HAL
open science

Prédiction de la qualité des bois de chêne pour l'élevage des vins et des alcools : comparaison des approches physicochimiques, sensorielles et moléculaires

Erwan Guichoux

To cite this version:

Erwan Guichoux. Prédiction de la qualité des bois de chêne pour l'élevage des vins et des alcools : comparaison des approches physicochimiques, sensorielles et moléculaires. Sciences du Vivant [q-bio]. Université des Sciences et Technologies (Bordeaux 1), 2011. Français. NNT : . tel-02805620

HAL Id: tel-02805620

https://hal.inrae.fr/tel-02805620v1

Submitted on 6 Jun 2020

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

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

THÈSE

Présentée à

L'UNIVERSITÉ DE BORDEAUX 1

Ecole Doctorale Sciences et Environnements

par Erwan GUICHOUX

pour obtenir le grade de

DOCTEUR

SPECIALITÉ : Ecologie évolutive, fonctionnelle et des communautés

Prédiction de la qualité des bois de chêne pour l'élevage des vins et des alcools : comparaison des approches physicochimiques, sensorielles et moléculaires

Soutenue le 6 avril 2011

Devant la commission d'examen formée de :

M. Jacques BONNET	Professeur, Université de Bordeaux 2	Président
Mme Frédérique PELSY	Directeur de Recherche, INRA, Colmar	Rapporteur
M. Christian LEXER	Professeur, Université de Fribourg	Rapporteur
M. Benoit COLONNA-CECCALDI	Chef du service Technologie du vivant , CRPR, Créteil	Co-directeur de thèse
M. Rémy PETIT	Directeur de Recherche, INRA, Bordeaux	Directeur de thèse

UMR BIOGECO 1202 – 69 route d'Arcachon 33612 CESTAS Cedex - France

REMERCIEMENTS

Je tiens à remercier en tout premier lieu mon directeur de thèse, Rémy Petit. Merci de m'avoir accompagné tout au long de ces années, depuis nos premières collaborations il y a maintenant sept ans. Merci pour ta disponibilité, ton aide précieuse et pour avoir encadré cette thèse avec juste ce qu'il faut de présence, et pour m'avoir laissé faire mes propres choix (même mauvais !). Merci pour toutes ces discussions conceptuelles qui m'ont fait prendre un peu de recul sur mon travail. J'espère que notre collaboration scientifique n'en ait qu'à ses débuts et je me réjouis à l'idée de continuer cette aventure sur la traçabilité du bois avec toi.

Je remercie également Frédérique Pelsy et Christian Lexer qui ont accepté d'être rapporteurs de ce travail de thèse, ainsi que Jacques Bonnet pour avoir accepté de présider ce jury.

Cette thèse est le fruit du travail d'un grand nombre de personnes et je tiens à en remercier certaines en particulier :

- Merci Lélia pour tout ce que tu as apporté à ce travail de thèse, en particulier toutes les données génétiques SSRs et SNPs. Partager le même bureau que toi pendant cette thèse est un plaisir et tu vas devoir encore me supporter un peu !
- Merci Steffi et Patrick pour votre aide importante lors de la mise au point du multiplexage SSRs.
- Merci à Fredo-le-magicien pour sa disponibilité de tous les instants et pour avoir trouvé une solution à tous mes problèmes techniques sur le bois.
- Merci à Pauline pour sa disponibilité et ses conseils précieux.
- Merci aux collègues du Centre de Recherche Pernod Ricard pour leur aide constante au cours de cette thèse, en particulier Nicolas pour la partie sensorielle.
- Merci à Alexis et Jean-Marc de m'avoir fait découvrir Pierroton il y a maintenant 11 ans.
- Merci à Henri-Gandalf-le-magicien de m'avoir prouvé qu'il y a pire à extraire que l'ADN de chêne. Merci aussi pour tout le reste.
- Un immense merci à Oliv' pour les innombrables coups de main lors de la rédaction des papiers et pour les années passées ensemble dans le même bureau.

Merci à mes amis François, Greg et Loïc pour ces longues discussions sur tout et sur rien.

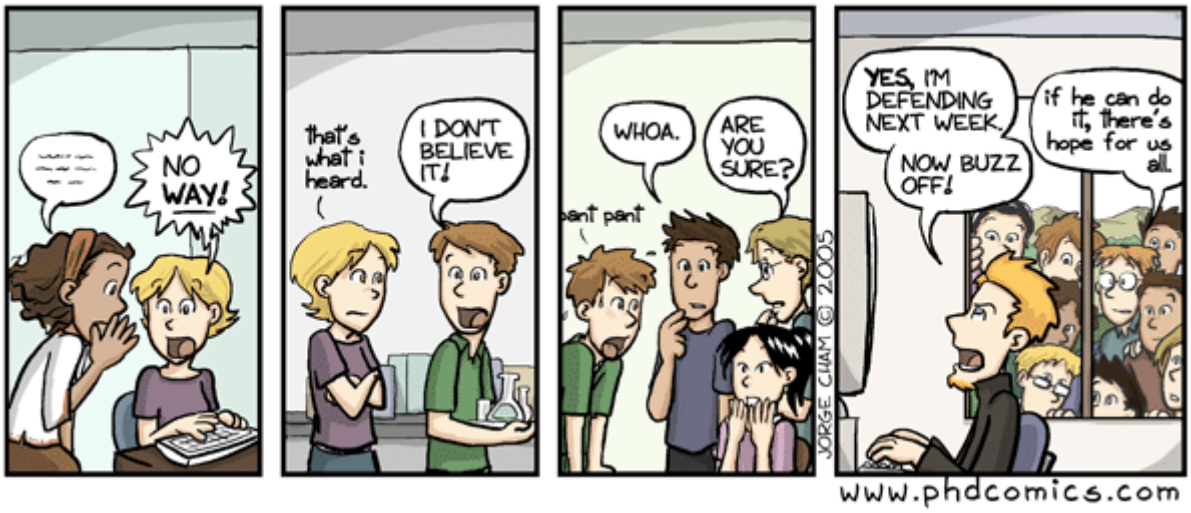
Merci à tous les Pierrotonnais qui rendent ce lieu de travail si agréable, les innombrables gâteaux de 10h y étant pour beaucoup ! Merci à tous d'avoir participé l'été dernier aux tests sensoriels sur copeaux et merci à tous ceux, qui de près ou de loin, ont contribué à ce travail de thèse.

Un immense merci à mes parents pour leur soutien de tous les instants. Merci de m'avoir laissé choisir ma voie, même si vous n'étiez pas forcément convaincus au départ. Merci de m'avoir guidé et soutenu depuis toujours, particulièrement dans les moments difficiles.

Merci à toute ma famille et à ma belle-famille pour leur affection et les bonnes parties de rigolade en toutes circonstances. Mention particulière à mon beauf pour ses relectures d'articles.

Merci à mon p'tit gars de m'avoir offert quelques pauses Playmobil ou pâte à modeler, particulièrement en cette fin de thèse. Rien de tel pour penser à autre chose et relativiser un peu.

Enfin, merci surtout à toi ma Julie pour tout ce que tu sais. Merci d'avoir supporté mes états d'âme en cette fin de thèse et d'avoir toujours été là.



SOMMAIRE

INTRODUCTION **1**
Objectifs de la thèse 11
Références 15

CHAPITRE 1 : Current trends in microsatellite genotyping **19**
Introduction 22
A review of current practices 24
SSR selection 25
Source of sequence data 25
From transcriptome to whole genome shotgun sequencing for SSR detection 26
Read-length 27
Advantages of next-generation sequencing 27
Choice of SSR type 28
Perfect or imperfect repeats 28
Size of repeat unit 28
Number of repeat units 29
Primer design 30
Primer validation in simplex 30
The multiplexing phase 32
Sizing precision 35
Allele calling and binning 35
Measuring and reporting error rates 37
Data management 38
Conclusions et perspectives 41
Box 1: SSRs versus SNPs 43
Advantages of SSRs over SNPs 44
Drawbacks of SSRs over SNPs 45
Box 2: Cost effectiveness of multiplex SSR typing 48
Other solutions to decrease costs 49
Box 3: Problems arising during SSR amplification 50
Acknowledgments 54
Supporting information 54
References 55

CHAPITRE 2 : Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.) **65**
Introduction 68
Material and Methods 69

Material 69
DNA isolation 70
Multiplex PCR optimization 70
Kit-1 70
Kit-2 70
Diversity analyses and assignment power 71
Microsatellites scoring 72
Error rate measurement 72
Results 73
Multiplex PCR optimization 73
SSR properties 73
Assignment power 75
SSR transferability 77
Microsatellites scoring and binning 77
Error rate measurement 77
Conclusion 78
Acknowledgements 78
Authors' contributions 79
Supporting information 80
References 82

CHAPITRE 3 : DNA-based identification of tree species from wood: application to oak staves 85
Introduction 88
Material and methods 91
Plant material 91
DNA isolation 91
Quantification of dsDNA 93
Real-time PCR to optimize DNA isolation 93
SSR genotyping 94
PCR inhibitor test 95
Species identification of wood samples 95
Results and discussion 95
Quantification of dsDNA 95
Real-time PCR to optimize DNA isolation 97
SSR genotyping 99
PCR inhibitor test 99
Species identification of wood samples 100
Conclusion and perspectives 101
Acknowledgments 102
References 103

CHAPITRE 4 : Genes under selection provide unique insights on oak trees demography 109
Introduction 112
Material & methods 116
Material 116
DNA isolation 116
SNP selection 116
SNP genotyping 117
Assignment methods for accurate species delimitation 117
Diversity analyses 118
Detection of outlier loci 119
Comparison of genotype likelihoods between species 119
Results 120
SNP genotyping 120
Species assignment 120
Genetic differentiation among populations 122
Genotype likelihoods and asymmetric introgression 124
Discussion 127
Species delimitation 127
Directional interspecific gene flow 128
Measuring intraspecific gene flow 129
Perspectives 129
Supporting information 130
References 133

CONCLUSIONS ET PERSPECTIVES 139
De l'importance du choix des marqueurs génétiques 140
Les contraintes liées à l'amplification d'ADN nucléaire à partir de bois 141
Les microsatellites supplantés par les SNPs ? 143
Perspectives appliquées pour la filière bois 145
L'apport significatif des marqueurs sous sélection 146
Références 148

ANNEXES 151

INTRODUCTION

Cette thèse CIFRE (Conventions Industrielles de Formation par la Recherche), sous l'impulsion de l'ANRT (Association Nationale de le Recherche et de la Technologie), a été financée par le CRPR (Centre de Recherche Pernod Ricard) et l'INRA (Institut National de la Recherche Agronomique). Elle a débuté en février 2008 pour une durée de trois ans. Elle s'intègre dans le projet « OakWood », coordonné par Daniel Derchue au sein du CRPR, qui vise à mieux caractériser les bois utilisés pour le vieillissement des vins et alcools du groupe Pernod. L'essentiel des travaux présentés ici a été réalisé au sein de l'UMR BIOGECO, dans l'équipe de Génétique, sous la direction de Rémy Petit. Les résultats de ces travaux sont présentés sous la forme d'une thèse sur articles. Après une introduction générale qui présente les objectifs de cette thèse, quatre chapitres sous forme d'articles en anglais sont présentés, suivis d'une conclusion générale et des perspectives possibles à ce travail.

