*-SNPs compared to *in vitro*-SNPs (30.9% versus 16.4%, respectively). Most of the 22 monomorphic *in vitro*-SNPs corresponded to either SNPs that were monomorphic in the Aquitaine sequences (10 SNPs), rare variants (3 SNPs with a minor allele frequency below 5% in the Aquitaine sequencing dataset), or were detected on alignments that did not include any sequences from South West of France (3 SNPs). Among the polymorphic SNPs, 35.7% of *in vitro* and 29.4% of *in silico* SNPs were rare variants (MAF ≤ 10%) (Figure 3).

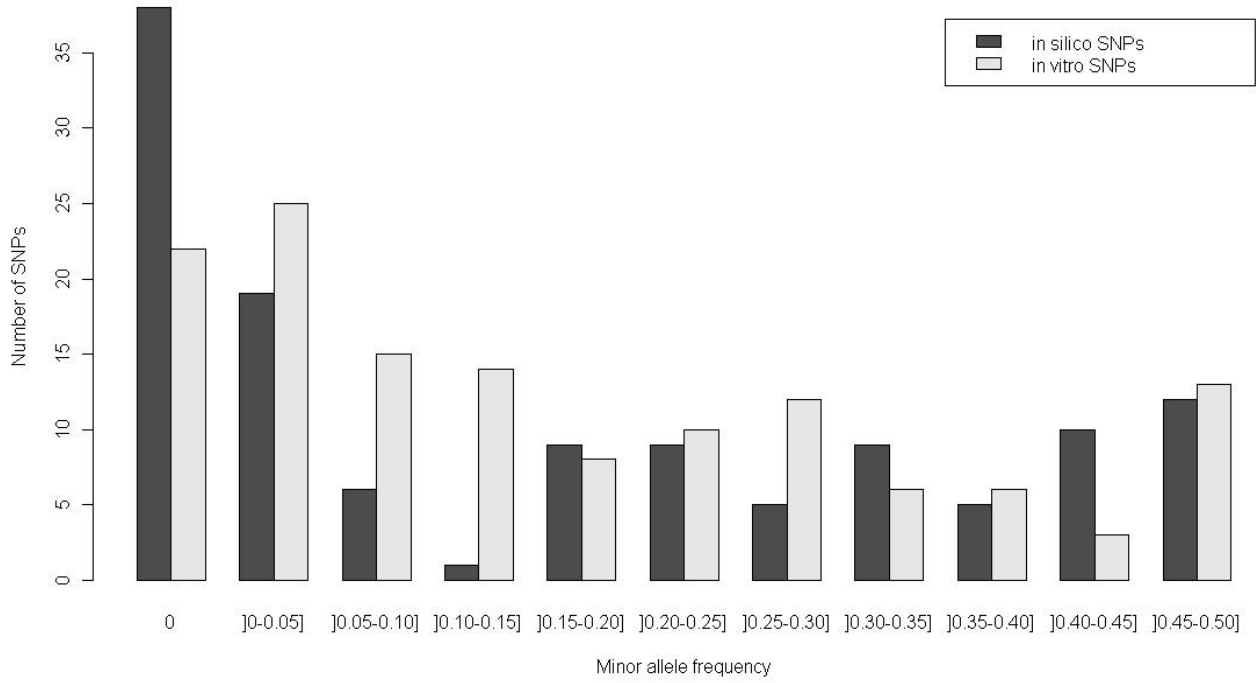


Figure 3: Allele frequency spectrum for 257 successfully genotyped *in vitro*- and *in silico*-SNPs.

SNP success rate according to a priori SNP functionality score

Prior to the construction of the SNP bead array, a functionality score was calculated for each candidate SNP using the Illumina Assay Design Tool. The higher the score, the more likely the SNP will to be successfully genotyped. We could not genotype any of the five SNPs with functionality scores below 0.5, and only 13 of the 27 SNPs with functionality scores between 0.5 and 0.6 (Figure 4). SNPs with a predicted functionality score above 0.6 had a much higher success rate than those below 0.6 (*P*-value = 0.0014 estimated using 10,000 permutations among the two classes of functionality score), as found in PAVY *et al*. (2008) for white and black spruce. This also agrees with Illumina's recommendations of using only SNPs with a functionality score above 0.6 to ensure a high success rate for the assay.

Comparison of allele frequency estimated by sequencing and genotyping

Among the 112 polymorphic *in vitro*-SNPs of the assay, 101 were discovered in alignments containing 10 sequences or more from the Aquitaine population and were used to assess the reliability of allele frequency estimates based on sequencing data. The correlation between marker allele frequencies determined by sequencing and genotyping was ~0.83 (considering only the 212 unrelated samples from the Aquitaine G0 breeding population) (Figure 5), showing that allelic frequencies estimated by genotyping were generally in the range of those estimated by sequencing. However, the confidence interval around frequency estimates was generally large for samples of the size of our sequencing panel (< 50 megagametophytes).

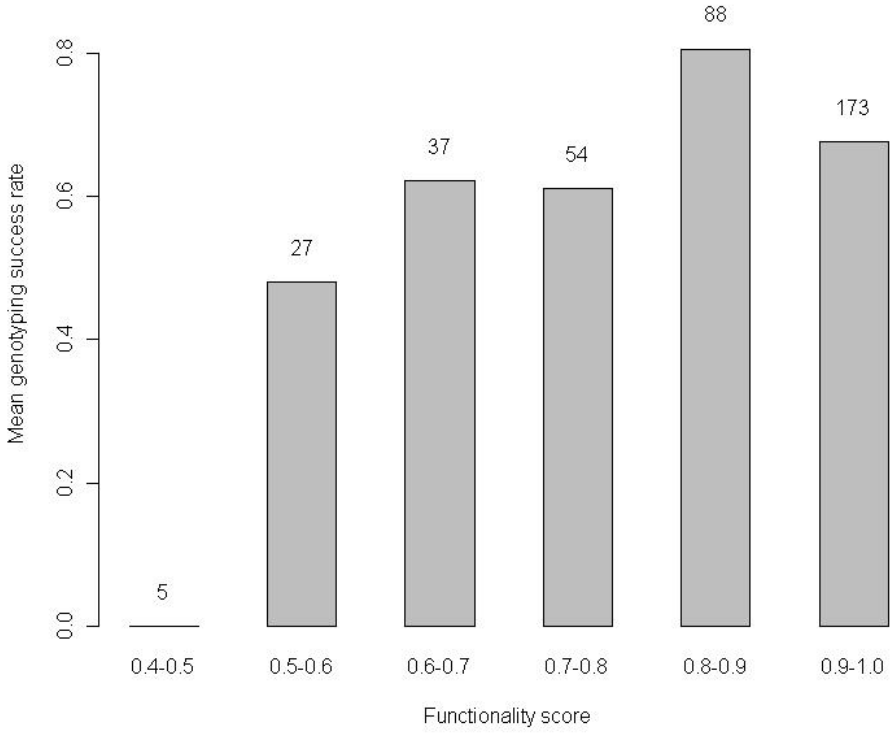


Figure 4: Genotyping success rate according to functionality score for the 384 SNPs of the assay. The number of SNPs in each functionality score class is indicated above each bar.

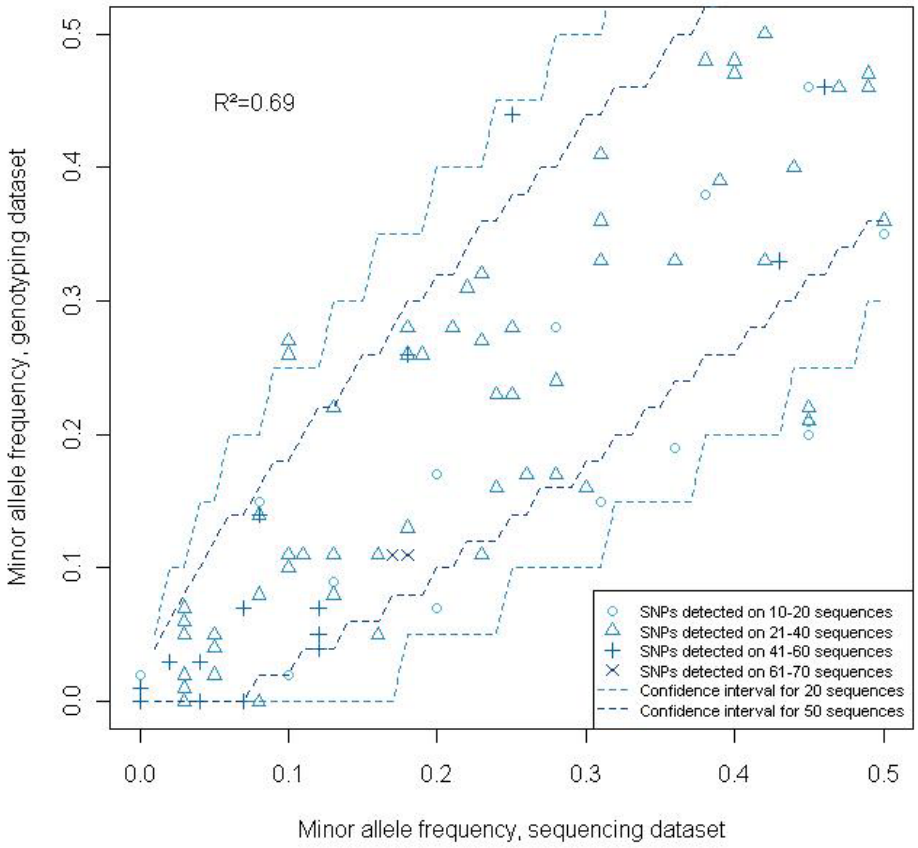


Figure 5: Correlation between allele frequencies estimated by sequencing and genotyping for 101 *in vitro*-SNPs. Dashed lines correspond to the 95% bootstrap confidence intervals for allele frequencies estimated on 20 (light blue lines) or 50 (dark blue lines) samples.

Measuring genotyping error rate with pedigree data

For 84 and 81 offsprings of the G1 population, either one or both parental G0 trees were genotyped in the assay. This dataset consisted in 36,991 genotyping datapoints corresponding to 222 samples (165 G1 and 57 G0 trees) genotyped for 188 polymorphic SNPs, after excluding 4,745 missing data. We found a total of 181 Mendelian Inconsistencies (MIs). Most of these errors (75%) appeared in only nine parents-offspring pairs for which the MI rate ranged from 4% to 17%, suggesting laboratory errors (either traceability errors during the controlled pollination, plant material sampling and handling, wet lab experiment, or DNA contamination) rather than genotyping errors. In six cases we assumed that the MIs originated from the offspring genotypes as the parents were involved in other crosses where no MI was found. For the other cases we could not tell apart parents and offspring MIs. Setting aside these possible human errors, 46 MIs were detected for 35,521 genotyping datapoints. To estimate Π, the genotyping error rate, we used a subset of 17 unrelated families corresponding to 75 G1 trees and their 26 G0 parents genotyped for 188 polymorphic SNPs (18,261 genotyping datapoints after removing 727 missing data). The observed MI count for this subset was 28, yielding a global mean genotyping error rate Π of 0.54%. At the SNP level, a total of 181 SNPs showed no MIs and thus a null *per* SNP genotyping error rate. Among the seven remaining markers, three showed error rates ranging from 2.9% to 3.3%, and four (two *in vitro*-SNPs and two *in silico*-SNPs) showed particularly elevated error rates (Π between 16% and 70%). In these cases (distribution of error rates skewed owing to a few SNPs with very high error rates), the estimate of the mean error rate tends to be biased upwards (SAUNDERS *et al.* 2007). When removing these four SNPs, the observed MI count dropped to 3, leading to a mean error rate Π of 0.06%.

Discussion

Data summary

A 384-SNPs GoldenGate genotyping array for *Pinus pinaster* was built from i/ 448 SNPs originally detected in a set of 41 re-sequenced candidate genes (*in vitro*-SNPs) and ii/ 9,364 SNPs screened from ESTs (*in silico*-SNPs). Two different SNP selection strategies were followed, "depth *vs.* breath of SNP coverage". For *in vitro*-SNPs we aimed at validating as many polymorphisms as technically possible for each fragment (depth), whereas for *in silico*-SNPs we aimed to validate few SNPs per unigene in a large number of unigenes (breath). A total of 184 *in vitro*-SNPs were chosen on the basis of functionality scores, presence of neighboring polymorphisms, minor allele frequencies and linkage disequilibrium. Moreover, 200 *in silico*-SNPs were selected based on three parameters that proved critical for high

validation rate of EST derived SNPs (WANG *et al.* 2008): the number of ESTs used for SNP detection, the SNP minor allele frequency and the quality of SNP flanking sequences. The global success rate of the assay was 66.9% (considering monomorphic and polymorphic SNPs), and a conversion rate of 51% was achieved (considering only polymorphic SNPs). *In vitro*-SNPs showed significantly higher genotyping-success (+11.5%, *P*-value 0.0115) and conversion (+18.5%, *P*-value 3.10^{-4}) rates than *in silico*-SNPs. The functionality score estimated for each SNP, which in our case could not account for sequence redundancy in the genome, showed a significant relationship with success of genotyping. The reproducibility of the assay was very good (100%, based on 19 replicated genotypes), and the genotyping error rate very low (0.54%, dropping down to 0.06% when removing four SNPs showing elevated error rates).

Conversion rates of in vitro- and in silico-SNPs for Pinus pinaster

Data obtained from the GoldenGate assay reported in this paper suggest that the bead array technology is suitable for the complex and large genome of *P. pinaster*: 66.9% of the SNPs were translated into easily interpreted genotypic clusters. This success rate is similar to that observed for *Pinus taeda* (66.9%, ECKERT *et al.* 2009), but lower than that observed for *Picea glauca* or *Picea mariana* (78.5% and 81.1% respectively when considering polymorphic and monomorphic SNPs, PAVY *et al.* 2008). So far, two main causes have been invoked in the literature for explaining genotyping failures in GoldenGate assays for non-model species. First, the partial knowledge of large and redundant genomes can be a limiting factor to design an efficient SNP genotyping assay. Indeed, flanking sequences cannot be fully validated for locus specificity and the possible presence of repetitive elements (FAN *et al.* 2003; SHEN *et al.* 2005; PAVY *et al.* 2008). Secondly, the sample size used for SNP discovery in species presenting high levels of nucleotide diversity may be too small, possibly leading to the presence of undetected SNPs within priming sites when larger sample of trees are genotyped (ECKERT *et al.* 2009). In the case of *Pinus pinaster*, both hypotheses can be examined: we reached a 79.6% success rate when considering only 103 *in vitro*-SNPs that were detected on more than 30 individuals from the Aquitaine population, which is similar to that observed in *Picea* species (78.5% and 81.1% in *P. glauca* and *P. mariana*, respectively, PAVY *et al.* 2008). The rate dropped to 55.8% for the 43 *in vitro*-SNPs detected on 10 to 30 samples. We checked that this was not due to differences in allelic frequency distribution or classes for SNPs detected using sample sizes (data not shown). This significant difference (*P*-value = 0.003 obtained by 10,000 permutations between the two classes) suggests that the sample size of the SNP discovery panel has a large impact on the conversion rate. However, the high conversion rate achieved using SNPs from well

characterized DNA regions (79.6%) still does not reach that reported for human (> 91% in FAN *et al.* 2003; MONTPETIT *et al.* 2006; GARCÍA-CLOSAS *et al.* 2007; CUNNINGHAM *et al.* 2008). As discussed in PAVY *et al.* (2008), the megagenome of conifers may hinder the development of specific probes for the assay. The nine cases of cluster compression detected in our assay support this hypothesis. The shift of a homozygous cluster toward the other one has previously been observed for a SNP in a gene presenting a nearly identical paralog in soybean, and is likely the sign of the targeted-sequence redundancy (HYTEN *et al.* 2008).

We found a significant difference between *in vitro*-SNP and *in silico*-SNP conversion rates (61% and 42.5%, respectively, *P*-value = 3.10^{-4}). According to WANG *et al.* (2008), genotyping failures in ESTs-derived SNPs may come either from sequencing errors that lead to the identification of false-positive SNPs (pseudo-SNPs), from low quality of the SNP flanking sequences, or from the presence of an exon-intron junction near the site of interest. In our study, the selection of false-positive SNPs should have been prevented by the use of trace data for SNP detection (MARTH *et al.* 1999), and a set of stringent criteria including MAF and contig size. Indeed, WANG *et al.* (2008) achieved a 70.9% conversion rate for catfish *in silico*-SNPs detected on at least four sequences and with a minor allele present twice, against a rate of 33.3% for SNPs detected on four or fewer sequences with minor allele present only once. In our case, chromatograms have also been checked to ensure high-quality of flanking sequences for primer design, but the presence of undetected polymorphisms in these regions is likely as most SNPs were detected on only ten ESTs or less (Figure 2). We could not confirm whether or not *in silico*-SNPs were located at exon-intron borders, as we lack a fully sequenced conifer genome to compare with. The presence of introns has been identified as a major cause for *in silico*-SNP genotyping failures (WANG *et al.* 2008), and may explain the conversion rate difference between *in vitro*- (revealed from genomic DNA sequences) and *in silico*- (discovered from mRNA sequences) SNPs.

Surprisingly, 6 *in vitro*-SNPs were found monomorphic on the genotyped trees, while they were detected as polymorphic loci with intermediate frequency estimates in the re-sequenced haploid panel from the Aquitaine population. Given that we are confident that they were not sequencing artifacts, this observation could be explained by either the lack of amplification of one allele due to polymorphism in the priming site, the presence of gametophytic selection against deleterious mutations (as sequences were obtained from haploid megagametophytes while genotyping was performed on diploid DNA), or the general complexity of the *pine* genome as previously discussed. In the latter case, the distinction between genotyping reaction failures and monomorphic SNPs is not obvious. In this study we decided to discard nine

monomorphic SNPs with acceptable GenTrain scores but showing either subgroups in the homozygous cluster, or normalized theta values departing from the classical 0/1 values for an homozygous locus. These patterns might be particular forms of cluster compression (shift of the BB cluster toward the AA cluster as illustrated in Figure 1D, or putative shift of the AA and AB clusters toward the BB cluster, Figure 1E). The main quality metrics for SNP assays (GenCall and GenTrain scores) measure the capacity to group samples into genotypic clusters, but to our knowledge no study have established yet the ability of genotype calling algorithms to tell apart failed reactions from monomorphic markers, or to detect cluster compression. Even if geneticists are generally not interested in failed or monomorphic markers, as they do not carry any information, detecting cluster compression would be very useful for non-model species. Markers presenting such patterns should not be used in highly heterozygous populations such as mapping pedigrees, as the heterozygous cluster is often indistinguishable from one or both homozygous ones (HYTEN *et al.* 2008).

Genotyping error rate

All large genotype datasets have errors that can be either due to sample mishandling, failures of analysis algorithms, or simply biochemical anomalies. Inclusion of incorrect data in genetic analysis can lead to an inflation in genetic map distances (HACKETT and BROADFOOT 2003), an increase in type I error and / or a decrease in statistical power in association studies (ABECASIS *et al.* 2001; GORDON *et al.* 2002), or to biased estimates of linkage disequilibrium (AKEY *et al.* 2001) and other allele-frequency related parameters (POMPANON *et al.* 2005). Errors in a dataset can be detected either by comparing genotypic information obtained from different technologies or by using Mendelian Inconsistencies (MIs) in family-based samples. In this study, we identified nine samples that concentrated 75% of all the observed MIs, which was interpreted as human errors. Sample mishandling has already been identified as a main issue during the genotyping process (BONIN *et al.* 2004; POMPANON *et al.* 2005), and could be reduced by the use of traceability systems such as Laboratory Information Management Systems (LIMs), quality insurance standards, and reduced human manipulation, according to the automation possibilities.

Using pedigree information of unrelated families, we also estimated a *per* SNP genotyping error-rate (SAUNDERS *et al.* 2007), which provides complementary information and helps to identify error-prone loci that can be removed from the study to increase its reliability. For example, the mean error rate per locus dropped from 0.54% to 0.06% when removing the four (out of 188) polymorphic loci that had the highest error rate. These genotyping error-rates are in the range of those recently reported for tetraploid and hexaploid wheat (0% and 1%,

respectively, AKHUNOV *et al.* 2009). Unfortunately, genotyping error-rates have seldom been reported for GoldenGate assays in non-model species. While this technique already proved accurate for human, the species for which it was developed (FAN *et al.* 2003), its reliability in the complex genomes of plants should be estimated before extensive use. If moderate error rates can be tolerated in cases such as QTL studies involving frequent alleles (ABECASIS *et al.* 2001), or identical by descent-based analyses when considering a large number of markers (SAUNDERS *et al.* 2007), conversely low error rates can be dramatic in association-mapping studies (KANG *et al.* 2004). Once the genotyping error-rate has been estimated, statistical tools that account for it have been developed for linkage analysis (GÖRING and TERWILLIGER 2000), family or population-based association mapping (GORDON *et al.* 2001; SOBEL *et al.* 2002; RICE and HOLMANS 2003).

Conclusion and perspectives

In this project, we demonstrated that ESTs provide a resource for SNP identification in non-model species, which do not require any additional bench work and little bio-informatics analysis. However, the time and cost benefits of *in silico*-SNPs are counterbalanced by a lower conversion rate than *in vitro*-SNPs. This drawback is acceptable for population-based experiments (in our study, a 42.5% conversion rate was achieved for *in silico*-SNPs, compared to 61% for *in vitro*-SNPs), but could be dramatic in experiments involving samples from narrow genetic backgrounds. For example, ECKERT *et al.* (2009) only reached an 18.2% conversion rate in a *P. taeda* mapping pedigree, using *in vitro*-SNPs from a database that did not include any sequences of the parental lines of the mapping population. In addition, we showed that both the visual inspection of genotyping clusters and the estimation of a *per* SNP error rate should help identify markers that are not suitable to the GoldenGate technology in species characterized by a large and complex genome.

Recently, a larger-scale SNP-array was designed for maritime pine, comprising 1,536 SNPs (826 *in vitro* SNPs, including 560 SNPs detected from re-sequenced amplicons provided by David Neale, UC Davis, CA, USA, http://dendrome.ucdavis.edu/crsp/, and 710 *in silico* SNPs selected with the same criteria as in this study). This second generation SNP-array will be used to establish a species consensus map based on the analysis of seven pedigrees, and for association mapping for a series of traits (biomass production, wood and end-use properties, drought stress resistance) measured on clonal and progeny tests on the first and second breeding populations. In parallel, a 454 pyrosequencing strategy of pooled cDNA samples (BARBAZUK *et al.* 2007) is being performed to enrich the SNP catalog based on high coverage sequence reads.

Acknowledgements

We would like to acknowledge the staff of the "Genome and Transcriptome" facility of Bordeaux (France) and the GenoToul facility of Toulouse (France) for their help in sequencing and genotyping, respectively. We also thank the Experimental Unit of Pierroton (UE570) and Pierre Alazard from FCBA for collecting the samples. This research was supported by grants from ANR Genoplante (GenoQB, GNP05013C) and PFTV (BOOST-SNP, 07PFTV002), the Aquitaine Region and the EU (alpha project GEMA, and Seventh Framework Programme FP7/2007-2013 under the grant agreement n°211868). C. Lepoittevin was supported by CIFRE contract between FCBA and INRA.

Author's contribution

CP and LH organized the funding of the study. CL extracted the DNA. FS performed the genotyping. PGG, CL, FB, EE, MTC, BV, LC and FH provided the candidate gene sequences. JMF assembled the EST data with the help of PGG and wrote the *snp2illumina* Perl script with the help of CL. CL and PGG designed the array. CL analyzed the data. CL and CP wrote the paper, with the helpful comments of PGG.

References

ABECASIS, G., S. S. CHERNY and L. R. CARDON, 2001 The impact of genotyping error on family-based analysis of quantitative traits. Journal of Human Genetics **9:** 130-134.

AKEY, J. M., K. ZHANG, M. XIONG, P. DORIS and L. JIN, 2001 The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. The American Journal of Human Genetics **68:** 1447-1456.

AKHUNOV, E., C. NICOLET and J. DVORAK, 2009 Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theoretical and Applied Genetics **119:** 507-517.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. Journal of Molecular Biology **215:** 403-410.

BARBAZUK, W. B., S. J. EMRICH, H. D. CHEN, L. LI and P. S. SCHNABLE, 2007 SNP discovery via 454 transcriptome sequencing. The Plant Journal **51:** 910.

BATLEY, J., G. BARKER, H. O'SULLIVAN, K. J. EDWARDS and D. EDWARDS, 2003 Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiology **132:** 84-91.

BONIN, A., E. BELLEMAIN, P. BRONKEN EIDESEN, F. POMPANON, C. BROCHMANN *et al.*, 2004 How to track and assess genotyping errors in population genetics studies. Molecular Ecology **13:** 3261-3273.

CHAGNÉ, D., J. BATLEY, D. EDWARDS and J. W. FORSTER, 2007 Single nucleotide polymorphisms genotyping in plants, pp. 77–94 in *Association Mapping in Plants*, edited by N. ORAGUZIE, RIKKERINK, EHA, GARDINER, SE AND, DE SILVA, HN. Springer, New York.

CUNNINGHAM, J. M., T. A. SELLERS, J. M. SCHILDKRAUT, Z. S. FREDERICKSEN, R. A. VIERKANT *et al.*, 2008 Performance of amplified DNA in an Illumina GoldenGate BeadArray Assay. Cancer Epidemiology, Biomarkers & Prevention **17:** 1781-1789.

DELEU, W., C. ESTERAS, C. ROIG, M. GONZALEZ-TO, I. FERNANDEZ-SILVA *et al.*, 2009 A set of EST-SNPs for map saturation and cultivar identification in melon. BMC Plant Biology **9:** 90.

ECKERT, A. J., B. PANDE, E. S. ERSOZ, M. H. WRIGHT, V. K. RASHBROOK *et al.*, 2009 High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genetics & Genomes **5:** 225-234.

EDWARDS, D., J. BATLEY, N. O. I. COGAN, J. W. FORSTER and D. CHAGNÉ, 2007 Single nucleotide polymorphism discovery, pp. 53–76 in *Association Mapping in Plants*, edited by N. ORAGUZIE, RIKKERINK, EHA, GARDINER, SE, AND DE SILVA, HN. Springer, New York.

FAN, J. B., A. OLIPHANT, R. SHEN, B. G. KERMANI, F. GARCIA *et al.*, 2003 Highly parallel SNP genotyping, pp. 69-78 in *Cold Spring Harbor Symposia on Quantitative Biology*, edited by C. S. H. L. PRESS, Cold Spring Harbor, New York.

GANAL, M. W., T. ALTMANN and M. S. RÖDER, 2009 SNP identification in crop plants. Current Opinion in Plant Biology **12:** 211-217.

GARCÍA-CLOSAS, M., N. MALATS, F. X. REAL, M. YEAGER, R. WELCH *et al.*, 2007 Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. PLoS Genetics **3:** e29.

GIBBS, J. R., and A. SINGLETON, 2006 Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS genetics **2:** e150.

GORDON, D., S. J. FINCH, M. NOTHNAGEL and J. OTT, 2002 Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Human Heredity **54:** 22-33.

GORDON, D., S. C. HEATH, X. LIU and J. OTT, 2001 ATransmission/Disequilibrium Test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. The American Journal of Human Genetics **69:** 371-380.

GÖRING, H. H. H., and J. D. TERWILLIGER, 2000 Linkage Analysis in the Presence of Errors IV: Joint Pseudomarker Analysis of Linkage and/or Linkage Disequilibrium on a Mixture of Pedigrees and Singletons When the Mode of Inheritance Cannot Be Accurately Specified. The American Journal of Human Genetics **66:** 1310-1327.

GUPTA, P. K., J. K. ROY and M. PRASAD, 2001 Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Current Science **80:** 524-535.

HACKETT, C. A., and L. B. BROADFOOT, 2003 Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity **90:** 33-38.

HAO, K., C. LI, C. ROSENOW and W. HUNG WONG, 2004 Estimation of genotype error rate using samples with pedigree information--an application on the GeneChip Mapping 10K array. Genomics **84:** 623-630.

HENIKOFF, S., and L. COMAI, 2003 Single-nucleotide mutations for plant functional genomics. Annual Review of Plant Biology **54:** 375-401.

HENRY, R. J., 2008 *Plant Genotyping II: SNP Technology*. Oxford University Press.

HYTEN, D., Q. SONG, I.-Y. CHOI, M.-S. YOON, J. SPECHT *et al.*, 2008 High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theoretical and Applied Genetics **116:** 945-952.

KANG, S. J., D. GORDON and J. S. FINCH, 2004 What SNP genotyping errors are most costly for genetic association studies? Genetic Epidemiology **26:** 132-141.

LABATE, J. A., and A. BALDO, 2005 Tomato SNP discovery by EST mining and resequencing. Molecular Breeding **16:** 343-349.

LE DANTEC, L., D. CHAGNE, D. POT, O. CANTIN, P. GARNIER-GERE *et al.*, 2004 Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Molecular Biology **54:** 461-470.

LIN, C. H., J. M. YEAKLEY, T. K. MCDANIEL and R. SHEN, 2009 Medium- to high-throughput SNP genotyping using VeraCode Microbeads, pp. 129-142 in *DNA and RNA Profiling in Human Blood*, edited by P. BUGERT. Humana Press, New York.

MARTH, G. T., I. KORF, M. D. YANDELL, R. T. YEH, Z. J. GU *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nature Genetics **23:** 452-456.

MONTPETIT, A., M. NELIS, P. LAFLAMME, R. MAGI, X. KE *et al.*, 2006 An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genetics **2:** e27.

MURRAY, B., 1998 Nuclear DNA amounts in gymnosperms. Annals of Botany **82:** 3-15.

O'CONNELL, J. R., and D. E. WEEKS, 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. The American Journal of Human Genetics **63:** 259-266.

OLIPHANT, A., D. L. BARKER, J. R. STUELPNAGEL and M. S. CHEE, 2002 BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques**:** 56-61.

PAVY, N., L. PARSONS, C. PAULE, J. MACKAY and J. BOUSQUET, 2006 Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. BMC Genomics **7:** 174.