Il y a plus de 6000 ans, l'homme cultivait déjà le raisin pour produire du vin. La découverte du plus vieux pressoir connu à ce jour, découvert en 2010 en Arménie dans la région du Vayots Dzor, nous renseigne sur les procédés de vinification utilisés alors (Barnard *et al.*, 2011). A l'époque, le vin était stocké dans des jarres en terre et il faudra attendre l'époque gallo-romaine pour voir apparaitre les premiers tonneaux en bois, au départ destinés à stocker la cervoise et l'eau. Ce sont les Romains qui, dès le IIIème siècle, ont commencé à stocker le vin dans des tonneaux en bois de chêne, en remplacement des amphores jugées trop fragiles. Au cours du XXème siècle, les viticulteurs ont découvert qu'au-delà d'être un moyen de stockage, les tonneaux en bois de chêne pouvaient améliorer les propriétés aromatiques du vin (Chatonnet, 1995a). Aujourd'hui encore, des vins produits dans le monde entier sont stockés dans des tonneaux en chêne, pour assurer leur conservation, mais surtout pour développer tout leur potentiel aromatique.

Car bien que d'autres essences aient été envisagées pour le vieillissement du vin (Young *et al.*, 2010), les fûts utilisés de nos jours demeurent quasi exclusivement réalisés avec du bois de chêne. Au cours du vieillissement, le bois de chêne apporte au vin des notes sensorielles désignées sous le vocable général de boisé qui renforce les qualités intrinsèques du vin et apporte un supplément aromatique appréciable. Près de la moitié des composés aromatiques du vin seraient ainsi directement liés à la maturation au contact du bois de

chêne (Boidron *et al.*, 1988 ; Jarauta *et al.*, 2005). Parmi les composés les plus aromatiques transmis par le bois de chêne au vin, la whisky-lactone (β-methyl-γ-octalactone) est le composé le plus cité, devant la vanilline, l'eugénol et le gaïacol (Sauvageot & Feuillat, 1999; Doussot *et al.*, 2002; Prida & Chatonnet, 2010). Cette molécule, qui dépasse largement les seuils de perception, est caractérisée par des notes intenses de noix de coco et de bois frais.

Le bois de chêne utilisé pour le vieillissement des vins peut être utilisé sous plusieurs formes, du classique tonneau aux alternatifs que sont les planches, copeaux ou poussières. Ces alternatifs, autorisés en vinification depuis 2009 au sein de l'Union Européenne, sont généralement ajoutés après la fermentation alcoolique pour des durées courtes (en moyenne un à six mois contre un à deux ans pour l'élevage en fûts). Ils permettent d'accélérer le processus de vieillissement pour des coûts inférieurs (en moyenne 88 €/hL de vin avec un fût contre 6 €/hL de vin avec des copeaux, source Institut Français de la Vigne et du Vin).

La filière bois représente un chiffre d'affaire annuel de 60 milliards d'euros et emploie plus de 400.000 personnes en France (source : Office National des Forêts). Directement lié à cette filière, au travers de l'industrie de la tonnellerie, le marché des vins et spiritueux représente en France plus de 80.000 emplois directs, pour un chiffre d'affaire annuel de près de 20 milliards d'euros, dont la moitié à l'export (source : Fédération des exportateurs de vins et spiritueux). Les tonneaux produits en France (pour une valeur estimée à 400 millions d'euros en 2006) fournissent 75% de la demande mondiale. Environ 80 % de ces tonneaux sont exportés vers les « nouveaux » pays producteurs de vin produits en fûts (à plus de 50 % vers les Etats-Unis, puis vers l'Australie, le Chili et l'Afrique du Sud). En 2006, environ 300.000 m^3 de bois à merrains (destinés à la tonnellerie) ont été commercialisés en France (soit plus de 10% du volume de bois d'œuvre de chêne mis sur le marché correspondant à 30% des revenus en valeur), pour un montant estimé à 120 millions d'euros/an. Au sein du groupe Pernod, 7.5 millions de fûts de chêne sont actuellement en vieillissement, et tous les ans ce sont près de 300.000 fûts qui sont achetés, ainsi que plusieurs dizaine de tonnes de copeaux et près de 40.000 planches.

Pour la fabrication des fûts et autres alternatifs destinés au vieillissement des vins, les espèces les plus recherchées sont le chêne sessile - *Quercus petraea* (Matt.) Liebl. et le chêne pédonculé - *Quercus robur* L. (Boidron *et al.*, 1988). Les chênes nord-américains (le chêne rouge (*Quercus rubra*) et le chêne blanc d'Amérique (*Quercus alba*)), bien que de plus en plus

utilisés, sont encore souvent réservés au vieillissement des alcools forts, car très riches en tanins. Le chêne sessile et le chêne pédonculé ont une large aire de répartition en Europe qui s'étend du nord de la péninsule ibérique jusqu'à la Russie, l'aire du chêne sessile étant limitée au nord et à l'est par rapport au chêne pédonculé à cause de sa sensibilité au froid (Figure 1). Les chênes pédonculés et sessiles représentent à eux deux plus de 40% de la forêt française.

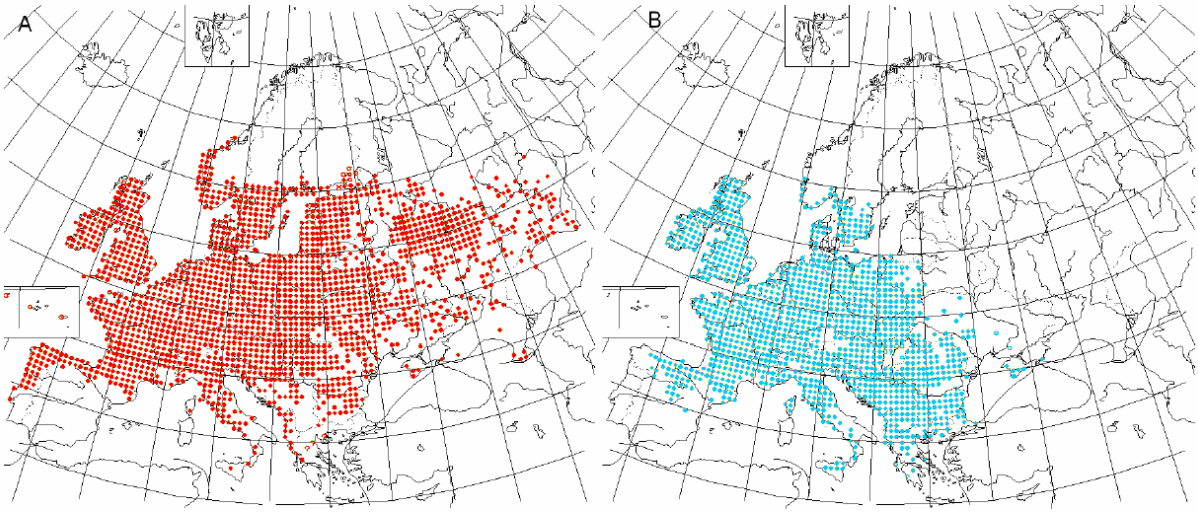


Figure 1 : Aires de répartition européenne du chêne pédonculé – *Q. robur* (A) et du chêne sessile – *Q. petraea* (B). Adapté de Atlas Florae Europaeae (1999) et de Lepais (2008).

Ces deux espèces, outre leur fort intérêt économique pour la filière des vins et spiritueux, ont été très étudiées depuis de nombreuses années comme modèle d'évolution, d'adaptation et de spéciation (Bodénès *et al.*, 1997 ; Streiff *et al.*, 1998; Streiff *et al.*, 1999; Muir *et al.*, 2000; Petit *et al.*, 2002; Petit *et al.*, 2004; Lepais *et al.*, 2009 voir **Annexe 1**; Lepais & Gerber, 2010). Des marqueurs génétiques de différents types (isozymes, RAPD, SCAR et microsatellites) ont montré que les chênes pédonculés et sessiles sont peu différenciés génétiquement (Bodénès *et al.*, 1997; Muir *et al.*, 2000; Gömöry *et al.*, 2001 ; Coart *et al.*, 2002; Mariette *et al.*, 2002). Aucun marqueur diagnostique n'a été identifié à ce jour et les espèces diffèrent seulement par leurs fréquences alléliques. Cette différenciation faible, malgré de relativement fortes différences phénotypiques (morphologiques et écologiques), pourrait être liée aux flux de gènes importants qui existent entre ces deux espèces (Bacilieri *et al.*, 1993; Jensen *et al.*, 2009; Lepais & Gerber, 2010).

Aujourd'hui encore, une grande partie des bois de chêne destinés à la tonnellerie sont sélectionnés sur la base du grain (largeur des cernes de croissance annuelle) ou de l'origine géographique (Chatonnet, 1995a; Chatonnet, 1995b ; Spillman *et al.*, 2004). Les bois à grain fin (<2mm) sont particulièrement recherchés car ils sont réputés pour être plus riches en arômes, alors que les bois à gros grain (>2mm) seraient plus riches en tanins (Mosedale *et al.*, 1996). Le gros grain est habituellement associé à *Q. robur* tandis que le grain fin est associé à *Q. petraea* (Vivas *et al.*, 1997; Feuillat *et al.*, 1998), bien que cela soit très dépendant des populations étudiées (Doussot *et al.*, 2002).

Pourtant, de nombreuses études ont confirmé que c'est l'effet « espèce » qui explique le mieux les variabilité des qualités aromatiques conférées aux vins et aux alcools, bien plus que l'effet « grain » ou « origine géographique » (Auer *et al.*, 2006; Guchu *et al.*, 2006; Prida *et al.*, 2006; Prida & Puech, 2006; Prida *et al.*, 2007). Il est d'ailleurs possible de différencier les deux espèces sur la seule base des tanins et composés volatiles du bois (Mosedale *et al.*, 1998). Dans une étude publiée en 2006, Auer *et al.* ont démontré que les différences aromatiques les plus significatives entre lots de bois concernaient les deux espèces et principalement la whisky-lactone (Figure 2). L'effet « origine géographique » est lui plus secondaire. Les niveaux de whisky-lactones chez *Q. robur* sont très faibles alors qu'ils sont très élevés chez une majorité des *Q. petraea* (Prida *et al.*, 2007).

L'effet « espèce » est directement mesurable par flairage de copeaux, *Q. petraea* est caractérisé par des odeurs « noix de coco », « boisé », « céleri » et « vanille » plus intenses et une odeur « foin » moins intense que *Q. robur*, alors que les différences de grain n'ont guère d'effet (Sauvageot *et al.*, 2002; Prida *et al.*, 2007). Ces résultats sensoriels ont d'ailleurs été confirmés lors de tests par flairage de copeaux que j'ai mis en place au sein de notre unité (Figure 3 et **Annexe 2**).

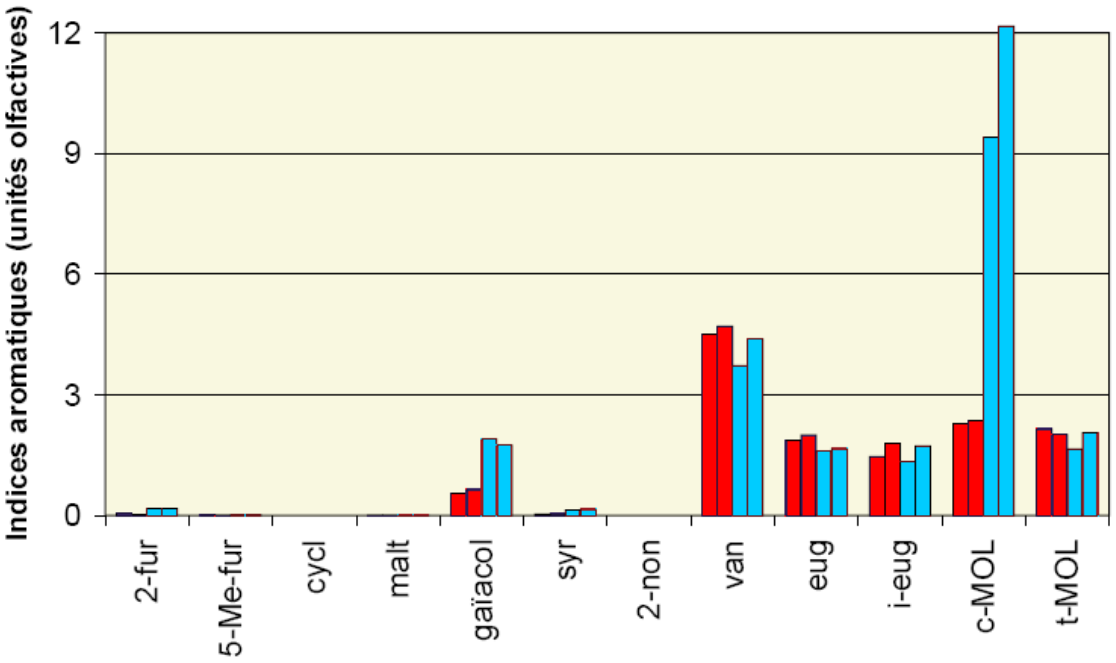


Figure 2 : Indices aromatiques (concentration pondérée par le seuil de détection de la molécule) pour 12 composés volatils aromatiques majeurs transmis au vin par le bois de chêne (Auer *et al.*, 2006). Deux lots de bois sont évalués pour chaque espèce (*Q. robur* en rouge et *Q. petraea* en bleu). Seule la whisky-lactone (isomère *cis* noté ici c-MOL) présente des différences très significatives entre espèces, et dans une moindre mesure le gaïacol (note « fumée »).

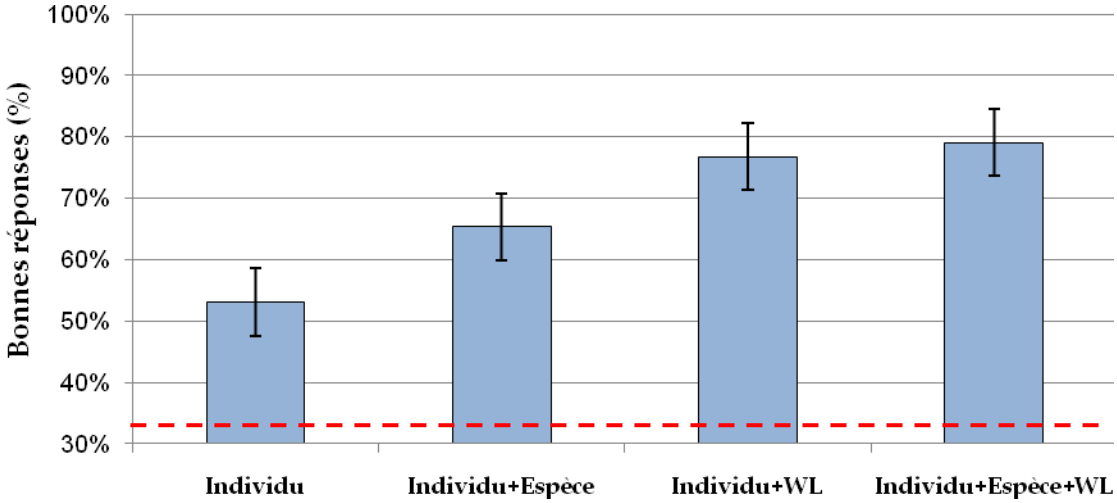


Figure 3 : Résultats des 2250 tests triangulaires sur copeaux de bois réalisés à Pierroton en juin 2010 (voir Annexe 2). Le taux de réponses correctes (i.e. reconnaitre l'échantillon différent des deux autres) est indiqué en ordonnée. Trois effets sont testés seuls ou en combinaison : individu, espèce et whisky-lactone (WL). La ligne rouge symbolise le taux attendu de bonnes réponses si les réponses sont données au hasard (33%).

Ces tests sensoriels mettent en avant plusieurs choses. D'une part, l'effet individuel est non-négligeable, et malgré des profils aromatiques quasi identiques (pour les whisky-lactones), plus de la moitié des testeurs identifient correctement l'échantillon qui diffère des deux autres. D'autre part, il existe un fort effet lié à l'espèce et un très fort effet lié aux différences de teneurs en whisky-lactones. L'effet « espèce » (et donc indirectement l'effet whisky-lactone, comme nous l'avons vu plus haut) persiste même au-delà des étapes de séchage et de chauffe (Doussot *et al.*, 2002). Le séchage à l'air libre, qui dure en moyenne entre 18 et 36 mois, a pour buts principaux d'abaisser l'hygrométrie des merrains à des valeurs comprises entre 14 et 18% d'humidité relative (assurant de bonnes propriétés des bois lors de la construction des fûts) et de modifier la composition physico-chimique du bois (diminution des teneurs en tanins et en vanilline en particulier). A l'issue de cette étape de séchage, le bois de chêne est appelé douelle. La chauffe est l'autre grande étape de transformation du bois de chêne avant son utilisation concrète pour le vieillissement des vins et alcools. Cette étape, dont l'intensité peut varier selon le résultat escompté, a un double objectif : permettre de cintrer les douelles afin d'assembler le tonneau (chauffe de cintrage) et développer certaines molécules aromatiques (principalement la vanilline et dans une moindre mesure les whisky-lactones).

Si les deux espèces (*Q. robur* et *Q. petraea*) préférentiellement utilisées sont désormais systématiquement différenciées en forêt par les sylviculteurs et dans la filière d'élevage des plants, la distinction taxonomique n'est pas maintenue dans la filière bois, faute de volonté, de critères de reconnaissance et de tests fiables sur bois. Pourtant, vu les différences importantes de composition chimique des bois des deux espèces, cette distinction serait doublement bénéfique à la filière. Elle permettrait d'obtenir des lots de bois plus homogènes du point de vue aromatique, facilitant ainsi le vieillissement des vins et alcools. Elle permettrait également de choisir l'espèce la plus adaptée au vieillissement en fonction des arômes recherchés.

Malheureusement, *Q. robur* et *Q. petraea* sont pratiquement indifférenciables sur la seule base de l'anatomie de leur bois (Schoch *et al.*, 2004). Les premiers essais réalisés à l'UMR Biogeco pour discriminer ces deux espèces par Spectrométrie Proche Infra-Rouge (SPIR), technique physique rapide et peu onéreuse déjà appliquée avec succès sur d'autres espèces d'arbres forestiers (Atkinson *et al.*, 1997; Humphreys *et al.*, 2008), se sont avérés non-

concluants (Camille Lepoittevin, communication personnelle). Les analyses chimiques demeurent les plus efficaces car plus qu'une simple différenciation d'espèce, elles permettent de quantifier directement les composés aromatiques d'intérêt (whisky-lactones par exemple). Malheureusement, ces techniques demeurent extrêmement coûteuses et sont fortement destructives (plusieurs grammes de bois nécessaire pour chaque analyse). Dans ce contexte, les analyses génétiques à partir de bois pour différencier les espèces d'arbres apparaissent comme une alternative pertinente (Eurlings *et al.*, 2010; Finkeldey *et al.*, 2010).