PAVY, N., B. PELGAS, S. BEAUSEIGLE, S. BLAIS, F. GAGNON *et al.*, 2008 Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics **9:** 17.

PICOULT-NEWBERG, L., T. E. IDEKER, M. G. POHL, S. L. TAYLOR, M. A. DONALDSON *et al.*, 1999 Mining SNPs from EST databases. Genome Research **9:** 167-174.

PINDO, M., S. VEZZULLI, G. COPPOLA, D. A. CARTWRIGHT, A. ZHARKIKH *et al.*, 2008 SNP high-throughput screening in grapevine using the SNPlex genotyping system. BMC Plant Biology **8:** 12.

POMPANON, F., A. BONIN, E. BELLEMAIN and P. TABERLET, 2005 Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics **6:** 847-846.

RAFALSKI, A., 2002a Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology **5:** 94-100.

RAFALSKI, J. A., 2002b Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Science **162:** 329-333.

RICE, K. M., and P. HOLMANS, 2003 Allowing for genotyping error in analysis of unmatched case-control studies. Annals of Human Genetics **67:** 165-174.

ROSTOKS, N., L. RAMSAY, K. MACKENZIE, L. CARDLE, P. R. BHAT *et al.*, 2006 Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proceedings of the National Academy of Sciences **103:** 18656-18661.

SAUNDERS, I. W., J. BROHEDE and G. N. HANNAN, 2007 Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. Genomics **90:** 291-296.

SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN *et al.*, 2003 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Research **13:** 1250-1257.

SHEN, R., J.-B. FAN, D. CAMPBELL, W. CHANG, J. CHEN *et al.*, 2005 High-throughput SNP genotyping on universal bead arrays. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis **573:** 70-82.

SLATE, J., J. GRATTEN, D. BERALDI, J. STAPLEY, M. HALE *et al.*, 2009 Gene mapping in the wild with SNPs: guidelines and future directions. Genetica **136:** 97-107.

SOBEL, E., J. C. PAPP and K. LANGE, 2002 Detection and integration of genotyping errors in statistical genetics. The American Journal of Human Genetics **70:** 496-508.

SOBRINO, B., M. BRIÓN and A. CARRACEDO, 2005 SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Science International **154:** 181-194.

SYVÄNEN, A. C., 2005 Toward genome-wide SNP genotyping. Nature genetics **37:** 5-10.

WANG, S., Z. SHA, T. S. SONSTEGARD, H. LIU, P. XU *et al.*, 2008 Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics **9:** 450.

Supplementary Table 1: List of the 41 candidate genes used for the *in vitro*-SNPs detection and associated projects.

Project	Gene	Description
BRG 2002	*AGP-like (AGP1)*	Arabinogalactan-like
ANR GenoQB	*BOTERO*	Katanin p60
BRG 2002 / Treesnips / Digenfor / Evoltree JERA 2	*C4H-2*	Trans-cinnamate-4-hydroxylase 2
BRG 2002 / Evoltree JERA 2	*CAD*	Cinnamyl-alcohol-dehydrogenase
BRG 2002 / Treesnips / Digenfor / Evoltree JERA 2	*CCoAOMT*	Caffeoyl-CoA 3-O-methyltransferase
GEMINI	*CesA3*	Cellulose Synthase
GEMINI	*CesA4*	Cellulose Synthase
GEMINI	*CesA7*	Cellulose Synthase
ANR GenoQB	*CHS*	Chalcone Synthase
ANR GenoQB	*Cobra1*	GPI-anchored protein
ANR GenoQB	*Cobra2*	GPI-anchored protein
ANR GenoQB	*Cobra3*	GPI-anchored protein
Treesnips / Digenfor	*Dhn1*	Dehydrin
Treesnips / Digenfor	*Dhn2*	Dehydrin
Treesnips / Digenfor	*ERD3*	Early Response to Drought 3
Treesnips / Digenfor	*Glucan*	Glucan-endo-1,3-beta-glucosidase precursor
Treesnips / Digenfor	*GRP3*	Putative Arabinogalactan / Glycin-Rich Protein
ANR GenoQB	*HDZ31*	Homeodomain Leucin Zipper transcription factor
ANR GenoQB	*Korrigan*	Endo-1,4-beta-glucanase
BRG 2002	*LAC-2*	Laccase 2
BRG 2002	*LAC-4*	Laccase 4
ANR GenoQB	*LIM2*	LIM domain protein
Treesnips / Digenfor	*LP3-1*	Water-stress-inducible protein
Treesnips / Digenfor	*LP3-2*	Water-stress-inducible protein
ANR GenoQB	*Myb1*	R2R3-Myb transcription factor
ANR GenoQB	*Myb14*	R2R3-Myb transcription factor
ANR GenoQB	*Myb2*	R2R3-Myb transcription factor
ANR GenoQB	*Myb4*	R2R3-Myb transcription factor
ANR GenoQB	*Myb8*	R2R3-Myb transcription factor
BRG 2002	*PAL1-P24*	Phenylalanine ammonia-lyase
Digenfor / Evoltree JERA 2	*PAL1-P31*	Phenylalanine ammonia-lyase
Digenfor	*PFK*	Phosphofructokinase
Digenfor	*PHY-N*	Phytochrome N
GEMINI	*Pp1*	Glycin rich protein homolog
Treesnips / Digenfor	*PP2C*	Protein phosphatase 2C
GEMINI	*Pp4*	ACC oxidase
Treesnips / Digenfor / Evoltree JERA 2	*PR-AGP4*	Putative Arabinogalactan / Proline Rich Protein
Treesnips / Digenfor	*RD21*	Cysteine protease
ANR GenoQB	*SCL1*	Scarecrow-like transcription factor
Digenfor	*SPS*	Sucrose Phosphate Synthase
ANR GenoQB	*THE1*	Protein kinase

Projects acronyms:

ANR GenoQB	*GNP05013C supported by the ANR Genoplante*
BRG 2002	*Bureau des Ressources Génétiques (French National Genetic Resources Institution)*
Digenfor	*Tri027 Trilateral project France (Genoplante) / Germany (GABI) / Spain (MCyT)*
Evoltree Jera 2	*Evoltree European funded Network of Excellence*
GEMINI	*QLRT-1999-00942 supported by the European Union*
Treesnips	*QLRT-2001-01973 supported by the European Union*

Supplementary Table 2: List of the 184 *in vitro*-SNPs and their frequencies in the total sequencing dataset (including Aquitaine), in the Aquitaine sequencing dataset, and in the genotyped samples.

			Total dataset used for *in vitro*-SNP detection (including Aquitaine sequences)		**Aquitaine dataset used for *in vitro*- SNP detection**		
Gene	SNP Name	Functionality Score	Number of sequences	MAF[a]	Number of sequences	MAF[a]	MAF[a] in the 456 genotyped samples
AGP1	AGP1.26	0.645	26	46%	13	38%	38%
BOTERO	BOTERO.202	0.97	33	24%	33	24%	23%
BOTERO	BOTERO.77	0.906	28	25%	28	25%	23%
C4H2	C4H2.158	0.511	35	8%	9	0%	Failed[b]
C4H2	C4H2.1782	0.793	10	42%	0	na[d]	5%
C4H2	C4H2.1860	0.737	10	42%	0	na[d]	6%
C4H2	C4H2.315	0.633	50	29%	23	17%	Failed[b]
C4H2	C4H2.399	0.904	50	24%	23	13%	Failed[b]
C4H2	C4H2.507	0.553	50	20%	23	4%	Failed[b]
C4H2	C4H2.570	0.665	50	17%	23	0%	Monomorphic[c]
C4H2	C4H2.684	0.42	50	42%	23	9%	Failed[b]
CAD	CAD.1046	0.816	60	46%	32	31%	36%
CAD	CAD.1114	0.948	60	15%	32	3%	5%
CAD	CAD.118	0.862	60	27%	0	na[d]	Failed[b]
CAD	CAD.1244	0.932	60	46%	32	34%	Failed[b]
CAD	CAD.1440	0.993	60	10%	18	0%	Failed[b]
CAD	CAD.259	0.895	60	7%	32	44%	40%
CAD	CAD.455	0.748	60	36%	32	3%	6%
CAD	CAD.660	0.909	60	43%	32	31%	36%
CCoAomt	CCoAomt1.1249	0.88	80	2%	55	0%	Monomorphic[c]
CCoAomt	CCoAomt1.1426	0.931	80	22%	55	2%	3%
CCoAomt	CCoAomt1.159	0.908	50	10%	23	0%	Monomorphic[c]
CCoAomt	CCoAomt1.2223	0.844	20	50%	0	na[d]	28%
CCoAomt	CCoAomt1.2415	0.62	20	40%	0	na[d]	23%
CCoAomt	CCoAomt1.244	0.654	50	10%	24	42%	33%
CCoAomt	CCoAomt1.2972	0.96	20	50%	0	na[d]	18%
CCoAomt	CCoAomt1.404	0.904	50	20%	25	8%	14%
CCoAomt	CCoAomt1.66	0.615	30	10%	23	39%	39%
CCoAomt	CCoAomt1.806	0.501	60	3%	35	0%	Monomorphic[c]
CCoAomt	CCoAomt1.873	0.944	60	23%	33	24%	16%
CCoAomt	CCoAomt1.980	0.916	60	5%	35	0%	Monomorphic[c]
CesA3	CesA3.190	0.53	47	21%	18	28%	28%
CesA3	CesA3.54	0.646	40	5%	15	13%	9%
CesA4	CesA4.319	0.921	28	4%	7	0%	Failed[b]
CesA7	CesA7.493	0.651	na[d]	na[d]	na[d]	na[d]	2%
CHS	CHS.252	0.991	37	3%	37	3%	Monomorphic[c]
CHS	CHS.651	0.733	46	39%	46	39%	Monomorphic[c]
CHS	CHS.99	0.902	32	31%	32	31%	41%
COBRA1	COBRA1.1054	0.925	39	5%	39	5%	2%
COBRA1	COBRA1.184	0.989	40	13%	40	13%	11%
COBRA1	COBRA1.413	0.973	40	28%	40	28%	17%
COBRA1	COBRA1.477	0.844	40	30%	40	30%	16%
COBRA1	COBRA1.595	0.883	37	16%	37	16%	5%
COBRA1	COBRA1.749	0.614	35	23%	35	23%	11%
COBRA2	COBRA2.35	0.995	31	16%	31	16%	11%
COBRA2	COBRA2.399	0.861	32	16%	32	16%	11%
COBRA2	COBRA2.501	0.73	32	31%	32	31%	33%
COBRA2	COBRA2.603	0.831	32	16%	32	16%	11%
COBRA3	COBRA3.142	0.993	37	27%	37	27%	Failed[b]
COBRA3	COBRA3.202	0.674	37	27%	37	27%	Failed[b]
COBRA3	COBRA3.309	0.937	37	11%	37	11%	11%
COBRA3	COBRA3.423	0.894	39	10%	39	10%	10%
COBRA3	COBRA3.505	0.995	39	10%	39	10%	11%
COBRA3	COBRA3.583	0.79	39	26%	39	26%	17%
COBRA3	COBRA3.644	0.81	39	13%	39	13%	8%
Dhn1	Dhn1.187	0.727	135	36%	50	25%	44%
Dhn1	Dhn1.304	0.49	135	35%	50	25%	Failed[b]
Dhn1	Dhn1.413	0.675	135	2%	50	0%	1%

Dhn1	Dhn1.612	0.729	135	40%	49	43%	33%
Dhn1	Dhn1.689	0.718	135	41%	43	25%	Failed[b]
Dhn2	Dhn2.191	0.435	56	43%	11	45%	Failed[b]
Dhn2	Dhn2.264	0.981	56	43%	11	45%	21%
Dhn2	Dhn2.321	0.981	56	43%	11	45%	20%
Dhn2	Dhn2.410	0.704	56	13%	11	27%	Failed[b]
Dhn2	Dhn2.479	0.704	56	13%	11	27%	Failed[b]
ERD3	ERD3.170	0.907	110	11%	11	18%	Monomorphic[c]
ERD3	ERD3.42	0.974	110	16%	11	0%	Failed[b]
Glucan	Glucan.172	0.916	133	37%	35	11%	Failed[b]
Glucan	Glucan.900	0.848	133	2%	29	0%	Monomorphic[c]
GRP3	GRP3.162	0.542	70	23%	38	3%	2%
GRP3	GRP3.304	0.903	70	35%	42	24%	Failed[b]
GRP3	GRP3.402	0.946	65	8%	43	0%	Monomorphic[c]
HDZ31	HDZ31.1324	0.971	40	3%	40	3%	1%
HDZ31	HDZ31.136	0.937	40	40%	40	40%	47%
HDZ31	HDZ31.1471	0.934	40	40%	40	40%	47%
HDZ31	HDZ31.1632	0.933	40	3%	40	3%	0.1%
HDZ31	HDZ31.2138	0.941	40	3%	40	3%	1%
HDZ31	HDZ31.2268	0.831	40	40%	40	40%	47%
HDZ31	HDZ31.2776	0.973	40	5%	40	5%	Monomorphic[c]
HDZ31	HDZ31.287	0.851	40	40%	40	40%	48%
HDZ31	HDZ31.3140	0.631	40	40%	40	40%	47%
HDZ31	HDZ31.3292	0.801	39	38%	39	38%	48%
HDZ31	HDZ31.409	0.755	40	40%	40	40%	47%
Korrigan	Korrigan.1205	0.95	36	25%	36	25%	28%
Korrigan	Korrigan.1272	0.871	36	47%	36	47%	46%
Korrigan	Korrigan.1395	0.937	39	23%	39	23%	27%
Korrigan	Korrigan.1499	0.935	50	18%	50	18%	26%
Korrigan	Korrigan.1687	0.954	53	0%	53	0%	0.4%
Korrigan	Korrigan.1801	0.919	49	8%	49	8%	14%
Korrigan	Korrigan.2019	0.72	51	4%	51	4%	0.1%
Korrigan	Korrigan.2103	0.663	50	46%	50	46%	46%
Korrigan	Korrigan.2387	0.689	40	48%	40	48%	Failed[b]
Korrigan	Korrigan.255	0.946	24	8%	24	8%	0.2%
Korrigan	Korrigan.2646	0.565	39	18%	39	18%	28%
Korrigan	Korrigan.2722	0.891	39	49%	39	49%	47%
Korrigan	Korrigan.2892	0.409	35	46%	35	46%	Failed[b]
Korrigan	Korrigan.418	0.768	39	49%	39	49%	46%
Korrigan	Korrigan.546	0.934	39	18%	39	18%	26%
Korrigan	Korrigan.635	0.973	39	21%	39	21%	28%
Korrigan	Korrigan.758	0.853	38	3%	38	3%	0.2%
Korrigan	Korrigan.83	0.921	7	29%	7	29%	28%
Korrigan	Korrigan.987	0.968	37	19%	37	19%	26%
LAC2	LAC2.635	0.622	15	7%	6	17%	Monomorphic[c]
LAC2	LAC2.816	0.89	15	7%	6	0%	Monomorphic[c]
LAC4	LAC4.189	0.813	68	10%	33	12%	Monomorphic[c]
LAC4	LAC4.496	0.971	68	12%	33	18%	Failed[b]
LIM2	LIM2.1100	0.806	39	36%	39	36%	33%
LIM2	LIM2.1616	0.618	39	23%	39	23%	Failed[b]
LIM2	LIM2.2005	0.887	39	23%	39	23%	32%
LIM2	LIM2.2372	0.998	27	22%	27	22%	31%
LIM2	LIM2.933	0.441	39	26%	39	26%	Failed[b]
LP3-1	LP3.1.176	0.958	111	61%	11	27%	Failed[b]
LP3-1	LP3.1.250	0.842	111	71%	11	45%	Failed[b]
LP3-1	LP3.1.324	0.937	111	70%	11	45%	46%
LP3-3	LP3.3.298	0.845	105	11%	11	0%	2%
LP3-3	LP3.3.43	0.764	105	16%	11	9%	Failed[b]
Myb1	Myb1.1189	0.685	40	45%	40	45%	22%
Myb1	Myb1.294	0.96	40	45%	40	45%	22%
Myb1	Myb1.710	0.946	40	45%	40	45%	21%
Myb1	Myb1.906	0.949	40	45%	40	45%	21%
Myb14	Myb14.136	0.957	29	24%	29	24%	Failed[b]
Myb14	Myb14.258	0.996	29	3%	29	3%	2%
Myb14	Myb14.373	0.992	29	3%	29	3%	Failed[b]
Myb14	Myb14.449	0.727	29	10%	29	10%	26%
Myb14	Myb14.587	0.675	29	24%	29	24%	Failed[b]

Myb14	Myb14.74	0.818	29	10%	29	10%	27%
Myb14	Myb14.822	0.583	29	10%	29	10%	Failed[b]
Myb2	Myb2.111	0.914	40	13%	40	13%	22%
Myb2	Myb2.1136	0.691	40	5%	40	5%	5%
Myb2	Myb2.1553	0.912	39	5%	39	5%	4%
Myb2	Myb2.1811	0.942	39	5%	39	5%	5%
Myb2	Myb2.194	0.977	40	18%	40	18%	13%
Myb2	Myb2.337	0.702	40	8%	40	8%	8%
Myb2	Myb2.526	0.953	40	8%	40	8%	8%
Myb2	Myb2.784	0.924	40	8%	40	8%	Failed[b]
Myb2	Myb2.844	0.972	39	28%	39	28%	24%
Myb2	Myb2.942	0.827	40	8%	40	8%	Failed[b]
Myb4	Myb4.119	0.971	10	10%	10	10%	2%
Myb8	Myb8.786	0.805	10	20%	10	20%	17%
PAL1-P24	PAL1.P24.220	0.504	51	0%	39	0%	Failed[b]
PAL1-P24	PAL1.P24.403	0.992	51	49%	39	41%	Failed[b]
PAL1-P24	PAL1.P24.535	0.755	51	18%	39	21%	Failed[b]
PAL1-P31	PAL1.P31.1118	0.599	18	11%	12	8%	15%
PAL1-P31	PAL1.P31.1508	0.898	18	39%	12	50%	35%
PAL1-P31	PAL1.P31.1691	0.982	18	44%	12	34%	Failed[b]
PAL1-P31	PAL1.P31.1919	0.992	19	32%	12	34%	Failed[b]
PAL1-P31	PAL1.P31.2090	0.773	19	26%	12	34%	Failed[b]
PAL1-P31	PAL1.P31.2264	0.931	7	14%	0	na[d]	Failed[b]
PAL1-P31	PAL1.P31.256	0.674	7	29%	0	na[d]	0.4%
PAL1-P31	PAL1.P31.36	0.65	4	25%	0	na[d]	Monomorphic[c]
PAL1-P31	PAL1.P31.371	0.939	7	29%	0	na[d]	Monomorphic[c]
PAL1-P31	PAL1.P31.704	0.892	19	16%	0	na[d]	1%
PFK	PFK.203	0.602	51	10%	na[d]	na[d]	4%
PFK	PFK.267	0.707	51	2%	0	na[d]	Monomorphic[c]
PFK	PFK.39	0.874	51	6%	na[d]	na[d]	4%
PHY-N	PHY.N.25	0.976	99	27%	10	20%	7%
PHY-N	PHY.N.258	0.935	99	27%	10	20%	Monomorphic[c]
Pp1	Pp1.104	0.704	22	5%	12	0%	Failed[b]
Pp1	Pp1.312	0.621	23	39%	13	31%	15%
Pp4	Pp4.233	0.991	32	9%	10	20%	Failed[b]
PR-AGP4	PR.AGP4.1021	0.844	183	2%	55	4%	3%
PR-AGP4	PR.AGP4.1165	0.568	150	28%	59	12%	7%
PR-AGP4	PR.AGP4.1303	0.743	175	11%	60	7%	Failed[b]
PR-AGP4	PR.AGP4.132	0.757	175	8%	63	0%	Monomorphic[c]
PR-AGP4	PR.AGP4.1376	0.953	150	40%	62	18%	11%
PR-AGP4	PR.AGP4.34	0.618	117	4%	59	2%	Failed[b]
PR-AGP4	PR.AGP4.802	0.951	175	19%	63	17%	11%
PR-AGP4	PR.AGP4.913	0.772	188	5%	59	2%	Monomorphic[c]
RD21	RD21.123	0.893	103	26%	11	0%	Monomorphic[c]
RD21	RD21.566	0.891	103	15%	11	36%	19%
RD21	RD21.630	0.598	103	39%	11	9%	Failed[b]
SCL1	SCL1.1434	0.984	40	13%	40	13%	Monomorphic[c]
SCL1	SCL1.1749	0.957	40	3%	40	3%	7%
SCL1	SCL1.228	0.966	41	7%	41	7%	0.3%
SCL1	SCL1.318	0.86	41	12%	41	12%	4%
SCL1	SCL1.431	0.895	41	7%	41	7%	7%
SCL1	SCL1.570	0.886	41	12%	41	12%	Failed[b]
SCL1	SCL1.663	0.972	41	12%	41	12%	5%
SCL1	SCL1.823	0.989	41	12%	41	12%	Failed[b]
SPS	SPS.258	0.628	38	34%	na[d]	na[d]	Failed[b]
SPS	SPS.407	0.722	38	8%	na[d]	na[d]	Failed[b]
THE1	THE1.271	0.859	33	42%	33	42%	Failed[b]
THE1	THE1.390	0.914	33	42%	33	42%	50%
THE1	THE1.74	0.684	32	50%	32	50%	36%

[a] Minor Allele Frequency
[b] SNP that failed to be genotyped
[c] Monomorphic SNP
[d] Data not available

Genetic parameters of growth and wood chemical-properties in *Pinus pinaster*

Camille Lepoittevin[*], Jean-Pierre Rousseau[†], Audrey Guillemin[‡], Christophe Gauvrit[§], Thomas Sanchez[†], François Besson[†], Denilson Da Silva Perez[‡], François Hubert[*], Luc Harvengt[**], Christophe Plomion[*]

[*] INRA, UMR1202 Biodiversité Gènes & Communautés, F-33610 Cestas, FRANCE

[†] FCBA, Station Sud-Ouest, F-33480 Moulis-en-Médoc, FRANCE

[‡] FCBA, InTechFibres Pôle Nouveaux Matériaux, F-38044 Grenoble, FRANCE

[§] INRA, UE570 Domaine Expérimental de l'Hermitage, F-33610 Cestas, FRANCE

[**] FCBA, Pôle Biotechnologies Sylviculture, F-77370 Nangis, FRANCE

Introduction

Lignin and cellulose are major constituents of wood, and the most abundant biopolymers on Earth. Lignin content largely determines the calorific value of wood as a fuel, but lignin is undesirable in the conversion of wood into pulp, as it causes discoloration and reduces paper brightness upon thermal or light exposure (CHIANG *et al.* 1988). Its removal is a major step in the papermaking process for the so-called "chemical pulps", but its extraction consumes large quantities of chemicals. Conversely, cellulose content is favorably correlated with pulp yield (WALLIS *et al.* 1996; KUBE and RAYMOND 2002). Breeders are thus seeking ways to increase the cellulose to lignin content ratio. Previous studies in various forest tree species showed that lignin and cellulose contents have low to moderate heritabilities, and are strongly negatively correlated at both phenotypic and genetic levels (BAILLERES *et al.* 2002; POT *et al.* 2002; SYKES *et al.* 2006; DA SILVA PEREZ *et al.* 2007), highlighting the possibility to obtain significant genetic gains for the cellulose to lignin content ratio by classical breeding methods. However, chemical-related traits have generally been evaluated in narrow genetic backgrounds such as diallel and factorial mating designs involving a low number of genotypes. Moreover, little is known about their correlations with mandatory selection criteria such as growth and straightness, or with other possible targets for wood end use properties such as wood density, modulus of elasticity or fiber related traits (POT *et al.* 2002; but see DA SILVA PEREZ *et al.* 2007).

In the last decade, wood phenotyping has been enhanced by the development of indirect tools such as the Pilodyn penetrometer and the Resistograph for measuring wood density (WANG *et al.* 1999; BOUFFIER *et al.* 2008), or Near Infrared Spectroscopy (NIRS) for predicting wood chemical and physical properties of a large number of samples by fast optical measurements (reviewed by TSUCHIKAWA 2007). NIRS have been widely used in industrial fields such as food, agriculture, pharmaceutical or chemical. However, its application in the wood and paper science is recent (most of the results have been published in the 90ies) and provides an efficient alternative to costly and time-consuming wet chemistry methods. On the other hand, due to the complexity of NIR spectra, all measurements are calibration-dependent, which prevent generic application (DA SILVA PEREZ *et al.* 2008).

In this paper, we first investigate the performance of NIRS combined to a non-destructive sampling method to assess wood chemical properties in maritime pine (*Pinus pinaster* Ait.), the first conifer species used for reforestation in the southwestern Europe. We then estimate genetic parameters of growth, stem-form and wood chemical-related traits using a progeny-

trial and clonally replicated progenies derived from control crosses, with the aim to assess the relative importance of additive, dominance and epistatic effects for these traits in a material representing a large genetic background. Finally, we estimate the phenotypic and genetic correlations between these traits and discuss the potential impact of breeding on wood chemical properties.

Material and Methods

Plant material

Two experimental trials, Hermitage (trial number 2-44-17) and Vaquey (trial number 33802), located in the Aquitaine region (southwestern France) were examined in this study. Hermitage is a 31-year-old progeny trial of plus-trees phenotypically selected for overall good growth and form in the Aquitaine forest. This material is commonly referred as the first generation breeding population (G0). A total of 261 G0 mother trees were crossed with a pollen mix collected from 28 unrelated G0 trees, resulting in 8,667 individuals distributed in families of 12 to 36 half-sibs. This trial was installed on a humid sandy moor site, using randomized complete block design with 3, 6 or 9 tree row-plots per family. Vaquey is a 13-year-old clonal trial of 189 trees from the second generation breeding population (G1) which were individually selected within full-sib progenies of 77 G0 trees, based on their genetic value for growth and stem straightness. These 189 G1 clones belonged to 78 full-sib families (some G0 parents were used in several crosses) and were replicated three to five times, leading to a total number of 892 trees. This trial was established on a humic podzol moor site using randomized single-tree plots. No field-blocks were *a priori* defined. These two trials will be referred as the Half-Sib Trial for Hermitage (HST) and the Clonal Trial for Vaquey (CT).

Data measurement

All the trees from the two trials were measured for growth related traits, *i.e.* for the HST total height (*H*) and girth at breast height (G*ir*) at 8 years, and for the CT total height (*H*) at 8 years and diameter at breast height (*Diam*) at 13 years. Trees from the HST were also evaluated for stem deviation from verticality (*Str*) at 8 years (this data is given in cm, and increases with the deviation of the tree from verticality).

Chemical characterization of the samples was conducted separately for the two tests by Near Infrared Spectroscopy (NIRS). For the HST, the sampling was carried out at 31 years on 993 trees representing 105 different half-sib families, with 7 to 12 trees per family. We collected

shavings by drilling a 2 cm large and 5 cm deep hole into the tree at breast height. The shavings were dried for 24 hours at 60°C, ground in a SM-100 three-knife mill (Retsch, Haan, Germany) and sieved. Near infrared spectra acquisition was then carried out on the 40-60 mesh sawdust fraction using a MPA spectrometer with integration sphere (Bruker Optics, Ettlingen, Germany). NIRS-partial least square (NIRS-PLSR) calibrations were build using a subset of 98 samples measured for i/ extractives content (*Ext*) using an automatic SoxTec extractor (Foss, Hillrod, Denmark) and an acetone-water extraction sequence (DA SILVA PEREZ *et al.* 2005), ii/ lignin content (*Lign*) based on the Klason method (SCHWANNINGER and HINTERSTOISSER 2002), and iii/ cellulose (*Cell*) and hemicellulose (*Hemi*) contents by HPLC analysis of monosugars (*Mann* for mannose, *Gal* for galactose and *Xyl* for xylose content) after acidic hydrolysis (PULS *et al.* 1995). PLSR was performed according to WORKMAN *et al.* (1996) and MARTENS and NAES (1989) using the OPUS Quant software (Bruker Optics).

The sampling and NIRS-PLSR evaluation in the CT for Klason lignin (*Lign*), cellulose (*Cell*), hemicellulose (*Hemi*), mannose (*Mann*), galactose (*Gal*), and xylose (*Xyl*) contents at 13 years was carried out on extractive-free sawdust obtained from whole disks as described in DA SILVA PEREZ *et al.* (2005).

For both trials, the quality of NIRS-PLSR models was assessed using the rank (number of PLS eigen-vectors used for the regression) and cross-validation results, *i.e.* coefficient of determination (R^2), and root mean square of error of prediction (*RMSCEV*).

Statistical models for genetic parameter estimation

The individual model was used to partition the phenotypic value of each tree in its genetic and environmental components.

The following model was used for the HST:

$$y = Xb + Z_1 a + Z_2 v + e \qquad (1)$$

where $\boldsymbol{y}$ is a vector of observations on a trait, $\boldsymbol{b}$ is a vector of fixed block effects, $\boldsymbol{a}$ is a vector of random genetic effects of individual genotypes, $\boldsymbol{v}$ is a vector of random plot effect (block × half-sib family interactions), $\boldsymbol{e}$ is the vector of residuals, $\boldsymbol{X}$, $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are the incidence matrices linking observations to the effects. The random effects in model (1) were assumed to follow a normal distribution with means and variances defined by:

$$\begin{bmatrix} a \\ v \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A\sigma_a^2 & 0 & 0 \\ 0 & I\sigma_v^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix} \right) \qquad (2)$$

where **0** is a null matrix; ***A*** is the genetic relationship matrix (computed from a pedigree that takes into account all the relationships among individual genotypes) ; ***I*** is the identity matrix, σ_a^2 is the additive genetic variance, σ_v^2 the plot variance, and σ_e^2 is the residual variance. As the variances are assumed to be independent, the phenotypic variance σ_P^2 is defined as:

$$\sigma_P^2 = \sigma_a^2 + \sigma_v^2 + \sigma_e^2 \qquad (3)$$

For the CT, two different models named “full model” and “simplified model” were used. The full model is as follows:

$$y = Z_1 a + Z_2 f + Z_3 c + e \qquad (4)$$

where ***a*** and ***e*** are defined as above, ***f*** is a vector of random full-sib family effects, ***c*** is a vector of random effect of clones within full-sib families, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ and $\boldsymbol{Z_3}$ are the incidence matrices linking the observations in ***y*** to the effects in ***a***, ***f***, and ***c***, respectively. The random effects in the model defined in (4) were assumed to follow normal distributions with means and variances defined by:

$$\begin{bmatrix} a \\ f \\ c \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A\sigma_a^2 & 0 & 0 & 0 \\ 0 & I\sigma_f^2 & 0 & 0 \\ 0 & 0 & I\sigma_c^2 & 0 \\ 0 & 0 & 0 & I\sigma_e^2 \end{bmatrix} \right) \qquad (5)$$

where **0**, **I**, σ_a^2 and σ_e^2 are defined as above (observations on different ramets of a clone were treated as repeated measurements on a single genotype, therefore ***A*** is the matrix of relationships among individual genotypes as mentioned before); σ_f^2 is the non-additive variance between full-sib families, and $\sigma^2{}_c$ is the non-additive variance among clones within full-sib families. This full model was used after excluding from the dataset the families with only one genotype, as they do not allow the estimation of dominance effects (COSTA E SILVA *et al.* 2004). Thus the full model was fitted for 36 families representing 143 clones and 705 phenotypes. The phenotypic variance σ_P^2, the total genetic variance σ_G^2, the dominance genetic variance σ_D^2 and the epistatic genetic variance σ_I^2 were defined as follows:

$$\sigma_P^2 = \sigma_a^2 + \sigma_f^2 + \sigma_c^2 + \sigma_e^2 \qquad (6)$$

$$\sigma_G^2 = \sigma_a^2 + \sigma_f^2 + \sigma_c^2 = \sigma_a^2 + \sigma_D^2 + \sigma_I^2 \qquad (7)$$

$$\sigma_D^2 = 4 \times \sigma_f^2 \qquad (8)$$

$$\sigma_I^2 = \sigma_c^2 - 3 \times \sigma_f^2 \qquad (9)$$

For (8) and (9) we assumed a large, random mating parental population with diploid inheritance and near linkage equilibrium at gene loci affecting the observed traits (COMSTOCK *et al.* 1958; FOSTER and SHAW 1988). As σ_a^2 and σ_f^2 contain portions of epistasis with successively decreasing contributions of interactions involving larger groups of loci, the unbiased estimation of additive and dominance variances assumes that lower-order interloci interactions only represent a small portion of the total epistasis (MULLIN and PARK 1992; WU 1996; COSTA E SILVA *et al.* 2004). Similarly, interactions involving groups of more than two or three loci are assumed in (9), since σ_I^2 contains only a fraction of the total epistasis with a major contribution of high-order interactions. In addition, non-genetic effects introduced by cloning ("C effects") were assumed to be negligible or absent (FOSTER and SHAW 1988; COSTA E SILVA *et al.* 2004).

The simplified model consisted in dropping the family and clonal effects from the full model (4) described above, so that we could use the total dataset to estimate the variance components (78 families, 189 genotypes and 892 phenotypes). In this model, the family structure is ignored although the kinship is accounted for. Dominance and epistatic genetic effects cannot be estimated anymore, therefore only the total genetic variance σ^2_G was estimated as described in (7).

The estimates of the fixed and random effects were obtained by solving Henderson's mixed model equations (HENDERSON 1975) using the average information REML algorithm (GILMOUR *et al.* 1995) implemented in the ASReml v2.0 software (GILMOUR *et al.* 2006). Wald tests as implemented in ASReml (GILMOUR *et al.* 2006) and Likelihood Ratio Tests (LRT) were used to assess the statistical significance of the fixed and random effects, respectively.

Analyses were performed for each trait separately. Phenotypic, genetic and additive coefficients of variation (CV_P, CV_G and CV_A) as well as broad and narrow-sense heritabilities (h^2_{bs} and h^2_{ns}) were defined as follows:

$$CV_P = \frac{\sigma_P(X)}{\overline{X}} \qquad CV_G = \frac{\sigma_G(X)}{\overline{X}} \qquad CV_A = \frac{\sigma_a(X)}{\overline{X}}$$

$$h^2_{bs} = \frac{\sigma_G^2}{\sigma_P^2} \qquad h^2_{ns} = \frac{\sigma_a^2}{\sigma_P^2}$$

where $\overline{X}$ is the mean of the studied trait for the trial considered.

The estimates of the phenotypic (r_P), genetic (r_G) and additive (r_A) correlations between traits were obtained using bivariate extensions of the models previously described, for each pair of traits and each trial. They were evaluated as follows:

$$r_P = \frac{Cov_P(x,y)}{\sqrt{\sigma^2_{Px} \times \sigma^2_{Py}}} \qquad r_G = \frac{Cov_G(x,y)}{\sqrt{\sigma^2_{Gx} \times \sigma^2_{Gy}}} \qquad r_A = \frac{Cov_A(x,y)}{\sqrt{\sigma^2_{ax} \times \sigma^2_{ay}}}$$

where $Cov_P(x,y)$, $Cov_G(x,y)$ and $Cov_A(x,y)$ are the phenotypic, genetic and additive genetic covariances between traits *x* and *y*, respectively. For the CT, the simplified model was used to estimate a total genetic correlations (r_G), including additive, dominance and epistatic effects. The standard errors for variances, heritabilities and correlation coefficients were calculated with ASReml using a standard Taylor series expansion (SORENSEN and GIANOLA 2002; GILMOUR *et al.* 2006).

Results

Near infrared spectroscopy calibrations

The performances of NIRS-PLSR calibrations for the prediction of chemical composition of maritime pine wood were assessed by cross-validations and are shown in Table 1. Extractive, lignin, cellulose, mannose and galactose contents could be predicted by NIRS with a quite good accuracy in one or both trials, with low *RMSECV* and determination coefficients (R^2) ranging from 0.7 to 0.96. The calibrations built for the CT generally performed better than those built for the HST, displaying higher R^2 and lower rank and *RMSECV*. Calibrations showing a cross-validation R^2 below 0.7 (*i.e.* cellulose and mannose contents for the HST, and hemicellulose and xylose contents of both trials) were not used for genetic parameters estimations because of the low accuracy of their predictions.

Table 1: Cross-validation performances of the NIRS-PLSR calibrations for the prediction of wood chemical composition. Data for the CT taken from DA SILVA PEREZ *et al.* (2005). Abbreviations as described in the Material and Methods section.

Trait	Trial	Range (%)		Rank	Cross validation	
		Min	Max		R^2	RMSECV (%)
Ext	Hermitage	3.3	10.3	4	0.70	0.76
Lign	Hermitage	21.7	30.6	4	0.91	0.40
	Vaquey	25.8	32.7	3	0.96	0.40
Cell	Hermitage	36.8	47.4	3	0.53	1.51
	Vaquey	39.9	51.1	2	0.92	0.80
Hemi	Hermitage	20.1	28.2	8	0.34	1.22
	Vaquey	23.3	28.4	5	0.53	0.73
Mann	Hermitage	8.6	13.8	7	0.60	0.62
	Vaquey	14.5	20	1	0.87	0.55
Gal	Hermitage	1.4	6.1	10	0.82	0.49
	Vaquey	2	10.9	7	0.94	0.72
Xyl	Hermitage	3.7	8.2	8	0.30	0.86
	Vaquey	9.1	13.9	5	0.63	0.50

Genetic parameters

Estimates of variance components and genetic parameters obtained from the analysis of the two trials are presented in Table 2. Growth and stem straightness, the main selection criteria of the maritime pine breeding program, showed moderate to high coefficients of phenotypic variation (CV_P=15-25% for growth and 54% for straightness), in agreement with previous results (COSTA and DUREL 1996; POT *et al.* 2002). With the exception of *Ext* and *Gal*, CV_P for chemical-related traits were low (< 7%), consistently with most of the results reported in the literature (SILVA *et al.* 1998; NYAKUENGAMA *et al.* 1999; POT *et al.* 2002; HANNRUP *et al.* 2004). The fixed block effect and random plot effect in the HST were highly significant for all traits except for *Lign*, for which only the plot effect was significant. For both trials, σ_a^2 and σ_G^2 were highly significant with very low *P*-values (< 0.1%).

In the CT, clonal progenies of controlled crosses were used to estimate non-additive genetic variances. For *H*, σ_f^2 was accurately estimated and much lower than σ_a^2, suggesting a low contribution of dominance effects on this trait. For *Diam*, both $\sigma^2{}_f$ and $\sigma^2{}_c$ were accurately estimated and again non-significant and much lower than $\sigma^2{}_a$, also indicating strong additive control. For the other traits, σ_f^2 and σ_c^2 were non-significant, with values similar to or greater than σ_a^2, and standard errors of the same magnitude as the estimates. In these cases, our data-set did not allow obtaining accurate estimates of non-additive variances, thus the simplified model was preferable to estimate heritabilities and correlations.

Broad-sense heritabilities in the CT were low to moderate, ranging from 0.17-0.23 for chemical-related traits to 0.24-0.38 for growth. Narrow-sense heritabilities estimated in the HST ranged from 0.15 to 0.55, and were generally higher than broad-sense heritabilities estimated on the CT for the same traits using the full and simplified models (except *Gir* which was twice lower).

Genetic correlations

Phenotypic and genetic correlations between traits are shown in Table 3 for the HST and Table 4 for the CT. In the HST, *H* and *Gir* were strongly correlated at the phenotypic and genetic levels (r_P = 0.82, r_A = 0.86), as expected, but weakly correlated to stem straightness at the genetic level (r_A = 0.15 and 0.24, respectively). In both trials, chemical-related traits were significantly correlated, showing in most cases higher values for genetic than phenotypic correlations. The highest correlation observed was between *Lign* and *Cell* in the CT (r_G = -0.98 with a standard error of 0.01). We did not observe any significant genetic correlation between growth and chemical-related traits, although the phenotypic correlations were generally significant in the HST.

Table 2: Variance components and genetic parameters estimated for Hermitage (HST) and Vaquey (CT) trials. Standard errors are indicated in brackets. The significance of random effects is indicated after each variance estimator: ^{ns} not significant at the 5% threshold; * significant at the 5% threshold; ** significant at the 1% threshold; *** significant at the 0.1% threshold.

Traits		Trials	Variance components						Genetic parameters				
			σ^2_v	σ^2_a	σ^2_f	σ^2_c	σ^2_G	σ^2_e	h^2_{bs}	h^2_{ns}	CV_P	CV_G	CV_A
Growth	*H*	Hermitage	2640 (172)***	2426*** (541)	-	-	-	3344	-	0.29 (0.06)	15.80%	-	8.50%
	H	Vaquey (full model)	-	35.09*** (15.44)	5.41E-4^{ns} (1.2E-4)	3.88^{ns} (9.86)	38.97*** (8.24)	160.18	0.20 (0.03)	0.18 (0.07)	21.36%	9.45%	8.97%
	H	Vaquey (simplified model)	-	-	-	-	52.58*** (7.94)	162.4	0.24 (0.03)	-	22.50%	11.13%	-
	Gir	Hermitage	6.6 (0.63)***	7.79*** (1.89)	-	-	-	36.05	-	0.15 (0.04)	25.20%	-	9.90%
	Diam	Vaquey (full model)	-	435.2*** (85.6)	2.86E-6^{ns} (1.7E-7)	2.63E-6^{ns} (1.57E-7)	435.2*** (85.6)	925.1	0.32 (0.05)	0.32 (0.05)	18.10%	10.24%	10.24%
	Diam	Vaquey (simplified model)	-	-	-	-	569.7*** (90.5)	913	0.38 (0.04)	-	19.19%	11.89%	-
Form	*Str*	Hermitage	0.72 (0.19)***	7.03*** (0.96)	-	-	-	16.49	-	0.29 (0.04)	54%	-	29.10%
Chemical composition	*Ext*	Hermitage	0.33 (0.07)***	0.58*** (0.21)	-	-	-	0.72	-	0.35 (0.12)	19%	-	11.30%
	Lign	Hermitage	ns	0.16*** (0.06)	-	-	-	0.5	-	0.25 (0.09)	3%	-	1.51%
	Lign	Vaquey (full model)	-	0.05*** (0.15)	0.08^{ns} (0.07)	0.08^{ns} (0.08)	0.21*** (0.05)	0.8	0.21 (0.04)	0.05 (0.15)	3.33%	1.52%	0.75%
	Lign	Vaquey (simplified model)	-	-	-	-	0.25*** (0.05)	0.84	0.23 (0.04)	-	3.45%	1.65%	-
	Cell	Vaquey (full model)	-	0.18*** (0.34)	0.17^{ns} (0.16)	0.05^{ns} (0.19)	0.39*** (0.11)	1.84	0.17 (0.04)	0.08 (0.15)	3.45%	1.44%	0.97%
	Cell	Vaquey (simplified model)	-	-	-	-	0.42*** (0.10)	1.92	0.18 (0.04)	-	3.53%	1.50%	-
	Mann	Vaquey (full model)	-	0.09*** (0.11)	0.05^{ns} (0.05)	0.04^{ns} (0.06)	0.17*** (0.04)	0.66	0.21 (0.04)	0.10 (0.13)	6.93%	3.14%	2.24%
	Mann	Vaquey (simplified model)	-	-	-	-	0.15*** (0.04)	0.68	0.19 (0.04)	-	6.95%	3.00%	-
	Gal	Hermitage	0.15*** (0.05)	0.64*** (0.18)	-	-	-	0.37	-	0.55 (0.14)	33.09%	-	24.60%
	Gal	Vaquey (full model)	-	0.02*** (0.23)	0.18* (0.13)	0.10* (0.13)	0.31*** (0.08)	1.27	0.20 (0.04)	0.01 (0.15)	30.63%	13.55%	3.59%
	Gal	Vaquey (simplified model)	-	-	-	-	0.28*** (0.07)	1.35	0.17 (0.04)	-	31.04%	12.90%	-

- : not available

ns: the effect was not significant and not considered in the model.

Table 3: Phenotypic and genetic correlations in the HST (r_P below and r_A above the diagonal, respectively). Standard errors are indicated in brackets below each coefficient. The significance of correlation coefficients is indicated after each estimator: ^{ns} not significant at the 5% threshold; * significant at the 5% threshold; ** significant at the 1% threshold; *** significant at the 0.1% threshold.

	H	***Gir***	***Str***	***Ext***	***Lign***	***Gal***
H	-	0.86*** (0.04)	0.15^{ns} (0.12)	-0.46^{ns} (0.26)	-0.16^{ns} (0.25)	0.12^{ns} (0.23)
Gir	0.82*** (0.004)	-	0.24* (0.12)	-0.10^{ns} (0.28)	-0.26^{ns} (0.26)	0.18^{ns} (0.25)
Str	0.27*** (0.01)	0.36*** (0.01)	-	-0.01^{ns} (0.21)	0.23^{ns} (0.19)	0.20^{ns} (0.17)
Ext	0.09* (0.04)	0.14*** (0.03)	0.10*** (0.03)	-	-0.57** (0.20)	-0.96*** (0.13)
Lign	-0.26*** (0.03)	-0.26*** (0.03)	-0.02^{ns} (0.03)	-0.32*** (0.03)	-	0.85*** (0.12)
Gal	-0.19*** (0.04)	-0.19*** (0.03)	0.04^{ns} (0.04)	0.03^{ns} (0.03)	0.57*** (0.02)	-

Table 4: Phenotypic and genetic correlations in the CT (r_P below and r_G above the diagonal, respectively). Standard errors are indicated in brackets below each coefficient. The significance of correlation coefficients is indicated after each estimator: ^{ns} not significant at a 5% threshold; * significant at a 5% threshold; ** significant at a 1% threshold; *** significant at a 0.1% threshold.

	Diam	***Lign***	***Cell***	***Mann***	***Gal***
Diam	-	-0.007^{ns} (0.13)	-0.12^{ns} (0.14)	0.09^{ns} (0.14)	0.18^{ns} (0.14)
Lign	-0.04^{ns} (0.04)	-	-0.98*** (0.01)	-0.87*** (0.05)	0.64*** (0.08)
Cell	-0.02^{ns} (0.04)	-0.94*** (0.004)	-	0.89*** (0.06)	-0.79*** (0.06)
Mann	0.05^{ns} (0.04)	-0.66*** (0.02)	0.49*** (0.03)	-	-0.75*** (0.09)
Gal	0.05^{ns} (0.04)	0.82*** (0.01)	-0.86*** (0.01)	-0.61*** (0.03)	-

Discussion

General considerations

Genetic parameters of growth, stem-form and wood chemical-properties as well as genetic correlations between these traits were estimated in two *Pinus pinaster* trials located in the SouthWest of France, a half-sib trial (HST) and a clonal trial (CT). Near Infrared Spectrometry (NIRS) calibrations were built to rapidly assess wood-chemical properties, either at 31-year old from raw sawdust obtained from drill shavings for the HST, or at 13-year old from extractive-free sawdust obtained from whole-disk samples in the CT. Lignin content was predicted by NIRS with a quite good precision in both trials. While NIRS-PLSR models for cellulose and most of monosugars contents were successfully obtained in the CT, no calibration could be achieved using the samples of the HST. We used the individual model to subdivide the phenotypic value of each tree in its genetic additive, genetic non-additive and environmental components. Additive variances were significantly different from zero, leading to low to moderate heritabilities (0.15-0.55). For growth traits non-additive variances were accurately estimated and much lower than the additive component, suggesting low dominance and / or epistatic effects. For most wood quality traits, non-additive genetic variances were not significant and their standard errors were of the same magnitude as variance estimates, which suggest that our dataset was not large enough to accurately estimate non-additive effects. Narrow-sense heritabilities in the HST were generally higher than broad-sense heritabilities in the CT for the same traits. In both trials, chemical-related traits were significantly genetically correlated, but not correlated to growth at the genotypic level.

Rapid wood-quality assessment techniques

A prerequisite of a tree-breeding program focusing on wood quality resides in the ability to measure whole-tree properties, which often implies the destructive sampling of disks to provide a bulk sample that represents the whole tree. Destructive sampling is extremely time-consuming and owing to practical constraints, the number of trees that can be sampled is limited. An alternative to destructive sampling is to take an increment core or, as done in this study for the HST, to collect shavings from drilling a hole in the stem of a standing tree and assume that the results are indicative of the whole-tree properties. However, this assumption can lead to erroneous results: indeed, we know that wood chemical composition shows significant radial variations, with mature wood containing more cellulose and less lignin than

juvenile wood (ZOBEL and SPRAGUE 1998). The proportions of mature wood and juvenile wood in the powder obtained from increment cores or drill shavings are different from the true proportions of a stem section, and thus do not represent the whole tree. To illustrate this sampling bias, AUGUSTO and BERT (2005) showed that using increment cores for determining nutrient contents generally leads to significant underestimations in sapwood and overestimations in heartwood. Since the error associated to the drilling sampling method increases with the drilling-depth, restricting the sampling to the last rings should reduce this bias, but with the drawback of measuring only mature wood properties. According to micro-density profiles of the HST (data from BOUFFIER *et al.* 2008), 5 cm deep drillings corresponded to 9 (for fast growing-trees) to 24 (for slow-growing trees) growth rings (Figure 1). Assuming a transition between juvenile and mature wood in maritime pine to occur around the 10th to the 12th growth ring (ZOBEL *et al.* 1972; DUMAIL *et al.* 1998), ~75% of the samples should mainly contain mature wood.

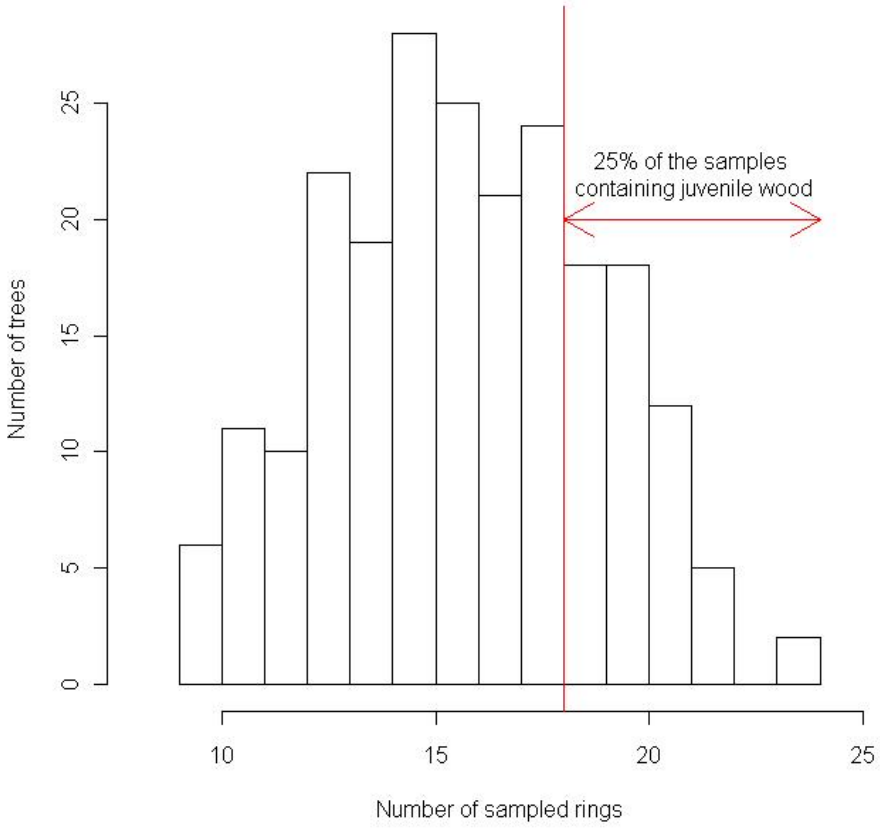


Figure 1: Distribution of the number of rings sampled for a 5 cm depth drilling, estimated from X-ray profiles of 221 trees sampled at 29 years old on the HST (from data provided by Laurent Bouffier). Ring-widths are larger for fast-growing trees than for slow-growing trees. A 5 cm depth drilling corresponds to 9 rings for the larger tree (on the left of the *x*-axis), *versus* 24 rings for the smaller one (on the right of the *x*-axis).

Even if we roughly avoided the sampling bias due to radial ontogenic effects, the number of rings sampled greatly varied depending on the trees, thus a bias due to different annual climatic conditions is not to be excluded. To our knowledge, the only published study using drill shavings to estimate conifer-wood chemical properties concluded that this sampling method gave errors that were too large for practical purpose (JONES *et al.* 2008). These authors used young trees (13-year old loblolly pines), drilled holes from the pith to the cambium, and obtained poor NIRS-PLSR calibrations. Our sampling method was quite different and gave rather good calibrations, especially for lignin content. Further investigations are nevertheless required to assess the representativeness of our samples.

NIRS analysis relies on developing a calibration that relates the NIR spectra of a large number of samples to their known chemical or physical properties. Once the calibration has been set, NIR spectra can be used to predict these properties with the advantages of minimal sample preparation, rapid acquisition times, non-contact and non-destructive spectral acquisition (SO *et al.* 2004). In the present study, two different types of samples were used for NIR spectra acquisition: raw sawdust for the HST samples, and extractive-free sawdust for the CT samples. They led to different calibration performances, the latter giving higher NIRS-PLSR prediction accuracy for all the analyzed traits. Such result has previously been reported for eucalyptus, with the improvement of calibration models for lignin content and sugar monomer composition after elimination of wood extractives (BAILLERES *et al.* 2002; DA SILVA PEREZ *et al.* 2008). To explain such difference these authors hypothesized that polyphenolic compounds of wood extracts alter the lignin absorption bands located in the same spectral zones. Despite this result, NIRS calibration for lignin content based on non-extracted powder was of high quality in the present study, as it provided good predictions (*RMSECV* of 0.4%). However, this was not the case for cellulose and most of monosugars calibrations, which showed very low performances on raw sawdust. Extractives are a complex group of cell wall chemicals mainly consisting of fats, fatty acids, fatty alcohols, phenols, terpenes, steroids, resin acids, waxes and many other minor organic compounds (ROWELL 2005). Their interaction with cellulose and monosugars absorption bands is likely, although to our knowledge it has never been explicitly reported. The differences between calibrations based on raw or extractive-free sawdust could also be due to other factors such as the sampling age. Indeed, the CT was sampled at 13-year old while the HST was sampled at 31-year old. Sampling method (disks *versus* drill shavings) may also be involved, but this seems unlikely as poor sugars calibrations have already been reported for young loblolly pine raw-sawdust ground from disks, while lignin calibration performed well (JONES *et al.* 2008),

or for young maritime pine raw sawdust ground from whole stems (data not shown). Actually, these technical issues for assessing cellulose content can be easily avoided providing that lignin content is accurately predicted, because lignin and cellulose contents are strongly negatively correlated at both phenotypic and genetic levels, as found in the CT (r_P=-0.94 and r_G=-0.98, with standard errors of 0.004 and 0.01, respectively). This cellulose-lignin balance has been largely studied in conifers (CHANTRE *et al.* 2002; POT *et al.* 2002; SEWELL *et al.* 2002; POT *et al.* 2006; DA SILVA PEREZ *et al.* 2007), and confirmed by transgenic studies in various plant species where a cross-talk between both pathway has been suggested (HU *et al.* 1999; BAUCHER *et al.* 2003; PARK *et al.* 2004).

Genetic effects and heritabilities

Genetic variance among individuals is usually partitioned into additive and non-additive components, the latter being a generic term including allelic interactions within gene (dominance) and interactions between genes at two or more loci (epistasis) (FALCONER *et al.* 1996). Selection response for traits mainly controlled by non-additive effects is not predictable from parental performance and non-cumulative over generations. Although they have been intensively studied in the last decades, contradictory results have been published for growth and stem form, showing either low (FOSTER and SHAW 1988; KUSNANDAR *et al.* 1998; ISIK *et al.* 2003; BALTUNIS *et al.* 2007; DA SILVA PEREZ *et al.* 2007) or moderate (COTTERILL *et al.* 1987; POT *et al.* 2002; HANNRUP *et al.* 2004) non-additive control. This heterogeneity could be due to the low sample sizes in the diallel and factorial designs employed. While these mating schemes are useful for the partitioning of the genetic variance into its sub-components, they generally involve a low number of parents (the number of parents ranged from 6 to 38 in the above mentioned studies). However, large samples are generally needed to accurately estimate non-additive genetic variances (FOSTER and SHAW 1988; COSTA E SILVA *et al.* 2004). Recently, COSTA E SILVA *et al.* (2004) provided an elegant model to estimate dominance and epistatic effects using clonally replicated progenies of eucalyptus: with 153 full-sib families originating from 79 parents, they showed that dominance and epistatic effects accounted for 0 to 4% and 0.4% of the phenotypic variance in stem diameter, and for 0% and 5% of the phenotypic variance for pilodyn penetration (an indirect measure of wood density), respectively. In the present study, using the same model but a lower number of parents and families (47 and 36, respectively), we showed that non-additive effects for *H* as well as dominance effects for *Diam* were accurately estimated and not significantly different from 0. However, our experimental design did not allow to

accurately estimate non-additive variances for chemical-related traits, which again shows that large data sets are required to reduce the standard errors of these components.

Interestingly, for the same traits narrow-sense heritabilities (h^2_{ns}) estimated on the HST were generally higher than broad-sense heritabilities (h^2_{bs}) estimated on the CT, whereas the opposite was expected, or at least equality between both estimates if non-additive effects and genotype-by-environment interactions were low. A first hypothesis is that the error variance in the CT is larger than in the HST, as it includes a part of environmental variance due to the absence of blocks in the trial, thus h^2_{bs} may be under-estimated. For example, when the block effect was omitted from model (1) in the HST, h^2_{ns} was under-estimated by ~ 10% for form and chemical-related traits and not significantly different from 0 for growth, because error and plot variances were greatly inflated. This illustrates the importance of controlling for environmental effects in large forestry trials, and also shows that growth traits are more sensitive to micro-environmental conditions than chemical-related traits. A spatial analysis of the CT would certainly improve heritability estimates. A second hypothesis is that ontogenic effects could also be partly responsible for heritability variations between both trials, as chemical-related traits were measured at different ages: 13- *versus* 31-year old for the CT and the HST, respectively. This bias could even be amplified by our sampling method in the HST, as previously discussed. COSTA and DUREL (1996) and DANJON (1994) showed that heritability of growth-related traits increased with age in maritime pine, and wood density also seems to follow this trend (BOUFFIER *et al.* 2008; GASPAR *et al.* 2008). However, to our knowledge time trend analysis of chemical-related traits heritability has not yet been reported. Further analyses will be necessary to confirm or reject this hypothesis. A third hypothesis is that half-sib families in the HST included full-sib pairs of genotypes that were not accounted for in our pedigree matrix, leading to an over-estimation of h^2_{ns}. Recently, JOAO GASPAR *et al.* (2009) showed using molecular markers that the coancestry coefficient of maritime pine open-pollinated families was 0.130, instead of the 0.125 value usually assumed for half-sib families. This under-estimation of offspring relationships did not dramatically change heritability values in the studied trial, but simulations showed that the presence of around 10% of full-sibs in half-sib families is enough to produce heritability overestimation by ~10% (JOÃO GASPAR *et al.* 2009). In our case, the progenies of the HST were not obtained by open pollination but by a controlled cross involving a pollen mix collected from 28 trees (*i.e.* a “polycross” trial), thus the family coancestry coefficient could be higher than that of open-pollinated families, resulting in significant bias in heritabilities. Only the genotyping of the half-sib families and their parents with highly polymorphic

markers such as microsatellites, as in JOAO GASPAR *et al.* (2009), would allow quantifying this bias and obtaining more accurate variance estimates.