Dans le cas d'espèces génétiquement proches, les méthodes les plus efficaces et les plus répandues pour identifier les espèces uniquement sur la base de données génétiques sont les méthodes d'affectation (Pritchard *et al.*, 2000; Manel *et al.*, 2005). Ces méthodes sans a priori sont basées sur les modèles classiques de génétique des populations et prennent en compte la structure des données génétiques dans le but d'identifier des groupes homogènes.

La présence de deux espèces crée un signal se traduisant par un écart à l'équilibre de Hardy-Weinberg et un déséquilibre de liaison entre marqueurs génétiques. Ces méthodes utilisent ce signal pour affecter les individus aux espèces en minimisant ces deux déséquilibres. Ces approches sont très efficaces pour affecter des individus, même dans le cas d'espèces ou de populations très peu différenciées génétiquement (Hausdorf & Hennig, 2010). Elles seront d'autant plus efficaces que les marqueurs utilisés pour les analyses sont nombreux et informatifs (Banks *et al.*, 2003) et que le nombre d'individus génotypés est important (Manel *et al.*, 2005). De nombreux travaux ont déjà permis de différencier ces deux espèces à l'aide des méthodes d'affectation (Lepais *et al.*, 2009; Lepais & Gerber, 2010; Neophytou *et al.*, 2010; Penaloza-Ramirez *et al.*, 2010). Mais appliquer ces méthodes sur des échantillons de bois, avec toutes les contraintes techniques qui y sont liées, n'a à ce jour pas encore été testé.

S'il est relativement facile d'extraire de l'ADN à partir de matériel végétal frais comme les feuilles, les bourgeons ou le cambium (Doyle & Doyle, 1990; Lin & Walker, 1997; Csaikl *et al.*, 1998), en obtenir en qualité et quantité suffisante à partir du bois demeure complexe, car l'ADN y est fragmenté et dégradé (Bär *et al.*, 1988; Lindahl, 1993 ; Deguilloux *et al.*, 2002; Rachmayanti *et al.*, 2009). Il sera donc plus difficile d'extraire de l'ADN exploitable à partir de douelles qui auront séché pendant 18 à 36 mois, compromettant l'identification de tels échantillons à partir de leur ADN.

OBJECTIFS DE LA THESE

L'objectif principal de cette thèse est de développer des outils génétiques pour permettre d'identifier les deux espèces de chêne à partir de bois, afin de fournir un outil de valorisation ou de contrôle des lots de bois, au regard de leur potentiel aromatique. Cela implique la mise au point de marqueurs génétiques suffisamment discriminants. Ceux-ci seront dans un premier temps testés sur matériel végétal frais (feuilles et bourgeons), avant d'être transférés sur bois. Cette étape de validation permet également de développer une base de données génétique importante, indispensable pour pouvoir identifier avec précision l'espèce des échantillons de bois. Ce travail sur le développement de marqueurs génétiques différenciant efficacement les deux espèces permet également d'approfondir nos connaissances sur la démographie du chêne sessile et du chêne pédonculé et sur le maintien de ces espèces en dépit de flux de gènes interspécifiques importants (Dering & Lewandowski, 2007; Lepais & Gerber, 2010). Le deuxième axe de travail, plus appliqué et répondant directement aux attentes du Centre de Recherche Pernod Ricard, consiste à optimiser les méthodes d'extraction d'ADN à partir de bois (merrains et douelles) pour obtenir de l'ADN en qualité et quantité suffisante afin d'y appliquer les méthodes de génotypage développées en amont et ainsi caractériser les bois utilisés pour le vieillissement des vins et des alcools.

Dans le **Chapitre 1 – « Current trends in microsatellite genotyping »** (en révision pour *Molecular Ecology Resources* – Invited Review), je réalise une revue des méthodes de génotypage à l'aide de marqueurs microsatellites (ou SSRs pour *Simple Sequence Repeats*). Ce sont aujourd'hui les marqueurs moléculaires les plus utilisés en biologie. L'avènement des méthodes de séquençage dites de nouvelle génération, plus rapide et moins couteuses, permettent d'obtenir une grande quantité de marqueurs microsatellites sur des espèces non-modèles. Cela fournit une bonne base pour le développement de méthodes de génotypage fiables et haut-débit. La fiabilité concerne principalement la lutte contre les erreurs de génotypage, très fréquentes avec ce type de marqueur (Hoffman & Amos, 2005; Pompanon *et al.*, 2005). Je synthétise dans ce chapitre toutes les avancées récentes dans ce domaine et propose de nouvelles perspectives. Le deuxième point concerne le débit de ces analyses,

jusque là très modéré. L'émergence depuis quelques années du multiplexage (plusieurs paires d'amorces dans une unique réaction de PCR) a permis d'augmenter considérablement le débit des analyses de génotypage SSRs (Edwards & Gibbs, 1994; Markoulatos *et al.*, 2002; Butler, 2005; Hayden *et al.*, 2008). Mais cette technique demeure encore sous-utilisée, comme je le démontre dans une analyse de 100 travaux publiés en 2009-2010 dans la revue *Molecular Ecology*. J'étudie les points critiques liés au multiplexage et propose une stratégie globale de mise au point de marqueurs microsatellites multiplexés adaptée à des espèces non-modèles, en apportant une attention particulière aux coûts de développement et à la limitation des erreurs de génotypage.

Le **Chapitre 2 – «Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.)** » (publié en 2011 dans *Molecular Ecology Resources* – Molecular Diagnostics and DNA Taxonomy) est l'application au genre *Quercus* des recommandations mis en avant dans le **Chapitre 1**. J'ai ainsi développé et validé deux kits multiplex de microsatellites : un premier kit (8-plex) composé de marqueurs génomiques, un deuxième (12-plex) basé sur des marqueurs issus de banques d'ESTs (Expressed Sequence Tags), préalablement développés dans notre laboratoire (Durand *et al.*, 2010, voir **Annexe 3)**. Le choix des marqueurs génétiques utilisés pour différencier des espèces proches est crucial, la première étape de ce travail a donc été de choisir les meilleurs marqueurs possibles pour cela. La mise au point de ces deux kits multiplex dans une optique de haut-débit m'a également permis de développer une base de données génétiques de référence de plus de 3500 individus pour le kit 12-plex. Un effort particulier a été mené sur les taux d'erreur de génotypage inhérents à ce type d'analyses et je propose plusieurs stratégies pour les limiter (validation en descendances des marqueurs en simplex et analyses des allèles en tailles réelles).

Dans le **Chapitre 3 – « DNA-based identification of tree species from wood: application to oak staves** » (article en préparation pour un journal spécialisé en œnologie), je me concentre sur le typage d'échantillons de bois afin de différencier efficacement les deux espèces de chêne (*Q. robur* ou *Q. petraea*). Ce chapitre détaille les contraintes liées aux analyses génétiques sur bois, particulièrement sur le génome nucléaire. J'y propose une

méthode innovante de PCR en temps réel ciblant le génome chloroplastique pour tester et valider différents protocoles d'extraction et de purification de l'ADN dégradé présent dans le bois. Les analyses génétiques sur des échantillons de bois secs ont impliqué un nouveau travail de mise au point et de validation des marqueurs microsatellites développés dans le **Chapitre 2**. Grâce à des protocoles d'extraction et de purification d'ADN optimisés, combinés à des marqueurs microsatellites spécifiques de l'ADN dégradé, j'ai pu identifier avec succès l'espèce de chêne sur des douelles séchées depuis 18 mois, fournies par le Centre de Recherche Pernod Ricard.

Le **Chapitre 4 – «Genes under selection provide unique insights on oak trees demography»** (article en préparation pour une revue généraliste en biologie) se focalise sur l'apport des marqueurs de type SNPs (*Single Nucleotide Polymorphisms*) pour délimiter efficacement les deux espèces de chênes mais également pour étudier la dynamique évolutive chez ces espèces. Comme souligné dans le **Chapitre 1**, les marqueurs SNPs commencent à remplacer progressivement les marqueurs microsatellites dans les études de génétique des populations. Même si ces marqueurs souffrent de quelques inconvénients (voir Box 1 du **Chapitre 1**), leur identification à moindre coût, combinée à leur facilité d'analyse, les rendent aujourd'hui incontournables, en particulier pour les espèces non-modèles (Helyar *et al.*, 2011). Dans ce chapitre, je démontre que les marqueurs SNPs soumis à une forte sélection divergente (on parle alors de loci « outliers ») permettent de détecter des processus démographiques qui n'auraient pas été visibles si seuls des marqueurs neutres avaient été utilisés. Ces résultats vont à l'encontre de tout ce qui était recommandé jusqu'alors pour étudier les processus démographiques, en particulier du fait que ces marqueurs ne devraient pas être affectés par la sélection (Luikart *et al.*, 2003; Beaumont, 2005; Helyar *et al.*, 2011). Grâce à près de 300 marqueurs SNPs (détectés au cours d'un projet de reséquençage réalisé au sein de notre unité, coordonné par Pauline Garnier-Géré), dont une partie est très différenciée entre les deux espèces de chênes, j'ai pu délimiter très précisément ces espèces et observer les différences de structuration génétique. Seuls les « outliers » permettent de visualiser l'existence passée d'échanges génétiques asymétriques entre ces deux espèces, confirmant ainsi les prédictions du modèle d'invasion par hybridation développé par Petit *et al.* (2004) et précisé par Currat *et al.* (2008). De même, seuls ces

marqueurs fortement sélectionnés permettent de visualiser la différenciation plus marquée entre les populations de chêne sessile comparativement aux populations de chêne pédonculé.

REFERENCES

Atkinson MD, Jervis AP, Sangha RS (1997) Discrimination between *Betula pendula*, *Betula pubescens*, and their hybrids using near-infrared reflectance spectroscopy. *Canadian Journal of Forest Research* 27, 1896-1900.

Auer J, Rawlyer A, Dumont-Beboux N (2006) Elevage des vins du terroir en fûts de chêne du terroir. *Revue suisse de Vitic. Arboric. Hortic* 38, 379-387.

Bacilieri R, Roussel G, Ducousso A (1993) Hybridization and mating system in a mixed stand of sessile and pedunculate oak. *Annals of Forest Science* 50, 122-127.

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics* 19, 1436-1438.

Bär W, Kratzer A, Mächler M, Schmid W (1988) Postmortem stability of DNA. *Forensic Science International* 39, 59-70.