Perspectives for breeding applications

We found low to moderate heritabilities for all the studied traits, and the phenotypic variation observed for growth and stem form were generally higher than that observed for chemical-related traits, suggesting that higher genetic gains can be expected for growth and straightness than for chemical properties, in agreement with previous findings (POT *et al.* 2002; DA SILVA PEREZ *et al.* 2007). However, given the volume of wood processed each year by the pulp industry and its predicted increase, even slight modifications of chemical-related trait performances would be of commercial value. Indeed, an increase in cellulose to lignin ratio would be advantageous in decreasing energy and chemical consumptions (RYDHOLM 1965) as well as increasing pulp yield. A recent economic study on loblolly pine showed that a 4% decline in lignin content could yield a 4% increase in mill profits, but the advantages of further decline were mitigated substantially by the loss of mill biopower generation and the consequent need to purchase more power from the open market (PETER *et al.* 2007). Following POT *et al.* (2002) and DA SILVA PEREZ *et al.* (2007), a ~1% decrease in lignin content could be achieved in one maritime pine breeding generation with a 5% selection rate. This breeding program is in its third generation of selection, therefore significant genetic gain can be expected by classical breeding. However, genetic gains are inevitably slow because of the long generation time and the fact that many traits can only be scored at rotation age. Further analyses of chemical-related traits are thus needed to assess the extent of age-age correlations and genotype-by-environment interactions, to estimate their trade-offs with other wood quality-related traits or the relative importance of additive and non-additive effects in their genetic control. In this context, the sampling method and NIRS technique used in the present study may offer some advantages when compared with conventional destructive sampling and wet chemistry methods. This rapid assessment technique gives estimates with reasonable precision for the lignin content, hence contributing to a reduction in measurement costs.

Acknowledgements

We thank Laurent Bouffier and Pauline Garnier-Géré for their help using the individual model and their helpful comments on the manuscript. We also thank Pierre Gardère, Maëlys Kerdraon and Guillaume Kubinski for their help in processing the wood samples. This research was supported by grants from ANR Genoplante (GenoQB, GNP05013C), from the European Union (GEMINI, QLRT-1999-00942) and from the Aquitaine Region. Phenotyping of the half-sib trial was performed at the GenoBois Facility of Pierroton (Cestas). C. Lepoittevin was supported by CIFRE contract between FCBA and INRA. F. Hubert was funded by the EVOLTREE Network of Excellence (http://www.evoltree.org).

Author's contributions

CP and LH organized the funding of the study. The wood sampling was performed by CL, CG, JPR, TS. Samples processing and NIRS acquisition were performed by CL, JPR, FB and FH. Chemical analyses were performed by AG and DDSP. NIRS calibrations were built by CL, AG and DDSP. CL performed the statistical analyses and wrote the paper with the help of CP.

References

AUGUSTO, L., and D. BERT, 2005 Estimating stemwood nutrient concentration with an increment borer: a potential source of error. Forestry **78:** 451-455.

BAILLERES, H., F. DAVRIEUS and F. H. PICHAVANT, 2002 Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. Annals of Forest Science **59:** 479-490.

BALTUNIS, B. S., D. A. HUBER, T. L. WHITE, B. GOLDFARB and H. E. STELZER, 2007 Genetic analysis of early field growth of loblolly pine clones and seedlings from the same full-sib families. Canadian Journal of Forest Research **37:** 195-205.

BAUCHER, M., C. HALPIN, M. PETIT-CONIL and W. BOERJAN, 2003 Lignin: Genetic engineering and impact on pulping. Critical Reviews in Biochemistry and Molecular Biology **38:** 305-350.

BOUFFIER, L., C. CHARLOT, A. RAFFIN, P. ROZENBERG and A. KREMER, 2008 Can wood density be efficiently selected at early stage in maritime pine (*Pinus pinaster* Ait.)? Annals of Forest Science **65:** 106-106.

CHANTRE, G., P. ROZENBERG, V. BAONZA, N. MACCHIONI, A. LE TURCQ *et al.*, 2002 Genetic selection within Douglas fir (*Pseudotsuga menziesii*) in Europe for papermaking uses. Annals of Forest Science **59:** 583-593.

CHIANG, V. L., R. J. PUUMALA, H. TAKEUCHI and R. E. ECKERT, 1988 Comparison of softwood and hardwood kraft pulping. Tappi journal **71:** 173-176.

COMSTOCK, R. E., T. KELLEHER and E. B. MORROW, 1958 Genetic variation in an asexual species, the garden strawberry. Genetics **43:** 634-646.

COSTA E SILVA, J., N. M. G. BORRALHO and B. M. POTTS, 2004 Additive and non-additive genetic parameters from clonally replicated and seedling progenies of *Eucalyptus globulus*. Theoretical and Applied Genetics **108:** 1113-1119.

COSTA, P., and C. E. DUREL, 1996 Time trends in genetic control over height and diameter in maritime pine. Canadian Journal of Forest Research **26:** 1209-1217.

COTTERILL, P. P., C. A. DEAN and G. VAN WYK, 1987 Additive and dominance genetic effects in *Pinus pinaster*, *P. radiata* and *P. elliottii* and some implications for breeding strategy. Silvae Genetica **36:** 221-231.

DA SILVA PEREZ, D., A. GUILLEMAIN, P. ALAZARD, C. PLOMION, P. ROZENBERG *et al.*, 2007 Improvement of *Pinus pinaster* Ait elite trees selection by combining near infrared spectroscopy and genetic tools. Holzforschung **61:** 611-622.

DA SILVA PEREZ, D., A. GUILLEMAIN, G. CHANTRE, P. ALAZARD, A. ALVES *et al.*, 2005 Improvement of wood, pulp and paper quality of maritime pine (*Pinus pinaster* Ait) by combining rapid assessment techniques and genetics., pp. 207:214 in *International symposium on wood, fibre and pulping chemistry*. Appita Inc, Auckland.

DA SILVA PEREZ, D., A. GUILLEMAIN and M. PETIT CONIL, 2008 Some factors influencing the prediction of wood and pulp properties by near infrared spectroscopy. O Papel **69:** 60-75.

DANJON, F., 1994 Heritabilities and genetic correlations for estimated growth curve parameters in maritime pine. Theoretical and Applied Genetics **89:** 911-921.

DUMAIL, J. F., P. CASTÉRA and P. MORLIER, 1998 Hardness and basic density variation in the juvenile wood of maritime pine. Annals of Forest Science **55:** 911-923.

FALCONER, D. S., T. F. C. MACKAY and M. BULMER, 1996 *Introduction to quantitative genetics*. Longman New York.

FOSTER, G. S., and D. V. SHAW, 1988 Using clonal replicates to explore genetic variation in a perennial plant species. Theoretical and Applied Genetics **76:** 788-794.

GASPAR, M. J., J. L. LOUZADA, M. E. SILVA, A. AGUIAR and M. H. ALMEIDA, 2008 Age trends in genetic parameters of wood density components in 46 half-sibling families of *Pinus pinaster*. Canadian Journal of Forest Research **38:** 1470-1477.

GILMOUR, A. R., B. J. GOGEL, B. R. CULLIS and R. THOMPSON, 2006 ASReml user guide release 2.0. VSN International Ltd., Hemel Hempstead, UK.

GILMOUR, A. R., R. THOMPSON and B. R. CULLIS, 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51:** 1440-1450.

HANNRUP, B., C. CAHALAN, G. CHANTRE, M. GRABNER, B. KARLSSON *et al.*, 2004 Genetic parameters of growth and wood quality traits in *Picea abies*. Scandinavian Journal of Forest Research **19:** 14-29.

HENDERSON, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics **31:** 423-447.

HU, W. J., S. A. HARDING, J. LUNG, J. L. POPKO, J. RALPH *et al.*, 1999 Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnology **17:** 808-812.

ISIK, F., B. LI and J. FRAMPTON, 2003 Estimates of additive, dominance and epistatic genetic variances from a clonally replicated test of loblolly pine. Forest Science **49:** 77-88.

JOÃO GASPAR, M., A. I. DE-LUCAS, R. ALÍA, J. ALMIRO PINTO PAIVA, E. HIDALGO *et al.*, 2009 Use of molecular markers for estimating breeding parameters: a case study in a *Pinus pinaster* Ait. progeny trial. Tree Genetics & Genomes **5:** 609-616.

JONES, P. D., L. R. SCHIMLECK, R. F. DANIELS, A. CLARK and R. C. PURNELL, 2008 Comparison of *Pinus taeda* L. whole-tree wood property calibrations using diffuse reflectance near infrared spectra obtained using a variety of sampling options. Wood Science and Technology **42:** 385-400.

KUBE, P. D., and C. A. RAYMOND, 2002 Prediction of whole-tree basic density and pulp yield using wood core samples in *Eucalyptus nitens*. Appita journal **55:** 43-48.

KUSNANDAR, D., N. W. GALWEY, G. L. HERTZLER and T. B. BUTCHER, 1998 Age trends in variances and heritabilities for diameter and height in maritime pine (*Pinus pinaster* Ait.) in Western Australia. Silvae Genetica **47:** 136-141.

MARTENS, H., and T. NAES, 1989 *Multivariate calibration*. John Wiley & Sons.

MULLIN, T. J., and Y. S. PARK, 1992 Estimating genetic gains from alternative breeding strategies for clonal forestry. Canadian Journal of Forest Research **22:** 14-23.

NYAKUENGAMA, J. G., R. EVANS, C. MATHESON, D. SPENCER and P. VINDEN, 1999 Wood quality and quantitative genetics of *Pinus radiata* D. Don: fibre traits and wood density. Appita Journal **52:** 348-350.

PARK, Y. W., K. I. BABA, Y. FURUTA, I. IIDA, K. SAMESHIMA *et al.*, 2004 Enhancement of growth and cellulose accumulation by overexpression of xyloglucanase in poplar. FEBS Letters **564:** 183-187.

PETER, G. F., D. E. WHITE, R. D. L. TORRE and R. SINGH, 2007 The value of forest biotechnology: a cost modelling study with loblolly pine and kraft linerboard in the southeastern USA. International Journal of Biotechnology **9:** 415-435.

POT, D., G. CHANTRE, P. ROZENBERG, J. C. RODRIGUES, G. L. JONES *et al.*, 2002 Genetic control of pulp and timber properties in maritime pine (*Pinus pinaster* Ait.). Annals of Forest Science **59:** 563-575.

POT, D., J.-C. RODRIGUES, P. ROZENBERG, G. CHANTRE, J. TIBBITS *et al.*, 2006 QTLs and candidate genes for wood properties in maritime pine (*Pinus pinaster* Ait.). Tree Genetics & Genomes **2:** 10-24.

PULS, J., T. D. GLAWISCHNIG, A. HERRMANN, A. BORCHMANN and B. SAAKE, 1995 Comparative investigations for quantitative determinations of wood sugars, pp. 503–

510 in *8th International Symposium on Wood and Pulping Chemistry*, Helsinki, Finland.

ROWELL, R. M., 2005 *Handbook of wood chemistry and wood composites*. Taylor & Francis; CRC Press.

RYDHOLM, S. A., 1965 *Pulping processes*. Interscience New York.

SCHWANNINGER, M., and B. HINTERSTOISSER, 2002 Klason lignin: modifications to improve the precision of the standardized determination. Holzforschung **56:** 161-166.

SEWELL, M. M., M. F. DAVIS, G. A. TUSKAN, N. C. WHEELER, C. C. ELAM *et al.*, 2002 Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). II. Chemical wood properties. Theoretical and Applied Genetics **104:** 214-222.

SILVA, J. C. E., H. WELLENDORF and H. PEREIRA, 1998 Clonal variation in wood quality and growth in young Sitka spruce (*Picea sitchensis* (Bong.) Xarr.): Estimation of quantitative genetic parameters and index selection for improved pulpwood. Silvae Genetica **47:** 20-33.

SO, C. L., B. K. VIA, L. H. GROOM, L. R. SCHIMLECK, T. F. SHUPE *et al.*, 2004 Near infrared spectroscopy in the forest products industry. Forest Products Journal **54:** 6-16.

SORENSEN, D., and D. GIANOLA, 2002 *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer.

SYKES, R., B. LI, F. ISIK, J. KADLA and H. M. CHANG, 2006 Genetic variation and genotype by environment interactions of juvenile wood chemical properties in *Pinus taeda* L. Annals of Forest Science **63:** 897-904.

TSUCHIKAWA, S., 2007 A review of recent near infrared research for wood and paper. Applied Spectroscopy Reviews **42:** 43-71.

WALLIS, A. F. A., R. H. WEARNE and P. J. WRIGHT, 1996 Analytical characteristics of plantation eucalypt woods relating to kraft pulp yields. Appita Journal **49:** 427-432.

WANG, T., S. N. AITKEN, P. ROZENBERG and M. R. CARLSON, 1999 Selection for height growth and Pilodyn pin penetration in lodgepole pine: effects on growth traits, wood properties, and their relationships. Canadian Journal of Forest Research **29:** 434-445.

WORKMAN, J. J., P. R. MOBVLEY, B. R. KOWALSKI and R. BRO, 1996 Review of chemometrics applied to spectroscopy: 1985-95. Applied Spectroscopy Reviews **31:** 73-124.

WU, R. L., 1996 Detecting epistatic genetic variance with a clonally replicated design: models for low vs high-order nonallelic interaction. Theoretical and Applied Genetics **93:** 102-109.

ZOBEL, B. J., R. C. KELLISON, M. F. MATTHIAS and A. V. HATCHER, 1972 Wood density of the southern pines. North Carolina Agricultural Experiment Station, Tech. Bul **208:** 56.

ZOBEL, B. J., and J. R. SPRAGUE, 1998 *Juvenile wood in forest trees*. Springer.

Association mapping for growth and wood chemical-properties in the *Pinus pinaster* Aquitaine breeding population

Introduction

Research for dissecting molecular basis of quantitative variation have historically focused on quantitative trait locus (QTL) mapping, which allows detecting associations between phenotypic variability and genomic regions identified by molecular markers. QTL studies depend on the development of suitable segregating populations (*e.g.* F_2, backcrosses or near-isogenic lines), which can be a serious limitation for species showing high levels of inbreeding depression and long generation times such as forest trees. Moreover, QTLs effects often depend on the narrow genetic background in which they have been detected, which also severely limits their application in breeding. Association mapping (or linkage disequilibrium mapping) is a more recent approach which can be applied at the population-level. Taking advantage of the great number of recombination events that have accumulated across past generations, its resolution is much higher than in QTL studies since only markers in strong linkage disequilibrium (LD) with a causative allele will be declared significant. This allows for a mapping scale finer than 1 cM (XIONG and GUO 1997), and in some cases for the detection of causative polymorphisms (INGVARSSON *et al.* 2008; FOURNIER-LEVEL *et al.* 2009; THUMMA *et al.* 2009).

The efficiency of association studies largely depends on LD patterns and extent in the population considered, but also on the ability to distinguish LD due to physical linkage from LD due to other evolutionary forces (FLINT-GARCIA *et al.* 2003; ABECASIS *et al.* 2005; GUPTA *et al.* 2005). For example, population structure is seen as the most serious systematic bias producing false-positive associations (MARCHINI *et al.* 2004; HIRSCHHORN and DALY 2005). A large effort has thus been made to develop methods for detecting hidden population stratification (PRITCHARD *et al.* 2000a; FALUSH *et al.* 2003; WU *et al.* 2006; HUBISZ *et al.* 2009) and for dealing with any structure in association models (PRITCHARD *et al.* 2000b; PRICE *et al.* 2006; ZHAO *et al.* 2007). The principle is that only associations caused by physical linkage should remain after removing genetic correlations due to subpopulations effects. Similarly, models accounting for familial relatedness have been developed (YU *et al.* 2006; MALOSETTI *et al.* 2007; STICH and MELCHINGER 2009), which allow association mapping to be carried out on breeding populations showing variable inbreeding levels, another cause of spurious associations (VOIGHT and PRITCHARD 2005).

Most coniferous species are considered as good models for association mapping due to their generally high levels of genetic diversity and pollen flow and large population sizes (NEALE and SAVOLAINEN 2004). These life history and mating characteristics *a priori* lead to rapid

LD decay between genes or polymorphisms, low inbreeding and low population structure, which is a favourable situation to avoid false positive associations. In this chapter we report the results of the first association mapping study in maritime pine (*Pinus pinaster* Ait.). This conifer species is of economical importance for timber and pulp production in southwestern Europe (POT *et al.* 2005), where it occurs naturally with a fragmented distribution (GONZALEZ-MARTINEZ *et al.* 2002; BURBAN and PETIT 2003; BUCCI *et al.* 2007) probably leading to smaller effective population sizes than usually recognized for other conifers (see Chapter I). First, SNP data were used to test for putative stratification in the Aquitaine breeding population. Then, two different association mapping approaches using pedigree records to account for inbreeding were used to detect significant associations between SNPs and growth, stem straightness and wood chemical properties.

Materials and Methods

Plant material

Two different populations were examined in this study: i/ 160 plus-trees belonging to the first generation breeding population (G0), resulting from mass selection for overall good growth and form in the forest of South West of France (Aquitaine region), and ii/ 162 trees from the second generation breeding population (G1) resulting from biparental crosses between 77 G0 trees, and individually selected based on their genetic value for growth and stem straightness. G0 and G1 trees were evaluated in the Hermitage progeny trial and the Vaquey clonal trial, respectively, as described in Chapter III. Briefly, the 160 G0 trees were crossed with a pollen mix collected from 28 unrelated G0 trees, resulting in 5,080 progenies distributed in families of 12 to 36 half-sibs. They were installed in a randomized complete block design with 3, 6 or 9 tree plots per family. The 162 G1 trees were cloned and 3 to 5 replicates per clone (768 trees in total) were installed in randomized single-tree plots with no defined field-blocks.

Phenotypic data

G0 progenies in the Hermitage trial were measured for total height (*H*), girth at breast height (G*ir*) and deviation from verticality (*Str*) at 8 years. G1 clones in the Vaquey trial were measured for total height (*H*) at 8 years and diameter at breast height (*Diam*) at 13 years. Wood samples were collected from both trials and indirect chemical characterization was obtained by near infrared spectroscopy (see Chapter III). In the Hermitage trial, 958 G0 progenies (7 to 12 half-sibs by family in 101 families) were measured for extractives (*Ext*)

and lignin (*Lign*) content at 31 years old. In the Vaquey trial, all the G1 replicates were measured for lignin (*Lign*), cellulose (*Cell*), mannose (*Mann*) and galactose (*Gal*) content at 13 years old.

Genotypic data

DNA was extracted from needles of the 160 G0 and 162 G1 trees using Invisorb® Spin Plant Mini Kit (Invitek, Berlin, Germany). Genotyping was conducted with the Illumina GoldenGate technology (Illumina Inc., San Diego, CA, USA) using 184 *in vitro*-SNPs discovered in 40 candidate genes involved in plant cell wall formation or drought stress resistance, and 200 *in silico*-SNPs detected in 146 contigs from the maritime pine EST database, as described in Chapter II. Among these 384 SNPs, 192 (111 *in vitro* SNPs in 32 candidate genes and 81 *in silico* SNPs in 69 contigs) and 186 (106 *in vitro* SNPs in 31 candidate genes and 80 *in silico* SNPs in 69 contigs) were polymorphic in the G0 and G1 trees, respectively.

Population structure

Genetic structure was assessed using the Structure software v2.2 (PRITCHARD *et al.* 2000a; FALUSH *et al.* 2003; FALUSH *et al.* 2007) on SNP genotypic data, first for the 160 G0 trees, and second for 28 unrelated G1 trees. This method assumes that populations are at Hardy-Weinberg equilibrium and that markers come from unlinked or weakly linked loci (FALUSH *et al.* 2003). Prior to the structure analysis, we thus discarded the SNPs significantly departing from Hardy-Weinberg equilibrium, detected using the GenePop software (RAYMOND and ROUSSET 1995). In our dataset, several SNPs within the same fragments were at distance closer than 1000 base-pairs, and therefore could be in strong linkage disequilibrium. We determined a subset of informative or weakly linked SNPs using the H-clust method described in RINALDO *et al.* (2005) and implemented in an R (R_DEVELOPMENT_CORE_TEAM 2009) function available at http://www.wpic.pitt.edu/WPICCompGen/hclust/hclust.htm. This method is based on the squared correlation matrix of genotypic data and allows selection of either unlinked or weakly linked SNPs without prior estimation of the haplotype structure. The cut-off value, which is one minus the maximum level of squared pairwise correlation (r^2) allowed between SNPs, was set to 0.5, leaving a certain amount of association among SNPs which is accounted for in the Structure software method. For both panels (160 G0 and 28 unrelated G1 trees) we performed 10 independent runs of Structure for numbers of groups (*K*-parameter) varying from 1 to 10, with the correlated allele frequencies model, and with burn-

in and run-length periods of 10^6 iterations. The best number of groups *K* was then determined using both the mean likelihood *L(K)* over 10 runs for each *K*, and the Δ*K* criterion of EVANNO *et al.* (2005).

Statistical models

To decrease computation time, association tests were performed on another subset of non-redundant SNPs defined as above with a cut-off value of 0.2 (meaning that the maximum level of r^2 allowed between SNPs was 0.8), successively applied to the 160 G0 trees and the 28 unrelated G1 trees. Two different methods were used to test for genotype-phenotype associations: a two-stage association analysis for the G0 samples, and a one-stage association analysis for the G1 samples.

Two-stage analysis on the G0 samples

Best Linear Unbiased Predictors (BLUPs) were obtained for the 160 G0 trees using their half-sib progenies installed in the Hermitage trial, by analysing each trait separately with model (1) described in Chapter III. These BLUPs were then used in a second step to test their possible association with each SNP marker using the following ANOVA model:

$$B_i = \mu + s_i + e_i \qquad (1)$$

where B_i is the BLUP of the *i*th entry, s_i the effect of the genotypic class at the SNP locus considered, and e_i the residual. Two different models were tested, with s_i representing either three genotypic classes (2 homozygous and one heterozygous) in the codominant model, or an allelic dose effect in the additive model (using a numerical variable taking the values 0, 1 and 2 for the absence, the presence in one copy and the presence in two copies for one of the two alleles, respectively). We assumed that the 160 G0 were unrelated and came from an unstructured population, as shown by the Structure analysis (see Results section). This model was solved using the *SNPassoc* package (GONZALEZ *et al.* 2007) implemented in R. In this package, the statistical significance of a given SNP is obtained by a likelihood ratio test that compares model (1) with a null model which only includes the intercept. When significant associations were observed, the genetic variance associated to the SNP effect was estimated in the codominant model as:

$$\sigma^2_{snp} = \frac{MS_s - MS_e}{k} \qquad (2)$$

where MS_s is the mean square associated with the SNP effect, MS_e is the residual mean square, and k=(N-1)/2, where N is the number of genotypes, taking into account the random

sampling of genotypes in each genotypic class that follows a binomial distribution depending on allelic frequencies (GARNIER-GERE 1992; CHARCOSSET and GALLAIS 1996). Then, the broad-sense heritability of the SNP was estimated as:

$$h_{bs}^2 = \frac{\sigma_{snp}^2}{\sigma_{snp}^2 + \sigma_e^2} \qquad (3).$$

In the additive model, the SNP effect is only additive (FALCONER *et al.* 1996), thus the coefficient of determination (R^2) is also the fraction of additive variance associated to the SNP in the model, *i.e.* the narrow-sense heritability of the SNP (h_{ns}^2).

One-stage analysis on the G1 samples

For the G1 samples, the analyses of phenotypic variation and association were performed in one step using the following model:

$$y = Xs + Zg + e \qquad (4)$$

where ***y*** is a vector of observations on a trait, ***s*** is a vector of fixed SNP effects, ***g*** is a vector of random genetic effects of individual genotypes, ***e*** is the vector of residuals, ***X*** and ***Z*** are the incidence matrices linking observations to the effects. We assumed that the G1 individuals came from an unstructured population (see Results section). The random effects in model (3) were assumed to follow normal distributions with means and variances defined by:

$$\begin{bmatrix} g \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A\sigma_g^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix} \right) \quad (5)$$

where **0** is a null matrix; ***A*** is the genetic relationship matrix (computed from a pedigree that takes into account all the relationships among individual genotypes) ; ***I*** is the identity matrix, σ_g^2 is the genetic variance, and σ_e^2 is the residual variance. We considered both codominant and additive allelic models, as previously described in the two-stage approach. The estimates of the fixed and random effects were obtained by solving Henderson's mixed model equations (HENDERSON 1975) using the average information REML algorithm (GILMOUR *et al.* 1995) implemented in the ASReml v2.0 software (GILMOUR *et al.* 2006). The Wald test as implemented in ASReml was used to assess the statistical significance of SNP effects.

Multiple-testing corrections

To account for multiple testing, we first computed *q*-values (STOREY 2002; STOREY and TIBSHIRANI 2003; DABNEY *et al.* 2009), which measure the significance of an association in terms of the false discovery rate rather than in terms of false positive rate. This method

however assumes that statistical tests are independent, which is not the case if SNPs are in linkage disequilibrium (LD). In our study, the arbitrary choice of informative SNPs with $r^2 < 0.8$ still meant that some markers could be in moderate to high LD. Therefore, we also implemented a permutation test using R software with the *boot* (CANTY and RIPLEY 2009), *SNPassoc* (GONZALEZ *et al.* 2007) and *asreml* (GILMOUR *et al.* 2006) packages: BLUPs or phenotypic values were permuted among individuals while keeping their genotypes fixed, and models (1) and (4) were used to test for false-positive associations. The minimum *P*-value (*P*-value$_{min}$) obtained in each of 1,000 permutations was then used to estimate the *P*-value$_{min}$ empirical distribution. Finally, a *P'*-value was computed as $P(P\text{-value} \leq P\text{-value}_{min})$. This permutation procedure retains the LD structure and thus allows an estimation of the false-positive rate for non-independent observations (HIRSCHHORN and DALY 2005). However, it is computationally demanding, particularly when model (4) is used since solving mixed-model equations has to be repeated one thousand times, which took ~4 hours by trait on a 2.53GHz Intel Core 2 Duo processor.

Results

Population structure

Among the 192 SNPs that were polymorphic in the G0 samples and the 169 SNPs that were polymorphic in the non-inbred G1 samples, 109 and 98 informative SNPs which did not depart significantly from Hardy-Weinberg equilibrium were retained for the structure analysis, respectively. The cut-off value of 0.5 in the H-clust method allowed the selection of the most informative SNP in each fragment. We observed a typical pattern of unstructured population (PRITCHARD *et al.* 2007): plateaus in the estimate of log-likelihood of the data were not observed since the highest likelihood was for K=1 (Figure 1), the proportion of samples assigned to each subpopulation was roughly symmetric, and all individuals were admixed (Figure 2). The EVANNO criterion ΔK (EVANNO *et al.* 2005) was not pertinent as it can only be computed for $K \geq 2$, and thus does not allow comparing the results of K=1 (no stratification) with the other cases. Moreover, ΔK did not show a clear peak for a specific K value (Figures 1B & D).

Selection of markers for association tests

Using the H-clust method, we selected two subsets of 141 and 121 informative SNPs (pairwise $r^2 < 0.8$) for the G0 and G1 samples, respectively (Figure 3). In these subsets ~60%

of the pairwise correlations were below 0.5, indicating a low to moderate level of LD. The allele frequency spectrum for each of the two samples is shown in Figure 4. Minimum allele frequencies (MAFs) were strongly correlated in the G0 and G1 samples (r=0.93), and the 20 SNPs that were polymorphic in the G0 and not in the G1 samples corresponded to rare variants (frequency < 10%, as shown in Figure 4).

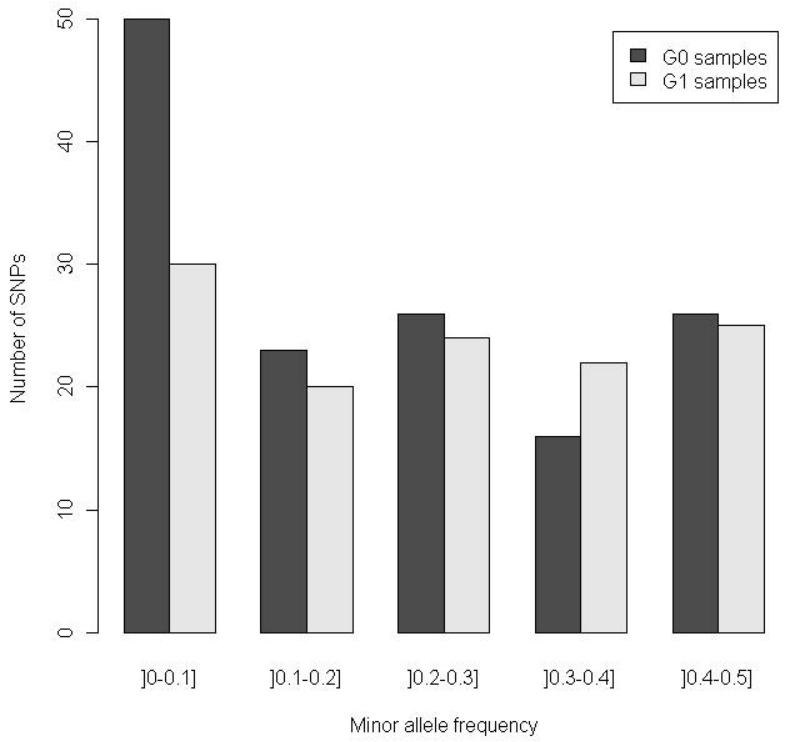


Figure 4: Allele frequency spectrum for the 141 and 121 informative SNPs in the G0 and G1 samples, respectively.

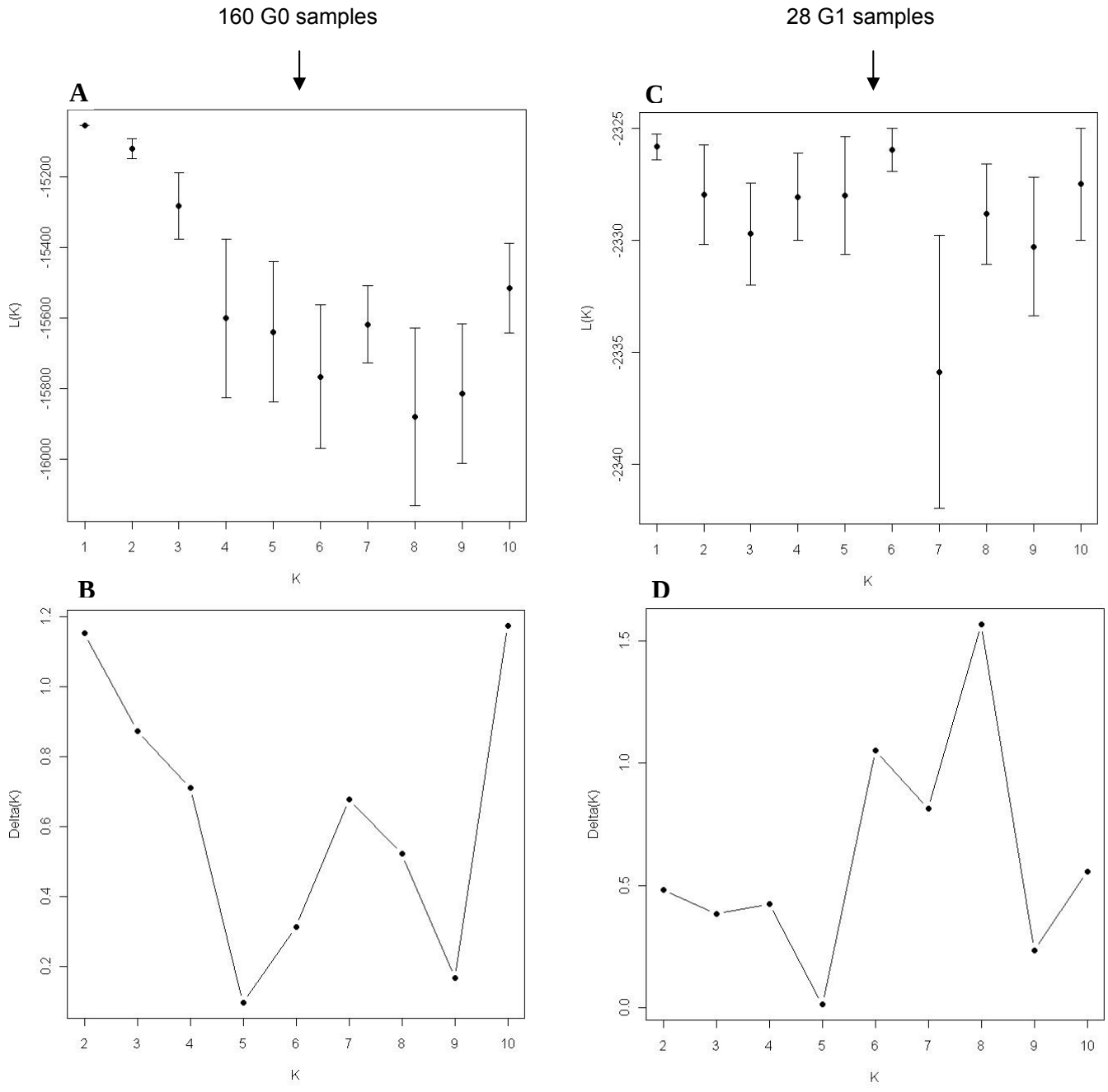


Figure 1: Results of the STRUCTURE analysis on the 160 G0 samples (**A and B**) and on the 28 unrelated G1 samples (**C and D**). **A and C*/*** Mean likelihood *L(K)* and its standard error over 10 runs of the STRUCTURE software for each number of subpopulations *K*. **B and D*/*** EVANNO criterion Δ*K* (EVANNO *et al.* 2005) for each *K* value.

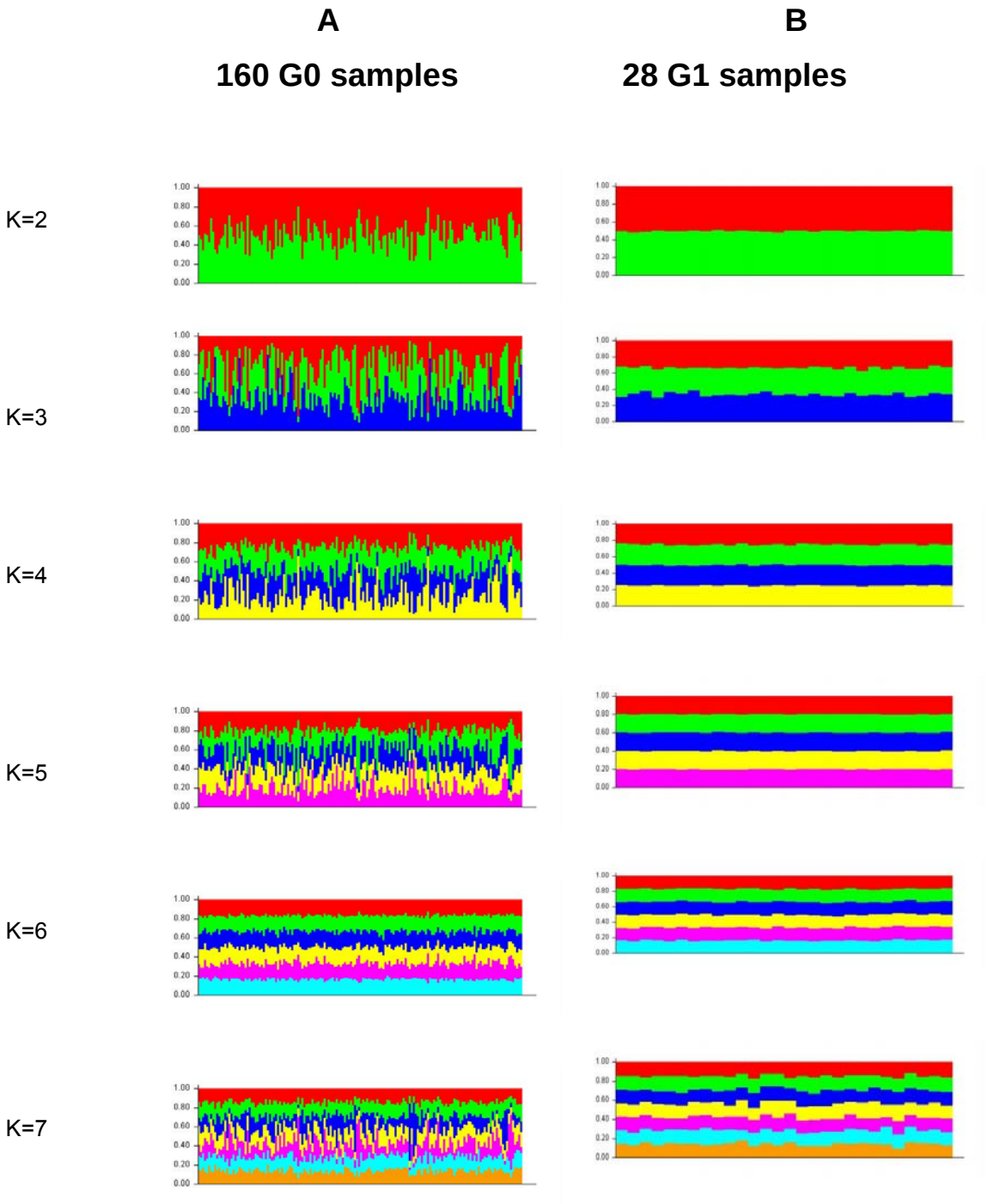


Figure 2: Population structure in the G0 (**A**) and G1 (**B**) non-inbred samples, as shown by the STRUCTURE clustering algorithm. Respectively 109 and 98 SNP informative markers non-departing from Hardy-Weinberg equilibrium were used for the G0 and G1 samples. Each individual is represented by a vertical line which is partitioned into *K* colour-scale segments (*K*=2 to *K*=7) that represent the individual's estimated membership fraction in *K* clusters.

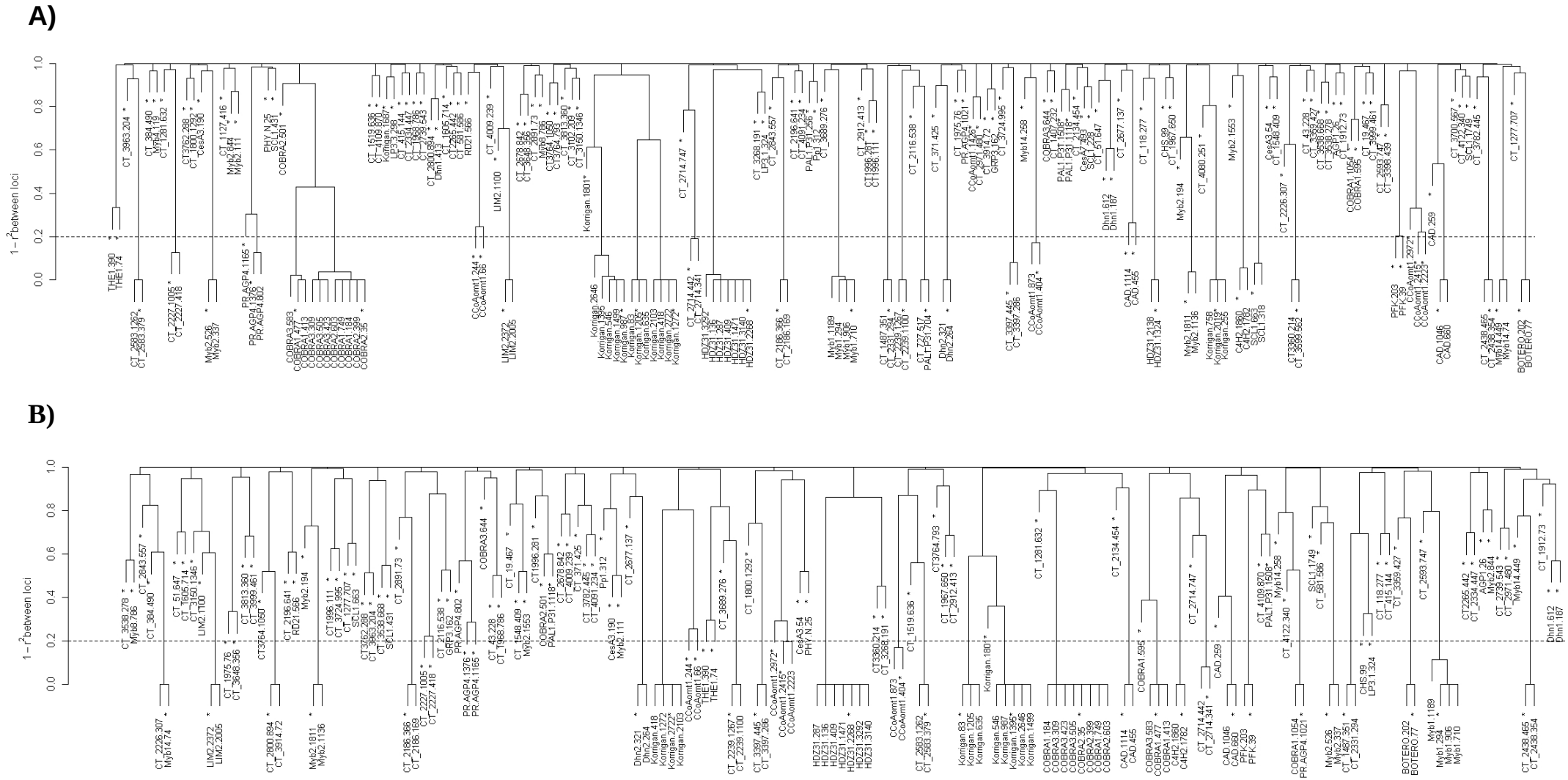


Figure 3: Hierarchical clustering dendrograms obtained by the H-clust method (RINALDO *et al.* 2005) for the 160 G0 samples (**A**) and the 28 unrelated G1 samples (**B**). The cut-off value is indicated by a dashed line and the informative SNPs by a star. In each block, the selected SNP is chosen as the SNP most correlated with all the other SNPs in the block. If several SNPs are equivalent, the one presenting the less missing data is selected.

Statistical tests

In the G0 samples only, the SNP CT_3782.445 was significantly associated with *H* and *Gir* with low *P*-values (< 2.10^{-4}), *q*-values (< 0.01) and *P'*-values (< 0.04) in both the codominant and additive models (Figures 5 and 6). The "T" allele was associated to a decrease in *Gir* (Figure 7) and *H* (data not shown). The association with *Gir* was slightly stronger than with *H* (for example, using the additive model the *P'*-value was 0.001 for *Gir* and 0.01 for *H*), but consistent with the strong correlation between these two traits (genetic and phenotypic correlation coefficients > 0.8, see Chapter III). Broad-sense and narrow-sense heritabilities of CT_3782.445 with *Gir* were similar (~11% and ~10%, respectively), indicating an effect with an additive mode of action. This *in silico*-SNP showed a MAF of 28.3% and no significant departure from Hardy-Weinberg equilibrium in the sample. It was randomly chosen in the maritime pine EST database among SNPs presenting high chances of genotyping success, but the selection was not based on its contig annotation (see Chapter II). It was the only SNP of the contig CT_3782 included in our study, and is likely synonymous according to open reading frame predictions. The other 140 SNPs were not significantly associated with any trait using both the codominant and additive models: the lower *q*-value and *P'*-value observed were above 0.5, indicating that associations based on *P*-values at a 5% threshold were likely due to type-I errors.

The results of association tests in the G1 samples are shown in Figures 8 and 9 for the codominant and additive models, respectively. We first observed that *q*-values for *Diam*, *Cell*, *Gal* and *Lign* in the codominant model did not give good estimates of the False Discovery Rate, since many SNPs showing high and thus non-significant *P*-values also showed very low and highly significant *q*-values (Figure 8). This was probably due to the *P*-value distribution among SNPs for these traits, which did not reach a plateau and thus did not allow for a good estimation of the proportion of null *P*-values (as illustrated for *Diam* in Figure 10) (STOREY and TIBSHIRANI 2003). In this particular case, the *P'*-value was a much better correction for multiple testing. The SNP HDZ31.2268 was significantly associated with *Cell* in the additive model (*P*-value = $2.3.10^{-4}$ and *P'*-value = 0.02), but not in the codominant model where the *P'*-value was much higher (0.11). This marker is a synonymous SNP located in the HDZ31 transcription factor, a candidate gene for wood formation (see Chapter I). Its MAF in the G1 samples was 36.8%, and its genotypic frequencies did not significantly depart from Hardy-Weinberg equilibrium. HDZ31.2268 was in complete linkage disequilibrium with 6 other synonymous SNPs of HDZ31 that were not included in the association analysis (Figure 3B).

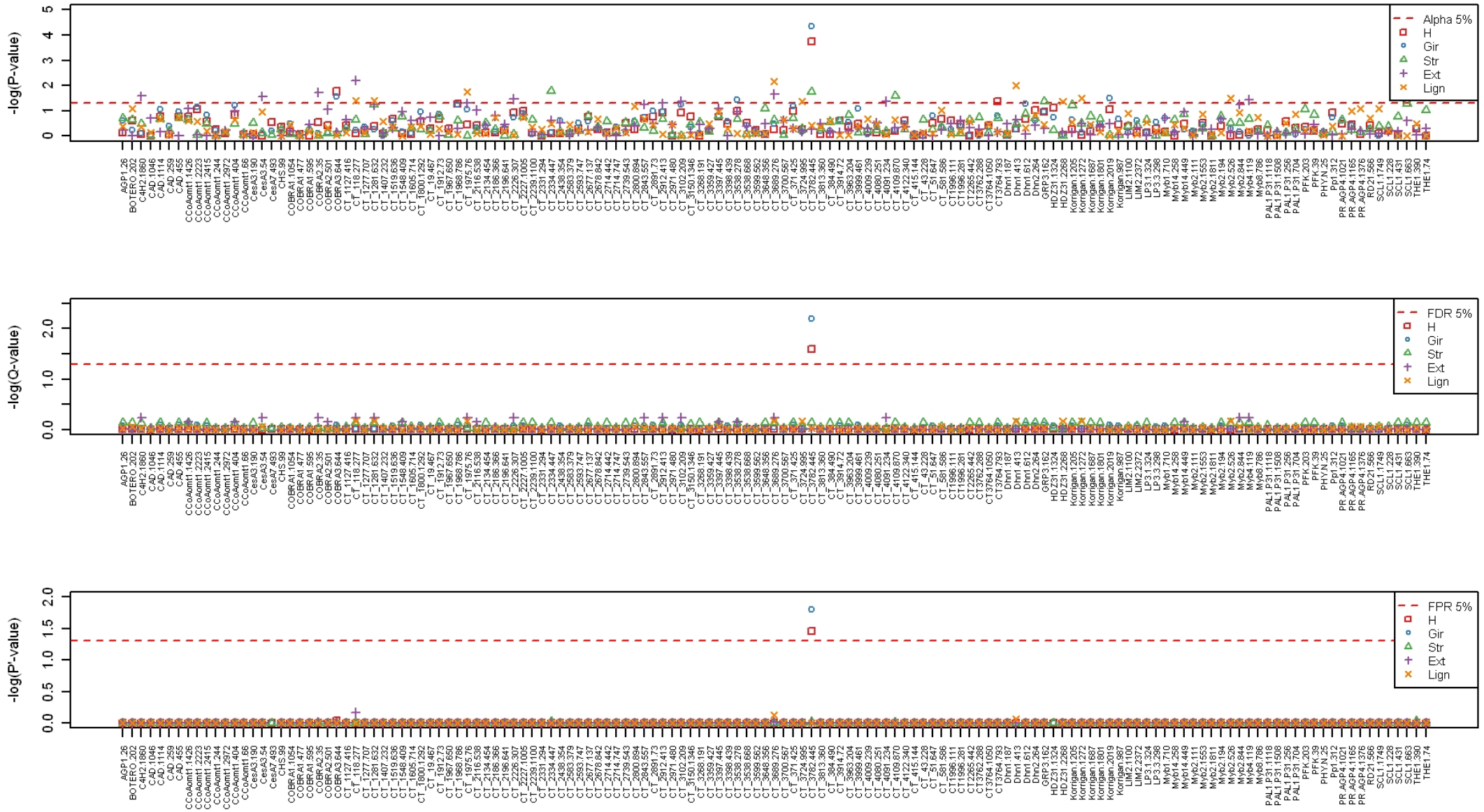


Figure 5: *P*-values, *q*-values and *P'*-values for the SNP effect in the codominant model, for 141 informative SNPs in the G0 samples.

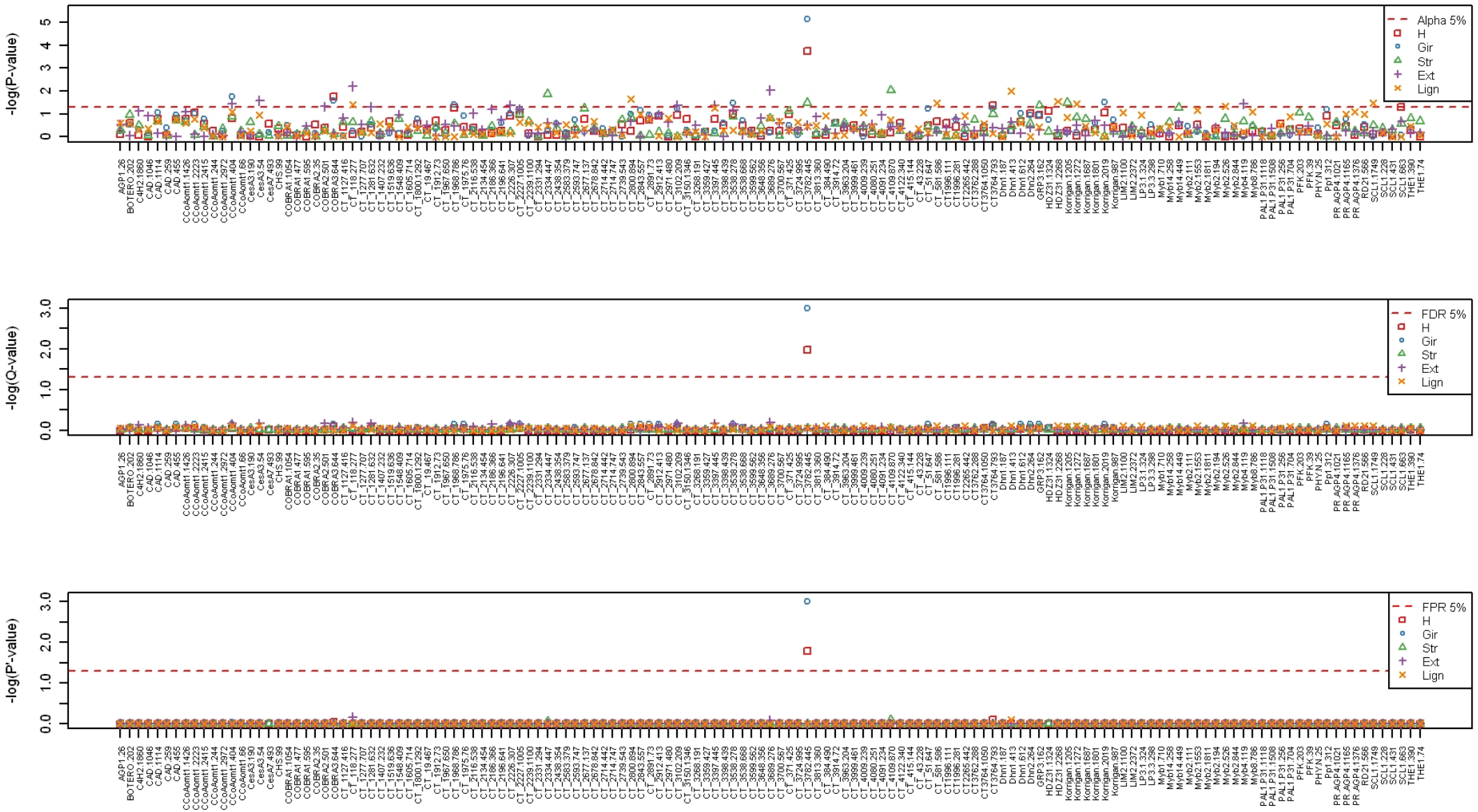


Figure 6: *P*-values, *q*-values and *P'*-values for the SNP effect in the additive model, for 141 informative SNPs in the G0 samples.

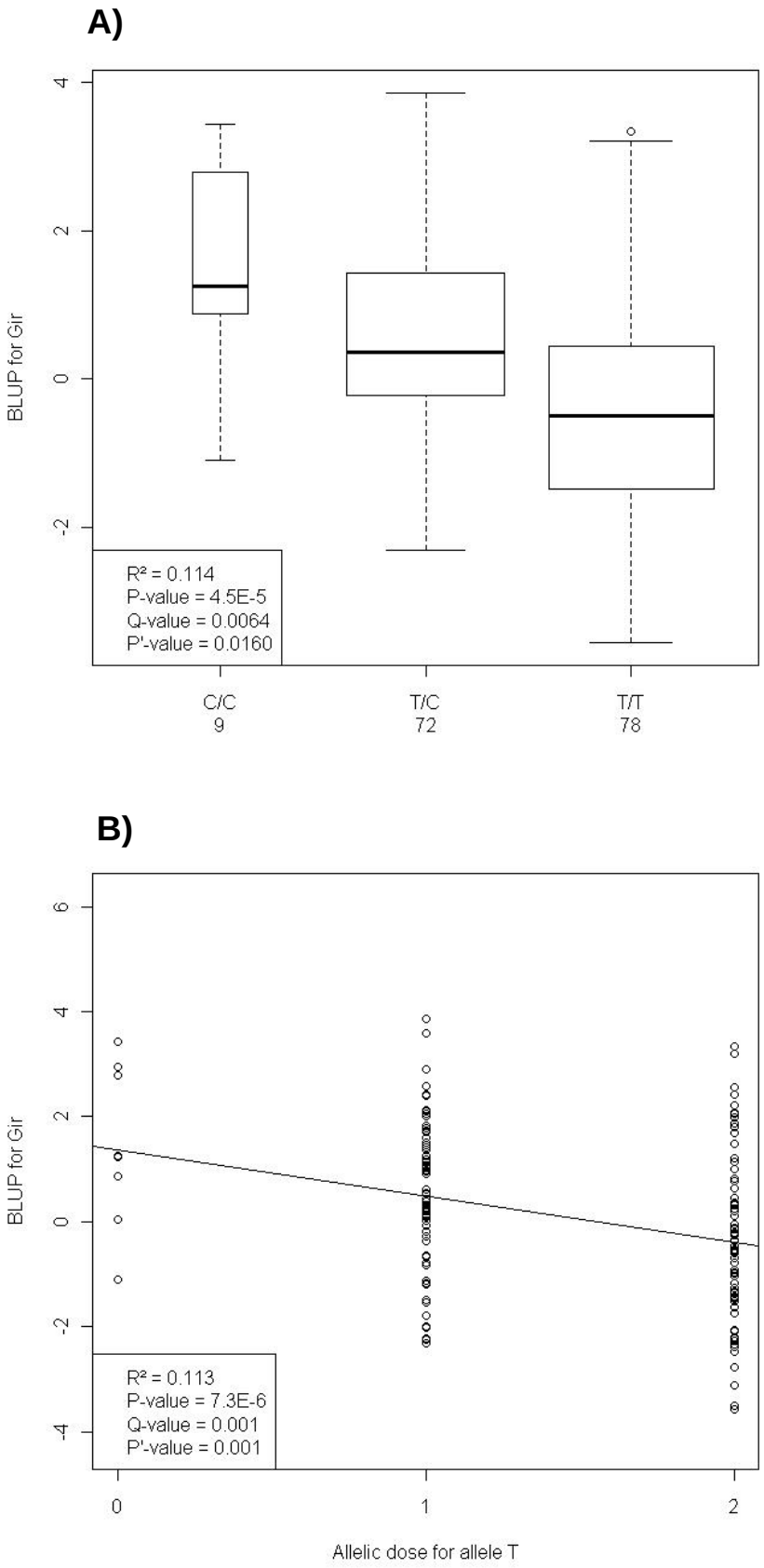


Figure 7: Genotypic effect of SNP CT_3782.445 on girth (Gir) in the G0 sample. **A*/*** Codominant model. The width of each boxplot is proportional to the number of observations, which are indicated below the *x*-axis. **B*/*** Additive model.

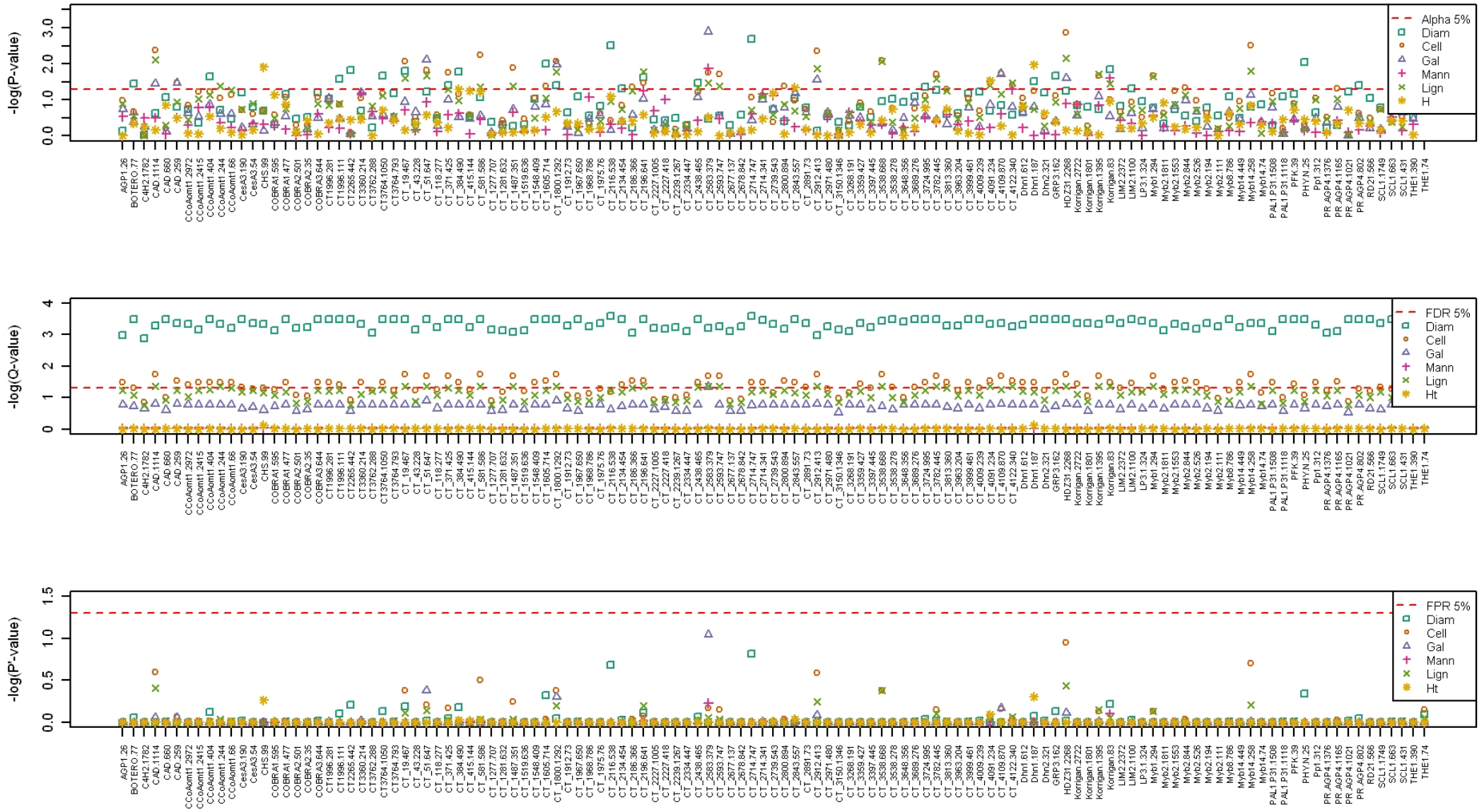


Figure 8: *P*-values, *q*-values and *P'*-values for the SNP effect in the codominant model, for 121 informative SNPs in the G1 samples.