Barnard H, Dooley AN, Areshian G, Gasparyan B, Faull KF (2011) Chemical evidence for wine production around 4000 BCE in the Late Chalcolithic Near Eastern highlands. *Journal of Archaeological Science* In Press, Corrected Proof.

Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? *Trends in Ecology & Evolution* 20, 435-440.

Bodénès C, Joandet S, Laigret F, Kremer A (1997) Detection of genomic regions differentiating two closely related oak species *Quercus petraea* (Matt) Liebl and *Quercus robur* L. *Heredity* 78, 433-444.

Boidron JN, Chatonnet P, Pons M (1988) Influence du bois sur certaines substances odorantes des vins. *Connaiss. Vigne Vin* 22, 275-294.

Butler JM (2005) Constructing STR multiplex assays. *Methods in Molecular Biology* 297, 53-65.

Chatonnet P (1995a) *Influence des procédés de tonnellerie et des techniques d'élevage sur la composition et la qualité des vins élevés en futs de chêne*, Thèse de Sciences - Université de Bordeaux.

Chatonnet P (1995b) Principales origines et caractéristiques des chênes destinés à l'élevage des vins. *Revue des Oenologues et des Techniques Vitivinicoles et Oenologiques* 75, 15-18.

Coart E, Lamote V, De Loose M, *et al.* (2002) AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl] in Flemish populations. *Theoretical and Applied Genetics* 105, 431-439.

Csaikl UM, Bastian H, Brettschneider R, *et al.* (1998) Comparative analysis of different DNA extraction protocols: A fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. *Plant Molecular Biology Reporter* 16, 69-86.

Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proceedings of the Royal Society B-Biological Sciences* 269, 1039-1046.

Dering M, Lewandowski A (2007) Unexpected disproportion observed in species composition between oak mixed stands and their progeny populations. *Annals of Forest Science* 64, 413-417.

Doussot F, De Jeso B, Quideau S, Pardon P (2002) Extractives content in cooperage oak wood during natural seasoning and toasting; Influence of tree species, geographic location, and single-tree effects. *Journal of Agricultural and Food Chemistry* 50, 5955-5961.

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus* 12, 13-15.

Durand J, Bodénès C, Chancerel E, *et al.* (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 11, 570.

Edwards MC, Gibbs RA (1994) Multiplex PCR: advantages, development, and applications. *PCR Methods and Applications* 3, 65-75.

Eurlings MCM, van Beek HH, Gravendeel B (2010) Polymorphic microsatellites for forensic identification of agarwood (*Aquilaria crassna*). *Forensic Science International* 197, 30-34.

Feuillat F, Keller R, Huber F (1998) Grain and quality of cooperage oak (*Quercus robur* L. and *Q. petraea* Liebl.): myth or reality? *Revue des Oenologues et des Techniques Vitivinicoles et Oenologiques* 87, 11-15.

Finkeldey R, Leinemann L, Gailing O (2010) Molecular genetic tools to infer the origin of forest plants and wood. *Applied Microbiology and Biotechnology* 85, 1251-1258.

Gömöry D, Yakovlev I, Zhelev P, Jedinakova J, Paule L (2001) Genetic differentiation of oak populations within the *Quercus robur*/*Quercus petraea* complex in Central and Eastern Europe. *Heredity* 86, 557-563.

Guchu E, Diaz-Maroto MC, Diaz-Maroto IJ, Vila-Lameiro P, Perez-Coello MS (2006) Influence of the species and geographical location on volatile composition of Spanish oak wood (*Quercus petraea* Liebl. and *Quercus robur* L.). *Journal of Agricultural and Food Chemistry* 54, 3062-3066.

Hausdorf B, Hennig C (2010) Species delimitation using dominant and codominant multilocus markers. *Systematic Biology* 59, 491-503.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics* 9, 80.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* 11, 1-14.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* 14, 599-612.

Humphreys JR, O'Reilly-Wapstra JM, Harbard JL, *et al.* (2008) Discrimination between seedlings of *Eucalyptus globulus*, *E. nitens* and their F-1 hybrid using near-infrared reflectance spectroscopy and foliar oil content. *Silvae Genetica* 57, 262-269.

Jarauta I, Cacho J, Ferreira V (2005) Concurrent phenomena contributing to the formation of the aroma of wine during aging in oak wood: An analytical study. *Journal of Agricultural and Food Chemistry* 53, 4166-4177.

Jensen J, Larsen A, Nielsen LR, Cottrell J (2009) Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark. *Annals of Forest Science* 66.

Lepais O (2008) *Dynamique d'hybridation dans le complexe d'espèces des chênes blancs européens*, Thèse de Sciences - Université de Bordeaux.

Lepais O, Gerber S (2010) Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution* 65, 156-170.

Lepais O, Petit RJ, Guichoux E, *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology* 18, 2228-2242.

Lin H, Walker MA (1997) Extracting DNA from cambium tissue for analysis of grape rootstocks. *Hortscience* 32, 1264-1266.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362, 709-715.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* 4, 981-994.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions techniques with appropriate. *Trends in Ecology & Evolution* 20, 136-142.

Mariette S, Cottrell J, Csaikl UM, *et al.* (2002) Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Quercus petraea* Liebl. and *Quercus robur* L. stands. *Silvae Genetica* 51, 72-79.

Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. *Journal of Clinical Laboratory Analysis* 16, 47-51.

Mosedale JR, Charrier B, Crouch N, Janin G, Savill PS (1996) Variation in the composition and content of ellagitannins in the heartwood of European oaks (*Quercus robur* and *Q. petraea*). A comparison of two French forests and variation with heartwood age. *Annals of Forest Science* 53, 1005-1018.

Mosedale JR, Feuillat F, Baumes R, Dupouey JL, Puech JL (1998) Variability of wood extractives among *Quercus robur* and *Quercus petraea* trees from mixed stands and their relation to wood anatomy and leaf morphology. *Canadian Journal of Forest Research* 28, 994-1006.

Muir G, Fleming CC, Schlötterer C (2000) Taxonomy: Species status of hybridizing oaks. *Nature* 405, 1016-1016.

Neophytou C, Aravanopoulos FA, Fink S, Dounavi A (2010) Detecting interspecific and geographic differentiation patterns in two interfertile oak species (*Quercus petraea* (Matt.) Liebl. and *Q. robur* L.) using small sets of microsatellite markers. *Forest Ecology and Management* 259, 2026-2035.

Penaloza-Ramirez JM, Gonzalez-Rodriguez A, Mendoza-Cuenca L, *et al.* (2010) Interspecific gene flow in a multispecies oak hybrid zone in the Sierra Tarahumara of Mexico. *Annals of Botany* 105, 389-399.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist* 161, 151-164.

Petit RJ, Brewer S, Bordacs S, *et al.* (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* 156, 49-74.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* 6, 847-846.

Prida A, Boulet JC, Ducousso A, Nepveu G, Puech JL (2006) Effect of species and ecological conditions on ellagitannin content in oak wood from an even-aged and mixed stand of *Quercus robur* L. and *Quercus petraea* Liebl. *Annals of Forest Science* 63, 415-424.

Prida A, Chatonnet P (2010) Impact of oak-derived compounds on the olfactory perception of barrel-aged wines. *American Journal of Enology and Viticulture* 61, 408-413.

Prida A, Ducousso A, Petit RJ, Nepveu G, Puech JL (2007) Variation in wood volatile compounds in a mixed oak stand: strong species and spatial differentiation in whisky-lactone content. *Annals of Forest Science* 64, 313-320.

Prida A, Puech JL (2006) Influence of geographical origin and botanical species on the content of extractives in American, French, and East European oak woods. *Journal of Agricultural and Food Chemistry* 54, 8115-8126.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.

Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2009) DNA from processed and unprocessed wood: Factors influencing the isolation success. *Forensic Science International-Genetics* 3, 185-192.

Sauvageot F, Feuillat F (1999) The influence of oak wood (*Quercus robur* L., *Q. petraea* Liebl.) on the flavor of Burgundy Pinot noir. An examination of variation among individual trees. *American Journal of Enology and Viticulture* 50, 447-455.

Sauvageot F, Tessier C, Feuillat F (2002) Variation (according to species, forest and tree) in the aroma of French oak cooperage evaluated by the sniffing of wood chips. *Annals of Forest Science* 59, 171-184.

Schoch W, Heller I, Schweingruber FH, Kienast F (2004) Wood anatomy of central European Species. Online version: www.woodanatomy.ch.

Spillman PJ, Sefton MA, Gawel R (2004) The effect of oak wood source, location of seasoning and coopering on the composition of volatile compounds in oak-matured wines. *Australian Journal of Grape and Wine Research* 10, 216-226.

Streiff R, Ducousso A, Lexer C*, et al.* (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Molecular Ecology* 8, 831-841.

Streiff R, Labbé T, Bacilieri R*, et al.* (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* 7, 317-328.

Vivas, Nicolas, Glories, Yves (1997) *Recherches sur la qualite du chene francais de tonnellerie (Q. petraea Liebl., Q. robur L.) et sur les mecanismes d'oxydoreduction des vins rouges au cours de leur elevage en barriques*, Thèse de Sciences - Université de Bordeaux.

Young OA, Kaushal M, Robertson JD, Burns H, Nunns SJ (2010) Use of species other than oak to flavor wine: an exploratory survey. *Journal of Food Science* 75, 490-S498.

CHAPITRE 1

Current trends in microsatellite genotyping

E. Guichoux[1,2,3], L. Lagache[1,2], S. Wagner[1,2,4], P. Chaumeil[1,2], P. Léger[1,2], O. Lepais[1,2,5], C. Lepoittevin[1,2], T. Malausa[6], E. Revardel[1,2], F. Salin[1,2], and R.J. Petit[1,2]

[1] INRA, UMR 1202 Biodiversity Genes & Communities, F- 33610 Cestas, France

[2] Univ. Bordeaux, UMR1202 Biodiversity Genes & Communities, Bordeaux, F-33400 Talence, France

[3] Pernod Ricard Research Center, F-94000 Creteil, France

[4] Univ. Bonn, Steinmann Institut, D-53115 Bonn, Germany

[5] School of Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, UK

[6]UMR 1301 INRA/CNRS/Université Nice-Sophia Antipolis. Equipe BPI, Sophia Antipolis, F-06560, France

In revision (*Molecular Ecology Resources*)

INTRODUCTION

At a time where radically new genome-wide approaches emerge to study genetic variation, it is important to recall that many questions in molecular ecology can be efficiently addressed with a limited number of highly polymorphic markers, such as microsatellites. Microsatellites, also known as SSRs (Simple Sequence Repeats) or STRs (Short Tandem Repeats), remain the most popular markers in population genetic studies (Chambers & MacAvoy, 2000). They consist of motifs of one to six nucleotides repeated several times that have a characteristic mutational behavior (Kelkar *et al.*, 2010). As a consequence of their elevated mutation rates, SSRs are typically highly polymorphic: different individuals exhibit variation manifested as repeat number differences. Microsatellites have been used increasingly since the late eighties for applications such as fingerprinting, parentage analyses, genetic mapping or genetic structure analyses (Ellegren, 2004). Their genomic distribution, evolutionary dynamics, biological function and practical utility have been the object of a very large body of research, as summarized in several review articles (Tautz & Schlötterer, 1994; Jarne & Lagoda, 1996; Schlötterer, 1998; Chambers & MacAvoy, 2000; Li *et al.*, 2002; Dieringer & Schlötterer, 2003; Ellegren, 2004; Buschiazzo & Gemmell, 2006; Chistiakov *et al.*, 2006; Oliveira *et al.*, 2006; Selkoe & Toonen, 2006). Their advantages over other types of molecular markers include high allelic diversity and relative ease of transfer between closely related species (Box 1). However, SSRs have some drawbacks: a lengthy and costly development phase and a relatively low throughput due to difficulties for automation and data management, especially when compared to Single Nucleotide Polymorphisms (SNPs), which tend to be used increasingly (Box 1). Hence, the continued use of microsatellites will likely depend on the possibility to overcome some of these limitations.

Recently, progresses in SSR development and genotyping have been made in several directions, suggesting that SSRs could remain relevant genetic markers, at least for specific applications. First, the emergence of next-generation sequencing technologies means that identifying SSRs has become cheaper and faster. This trend is very recent, with the first reports appearing only in 2009 (Abdelkrim *et al.*, 2009; Rasmussen & Noor, 2009; Santana *et al.*, 2009). Further improvements along these lines are therefore very likely. Second, multiplexing microsatellites has become much easier. It can be accomplished through the co-amplification of multiple microsatellites in a single PCR cocktail, a procedure called true

multiplexing. Alternatively, PCR products from multiple amplification reactions can be combined, a procedure referred to as pseudo-multiplexing or poolplexing (Ghislain *et al.*, 2004; Meudt & Clarke, 2007). A blend of the two approaches is also possible. In true multiplex PCR (henceforth simply called multiplex), more than one target sequence is amplified by including more than one pair of primers in the reaction. The first successful attempt to multiplex PCR took place more than 20 years ago (Chamberlain *et al.*, 1988). Since then, capillary electrophoresis equipments relying on automated laser-induced fluorescence DNA technology have facilitated the use of this technique (Butler *et al.*, 2001; Butler *et al.*, 2004). Loci with non-overlapping allele-size ranges are labeled with the same fluorescent dye, whereas those with overlapping allele-size ranges are labeled with different dyes and resolved individually due to the different characteristic emission spectrum of each dye, hence considerably expanding the multiplexing potential. In addition, one of the dyes is used as an in-lane size standard, greatly improving the sizing precision of alleles. Multiplex PCR now forms the basis for many studies, reducing very significantly the cost and time of genetic analyses (Box 2). Important progresses have also been made in SSR data scoring, a critical and time-limiting step.

In this paper, we survey a sample of the recent literature on SSR genotyping. We show that multiplexing many (≥ 8) SSRs is not yet commonplace, despite the potential for much higher levels of multiplexing (e.g., Hill *et al.*, 2009). We continue by outlining the key steps necessary to develop accurate SSR multiplex. This involves paying attention to the whole process, from microsatellite identification to primers selection, data scoring and associated bioinformatics. We consider genotyping accuracy and troubleshooting and discuss areas where technical improvements of SSR genotyping are already possible and other areas where new developments would be important. We rely on our recent efforts to develop SSR multiplexes in forest trees for parentage analyses and population genetic surveys, during which we have reconsidered most steps to obtain high quality datasets (Guichoux *et al.*, 2011). Although several review articles on multiplex development already exist (Edwards & Gibbs, 1994; Henegariu *et al.*, 1997; Elnifro *et al.*, 2000; Markoulatos *et al.*, 2002; Wallin *et al.*, 2002; Butler, 2005a; Cryer *et al.*, 2005), none of these papers has provided a complete overview of SSR identification, multiplex design and genotyping. In addition, the latest developments based on next-generation sequencing techniques postdate these studies. Here, we first review current practices in SSR genotyping studies and then consider the entire

process of SSR genotyping, which ranges from SSR selection to data scoring and managing, while paying special attention to methods that help improve throughput and workflow, such as multiplexing.

A review of current practices

We surveyed a subset of the recent literature to examine current practices in terms of SSR genotyping. We checked 100 original journal articles relying on SSRs that had been published recently (in 2009-2010, see File S1, Supporting Information) in the journal *Molecular Ecology*, along with associated primer notes, if needed. Among the 100 original studies, 69 deal with population structure and 31 with parentage or sibship analyses. The organisms studied were all diploid and involved vertebrates, invertebrates, fungi or plants (Table 2). On average, 564 individuals were surveyed at 11.6 nuclear SSR loci, with no major bias depending on the organism investigated. Most studies took advantage of an automatic capillary electrophoresis system (90%). Overall, less than half of the studies (42%) used true multiplexing. This result illustrates the still limited penetration of multiplexing technique in the field, despite the nearly universal availability of suitable equipment. Unfortunately, the frequency of pseudo-multiplexing could not be calculated as its use appears not to be systematically reported. The mean number of SSRs surveyed was 11.1 in studies without multiplexing and 12.3 in studies with multiplexing, with an average of 3.9 loci (2-12) per multiplex. For those studies that used a specialized multiplex PCR buffer (e.g. Qiagen PCR Multiplex Kit), the corresponding figures are 13.9 SSRs with 5.0 loci per multiplex. Therefore, researchers using multiplexing techniques tend to use more loci, either to address different questions requiring more markers or to produce higher quality datasets for similar applications. Still higher levels of multiplexing are possible in the context of studies of non-model species, as 11 studies among the 100 surveyed relied on ≥8-plex. In fact, a few recent SSR studies have relied on very large (>20) multiplexes (e.g. Hill *et al.*, 2009; Chen *et al.*, 2010), whereas simultaneous PCR amplification of 35-40 PCR products is routinely achieved in the case of SNPs (e.g. Gabriel *et al.*, 2009; Buggs *et al.*, 2010), demonstrating that problems of primer competition can be overcome. The poor penetration of multiplexing, despite considerable potential, might be caused by the persistent belief that multiplexing greatly increases complexity or costs of microsatellite development (e.g. Neff *et al.*, 2000), which dates from the early times of PCR multiplexing (Edwards & Gibbs, 1994). Further results

regarding the types of SSRs studied and the quality controls used (estimation of the frequency of null alleles and of error rates) are discussed below. In general, our survey illustrates the need for more standardized reporting of microsatellite studies. This would help monitor the developments in the field and better evaluate the quality of the datasets produced.

Organisms studied	
Mammals	18%
Other invertebrates	16%
Plants	15%
Arthropods	14%
Amphibian and reptiles	12%
Birds	11%
Fungi	8%
Fish	6%
Multiplexing	
1-4 markers	15%
5-8 markers	19%
> 8 markers	8%
No	**58%**

Size of repeat units	
Di-nucleotides	**46%**
Tri-nucleotides	13%
Tetra-nucleotides	14%
Imperfect	26%
Null alleles check	
Yes	**40%**
No	60%
Error-rate measurement	
Yes	**26%**
No	74%

Table 2: Characteristics of 100 original journal articles relying on SSRs published in the journal *Molecular Ecology* in 2009-2010. Values outlined in the text are in bold.

SSR selection

Source of sequence data

Microsatellite detection requires sequence data. Until recently, the only possibility to identify sequences harboring SSR motifs was the screening of size-fractionated genomic DNA or of EST (Expressed Sequence Tag) libraries (Zane *et al.*, 2002). EST-SSRs are often reported to be less variable than genomic SSRs, being found in selectively more constrained regions of the genome (Gupta *et al.*, 2003). They also have the disadvantage that amplicon sizes can differ from expectation, as a consequence of the undetected presence of introns in flanking regions (Varshney *et al.*, 2005). However, this is balanced by several important advantages over genomic SSRs: (i) they should detect variation in the expressed portion of the genome, which might be of interest for studies of marker-trait associations; (ii) they can be developed at no

cost from EST databases; and (iii) once developed, these markers, unlike genomic SSRs, may work across a number of related species, because primers designed in flanking coding sequences are more likely to be conserved across species, resulting in high levels of transferability (Gupta *et al.*, 2003; Pashley *et al.*, 2006), especially if efforts are made to target conserved regions by using multiple alignments to design primers (Dawson *et al.*, 2010).

Regardless of whether genomic or EST sequences are used for SSR detection, traditional laboratory methods involving cloning, cDNA library construction, and Sanger sequencing remain costly and time-consuming (Squirrell *et al.*, 2003; Pashley *et al.*, 2006; Parchman *et al.*, 2010). To remediate this, next-generation sequencing techniques have now started to be used to identify sequences harboring SSR motifs in non-model species (Allentoft *et al.*, 2009). The first successful attempts have allowed a two to five times cost reduction as well as a significant decrease in time expenditure compared to traditional microsatellites development (Abdelkrim *et al.*, 2009; Santana *et al.*, 2009; Castoe *et al.*, 2010; Csencsics *et al.*, 2010; Malausa *et al.*, 2011). Besides, these approaches generate millions of base pairs of genomic sequence that may be useful for both SSRs-related and -unrelated research.