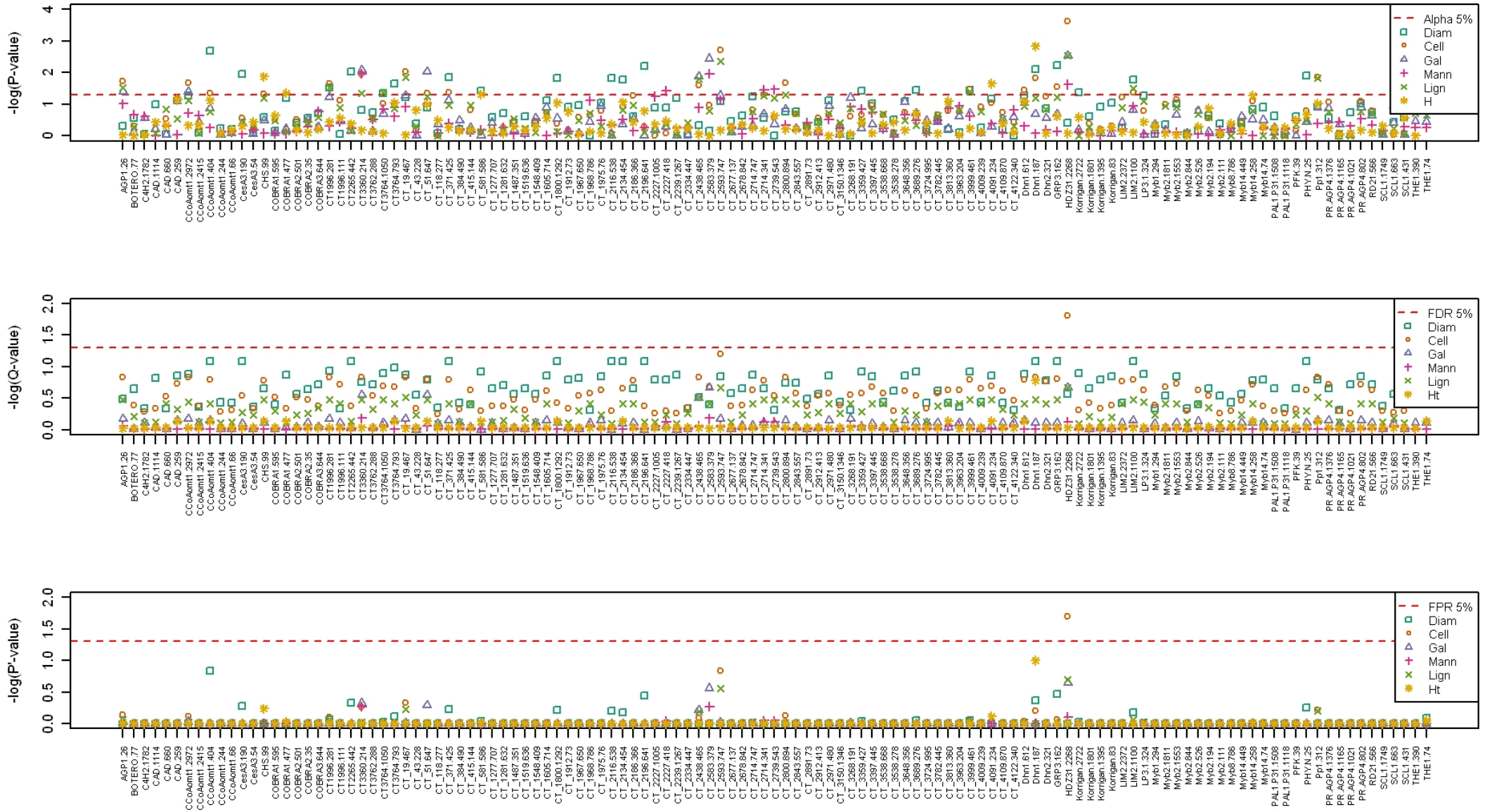


Figure 9: *P*-values, *q*-values and *P'*-values for the SNP effect in the additive model, for 121 informative SNPs in the G1 samples.

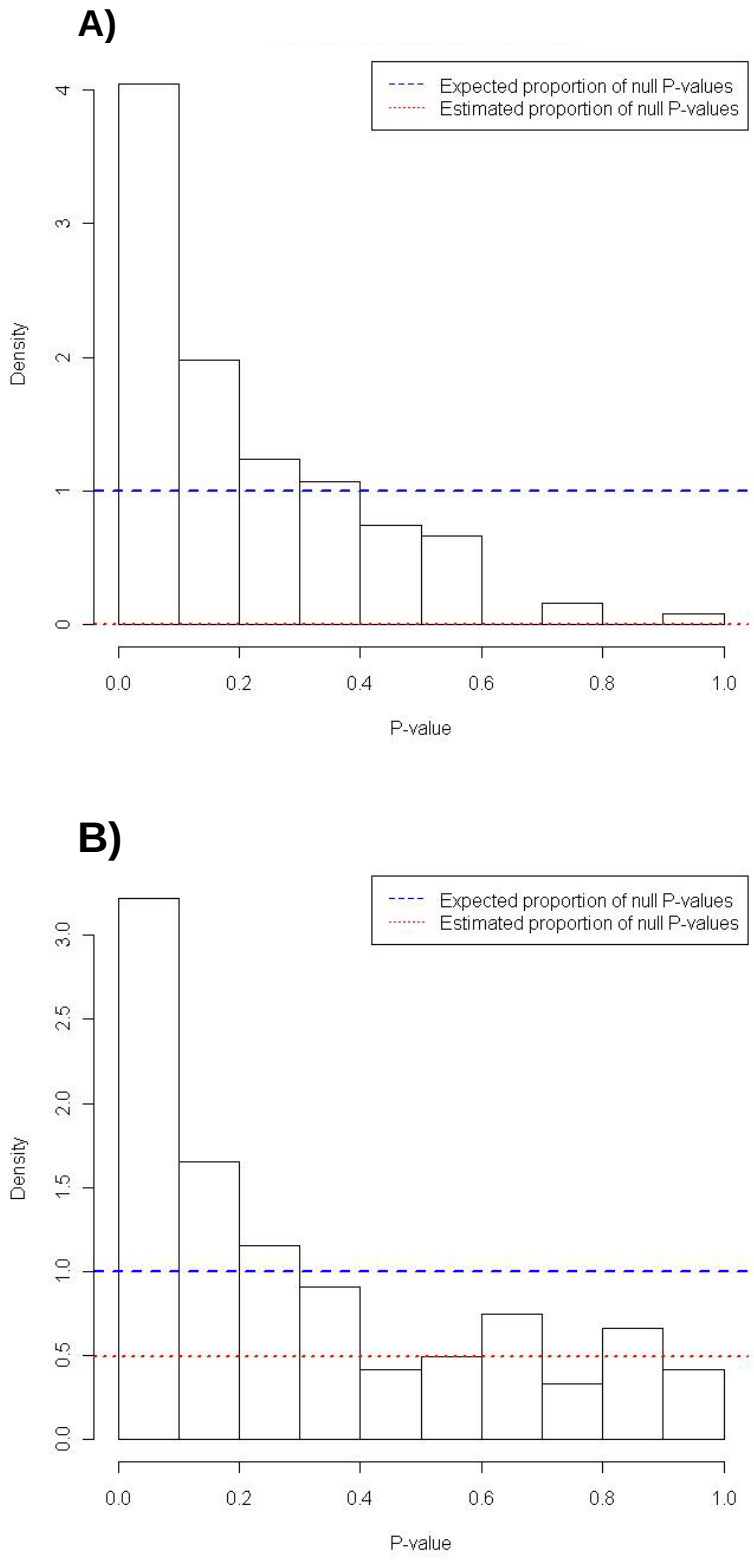


Figure 10: Distribution of *P*-values for the codominant (**A**) and additive (**B**) models for *Diam* in the G1 sample. **A**) There is no plateau in the distribution, thus the proportion of null *P*-values cannot be estimated by the empirical method (STOREY and TIBSHIRANI 2003) and the *ad hoc* estimation produces very low *q*-values (see Figure 8). **B**) There is a plateau in the distribution that allows estimating the proportion of null *P*-values, which is close to 0.5.

The other 120 informative SNPs were not significantly associated with any trait using both the codominant and additive models at a 5% threshold for *P'*-values.

Discussion

Population structure and familial relatedness

Linkage disequilibrium (LD) in association mapping populations can be the consequence of physical linkage, population structure, relatedness, genetic drift or selection (FLINT-GARCIA *et al.* 2003). Therefore, the success of an association mapping experiment largely depends on the ability to separate LD due to linkage from LD due to other causes (CARDON and BELL 2001; STICH and MELCHINGER 2009). In the present study, no population structure was found in the maritime pine Aquitaine G0 and G1 samples using ~100 SNP markers. This result suggests that the mass selection performed in the Aquitaine forest, which is generally considered as an unstructured population both at the phenotypic (DANJON 1994) and molecular (MARIETTE *et al.* 2001; DERORY *et al.* 2002; RIBEIRO *et al.* 2002; EVENO *et al.* 2008) levels, did not induce any artificial stratification in the breeding population for this group of SNPs. Further studies are however necessary to confirm this result, as our markers only covered a very small part of the pine genome, and finer substructure could exist at the phenotypic level.

Familial relatedness was accounted for in our association models, either in the first step of the two-stage analysis when estimating BLUPs for the G0 trees (see Chapter III), or directly in model (4) for the one-stage analysis of the G1 samples. The coancestry coefficient between individuals with unknown relationship was set to zero, thus assuming that the G0 trees were unrelated. Indeed, many arguments are in favour of low inbreeding in the Aquitaine population: maritime pine shows a high outcrossing rate and an extensive pollen flow (GONZÁLEZ-MARTÍNEZ *et al.* 2003; DE-LUCAS *et al.* 2008; JOÃO GASPAR *et al.* 2009), and probably a strong selection against inbreds, as reported for various pine species (reviewed in LEDIG 1998). Furthermore, the effective size of the Aquitaine population that we inferred from the molecular analysis on candidate genes (around a few thousands individuals, see Chapter I) still seems large enough to expect low inbreeding levels among the G0 trees. Recent studies proposed the use of marker-based kinship estimates to account for familial relatedness in association mapping experiments, which proved useful when pedigree records were incomplete or inaccurate (YU *et al.* 2006; STICH *et al.* 2008; STICH and MELCHINGER 2009). We did not use these approaches, as ~100 informative biallelic markers is probably too low to obtain accurate estimates of kinship coefficients. Moreover, we are confident, given the arguments listed above, that a very low level of inbreeding among the G0 trees can be

assumed. These methods will however be tested in forthcoming association studies involving more SNPs or SSRs, to ensure that the G0 trees can be considered as unrelated.

One-stage versus two-stage association mapping approaches

To date, the two-stage association-mapping approach is the most commonly used in plants (STICH *et al.* 2008). It first implies the analysis of phenotypic data and the calculation of BLUPs or entry means for each individual of the population, followed by a second step where these estimates are used for the association analysis. However, this two-stage procedure does not account for heterogeneity in experimental errors among the adjusted entry means. This problem can be overcome by applying a one-stage association approach in which the phenotypic and association analyses are performed together (STICH *et al.* 2008). It requires the availability of both phenotypes and genotypes, as for the 162 G1 clones. However, this approach could not be applied to the G0 samples for which only mother-trees were genotyped and half-sib progenies phenotyped. In YU *et al.* (2006), the heterogeneity of experimental errors among entry means in the two-stage approach was partially accounted for by specifying a structure for the residual variance such as $\sigma^2_e = R.V_R$, where R is a square matrix in which the off-diagonal elements are 0 and the diagonal elements are reciprocal of the number of phenotypic observations underlying each entry mean, and V_R is the residual variance. STICH *et al.* (2008) improved this method by replacing the diagonal elements of the R matrix by the square of the standard errors of the adjusted entry means, which led to lower type-I error rates. These "experimental wise error" corrections are not implemented in the software packages commonly used for association-mapping such as Tassel (BRADBURY *et al.* 2007) or SNPassoc (GONZALEZ *et al.* 2007). In the present study, we did not use any correction for the heterogeneity of experimental errors. The number of observations for each G0 sample was however either large (12 to 36 half-sibs for growth-related traits and stem straightness) or balanced (7 to 12 half-sibs for chemical-related traits, with 93% of the samples represented by 8 to 11 half-sibs), resulting in low coefficients of variation for BLUP standard errors (< 2.7% for all the traits except for *Str* for which it was higher, ~7.5%). Thus the correction would probably not have changed the results much, considering the high homogeneity of BLUP standard errors. Although it would certainly increase the computation time, it would nevertheless be interesting to program the correction method based on standard errors for upcoming association mapping studies, since unbalanced designs are frequent in forestry. For the G1 samples, the number of replicates by clone was low (3 to 5) and BLUP standard errors were more variable (coefficients of variation ranging from 6% to 12.5% for the different traits

considered), thus the one-stage approach was likely to be more appropriate since it accounts directly for heterogeneity in experimental errors.

Power, allele frequency and sample size

Power is the probability of detecting an effect when a statistic exceeds a pre-determined threshold. In the case of association-mapping, power largely depends on the sample size, allele frequency, marker heritability and level of linkage disequilibrium between the marker considered and the causative mutation (BALL 2005; WANG *et al.* 2005). Previous studies of statistical power for association using coalescence simulations showed that ~500 individuals typed for ~20 SNPs spaced throughout a candidate gene region are necessary for detecting a causative polymorphism of small effect (~5%), and that greater power is achieved more by increasing the sample size than by increasing the number of polymorphisms (LONG and LANGLEY 1999). In the present study we used two different samples of 160 unrelated and 162 related individuals, respectively, which is far less than the sample size recommended by LONG and LANGLEY (1999), but was nonetheless enough to detect two associations that remained significant after correction for multiple testing. Significant associations have similarly been reported for small samples, in *Eucalyptus* for wood microfibril angle (N=290 in THUMMA *et al.* 2005) and in *Populus* for timing of bud set (N=120 in INGVARSSON *et al.* 2008). In our case, the minor allele frequencies (MAFs) for the two SNPs significantly associated to either growth or cellulose content were rather high (28.3% and 36.8%, respectively), which is not surprising. Indeed, variants that contribute to complex traits are likely to have only modest phenotypic effects, and rare variants with modest effects are difficult to detect by any method because they explain only a very small fraction of the phenotypic variance in the population (HIRSCHHORN and DALY 2005). For example, the power to detect a significant association at a 10^{-5} alpha threshold for a SNP with a 5% heritability and with a sample size of 1,000 is near zero for MAFs below 25%, but rapidly increases for larger MAFs (assuming Hardy-Weinberg equilibrium for genotypic frequencies, see Figure 11A). With a sample size six times smaller (160 individuals, as in the present study), the power is null whatever the MAF for such small SNP effects (Figure 11B). Even at a low 10^{-5} alpha threshold, the power in our study was very low (around 10^{-3} for the SNP CT_3782.445 for example, see Figure 12), which probably explains the lack of repeatability (*i.e.* the significant association detected in the G0 sample was not significant in the G1 sample, and *vice versa*) (LONG and LANGLEY 1999; BALL 2005). The most obvious limitations of association mapping studies are the high genotyping and phenotyping costs, which hamper the use of large sample sizes crucial to detect more robust

and repeatable associations. To minimize the amount of genotyping required in association studies without sacrificing power, HIRSCHHORN and DALY (2005) proposed a multistage strategy: first a large number of SNPs is genotyped in a small sample, allowing for the detection of a subset of SNPs with putative associations at a high nominal *P*-value threshold. In the following stages, these SNPs are re-tested in similar or larger samples to distinguish the more likely to be true positive associations from the many false positive results. This strategy was successfully used to detect variants associated with autism in humans in a design involving ~500,000 SNPs and over one thousand families in the first step (WEISS and ARKING 2009). In our study, only ~10 SNPs showed significant associations at a very liberal 1% alpha threshold, and re-testing them in samples of a similar size is pointless as our design lack repeatability. Since the cost of field trials and phenotyping is certainly more important than that of genotyping for forest trees, we rather propose to extend the sample size by including more G0 and G1 trees, ideally taking advantage of the many field trials available through the breeding program. This approach however requires that either control families or clones or related individuals are installed in each trial to allow estimating environmental effects in a combined analysis of different experiments. Finally, if the genotyping capacities are limited, a SNP selection can be done based on the MAFs, as this criterion is directly linked to the power of detection (Figures 11 and 12).

Significant associations: which genes, which traits?

The significant genotype-phenotype association with growth involved one SNP located in contig CT_3782. This consensus group of ESTs theoretically corresponds to a gene, identified by a tBLASTx search (ALTSCHUL *et al.* 1997) against non-redundant GenBank database as a putative fasciclin-like arabinogalactan protein (FLA) for its high score and E-value (300 and 4e-126, respectively) with *Arabidopsis thaliana* FLA17 protein (accession number NM_120722). FLAs are a subclass of arabinogalactan proteins (AGPs) that have, in addition to predicted AGP-like glycosylated regions, putative cell adhesion domains known as fasciclin domains (JOHNSON *et al.* 2003). AGPs have been found in all organs and tissues of higher plants, they are components of the plasma membrane, the extracellular matrix and the cell wall (LIU *et al.* 2008). They are involved in many cellular processes such as cell proliferation, cell expansion and differentiation, cell-cell recognition or programmed cell death (ZHANG *et al.* 2003). If the specific functions of AGPs remain uncertain, their involvement in wood formation has been hypothesized through numerous expressional studies, for example in poplar (LAFARGUETTE *et al.* 2004), loblolly pine (NO and LOOPSTRA

2000; ZHANG *et al.* 2000; WHETTEN *et al.* 2001), radiata pine (LI *et al.* 2009) or maritime pine (GION *et al.* 2005; PAIVA *et al.* 2008). Immunolocalization studies also showed that AGPs are expressed during pattern formation in vascular tissues, giving stronger evidence for their role in xylem development (CASERO *et al.* 1998; ZHANG *et al.* 2003). Given the arguments above, the association of CT_3782 with growth is plausible. In upcoming studies, new markers could be developed around CT_3782.445 to assess the extent of LD in this region.

Another significant association was detected between SNP HDZ31.2268 and cellulose content. HDZ31 encodes a class III homeodomain-leucine zipper (HD-ZIPIII) transcription factor, a class of genes unique to plants (SESSA *et al.* 1993; PRIGGE and CLARK 2006) involved in tissue patterning and polarity (reviewed in DEMURA and FUKUDA 2007). Indeed, gain-of-function mutations in *A. thaliana* HD-ZIPIII genes resulted in adaxialized lateral organs and amphivasal (xylem surrounding phloem) vascular bundles (MCCONNELL *et al.* 2001; EMERY *et al.* 2003; JUAREZ *et al.* 2004; ZHONG and YE 2004). In addition, the functions of HD-ZIPIII genes in wood formation have been demonstrated by reverse genetic approaches in other plant species: in *Zinnia elegans*, four HD-ZIPIII genes have been found differentially expressed in vascular tissues (procambium, immature xylem and/or xylem parenchyma cells) (OHASHI-ITO and FUKUDA 2003), and mutant approaches suggested that these genes regulate xylem cell differentiation (OHASHI-ITO *et al.* 2005). In *Populus*, expressional studies showed that the PtaHB1 gene is closely associated with secondary growth and inversely correlated with the level of microRNA miR166 (KO *et al.* 2006). The function of HDZ31 in *P. pinaster* remains unknown, but its role in wood formation and thus its association with wood cellulose content are plausible. In the Aquitaine natural population, it showed a low level of nucleotide diversity, and neutrality statistics suggested that its diversity pattern could result from either a bottleneck or balancing selection (see Chapter I). It also showed a significantly high level of LD across more than 3.2 kilobase pairs, which was also detected in the G0 and G1 samples (Figure 3) and is quite unusual in conifer species (BROWN *et al.* 2004; GONZALEZ-MARTINEZ *et al.* 2006; HEUERTZ *et al.* 2006; PYHAJARVI *et al.* 2007; SAVOLAINEN and PYHAJARVI 2007). This high level of LD prevents the precise detection of mutations putatively associated with wood chemical properties in HDZ31. It would be interesting first to obtain more sequence data in the vicinity of HDZ31 to explore any potential drop in LD, and second to check if a similar LD pattern or association is found in other maritime pine populations.

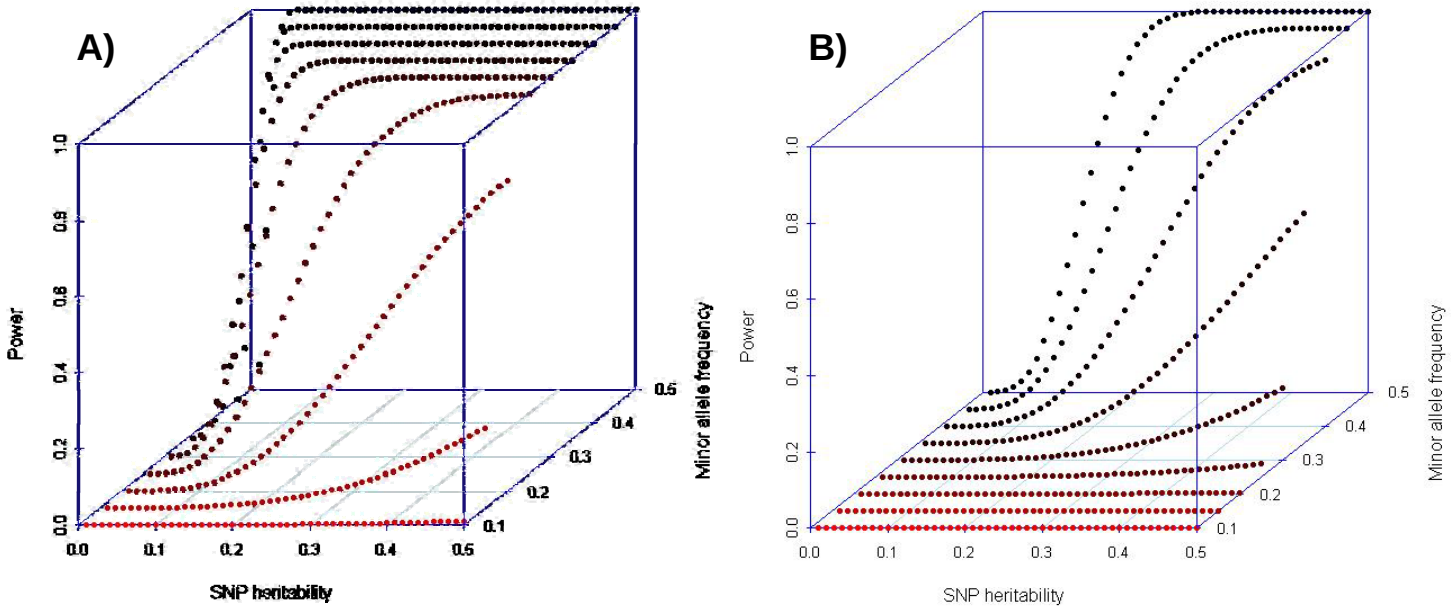


Figure 11: Power of single-marker based association tests as a function of SNP minor allele frequency and heritability, for a 10^{-5} alpha threshold and a sample size of 1000 (**A**) or 160 (**B**) individuals. Power was calculated using the *ldDesign* package (BALL 2004), assuming that the SNP considered was the causative mutation (implemented in the *ldDesign* package by setting the linkage disequilibrium *D* value to its maximum), and that the genotypic frequencies followed Hardy-Weinberg equilibrium.

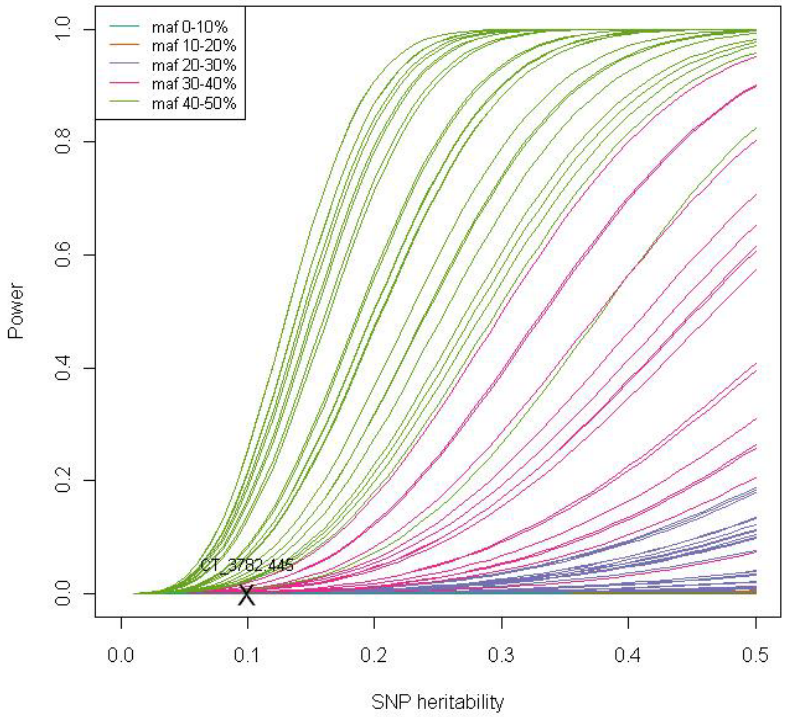


Figure 12: Power of association tests for the G0 sample as a function of SNP heritability at a 10^{-5} alpha threshold. Each curve represents one of the 141 SNPs of the study, with its particular genotypic frequencies. Colours correspond to SNP minor allele frequencies in the G0 sample. Power was calculated using the *ldDesign* package (BALL 2004), assuming that the SNPs considered were the causative mutations (as for Figure 11). The power of the association test for the SNP CT_3782.445, which is significantly associated with growth-traits with 11% heritability, is represented by a cross and is near zero.

Conclusion and perspectives

We have shown that association mapping is a valuable tool to detect genotype-phenotype associations in the maritime pine breeding population, and highlighted two genes significantly associated with growth and wood cellulose content, respectively. However, our power was low and increasing the sample size is required to detect repeatable associations. Correcting for heterogeneity of experimental errors, discarding low-MAF SNPs to reduce the genotyping cost and enriching the CT_3782 and HDZ31 regions with new markers have also been suggested for upcoming association studies. We showed that correcting for multiple testing through *P'*-values was in some cases more appropriate than using the positive FDR method, at the cost of longer computation time. These multiple testing corrections are absolutely necessary, as the probability that a SNP with a given *P*-value is truly associated with the phenotype depends not only on how unlikely that *P*-value is under the null hypothesis H_0, but also on how unlikely it is under the alternative hypothesis H_1. In this frame, Bayesian methods would provide an interesting alternative, as they do not depend on the number of tests performed (STEPHENS and BALDING 2009) and give a well-defined measure of strength of evidence, independent of the experimental design or sample size used (BALL 2005).

Acknowledgements:

This research was supported by grants from ANR: Genoplante (GenoQB, GNP05013C), PFTV (BOOST-SNP, 07PFTV002), the Aquitaine Region and the EU (GEMINI project, QLRT-1999-00942). C. Lepoittevin was supported by CIFRE contract between FCBA and INRA.

Contributions to the chapter:

C. Plomion and L. Harvengt organized the funding of the experiment and C. Lepoittevin's PhD grant. C. Plomion designed the experiment. For the contributions to the genotyping and phenotyping activities see Chapters II and III, respectively. C. Lepoittevin and P. Garnier-Géré analyzed the data. C. Lepoittevin wrote the chapter, with the helpful comments of P. Garnier-Géré and C. Plomion.

References

ABECASIS, G. R., D. GHOSH and T. E. NICHOLS, 2005 Linkage disequilibrium: Ancient history drives the new genetics. Human Heredity **59:** 118-124.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25:** 3389-3402.

BALL, R. D., 2004 ldDesign: Design of experiments for detection of linkage disequilibrium, http://cran.r-project.org/web/packages/ldDesign/index.html.

BALL, R. D., 2005 Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics **170:** 859-873.

BRADBURY, P. J., Z. ZHANG, D. E. KROON, T. M. CASSTEVENS, Y. RAMDOSS *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635.

BROWN, G. R., G. P. GILL, R. J. KUNTZ, C. H. LANGLEY and D. B. NEALE, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America **101:** 15255-15260.

BUCCI, G., S. C. GONZALEZ-MARTINEZ, G. LE PROVOST, C. PLOMION, M. M. RIBEIRO *et al.*, 2007 Range-wide phylogeography and gene zones in *Pinus pinaster* Ait. revealed by chloroplast microsatellite markers. Molecular Ecology **16:** 2137-2153.

BURBAN, C., and R. J. PETIT, 2003 Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. Molecular Ecology **12:** 1487-1495.

CANTY, A., and B. RIPLEY, 2009 boot: Bootstrap R (S-Plus) Functions, http://cran.r-project.org/web/packages/boot/.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nature Reviews Genetics **2:** 91-99.

CASERO, P. J., I. CASIMIRO and J. P. KNOX, 1998 Occurrence of cell surface arabinogalactan-protein and extensin epitopes in relation to pericycle and vascular tissue development in the root apex of four species. Planta **204:** 252-259.

CHARCOSSET, A., and A. GALLAIS, 1996 Estimation of the contribution of quantitative trait loci (QTL) to the variance of a quantitative trait by means of genetic markers. Theoretical and Applied Genetics **93:** 1193-1201.

DABNEY, A., J. D. STOREY and G. R. WARNES, 2009 qvalue: Q-value estimation for false discovery rate control, http://CRAN.R-project.org/package=qvalue.

DANJON, F., 1994 Stand features and height growth in a 36-year-old maritime pine (*Pinus-Pinaster* Ait) provenance test. Silvae Genetica **43:** 52-62.

DE-LUCAS, A. I., J. J. ROBLEDO-ARNUNCIO, E. HIDALGO and S. C. GONZALEZ-MARTINEZ, 2008 Mating system and pollen gene flow in Mediterranean maritime pine. Heredity **100:** 390-399.

DEMURA, T., and H. FUKUDA, 2007 Transcriptional regulation in wood formation. Trends in Plant Science **12:** 64-70.