From transcriptome to whole genome shotgun sequencing for SSR detection

To optimize SSR detection with next-generation sequencing techniques, several strategies can be adopted, depending on the species' genome size, the abundance and nature of SSR motifs, and the sequencing coverage that can be achieved. For species harboring large and complex genomes, such as conifers, direct approaches might be risky due to the large amount of repetitive sequences with no interest for SSR detection (Parchman *et al.*, 2010). In this case, focusing on transcriptome – with the advantages and drawbacks previously discussed – can be more appropriate than whole genome shotgun sequencing. For genomes with a low frequency of SSRs, SSR enrichment techniques should be considered. Pyrosequencing of enriched libraries has proved efficient and cost effective to isolate SSRs in non-model species (Santana *et al.*, 2009; Malausa *et al.*, 2011). Moreover, a test of this procedure on model species showed that distribution of isolated markers across the genome satisfactorily reflects the actual distribution of SSRs across the genome (Martin *et al.*, 2010). If possible, informed choices about the motifs to target should be made, as this can greatly increase the number of useful SSR loci eventually identified (Santana *et al.*, 2009; Dubut *et al.*, 2010; Lepais & Bacles, 2010; Techen *et al.*, 2010; Malausa *et al.*, 2011). To date, however, most

studies (12 out of the 15 articles relying on SSR detection with next-generation techniques that we identified, see File S2, Supporting Information) have relied on whole genome shotgun sequencing, even when genome coverage was low (0.1x in Rasmussen & Noor, 2009, 0.02x in Castoe *et al.*, 2010) or when the genomes studied were known to have a low frequency of SSRs (Abdelkrim *et al.*, 2009).

Read-length

Interestingly, in all 15 studies published to date, the only sequencing technology used was the 454 pyrosequencing method of Roche. This technology generates the longest read-length among the next-generation sequencing methods currently available. Hence, single reads can be used for SSR identification and primer design (Abbott *et al.*, 2010). By circumventing the need for sequence assembly, this saves researchers from time-consuming bioinformatics steps. Software, such as MSATCOMMANDER (Faircloth, 2008) or QDD (Meglécz *et al.*, 2010), have been created to identify SSRs from 454 sequence data, the first one being used in more than half of the studies. Despite this, read-length remains a limiting factor: when the average read-length is around 200 bp, up to 2/3rd of the SSRs detected are too close to either fragment end to enable design of flanking PCR primers (Abdelkrim *et al.*, 2009; Castoe *et al.*, 2010; Csencsics *et al.*, 2010; Lepais & Bacles, 2010; Parchman *et al.*, 2010). Such limitations should no longer be an issue since 454 technologies delivering >400bp reads have now become available (Schuster, 2008; Kircher & Kelso, 2010). Such read lengths, in combination with the sequencing depth of the 454 technology, allow the design of a medium number of markers at sizes > 300 bp (Malausa *et al.*, 2011).

Advantages of next-generation sequencing

Hundreds or even thousands of SSR loci can be identified from a fraction of a single next-generation sequencing run (Tang *et al.*, 2008; Boomer & Stow, 2010; Castoe *et al.*, 2010; Saarinen & Austin, 2010). Moreover, if coverage is sufficient, shotgun data can be used to identify SSRs with unique primer sequences, which have a higher probability of producing successful locus-specific PCR amplification products (Castoe *et al.*, 2010). Next generation sequencing also provides preliminary information on SSR polymorphism, in particular if more than one genotype is sequenced. In our survey, only one study reported the use of more than one genotype at the sequencing stage, but available polymorphism data were not used to select candidate SSRs (Parchman *et al.*, 2010). The low coverage attained in most of

the studies likely precludes reliable detection of polymorphism. However, the throughput of sequencing technologies increases constantly, so we can expect higher genome coverage in the near future. Potentially, SSR polymorphism data should therefore become available very early on, which should in turn greatly facilitate SSR selection and optimization, at least if the necessary bioinformatic tools are accessible to the research team.

Choice of SSR type

Once sequence data harboring candidate SSR loci have been obtained, a number of choices need to be made, as outlined below. Interestingly, the availability of large amounts of sequence data obtained from next-generation sequencing projects will allow stringent selection of the best markers, thereby greatly saving time in downstream optimizations.

Perfect or imperfect repeats

Microsatellites have been classified according to the type of repeat sequence as perfect (with simple repeats only) or imperfect (Urquhart *et al.*, 1994). A common characteristic of imperfect repeats is that there is no more equivalency between fragment length and amplicon sequence: several sequences can correspond to a given length variant (e.g. Estoup *et al.*, 1995). Hence, preference should be given to perfect motifs (Gusmão *et al.*, 2006). Yet, imperfect SSRs remain frequently used. In the 100 studies surveyed, 26% of the SSRs used were imperfect (Table 2).

Size of repeat unit

Microsatellite repeat units typically vary from one to six bases. Focusing on the shortest motifs (such as mono- or di-nucleotide repeats) rather than on longer ones (≥ tri-nucleotide repeats) should allow packing more loci on a given separation system, resulting in larger multiplexes. This can be important because sequencing machines used for SSR genotyping make use of no more than five fluorochromes, which severely limits the number of SSR loci that can be analyzed simultaneously, given that allelic range size often reaches up to 50 or 100 bp and that amplicons measuring over 300 bp are rarely used (e.g. Hill *et al.*, 2009; Chen *et al.*, 2010). However, mono-nucleotide repeat SSRs can be difficult to accurately assay (Sun *et al.*, 2006), so they are often eliminated at the outset (Kim *et al.*, 2008). Among the 100 studies we surveyed, there was not a single case of mono-nucleotide repeat SSRs (Table 2)

even if these markers have been used successfully in studies of chloroplast DNA variation in plants (Ebert & Peakall, 2009), SSR-poor fungi (Christians & Watt, 2009), or in other circumstances where mono-nucleotide repeats are of special interest. In contrast, di-nucleotide repeat SSRs were the most frequently used. Unfortunately, di-nucleotide repeats often show one or more 'stutter' bands (multiple PCR products from the same fragment that are typically shorter by one or a few repeats than the full length product, see Chambers & MacAvoy, 2000). This is attributed to enzyme slippage during amplification (slipped-strand mispairing), making allele designation difficult (Levinson & Gutman, 1987; Meldgaard & Morling, 1997), especially for heterozygotes with adjacent alleles. In contrast, tri-, tetra- or penta-nucleotide repeats appear to be significantly less prone to slippage (Edwards *et al.*, 1991). Hence, SSRs with core repeats three to five nucleotides long are sometimes preferred for forensic and parentage applications (Kirov *et al.*, 2000; Cipriani *et al.*, 2008). Note however that stutter bands, when not too strong, can be useful, by helping distinguish true alleles from artifacts (e.g. Schwengel *et al.*, 1994), and that solutions have been proposed to overcome stuttering problems (Box 3).

Number of repeat units

The number of repeats has a critical effect on mutation behavior, to the point that it helps define which sequences actually represent microsatellites (Kelkar *et al.*, 2010). As on average SSR loci with more repeats have higher mutation rates (Weber, 1990; Ellegren, 2000; Petit *et al.*, 2005; Kelkar *et al.*, 2008), selecting loci with sufficient number of repeats is necessary to ensure polymorphism. However, SSRs with numerous repeats have also some drawbacks, such as increased allele dropout (Kirov *et al.*, 2000; Buchan *et al.*, 2005) and increased stutter (Hoffman & Amos, 2005). Moreover, these loci with numerous repeats are characterized by large allelic range, so that fewer can be combined in a given multiplex. Hence, an intermediate number of repeats could represent a good compromise. For instance, van Asch *et al*. (2010) suggest to select tetra-nucleotide repeats having more than 11 but less than 16 repeats. The lower limit is based on reported higher mutation rate for alleles with ≥11 repeats, thus increasing the chance of identifying highly polymorphic loci. The upper limit was defined based on the assumption that alleles with >16 repeats have a higher probability of accumulating interrupted motifs that confound the interpretation of the results.

Primer design

Once the sequences harboring repeat motifs have been identified, suitable primers must be chosen. To develop high quality multiplexed SSRs, stringent selection of markers is necessary (Varshney *et al.*, 2005). Primer pairs that amplify fragments of contrasted sizes (e.g. about 100, 200 and 300 bp) should be chosen to permit amplification of several non-overlapping markers with a single dye. Computer programs that simultaneously identify SSRs and design primers for multiplex exist (Kaplinski *et al.*, 2005; Rachlin *et al.*, 2005; Kraemer *et al.*, 2009; Shen *et al.*, 2010). Some of them search for suitable combinations of primer pairs for multiplex PCR and handle large datasets automatically. To ensure the success of co-amplification, it is critical to eliminate primers with potential primer-dimer interactions (Vallone & Butler, 2004; van Asch *et al.*, 2010). A local blast or dedicated tools such as Multiplex Manager (Holleley & Geerts, 2009) or NetPrimer (Premier Biosoft International, USA) can be used for this purpose (Table 3).

For multiplexing, primer pairs should have similar annealing temperature range (58–60°C has been cited as being optimal (Butler, 2005a; Hill *et al.*, 2009)). If primers have been developed previously and have different melting temperatures, primer redesign should be considered before multiplexing. However, redesign should be restricted to specific cases, such as when available SSRs are in short supply or when the corresponding SSRs are of special interest. Another possibility to buffer annealing temperatures is to add some extra sequence to primers (e.g. 5′-ACGTTGGATG-3'), thereby bringing GC% closer to 50% (Ghebranious *et al.*, 2005). The presence of nanosatellites (i.e. low complexity sequences that are too short to qualify as microsatellites) in the amplicons should be avoided. Since nanosatellites are abundant, this reduces the size of flanking sequences available for design, which can be problematic when selecting primers that amplify longer amplicons. This has been taken into account in the computer program QDD designed to isolate microsatellite loci from libraries of thousands DNA fragments (Meglécz *et al.*, 2010).