DERORY, J., S. MARIETTE, S. C. GONZALEZ-MARTINEZ, D. CHAGNE, D. MADUR *et al.*, 2002 What can nuclear microsatellites tell us about maritime pine genetic resources conservation and provenance certification strategies? Annals of Forest Science **59:** 699-708.

EMERY, J. F., S. K. FLOYD, J. ALVAREZ, Y. ESHED, N. P. HAWKER *et al.*, 2003 Radial Patterning of Arabidopsis Shoots by Class III HD-ZIP and KANADI Genes. Current Biology **13:** 1768-1774.

EVANNO, G., S. REGNAUT and J. GOUDET, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology **14:** 2611-2620.

EVENO, E., C. COLLADA, M. A. GUEVARA, V. LEGER, A. SOTO *et al.*, 2008 Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Molecular Biology and Evolution **25:** 417-437.

FALCONER, D. S., T. F. C. MACKAY and M. BULMER, 1996 *Introduction to quantitative genetics*. Longman New York.

FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics **164:** 1567-1587.

FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2007 Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes **7:** 574.

FLINT-GARCIA, S. A., J. M. THORNSBERRY and E. S. BUCKLER, 2003 Structure of linkage disequilibrium in plants. Annual Review of Plant Biology **54:** 357-374.

FOURNIER-LEVEL, A., L. LE CUNFF, C. GOMEZ, A. DOLIGEZ, A. AGEORGES *et al.*, 2009 Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp sativa) berry: a QTL to QTN integrated study. Genetics **Published Articles Ahead of Print: August 31, 2009.**

GARNIER-GERE, P., 1992 Contribution à l'étude de la variabilité génétique inter et intra-population chez le maïs (*Zea mays* L.): valorisation d'informations agromorphologiques et enzymatiques, PhD Thesis, Institut National Agronomique Paris-Grignon.

GILMOUR, A. R., B. J. GOGEL, B. R. CULLIS and R. THOMPSON, 2006 ASReml user guide release 2.0. VSN International Ltd., Hemel Hempstead, UK.

GILMOUR, A. R., R. THOMPSON and B. R. CULLIS, 1995 Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51:** 1440-1450.

GION, J. M., C. LALANNE, G. LE PROVOST, H. FERRY-DUMAZET, J. PAIVA *et al.*, 2005 The proteome of maritime pine wood forming tissue. Proteomics **5:** 3731-3751.

GONZALEZ-MARTINEZ, S. C., R. ALIA and L. GIL, 2002 Population genetic structure in a Mediterranean pine (*Pinus pinaster* Ait.): a comparison of allozyme markers and quantitative traits. Heredity **89:** 199-206.

GONZALEZ-MARTINEZ, S. C., E. ERSOZ, G. R. BROWN, N. C. WHEELER and D. B. NEALE, 2006 DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics **172:** 1915-1926.

GONZÁLEZ-MARTÍNEZ, S. C., S. GERBER, M. T. CERVERA, J. M. MARTÍNEZ-ZAPATER, R. ALÍA *et al.*, 2003 Selfing and sibship structure in a two-cohort stand of maritime pine (*Pinus pinaster* Ait.) using nuclear SSR markers. Annals of Forest Science **60:** 115-121.

GONZALEZ, J. R., L. ARMENGOL, X. SOLE, E. GUINO, J. M. MERCADER *et al.*, 2007 SNPassoc: an R package to perform whole genome association studies. Bioinformatics **23:** 654-655.

GUPTA, P. K., S. RUSTGI and P. L. KULWAL, 2005 Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Molecular Biology **57:** 461-485.

HENDERSON, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics **31:** 423-447.

HEUERTZ, M., E. DE PAOLI, T. KALLMAN, H. LARSSON, I. JURMAN *et al.*, 2006 Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics **174:** 2095-2105.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6:** 95-108.

HUBISZ, M. J., D. FALUSH, M. STEPHENS and J. K. PRITCHARD, 2009 Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources **9:** 1322-1332.

INGVARSSON, P. K., M. V. GARCIA, V. LUQUEZ, D. HALL and S. JANSSON, 2008 Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics **178:** 2217-2226.

JOÃO GASPAR, M., A. I. DE-LUCAS, R. ALÍA, J. ALMIRO PINTO PAIVA, E. HIDALGO *et al.*, 2009 Use of molecular markers for estimating breeding parameters: a case study in a *Pinus pinaster* Ait. progeny trial. Tree Genetics & Genomes **5:** 609-616.

JOHNSON, K. L., B. J. JONES, A. BACIC and C. J. SCHULTZ, 2003 The fasciclin-like arabinogalactan proteins of Arabidopsis. A multigene family of putative cell adhesion molecules. Plant Physiology **133:** 1911-1925.

JUAREZ, M. T., J. S. KUI, J. THOMAS, B. A. HELLER and M. C. P. TIMMERMANS, 2004 microRNA-mediated repression of rolled leaf1 specifies maize leaf polarity. Nature **428:** 84-88.

KO, J. H., C. PRASSINOS and K. H. HAN, 2006 Developmental and seasonal expression of PtaHB1, a Populus gene encoding a class IIIHD-Zip protein, is closely associated with secondary growth and inversely correlated with the level of microRNA (miR166). New Phytologist **169:** 469-478.

LAFARGUETTE, F., J.-C. LEPLE, A. DEJARDIN, F. LAURANS, G. COSTA *et al.*, 2004 Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. New Phytologist **164:** 107-121.

LEDIG, F. T., 1998 Genetic variation in *Pinus*, pp. 251-280 in *Ecology and Biogeography of Pinus*, edited by D. M. RICHARDSON. Cambridge University Press, Cambridge.

LI, X., H. WU, S. DILLON and S. SOUTHERTON, 2009 Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata* D. Don. BMC Genomics **10:** 41.

LIU, D., L. TU, Y. LI, L. WANG, L. ZHU *et al.*, 2008 Genes encoding fasciclin-like arabinogalactan proteins are specifically expressed during cotton fiber development. Plant Molecular Biology Reporter **26:** 98-113.

LONG, A. D., and C. H. LANGLEY, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research **9:** 720-731.

MALOSETTI, M., C. G. VAN DER LINDEN, B. VOSMAN and F. A. VAN EEUWIJK, 2007 A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato. Genetics **175:** 879-889.

MARCHINI, J., L. R. CARDON, M. S. PHILLIPS and P. DONNELLY, 2004 The effects of human population structure on large genetic association studies. Nature Genetics **36:** 512-517.

MARIETTE, S., D. CHAGNE, C. LEZIER, P. PASTUSZKA, A. BAFFIN *et al.*, 2001 Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. Heredity **86:** 469-479.

MCCONNELL, J. R., J. EMERY, Y. ESHED, N. BAO, J. BOWMAN *et al.*, 2001 Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. Nature **411:** 709-713.

NEALE, D. B., and O. SAVOLAINEN, 2004 Association genetics of complex traits in conifers. Trends in Plant Science **9:** 325-330.

NO, E. G., and C. A. LOOPSTRA, 2000 Hormonal and developmental regulation of two arabinogalactan-proteins in xylem of loblolly pine (*Pinus taeda*). Physiologia Plantarum **110:** 524-529.

OHASHI-ITO, K., and H. FUKUDA, 2003 HD-Zip III homeobox genes that include a novel member, ZeHB-13 (Zinnia)/ATHB-15 (Arabidopsis), are involved in procambium and xylem cell differentiation. Plant and Cell Physiology **44:** 1350-1358.

OHASHI-ITO, K., M. KUBO, T. DEMURA and H. FUKUDA, 2005 Class III homeodomain leucine-zipper proteins regulate xylem cell differentiation. Plant and Cell Physiology **46:** 1646-1656.

PAIVA, J. A. P., P. H. GARNIER-GERE, J. C. RODRIGUES, A. ALVES, S. SANTOS *et al.*, 2008 Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. New Phytologist **179:** 1180-1194.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist **167:** 101-112.

PRICE, A. L., N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N. A. SHADICK *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38:** 904-909.

PRIGGE, M. J., and S. E. CLARK, 2006 Evolution of the class III HD-Zip gene family in land plants. Evolution & Development **8:** 350-361.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000a Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000b Association Mapping in Structured Populations. The American Journal of Human Genetics **67:** 170-181.

PRITCHARD, J. K., X. WEN and D. FALUSH, 2007 Documentation for *structure* software: Version 2.2. Department of Human Genetics, University of Chicago (http://pritch.bsd.uchicago.edu/software).

PYHAJARVI, T., M. R. GARCIA-GIL, T. KNURR, M. MIKKONEN, W. WACHOWIAK *et al.*, 2007 Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics **177:** 1713-1724.

R_DEVELOPMENT_CORE_TEAM, 2009 R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

RAYMOND, M., and F. ROUSSET, 1995 Genepop (Version-1.2) - Population-Genetics Software for Exact Tests and Ecumenicism. Journal of Heredity **86:** 248-249.

RIBEIRO, M. M., S. MARIETTE, G. G. VENDRAMIN, A. E. SZMIDT, C. PLOMION *et al.*, 2002 Comparison of genetic diversity estimates within and among populations of maritime pine using chloroplast simple-sequence repeat and amplified fragment length polymorphism data. Molecular Ecology **11:** 869-877.

RINALDO, A., S. A. BACANU, B. DEVLIN, V. SONPAR, L. WASSERMAN *et al.*, 2005 Characterization of multilocus linkage disequilibrium. Genetic epidemiology **28:** 193-206.

SAVOLAINEN, O., and T. PYHAJARVI, 2007 Genomic diversity in forest trees. Current Opinion in Plant Biology **10:** 162-167.

SESSA, G., G. MORELLI and I. RUBERTI, 1993 The Athb-1 and Athb-2 Hd-Zip domains homodimerize forming complexes of different DNA-binding specificities. Embo Journal **12:** 3507-3517.

STEPHENS, M., and D. J. BALDING, 2009 Bayesian statistical methods for genetic association studies. Nature Reviews Genetics **10:** 681-690.

STICH, B., and A. E. MELCHINGER, 2009 Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics **10:** 94.

STICH, B., J. MOHRING, H. P. PIEPHO, M. HECKENBERGER, E. S. BUCKLER *et al.*, 2008 Comparison of mixed-model approaches for association mapping. Genetics **178:** 1745.

STOREY, J. D., 2002 A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B, Statistical Methodology**:** 479-498.

STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences **100:** 9440-9445.

THUMMA, B. R., B. A. MATHESON, D. ZHANG, C. MEESKE, R. MEDER *et al.*, 2009 Identification of a *cis*-acting regulatory polymorphism in a eucalypt *Cobra*-like gene affecting cellulose content. Genetics **Published Articles Ahead of Print: September 7, 2009**.

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

VOIGHT, B. F., and J. K. PRITCHARD, 2005 Confounding from cryptic relatedness in case-control association studies. PLoS Genetics **1:** 302-311.

WANG, W. Y. S., B. J. BARRATT, D. G. CLAYTON and J. A. TODD, 2005 Genome-wide association studies: Theoretical and practical concerns. Nature Reviews Genetics **6:** 109-118.

WEISS, L. A., and D. E. ARKING, 2009 A genome-wide linkage and association scan reveals novel loci for autism. Nature **461:** 802-808.

WHETTEN, R., Y.-H. SUN, Y. ZHANG and R. SEDEROFF, 2001 Functional genomics and cell wall biosynthesis in loblolly pine. Plant Molecular Biology **47:** 275-291.

WU, B., N. LIU and H. ZHAO, 2006 PSMIX: an R package for population structure inference via maximum likelihood method. BMC Bioinformatics **7:** 317.

XIONG, M., and S.-W. GUO, 1997 Fine-Scale Genetic Mapping Based on Linkage Disequilibrium: Theory and Applications. The American Journal of Human Genetics **60:** 1513-1531.

YU, J. M., G. PRESSOIR, W. H. BRIGGS, I. V. BI, M. YAMASAKI *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38:** 203-208.

ZHANG, Y., G. BROWN, R. WHETTEN, C. A. LOOPSTRA, D. NEALE *et al.*, 2003 An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. Plant Molecular Biology **52:** 91-102.

ZHANG, Y., R. R. SEDEROFF and I. ALLONA, 2000 Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiology **20:** 457-466.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007 An *Arabidopsis* example of association mapping in structured samples. PLoS Genetics **3:** e4.

ZHONG, R., and Z.-H. YE, 2004 Amphivasal vascular bundle 1, a gain-of-function mutation of the IFL1/REV gene, is associated with alterations in the polarity of leaves, stems and carpels. Plant and Cell Physiology **45:** 369-385.

General discussion and perspectives

The GenoQB project was driven by our desire to understand the genetic basis of wood formation, a unique feature of trees, and by our ambition to exploit this knowledge to accelerate and improve tree breeding for wood properties using the natural variation at the gene level within species. In this frame, the objectives of this thesis were two-fold: i/ study the landscape of nucleotide diversity in previously identified candidate genes for wood formation, and ii/ test the relationships between naturally occurring nucleotide polymorphisms with the variability of wood quality related traits measured in experimental designs of the maritime pine breeding program.

Principal results obtained in this thesis

In **Chapter I**, we pursued our first objective by assessing nucleotide diversity patterns for nine transcription factors (TFs) putatively involved in wood formation in the large unstructured *Pinus pinaster* population of the French Atlantic coast. Three of these genes presented a very low nucleotide diversity, which might be a sign of strong purifying selection, whereas the other six showed a higher diversity, although lower than that observed for more than 20 structural genes in the same species. Given that TFs could interact with the promoters of tens to hundreds of genes (WRAY *et al.* 2003), a low diversity at TFs could be caused by the negative impact of their pleiotropic action, consistently with the *cis*-regulatory evolution theory (CARROLL 2005). Strong departures from the standard neutral model were observed for three TFs (*MYB1*, *LIM2* and *HDZ31*), with common patterns of excess of intermediate frequency variants, deficit in number of haplotypes, and high levels of linkage disequilibrium between sites within each gene. In order to distinguish between selection and demographic effects, we modeled a large range of bottleneck *scenarii* and explored their impact on the significance of various test statistics. Only a few of these *scenarii* could explain the patterns observed for all TFs. Using knowledge of past climatic events and *P. pinaster* recolonization history, we proposed that the population studied has undergone a size contraction that could have occurred during the last glacial period, and that its current effective population size could be of a few thousands individuals. This interpretation contrasts with that usually proposed for European conifer species, *i.e.* a very old bottleneck (more than 2 million years old) and the assumption of very large effective population sizes of several hundred thousands individuals. Our models also suggested possible balancing selection effects for the *MYB1* gene, which is a potential regulator of the phenylpropanoid metabolism in the lignin pathway. This gene is likely to be important for adaptation, and maintenance of alleles at intermediate

frequencies could result from a selective advantage for heterozygote genotypes at the locus, but also from the variation of selection pressures in space or time, as found in various case studies showing the maintenance of polymorphisms (HEDRICK 2006). Looking at the F_{is} coefficient from genotyping data at 4 SNPs within this gene, they all showed heterozygote excess (consistently with their quasi-complete LD), which would be in favor of the hypothesis of heterozygous-stage advantage.

In **Chapters II** and **III**, significant technical improvements were developed for high to medium throughput genotyping and phenotyping in maritime pine. First we showed that the GoldenGate genotyping technology is suitable for the complex and large genome of maritime pine, providing that the clusters of genotypic data points obtained are carefully examined to avoid the "cluster compression" phenomenon, and that genotyping error rates are estimated for each SNP. We also demonstrated that ESTs provide a cheap resource for SNP identification in non-model species for the purpose of large scale genotyping in population-based experiments. Secondly, we developed a non-destructive sampling method and near infrared spectroscopy calibrations to rapidly assess wood chemical properties on 30-year-old maritime pine trees. Although this method has to be further investigated (*i.e.* the samples used may not reflect the whole-tree properties), the gain in sampling and processing time was significant and lignin content was precisely measured. We found low heritabilities (0.17-0.23) for chemical properties and low to moderate heritabilities (0.15-0.55) for growth. The total genotypic variance that could be estimated in the clonal trial for growth was mainly due to an additive variance component. Moreover, growth and chemical related traits were not genetically correlated. If mass selection for growth proved efficient in the last two breeding cycles (+30% in volume and straightness), our results suggest that chemical properties can also be improved to a lesser extent. This would be profitable for the paper industry, as economic studies showed that a small decrease in lignin content could lead to substantial increases in mill profits (PETER *et al.* 2007).

Finally in **Chapter IV**, using an association mapping approach accounting for familial relatedness in the Aquitaine breeding population, we detected two mutations that were significantly associated, one with variation in growth and the other with variation in cellulose content. Additional studies that would examine the extent of linkage disequilibrium around these two polymorphisms seem however necessary to assess the involvement of the two genes identified, *i.e.* a fasciclin-like arabinogalactan protein and a *HD-ZIPIII* transcription factor, in trait variations. Looking further across the chromosome would confirm that other genes that could be in LD with the mutations detected are not causing the associations. If confirmed,

transformation studies can also be envisaged. Considering the quite low power of this experiment (see below), and in the absence of replicates in other trial sites, larger sample sizes seem necessary for upcoming association-mapping studies. We therefore propose to take advantage of the many field trials available from the maritime pine breeding program, and of rapid wood properties assessment techniques such as those described in **Chapter III**.

The candidate-gene approach in conifers

We previously saw (**Chapter IV**) that association mapping is *a priori* well suited for conifer species, since their high levels of genetic diversity and pollen flow as well as large population sizes are likely to lead to rapid LD decay between genes or polymorphisms, low inbreeding and low population structure. We have seen however that this general picture is certainly too simple: our results on TFs in **Chapter I** and other results on drought stress candidate genes in the Aquitaine population (EVENO 2008) both showed that some genes in *P. pinaster* can show very strong LD between SNPs with no decay. A more exhaustive characterization of the LD patterns would be needed across the genome of conifer species, but genome-wide association studies are not yet possible, because of the lack of a full genome sequence and the enormous marker density required to tag their large genome (NEALE and SAVOLAINEN 2004). Candidate gene approaches have therefore been favored in all previous conifer studies, whether in nucleotide diversity analyses (POT *et al.* 2005; EVENO *et al.* 2008; GRIVET *et al.* 2009; WACHOWIAK *et al.* 2009), or more recently in association studies (THUMMA *et al.* 2005; GONZALEZ-MARTINEZ *et al.* 2007; GONZALEZ-MARTINEZ *et al.* 2008; INGVARSSON *et al.* 2008; THUMMA *et al.* 2009). In the present study, candidate genes were chosen based on different criteria including i/ known or suspected role(s) in relevant biochemical pathways, ii/ orthology to genes in other plant species that presented a role in traits of interest, and iii/ transcript or proteome profiling experiment identifying them as having expressional patterns that were correlated with variation in targeted traits. Among the 70 successfully genotyped and non-redundant *in vitro*-SNPs located in more than 30 carefully chosen candidate genes, associations were tested with all traits and only one SNP (HDZ31.2268) showed a significant association with cellulose content. As discussed in **Chapter IV**, this is likely due to the lack of power of our experiment, but this could also result from a choice of candidate genes or SNPs that was not appropriate. A larger number of SNPs will be included in upcoming association studies in maritime pine (see the conclusion of **Chapter II**). The new 1536-plex Illumina array benefited from the pilot 384 array designed in **Chapter II**, in the sense that we included both already validated SNPs and new SNPs replacing those that failed in order to

recover haplotype information in the same genes. However the criteria of choice in a limited number of candidate genes were similar and the technical constraints of the Illumina technology were the same (inability to genotype neighboring SNPs). We may thus encounter similar drawbacks in terms of number of associations detected. This calls for a thorough examination of the strategy pursued to find relevant markers for large scale genotyping studies.

The first point to consider seems to be the targeted genes: are we targeting the relevant ones? Is it just a matter of the number of genes used? In this context, MORGANTE and SALAMINI (2003) and PARAN and ZAMIR (2003) highlighted the interest of regulatory mutations which affect both the level and pattern of gene expression, and could thus be involved with stronger effects in trait variation.

Despite their generally low level of nucleotide diversity (see **Chapter I**), TFs and their underlying mutations have been associated with a diverse set of diseases in humans (DARNELL 2002; LOPEZ-BIGAS *et al.* 2006; DIXON *et al.* 2007), with anthocyanin variation in grape (FOURNIER-LEVEL *et al.* 2009), with inflorescence structure in *Brassica* (PURUGGANAN *et al.* 2000), with plant architecture in maize (WANG *et al.* 2005a) or with carbon isotope discrimination in pine (GONZALEZ-MARTINEZ *et al.* 2008). Moreover, misregulation of TFs themselves also has important consequences on complex traits (VAQUERIZAS *et al.* 2009). TFs thus would appear as good candidates for association mapping studies. However, we have seen in **Chapter IV** that only one mutation from a HD-ZIP TF has been significantly associated to cellulose content. Although this lack of results may be due to many other reasons (see below), we should not ignore *cis*-regulatory mutations, which are in close proximity to the gene being regulated, and might directly affect the transcription initiation, transcription rate or transcript stability in an allele-specific manner (TAO *et al.* 2006). Indeed, *cis*-regulation generally accounts for more genetic variation than *trans*-regulation in humans (MORLEY *et al.* 2004; DIXON *et al.* 2007), mice (SCHADT *et al.* 2003), *Drosophila* (HUGHES *et al.* 2006) or *Arabidopsis* (KEURENTJES *et al.* 2007; WEST *et al.* 2007). This is consistent with the *cis*-regulatory evolution theory (summarized in the introduction of **Chapter I**), which stresses out the importance of *cis*-mutations for micro-evolution within species. In contrast, mutations in TFs, which are pleiotropic, are more likely to be deleterious than mutations in less interconnected genes (FRASER *et al.* 2002; YU *et al.* 2004; WITTKOPP 2005), and are thus more strongly counter-selected. In eucalypt, genes involved in the lignin biosynthetic pathway have been shown to be predominantly regulated through *cis*-acting effects (KIRST *et al.* 2004;

KIRST *et al.* 2005). A *cis*-regulatory SNP significantly associated to cellulose content was also recently identified (THUMMA *et al.* 2009), illustrating the involvement of *cis*-regulatory mutations in wood formation. Unfortunately, *P. pinaster* current EST databases contain few intron and promoter sequences, and *de novo* sequencing will be necessary to exploit the variability in these genomic regions. SNPs in exons might however also be of similar or greater interest, as shown by VEYRIERAS *et al.* (2008) in humans, where they are twice more likely to be expressional quantitative trait nucleotides (eQTNs) than SNPs in introns. Provided that the distribution of eQTNs is similar in plants, EST databases should already contain a large part of workable *cis*-regulatory mutations, but both coding and non-coding regions should be considered in the search for complex traits associated mutations.

Power of association studies

Over the last decade, a general pattern seems to have emerged from the few association mapping studies in tree species involving complex traits. This pattern can be summarized as one where a very few number of SNPs emerged with significant associations, even from strong putative candidate genes based on neutrality testing and expressional patterns, and moreover with a very small proportion of variation explained (THUMMA *et al.* 2005; GONZALEZ-MARTINEZ *et al.* 2007; GONZALEZ-MARTINEZ *et al.* 2008; INGVARSSON *et al.* 2008). One may argue as above that candidate gene approaches are necessarily limited due to an imperfect knowledge of complex biochemical pathways, and thus that genome-wide association studies would be more successful. The same pattern has however been observed in model species, where highly heritable traits like human height (80-90% heritable, reviewed in VISSCHER 2008) or maize flowering time (70% to over 80% heritable, BUCKLER *et al.* 2009) have been shown to be associated with many variants of very small effects. Additionally, the estimated effect of all the associated variants taken together barely reached 5% of the trait variation. These observations have been referred to recently as the "missing heritability" issue (MAHER 2008; BUCKLER *et al.* 2009). Although referring initially to high heritability traits for which a large proportion of variation remained to be explained, the same issue is now raised when searching for the genetic determinism of lower heritability traits (MANOLIO *et al.* 2009; MYLES *et al.* 2009). We examine below different factors likely to have an impact on the power of detection of phenotype-genotype associations. We won't discuss here the problem of genetic relatedness previously tackled in **Chapter IV** (see however the recent review of MYLES *et al.* 2009 on methods allowing to correct for it).

Importance of sample size

In **Chapter IV**, we estimated the power of our association mapping experiment and showed that larger sample sizes are essential to detect either small-effects variants, rare variants with large effects, or any of them across experiments. For example, GUDBJARTSSON *et al.* (2008) validated 3 SNPs that had been previously associated with human height (MAMADA *et al.* 2007; WEEDON *et al.* 2007; SANNA *et al.* 2008) by examining more than 30,000 people. However each of these studies highlighted about 40 previously unknown variants associated to height which did not overlap, probably because of their small effects (reviewed in VISSCHER 2008). The problem is also that many of these functional variants show a low frequency, and since the power to detect an association is a function of allele frequency, low sample sizes are not sufficient to detect those variants (LUO 1998; BALL 2005; WANG *et al.* 2005b). Even if the low-frequency variants are present in the sample, their signal is difficult to detect (see **Chapter IV** Figures 11 and 12, MYLES *et al.* 2009 Figure 2 reproduced below). For example, based on Figure 11A we can see that a sample size of 1,000 individuals is necessary to achieve a 50% power for a SNP with 10% heritability and a MAF of 0.25.

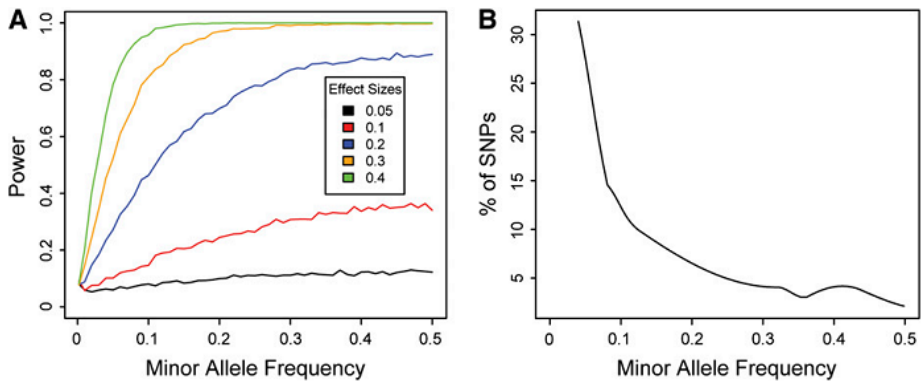


Figure 2. Important factors affecting the power of population mapping studies.

(A) The power of an association test is a function of the allele frequency and the effect size.
(B) The allele frequency spectrum from 3641 SNPs genotyped in 25 diverse maize inbred lines (www.panzea.org) demonstrates that most alleles in a population are rare. Therefore, if the frequency spectrum of functional alleles is similar to the frequency spectrum of random SNPs, most functional alleles will remain undetected through population mapping because of low power.
For **(A)**, phenotype data were simulated for 1000 haploid samples as a normal distribution with mean = 0 and SD = 1 for one allele and mean 0 + effect size and SD = 1 for the other allele. Effect size is therefore defined as the difference between the mean phenotypic values of the two alleles. Power is defined as the proportion of association tests (Pearson correlation) significant at $P < 0.05$ out of 5000 simulated data sets.

Figure 2 from MYLES *et al.* (2009)

The recent findings in humans and other model species also show that complex traits can be ruled by thousands of variants with tiny effects rather than by dozens of variants with moderate effects. According to VISSCHER (2008), upcoming association studies in humans will include roughly one hundred-thousand individuals, allowing for the detection of more

small-effects variants. Such huge sample sizes are unthinkable in forestry at the moment, mainly because of the high cost required to measure so many trees. The development of rapid and cost-effective phenotyping techniques as described in **Chapter III** should nevertheless help reaching larger sample sizes at similar costs. Large collaborative projects including many experimental trials would also have to be favored to combine data and analyses and reach a much larger scale.

Power of association studies and heritability

The statistical power to detect nucleotide variants affecting phenotypes also depends on the magnitude of traits heritability (VISSCHER *et al.* 2008). A large heritability estimate implies a strong correlation between phenotype and genotype, so one could think that loci with an effect on the trait should be more easily detected. However, the heritability value in itself does not provide information on the genetic architecture of the trait. For example, a trait with low heritability can depend on a limited number of loci, some with strong effects (BRENDEL *et al.* 2008), and a trait with high heritability can have hundreds of contributing loci with small effects (see examples cited above in humans and maize, and see SCOTTI-SAINTAGNE *et al.* 2004 in *Quercus*). So far, there does not seem to be any general trend characterizing the genetic architecture of complex traits, with large differences among species and traits for the number QTLs, the number of underlying genes and variants probably involved and the magnitude of their effects, as reviewed in animals (FLINT and MACKAY 2009), but this is likely to be the same across plant species (see the two contrasted examples cited above in *Quercus*). Low to moderate heritabilities have been found in trees for growth and wood quality related traits (see **Chapter III** and bibliography therein), indicating an important contribution of environmental variation to the phenotypic variance of the traits. This can be related to the observation that, despite the relative simplicity of xylem structure, wood is a highly variable material (ZOBEL and VAN BUIJTENEN 1989) that retains environmental and developmental signatures at both morphological and molecular levels (PLOMION *et al.* 2001; MARTINEZ-MEIER *et al.* 2008). For example, even within a single tree, five main types of wood characterized by contrasting chemical, physical and anatomical properties can be identified: i/ earlywood and latewood are formed at the beginning and end of each growing season, respectively, and depend on climatic seasonal variations in temperate regions (PAIVA *et al.* 2008b), ii/ juvenile wood and mature wood are formed at the beginning and end of the tree life, respectively (PAIVA *et al.* 2008a), and iii/ reaction wood is mainly formed in response to gravitational forces sensed in non-vertically growing stems (PAUX *et al.* 2005). In

Chapter III, we discussed the impact of our wood sampling strategy on heritability values: since a range of different growing seasons were sampled for each tree (9 to 24), heritability estimates could be affected by radial ontogenic or annual environmental effects. For industrial applications, much research has focused on measuring whole-tree properties (SCHIMLECK *et al.* 2006; JONES *et al.* 2008), but without considering the elevated intra-individual variability of trees. Focusing on less complex phenotypic variation could perhaps lead to higher trait heritabilities (RIPETTI *et al.* 2008), and thus to more powerful association studies. Developing ecophysiological models to better characterize the phenotypic response of each genotype to environmental variation also appears promising for dissecting complex traits architecture (REYMOND *et al.* 2003). For example, measuring separately the properties of juvenile or mature wood, or at a finer scale the part of a ring associated to earlywood and latewood in a particular growing season, could be of interest if they were showing higher heritability estimates. The drawback is that the relationships between these new "fine-scale" traits and wood end-use properties might not be clear and would also have to be examined.

Context-dependent mutations due to environmental variation

Focusing on less complex phenotypic variation may lead to increased trait heritabilities, although this may not be sufficient to increase power for detecting associations. Focusing on simpler traits could however bring valuable insights into the detection of context-dependent mutations. Indeed, in a specific environment or genetic background, only a part of all the components of genetic architecture is variable and has an impact on the phenotype, and therefore constitutes the visible variation. Other allelic effects are undetectable, either because they are not measurable at the phenotypic level or because they are only detectable under specific genetic or environmental conditions (LE ROUZIC and CARLBORG 2008). The part of genetic diversity that has the potential to affect phenotypic variation, but that is not expressed under the current genotypic or environmental conditions, is generally referred to as cryptic or context-dependent variation. Expressional studies showed that the variability between trees collected in different years was much higher than that found between trees sampled in the same year, reflecting a strong environmental effect on gene expression during xylogenesis (PAIVA *et al.* 2008a). We can thus expect that the effects of different quantitative trait nucleotides (QTNs) will differ in their magnitude or direction depending on the growing season. In that case, association studies will likely not highlight the same QTNs depending on the tree-ring observed due to genotype-by-season interactions. This could for example explain why the associations detected in **Chapter IV**, if they are real, are not repeatable, in addition

to low sample sizes. Although genetic values of individual trees can be technically easily estimated across different environments by progeny or clonal trials, experimental designs devoted to the study of genotype-by-environment interactions are more difficult to install and manage than one-location trials, and are not so common in forest tree species. However, previous results have shown that for traits such as growth and some stem properties, population-by-environment interactions represent a large part of the total phenotypic variance, and that it can be predicted by environmental covariates describing either the experimental sites or the populations original sampling sites (ALÍA *et al.* 1997; GARNIER-GÉRÉ and ADES 2001). This suggests that variants showing effects specific to particular locations are likely to exist and associations should be compared across these. The seasonal variation in wood rings, as biomarkers of environmental changes (MARTINEZ-MEIER *et al.* 2008), could also be used to estimate associations at a much finer level.

Context-dependent mutations due to interactions among genes

Context-dependent effects can be due to different environments, but also to sex in dioecious species, or to different genetic backgrounds due to the interaction among alleles at different loci, a phenomenon called epistasis (CARLBORG and HALEY 2004; MACKAY *et al.* 2009). Even though most biologists would agree that genetic interactions are common in the expression of complex traits, the importance of this phenomenon is not clear. Besides, association methods do not often explicitly test for epistatic effects (CARLBORG and HALEY 2004) as they require larger sample sizes that commonly available, and more complex models than the single-locus ones generally used. Epistasis was first defined by Fisher as the statistical deviation from the additive combination of single allelic effects at 2 or more loci (FISHER 1918). It is important to distinguish this simple definition from the more general physiological definition given above, the latter having an impact on the detection of the effects of particular mutations with different sets of genotypes. Indeed, a different sample of genotypes would represent a different genetic background (*i.e.* different alleles at other unknown loci that would not be accounted for), which could lead to the absence of repeatability of a detected association from one sample to another. In **Chapter III**, using clonally replicated progenies, we found that growth phenotypic variance was essentially due to an additive component. However, the apparent lack of dominance or epistatic variance in the total genotypic variance does not mean that physiological epistasis did not contribute to this additive variance component (CHEVERUD and ROUTMAN 1995). In case of relatedness between individuals, the additive-by-additive component of epistasis could also contribute to

additive variance (GALLAIS 1990). Our results have thus to be taken with caution when the potential impact of epistasis is concerned. HILL *et al.* (2008) also showed that most genetic variance appears to be additive, either because there is indeed little dominant or epistatic gene action, or because genetic variance becomes additive at extreme gene frequencies (which are common in a wide range of species, see for example the "U shape" spectrum of allelic frequencies expected under the standard neutral model, or the Figure 2B from MYLES *et al.* 2009 reproduced above). When modeled explicitly, statistical or physiological epistasis has however be shown to greatly vary among experiments, species and traits (CARLBORG and HALEY 2004). A recent simulation study focusing on gene regulatory networks has shown that applying complex genetic models to a reasonable number of individuals (N=2000) can potentially enhanced our understanding of different types of epistatic connections in gene networks (GJUVSLAND *et al.* 2007). The presence of epistasis can greatly obscure the detection of genotype-phenotype associations: interacting QTLs explaining a large part of the phenotypic variation have been identified for body size in ~750 F2 chickens (CARLBORG *et al.* 2006) or for obesity in mice (STYLIANOU *et al.* 2006), whereas underlying QTLs were barely or not significant when tested individually. Significant QTL or QTN interactions have also been suggested in shoot-branch architecture (EHRENREICH *et al.* 2007) or metabolome (ROWE *et al.* 2008) in *Arabidopsis*, in tomato-fruit quality (CAUSSE *et al.* 2007) or in drug-response in humans (LIN *et al.* 2007). In **Chapter IV**, we used single-marker based tests to detect genotype-phenotype associations, so epistasis was not explicitly modeled, and significant two or more loci interactions involved in phenotypic variation could have been missed. A range of methods have been proposed to test for interaction between loci in association mapping experiments (reviewed in MUSANI *et al.* 2007; CORDELL 2009), but often require extensive computation time and, above all, large sample sizes.

Within genes, we could however look at the problem of interaction among SNPs by testing for haplotype-based effects. Combining the information of several bi-allelic markers (more or less linked) can indeed emulate a new multi-allelic marker that is more informative (LU *et al.* 2003; BUNTJER *et al.* 2005). For example, in *Arabidopsis*, significant associations with flowering time were found with haplotypes that could not have been found using single markers (HAGENBLAD *et al.* 2004). More evolutionary inferences could also be made at the haplotype level in maritime pine, with an increased proportion of significant tests showing departures from neutrality in haplotype- versus single-SNP levels in a set of drought stress tolerance genes (EVENO *et al.* 2008). In our study (**Chapters II and IV**), those genes were included in the genotyping array, but their haplotype information was not recovered due to

technical constraints, probably because some belonged to multigene families. We however used haplotype-based tests in our association mapping experiments (data not shown) for the few genes for which two or more SNPs that were not in complete LD were available. These tests did not allow to detect significant associations, possibly because of low sample sizes (the power was lower than that of single-marker based approaches), but also because we did not have an exhaustive representation of all possible haplotypes in the material studied, so we may have missed the relevant ones. The overall low recovery rate of the SNPs tested with the Illumina array could here be a limiting factor. Employing a genotyping technology that would allow to obtain a more exhaustive haplotype representation of genes is likely to increase the power of association mapping studies.

What is not accounted for yet…

In **Chapter IV**, two association models were used, testing for either a codominant or an additive effect at each marker. Other models were also tested that considered dominant, recessive and over-dominant modes of action for the SNP effects. The results were the same than with the additive model. However, given the general lack of power of our experiments and since dominant QTNs have already been detected (FOURNIER-LEVEL *et al.* 2009; THUMMA *et al.* 2009), it appears important to test for such effects.

In this discussion, we successively examined the possible effects of choosing specific SNPs and candidate genes, sample size, trait heritability, genotype-by-environment interactions, epistasis and dominance on association mapping experiments. Other factors have not been considered that could potentially be of importance for the expression of complex traits: the effects of Copy Number Variations (CNVs) and epigenetic changes, which, although undetected since they don't alter SNP sequences, can play an important role in phenotypic variation. Recent studies in humans found strong associations between CNVs and schizophrenia (STEFANSSON *et al.* 2008), autoimmune diseases (reviewed in SCHASCHL *et al.* 2009) or autism (GLESSNER *et al.* 2009), bridging the gap between the small portion of heritability explained by QTNs and the high heritability of these diseases. However, there is no information available at the moment about the existence and putative role of CNVs in plants (GAUT and ROSS-IBARRA 2008). Epigenetics refers to changes in phenotype or gene expression caused by mechanisms other than changes in the DNA sequence. These mechanisms generally involve either transcriptional silencing, for example by chromatin remodeling (MEYER 2001) or DNA methylation (LAIRD 2003; KALISZ and PURUGGANAN 2004), or post-transcriptional silencing with different RNA degradation pathways (MEYER

2005). Epigenetic processes are essential for development and differentiation, but also arise in mature individuals, either by random changes or under the influence of environment (JAENISCH and BIRD 2003). Studies have shown the importance of epigenetic mechanisms on plant response to environmental stresses (reviewed in LUKENS and ZHAN 2007; see GOURCILLEAU *et al.* in press for an example on poplar). They are rarely accounted for when studying phenotype variation since classical genotyping methods are not designed to detect them. However, they belong to the cryptic variation that could play an important role in adaptive traits variation, and should be taken into account in future experiments. This is especially relevant for perennial tree species which are likely to be submitted to stronger selection pressures in the context of predicted global changes.

Conclusion

Association mapping is an integrative research field that benefits from i / the knowledge on the evolutionary history of populations, essential for understanding the interplay between forces that shape diversity, LD patterns and population structure, ii/ the predictive power of quantitative genetics, which allow to estimate phenotypic values accounting for familial relatedness and micro- or macro-environmental variation in appropriate designs, iii / the maze of genomic data, which is an important resource for choosing expressional and functional candidate genes potentially involved in the surveyed traits, and iv / the recent developments in molecular techniques, which enhance large-scale genotyping and sequencing. Association mapping potentially gives a much finer resolution than pedigrees used for QTL mapping. It is a method of choice for tree species, since it frees breeders from developing segregating populations in species showing high levels of inbreeding depression and long generation times. However, the theoretical finer resolution due to population recombination history comes at the cost of a larger number of markers and much larger sample sizes needed for detecting associations with enough statistical power. Since the number of markers is not a limiting factor nowadays, it is thus crucial to concentrate on experimental designs, so that the potential of molecular tools can be fully exploited. Progress could first be made on phenotype assessment. Compared to humans, trees have the striking advantage of allowing to control and assess genotype-by-environment variation through progeny or clonal trials installed in multiple sites. Moreover, wood tissue in itself is a very good biomarker of fine scale environmental changes, thus could be used for developing new target traits representing reaction norms that would integrate the response to those changes (REYMOND *et al.* 2003). Second, it seems important to invest in statistical methodology for better modeling this

phenotypic variation, in particular for understanding interactions between genes and regulatory networks that have been linked to additivity, dominance and epistasis (OMHOLT *et al.* 2000). Finally, we should carefully conceive appropriate designs with reasonably large sample sizes for both testing those models and reveal the cryptic variation underlying missing heritability.

References

ALÍA, R., J. MORO and J. B. DENIS, 1997 Performance of *Pinus pinaster* provenances in Spain: interpretation of the genotype by environment interaction. Canadian Journal of Forest Research **27:** 1548-1559.

BALL, R. D., 2005 Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics **170:** 859-873.

BRENDEL, O., D. LE THIEC, C. SCOTTI-SAINTAGNE, C. BODÉNÈS, A. KREMER *et al.*, 2008 Quantitative trait loci controlling water use efficiency and related traits in *Quercus robur* L. Tree Genetics & Genomes **4:** 263-278.

BUCKLER, E., J. B. HOLLAND, P. BRADBURY, C. B. ACHARYA and P. J. BROWN, 2009 The genetic architecture of maize flowering time. Science **325:** 714-718.

BUNTJER, J. B., A. P. SORENSEN and J. D. PELEMAN, 2005 Haplotype diversity: the link between statistical and biological association. Trends in Plant Science **10:** 466-471.

CARLBORG, O., and C. S. HALEY, 2004 Epistasis: too often neglected in complex trait studies? Nature Reviews Genetics **5:** 618-625.

CARLBORG, O., L. JACOBSSON, P. AHGREN, P. SIEGEL and L. ANDERSSON, 2006 Epistasis and the release of genetic variation during long-term selection. Nature **38:** 418-420.

CARROLL, S. B., 2005 Evolution at two levels: On genes and form. PLoS Biology **3:** 1159-1166.

CAUSSE, M., J. CHAÏB, L. LECOMTE, M. BURET and F. HOSPITAL, 2007 Both additivity and epistasis control the genetic variation for fruit quality traits in tomato. TAG **115:** 429-442.

CHEVERUD, J. M., and E. J. ROUTMAN, 1995 Epistasis and its contribution to genetic variance components. Genetics **139:** 1455-1461.

CORDELL, H. J., 2009 Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics **10:** 392-404.

DARNELL, J. E., 2002 Transcription factors as targets for cancer therapy. Nature Reviews Cancer **2:** 740-749.

DIXON, A. L., L. LIANG, M. F. MOFFATT, W. CHEN, S. HEATH *et al.*, 2007 A genome-wide association study of global gene expression. Nature Genetics **39:** 1202-1207.

EHRENREICH, I. M., P. A. STAFFORD and M. D. PURUGGANAN, 2007 The genetic architecture of shoot branching in *Arabidopsis thaliana*: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. Genetics **176:** 1223-1236.

EVENO, E., 2008 L'adaptation à la sécheresse chez le pin maritime (*Pinus pinaster* Ait.) : patrons de diversité et différenciation nucléotidiques de gènes candidats et variabilité de caractères phénotypiques. , PhD Thesis, University of Bordeaux 1 (France).

EVENO, E., C. COLLADA, M. A. GUEVARA, V. LEGER, A. SOTO *et al.*, 2008 Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Molecular Biology and Evolution **25:** 417-437.

FISHER, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh **52:** 399-433.

FLINT, J., and T. F. C. MACKAY, 2009 Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research **19:** 723.

FOURNIER-LEVEL, A., L. LE CUNFF, C. GOMEZ, A. DOLIGEZ, A. AGEORGES *et al.*, 2009 Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera* L. ssp sativa) berry: a QTL to QTN integrated study. Genetics **Published Articles Ahead of Print: August 31, 2009.**

FRASER, H. B., A. E. HIRSH, L. M. STEINMETZ, C. SCHARFE and M. W. FELDMAN, 2002 Evolutionary Rate in the Protein Interaction Network. Science **296:** 750-752.

GALLAIS, A., 1990 *Théorie de la sélection en amélioration des plantes*. Masson, Paris.

GARNIER-GÉRÉ, P. H., and P. K. ADES, 2001 Environmental surrogates for predicting and conserving adaptive genetic variability in tree species. Conservation Biology**:** 1632-1644.

GAUT, B. S., and J. ROSS-IBARRA, 2008 Selection on major components of angiosperm genomes. Science **320:** 484-486.

GJUVSLAND, A. B., B. J. HAYES, S. W. OMHOLT and O. CARLBORG, 2007 Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. Genetics **175:** 411-420.

GLESSNER, J. T., K. WANG, G. CAI, O. KORVATSKA, C. E. KIM *et al.*, 2009 Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature **459:** 569-573.

GONZALEZ-MARTINEZ, S. C., D. HUBER, E. ERSOZ, J. M. DAVIS and D. B. NEALE, 2008 Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity **101:** 19-26.

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

GOURCILLEAU, D., M.-B. BOGEAT-TRIBOULOT, D. LE THIEC, C. LAFON-PLACETTE, A. DELAUNAY *et al.*, in press DNA methylation and histone acetylation: genotypic variations in hybrid poplars, impact of water deficit and relationships with productivity. Journal of Experimental Botany.

GRIVET, D., F. SEBASTIANI, S. GONZÁLEZ-MARTÍNEZ and G. VENDRAMIN, 2009 Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine. New Phytologist **9999**.

GUDBJARTSSON, D. F., G. B. WALTERS, G. THORLEIFSSON, H. STEFANSSON, B. V. HALLDORSSON *et al.*, 2008 Many sequence variants affecting diversity of adult human height. Nature Genetics **40:** 609-615.

HAGENBLAD, J., C. TANG, J. MOLITOR, J. WERNER, K. ZHAO *et al.*, 2004 Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics **168:** 1627.

HEDRICK, P. W., 2006 Genetic polymorphism in heterogeneous environments: the age of genomics. Annual Review of Ecology, Evolution, and Systematics **37:** 67-93.

HILL, W. G., M. E. GODDARD and P. M. VISSCHER, 2008 Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genetics **4:** e1000008.

HUGHES, K. A., J. F. AYROLES, M. M. REEDY, J. M. DRNEVICH, K. C. ROWE *et al.*, 2006 Segregating Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics **173:** 1347-1355.

INGVARSSON, P. K., M. V. GARCIA, V. LUQUEZ, D. HALL and S. JANSSON, 2008 Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics **178:** 2217-2226.

JAENISCH, R., and A. BIRD, 2003 Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics **33:** 245-254.

JONES, P. D., L. R. SCHIMLECK, R. F. DANIELS, A. CLARK and R. C. PURNELL, 2008 Comparison of *Pinus taeda* L. whole-tree wood property calibrations using diffuse reflectance near infrared spectra obtained using a variety of sampling options. Wood Science and Technology **42:** 385-400.

KALISZ, S., and M. D. PURUGGANAN, 2004 Epialleles via DNA methylation: consequences for plant evolution. Trends in Ecology & Evolution **19:** 309-314.

KEURENTJES, J. J. B., J. FU, I. R. TERPSTRA, J. M. GARCIA, G. VAN DEN ACKERVEKEN *et al.*, 2007 Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proceedings of the National Academy of Sciences **104:** 1708.

KIRST, M., C. J. BASTEN, A. A. MYBURG, Z. B. ZENG and R. R. SEDEROFF, 2005 Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. Genetics **169:** 2295-2303.

KIRST, M., A. A. MYBURG, J. P. G. DE LEON, M. E. KIRST, J. SCOTT *et al.*, 2004 Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiology **135:** 2368-2378.

LAIRD, P. W., 2003 The power and the promise of DNA methylation markers. Nature Reviews Cancer **3:** 253-266.

LE ROUZIC, A., and Ö. CARLBORG, 2008 Evolutionary potential of hidden genetic variation. Trends in Ecology & Evolution **23:** 33-37.

LIN, M., H. LI, W. HOU, J. A. JOHNSON and R. WU, 2007 Modeling sequence sequence interactions for drug response. Bioinformatics **23:** 1251-1257.

LOPEZ-BIGAS, N., B. J. BLENCOWE and C. A. OUZOUNIS, 2006 Highly consistent patterns for inherited human diseases at the molecular level. Bioinformatics **22:** 269-277.

LU, X., T. H. NIU and J. S. LIU, 2003 Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms. Genome Research **13:** 2112-2117.

LUKENS, L. N., and S. ZHAN, 2007 The plant genome's methylation status and response to stress: implications for plant improvement. Current Opinion in Plant Biology **10:** 317-322.

LUO, Z. W., 1998 Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity **80:** 198-208.

MACKAY, T. F. C., E. A. STONE and J. F. AYROLES, 2009 The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics **10:** 565-577.

MAHER, B., 2008 Personal genomes: The case of the missing heritability. Nature **456:** 18-21.

MAMADA, M., T. YORIFUJI, J. YORIFUJI, K. KUROKAWA, M. KAWAI *et al.*, 2007 Fibrillin I gene polymorphism is associated with tall stature of normal individuals. Human Genetics **120:** 733-735.

MANOLIO, T. A., F. S. COLLINS, N. J. COX, D. B. GOLDSTEIN, L. A. HINDORFF *et al.*, 2009 Finding the missing heritability of complex diseases. Nature **461:** 747-753.

MARTINEZ-MEIER, A., L. SANCHEZ, M. PASTORINO, L. GALLO and P. ROZENBERG, 2008 What is hot in tree rings? The wood density of surviving Douglas-firs to the 2003 drought and heat wave. Forest Ecology and Management **256:** 837-843.

MEYER, P., 2001 Chromatin remodelling. Current Opinion in Plant Biology **4:** 457-462.

MEYER, P., 2005 *Plant Epigenetics*. Blackwell Publishing.

MORGANTE, M., and F. SALAMINI, 2003 From plant genomics to breeding practice. Current Opinion in Biotechnology **14:** 214-219.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

MUSANI, S. K., D. SHRINER, N. LIU, R. FENG, C. S. COFFEY *et al.*, 2007 Detection of Gene × Gene Interactions in Genome-Wide Association Studies of Human Population Data. Human Heredity **63:** 67-84.

MYLES, S., J. PEIFFER, P. J. BROWN, E. S. ERSOZ, Z. ZHANG *et al.*, 2009 Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell **21:** 2194-2202.

NEALE, D. B., and O. SAVOLAINEN, 2004 Association genetics of complex traits in conifers. Trends in Plant Science **9:** 325-330.

OMHOLT, S. W., E. PLAHTE, L. OYEHAUG and K. XIANG, 2000 Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. Genetics **155:** 969-980.

PAIVA, J. A. P., M. GARCES, A. ALVES, P. GARNIER-GERE, J. C. RODRIGUES *et al.*, 2008a Molecular and phenotypic profiling from the base to the crown in maritime pine wood-forming tissue. New Phytologist **178:** 283-301.

PAIVA, J. A. P., P. H. GARNIER-GERE, J. C. RODRIGUES, A. ALVES, S. SANTOS *et al.*, 2008b Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. New Phytologist **179:** 1180-1194.

PARAN, I., and D. ZAMIR, 2003 Quantitative traits in plants: beyond the QTL. Trends in Genetics **19:** 303-306.

PAUX, E., V. CAROCHA, C. MARQUES, A. M. DE SOUSA, N. BORRALHO *et al.*, 2005 Transcript profiling of Eucalyptus xylem genes during tension wood formation. New Phytologist **167:** 89-100.

PETER, G. F., D. E. WHITE, R. D. L. TORRE and R. SINGH, 2007 The value of forest biotechnology: a cost modelling study with loblolly pine and kraft linerboard in the southeastern USA. International Journal of Biotechnology **9:** 415-435.

PLOMION, C., G. LEPROVOST and A. STOKES, 2001 Wood formation in trees. Plant Physiology **127:** 1513-1523.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist **167:** 101-112.

PURUGGANAN, M. D., A. L. BOYLES and J. I. SUDDITH, 2000 Variation and selection at the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. Genetics **155:** 855-862.

REYMOND, M., B. MULLER, A. LEONARDI, A. CHARCOSSET and F. TARDIEU, 2003 Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiology **131:** 664-675.

RIPETTI, V., J. ESCOUTE, J. L. VERDEIL and E. COSTES, 2008 Shaping the shoot: the relative contribution of cell number and cell shape to variations in internode length between parent and hybrid apple trees. Journal of Experimental Botany.

ROWE, H. C., B. G. HANSEN, B. A. HALKIER and D. J. KLIEBENSTEIN, 2008 Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. Plant Cell **20:** 1199-1216.

SANNA, S., A. U. JACKSON, R. NAGARAJA, C. J. WILLER, W.-M. CHEN *et al.*, 2008 Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet **40:** 198-203.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHASCHL, H., T. J. AITMAN and T. J. VYSE, 2009 Copy number variation in the human genome and its implication in autoimmunity. Clinical & Experimental Immunology **156:** 12-16.

SCHIMLECK, L. R., G. D. S. P. REZENDE, B. J. DEMUNER and G. M. DOWNES, 2006 Estimation of whole-tree wood quality traits using near infrared spectra collected from increment cores. . Appita Journal **59:** 231-236.

STEFANSSON, H., D. RUJESCU, S. CICHON, O. P. H. PIETILAINEN, A. INGASON *et al.*, 2008 Large recurrent microdeletions associated with schizophrenia. Nature **455:** 232-236.

STYLIANOU, I., R. KORSTANJE, R. LI, S. SHEEHAN, B. PAIGEN *et al.*, 2006 Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. Mammalian Genome **17:** 22-36.

TAO, H., D. R. COX and K. A. FRAZER, 2006 Allele-specific *KRT1* expression is a complex trait. PLoS Genetics **2:** e93.

THUMMA, B. R., B. A. MATHESON, D. ZHANG, C. MEESKE, R. MEDER *et al.*, 2009 Identification of a *cis*-acting regulatory polymorphism in a eucalypt *Cobra*-like gene affecting cellulose content. Genetics **Published Articles Ahead of Print: September 7, 2009**.

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

VAQUERIZAS, J. M., S. K. KUMMERFELD, S. A. TEICHMANN and N. M. LUSCOMBE, 2009 A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics **10:** 252-263.

VEYRIERAS, J. B., S. KUDARAVALLI, S. Y. KIM, E. T. DERMITZAKIS, Y. GILAD *et al.*, 2008 High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genetics **4**.

VISSCHER, P. M., 2008 Sizing up human height variation. Nature Genetics **40:** 489-490.

VISSCHER, P. M., W. G. HILL and N. R. WRAY, 2008 Heritability in the genomics era - concepts and misconceptions. Nature Reviews Genetics **9:** 255-266.

WACHOWIAK, W., P. A. BALK and O. SAVOLAINEN, 2009 Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (Pinus sylvestris L.). Tree Genetics & Genomes **5:** 117-132.

WANG, H., T. NUSSBAUM-WAGLER, B. L. LI, Q. ZHAO, Y. VIGOUROUX *et al.*, 2005a The origin of the naked grains of maize. Nature **436:** 714-719.

WANG, W. Y. S., B. J. BARRATT, D. G. CLAYTON and J. A. TODD, 2005b Genome-wide association studies: Theoretical and practical concerns. Nature Reviews Genetics **6:** 109-118.

WEEDON, M. N., G. LETTRE, R. M. FREATHY and C. M. LINDGREN, 2007 A common variant of HMGA2 is associated with adult and childhood height in the general population. Nature Genetics **39:** 1245.

WEST, M. A. L., K. KIM, D. J. KLIEBENSTEIN, H. VAN LEEUWEN, R. W. MICHELMORE *et al.*, 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics **175:** 1441-1450.

WITTKOPP, P. J., 2005 Genomic sources of regulatory variation in cis and in trans. Cellular and Molecular Life Sciences **62:** 1779-1783.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER *et al.*, 2003 The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution **20:** 1377-1419.

YU, H., D. GREENBAUM, H. XIN LU, X. ZHU and M. GERSTEIN, 2004 Genomic analysis of essentiality within protein networks. Trends in Genetics **20:** 227-231.

ZOBEL, B. J., and J. P. VAN BUIJTENEN, 1989 *Wood variation: its causes and control*. Springer Verlag, Berlin.

Lepoittevin Camille, 2009.
UMR BIOGECO 1202, INRA, 69 route d'Arcachon, 33612 Cestas Cedex.
FCBA, Laboratoire de Biotechnologies, Domaine de l'Etançon, 77370 Nangis.

Résumé

Au cours des quarante dernières années, l'optimisation des méthodes sylvicoles et l'introduction de variétés améliorées ont permis d'accroître considérablement la productivité du pin maritime. Pour permettre à la filière bois de disposer d'une matière première de qualité sur ce matériel amélioré, un programme de recherches multidisciplinaire a été développé afin d'étudier le déterminisme génétique de la qualité du bois. Neuf facteurs de transcription potentiellement impliqués dans la xylogenèse et l'adaptation des arbres à leur milieu, ont tout d'abord été séquencés dans la population Aquitaine, et leurs patrons de diversité nucléotidique ont été étudiés. Ces patrons ont été comparés à l'attendu de modèles neutres d'évolution et s'en écartent par un niveau élevé de déséquilibre de liaison et l'excès de mutations en fréquences intermédiaires détectés pour trois de ces gènes (HDZ31, LIM2 et MYB1). Ces résultats suggèrent des changements de taille de population affectant l'ensemble du génome, et l'action de sélection balancée sur l'un d'entre eux (MYB1). Les géniteurs de la population d'amélioration Aquitaine ont ensuite été génotypés pour 384 marqueurs moléculaires et évalués pour la croissance et les propriétés chimiques du bois. Ces données moléculaires et phénotypiques ont permis de mettre en évidence des associations significatives entre la variation pour le diamètre du tronc ou la teneur en cellulose du bois et deux marqueurs situés respectivement dans un facteur de transcription HD-Zip (HDZ31) et dans un gène encodant une fascicline. La cohérence des résultats de génétique évolutive et de génétique d'association ouvre ainsi des perspectives encourageantes pour la compréhension de l'architecture génétique de la formation du bois chez cette espèce. Cependant, le faible nombre d'associations significatives pose de nombreux problèmes théoriques et méthodologiques qui sont discutés en vue d'améliorations pour de nouveaux designs expérimentaux.
Mots-clefs : Pin maritime, qualité du bois, diversité nucléotidique, bottleneck, sélection naturelle, gènes candidats, facteurs de transcription, génétique d'association

---

Abstract

During the last four decades, the optimization of silvicultural and tree breeding methods has contributed to improve growth and wood homogeneity of maritime pine. In order to provide the different actors of the forestry wood-chain with high quality raw material, the genetic determinism and chemical components of wood quality are being studied in the frame of a multidisciplinary research program. First, nine transcription factors putatively involved in wood formation have been sequenced in the Aquitaine population, and their nucleotide diversity pattern studied. Since these genes potentially play important roles in the adaptation of trees to their environment, their patterns have been compared to those expected under neutral evolution. Strong departures from neutrality were observed, with high levels of linkage disequilibrium and an excess of intermediate frequency variants for three of them (HDZ31, LIM2 and MYB1), which could be linked to population size changes that affected the whole genome, and to balancing selection effects at one of them (MYB1). Secondly, the genitors of the Aquitaine breeding population were genotyped for 384 markers and evaluated for growth and wood chemical properties. Significant associations were detected for two markers, one in a HD-Zip transcription factor (HDZ31) with growth, and the other in a gene coding for a fasciclin protein with cellulose content. The consistency of evolutionary and molecular genetics opens encouraging perspectives for understanding the genetic architecture of wood formation in this species. However, the low number of associations detected raises several theoretical and methodological issues which are discussed for the perspective of improving future experimental designs.
Keywords: Maritime pine, wood quality, nucleotide diversity, bottleneck, natural selection, candidate genes, transcription factors, association mapping