Primer validation in simplex

It is important to fully validate primer pairs early in the development process, so as to avoid losing time later with inefficient primers or uninformative loci (Figure 5). In particular, SSR loci presenting excessive stuttering, split peaks, null alleles, low heterozygote peak height ratios, and other artifacts should be identified early on and discarded or primers redesigned

(Box 3). For this purpose, SSRs need to be tested in simplex, e.g. using labeled M13-tails (Schuelke, 2000). Briefly, the primer mix contains a forward primer that has a specific sequence at its 5′ end (the M13-tail), a reverse primer and a universal fluorescent-labeled M13-tail. This technique is economic because the cost of direct fluorescent primer labeling is typically five to ten times higher than the cost of the synthesis of an unlabeled primer (Hayden *et al.*, 2008). However, the PCR conditions required for amplification using the M13-tailed primer method are often somewhat different from those optimal for amplification using standard length primers, which could create difficulties if the PCR protocol is tested in simplex with M13-tailed primers and then in multiplex with labeled primers but without M13-tail. In particular, M13-tails appear to decrease PCR efficiency, resulting in a need for additional PCR amplification cycles (de Arruda *et al.*, 2010). The samples used for validation of the primers should be representative of the genetic diversity (i.e. originating from different populations) to identify most alleles early on. This will minimize the risks to subsequently discover new alleles differing widely in size and overlapping with the allelic range of other loci labeled with the same fluorochrome, thereby compromising allele scoring. DNA pooling has been suggested as a cost-effective way to expedite this phase (Collins *et al.*, 2000; Cryer *et al.*, 2005).

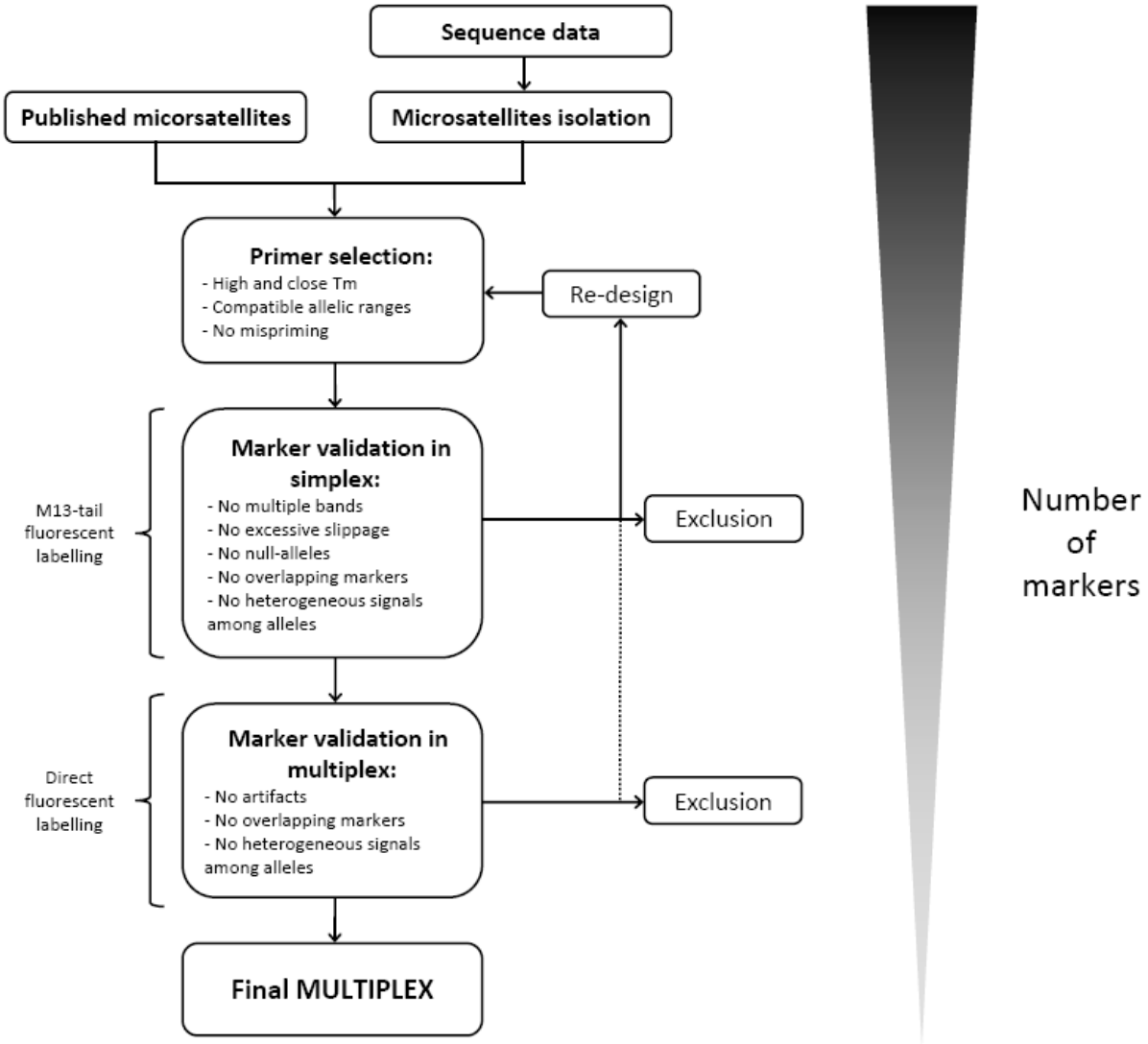


Figure 5: One possible strategy for the development of multiplex SSRs suitable for high-throughput genotyping.

The multiplexing phase

The throughput of standard (i.e. simplex) SSR analysis is low as it yields genotype information at only one locus per reaction. In contrast, multiplex PCR can boost genotyping by reducing laboratory work and consumption of expensive reagents without compromising test utility (Elnifro *et al.*, 2000; Lederer *et al.*, 2000; Galan *et al.*, 2003; Renshaw *et al.*, 2006 and see Box 3). Moreover, a reduced amount of DNA is needed to genotype a given number of loci (Karaiskou & Primmer, 2008), even if for high levels of multiplexing more DNA per reaction is necessary compared to standard simplex PCR (Chen *et al.*, 2010). Another advantage is that multiplex PCR provides better indications on template quantity and quality (Edwards & Gibbs, 1994). Potential problems in PCR include false negatives due to

reaction failure or false positives due to contamination. In particular, complete PCR failure can be more easily distinguished from an informative no-amplification. In view of these advantages, multiplexing SSRs should be a priority in all but the smallest SSR genotyping projects (Box 2).

The objective of the multiplexing phase is to combine all markers into the smallest number of reactions or select a subset of markers to design efficient and robust multiplexes, with each locus assigned a given florescent dye. A computer program (Multiplex Manager 1.0) has been developed to perform this task using prior marker information (Holleley & Geerts, 2009). It minimizes differences in annealing temperature and maximizes the spacing between markers, the heterozygosity, and the number of alleles (Figure 6).

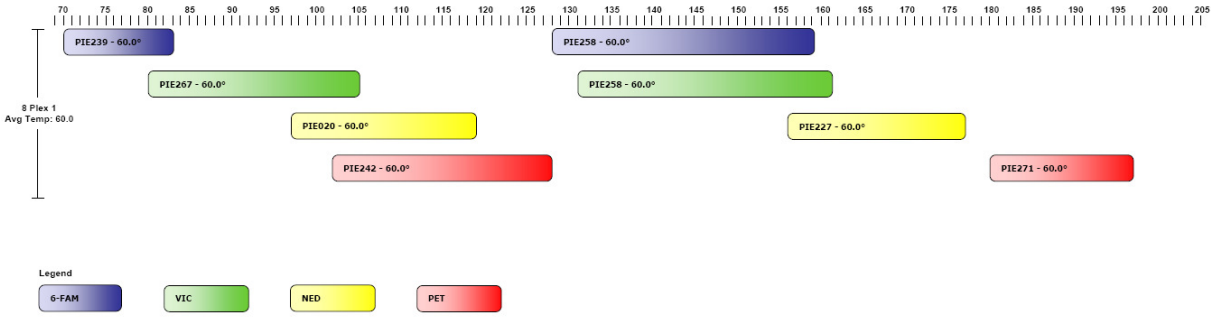


Figure 6: Example of output obtained with Multiplex Manager software (Holleley *et al.*, 2009). This software is used to identify combinations of markers suitable for multiplex reactions. In this example, for each of the eight SSRs, one of the four dyes (6-FAM, VIC, NED and PET) is assigned and the allele size range is provided along the main axis (in base pairs).

Multiplex PCR is a sensitive technique. To obtain repeatable results, careful standardization of all steps is needed. In particular, DNA concentration should be standardised (e.g. Livingstone *et al.*, 2009), if possible using automated pipetting robots. Although too little DNA can result in poor amplification, including imbalance among loci and allele dropout, too much DNA is generally more problematic. It can lead to off-scale fluorescent signal and to various PCR artifacts, such as imbalance among loci, incomplete adenylation of PCR products, and enhanced strand-slippage or "stutter" of various forms (Kline *et al.*, 2005). The use of specialized multiplex PCR buffer (e.g. Qiagen PCR Multiplex Kit) can help overcome some problems during PCR, particularly if a high level of multiplexing is targeted (Anonymous, 2002). In our survey, all studies with high level of multiplexing (≥8-plex) used the Qiagen PCR Multiplex Kit. This technique relies on a synthetic factor that allows efficient

primer annealing and extension irrespective of primer sequence, by increasing the local concentration of primers at the DNA template and stabilizing specifically bound primers (Anonymous, 2002). Whereas excellent results have been obtained without resorting to the use of specialized multiplex buffers, by stringent optimization of all parameters (e.g. Hill *et al.*, 2009), such buffers should be particularly useful when primers have different optimal annealing temperatures (Anonymous, 2002; Karaiskou & Primmer, 2008). Touchdown PCR protocols can also be used to amplify heterogeneous SSR sets via progressively reducing annealing temperature in successive annealing cycles, so that the optimal annealing temperature of every primer pairs is matched at some point during PCR (Rithidech & Dunn, 2003; Renshaw *et al.*, 2006).

Even when stringent selection of SSRs has been performed on the basis of simplex PCR, problems can occur during the multiplexing step, in particular heterogeneous amplification of the different SSR loci (i.e. locus-to-locus imbalance). To limit this problem, primers should have similar annealing temperatures, as pointed out before. If differences are nevertheless observed following multiplexing, a first possibility is to increase the primer concentration for the weakest markers or alternatively decrease primer concentration for the strongest ones, and repeat the process to adjust locus-to-locus balance. Obtaining uniform amplification signal facilitates automatic reading of the electropherograms.

To increase the consistency of genetic profiling protocols, testing the quantity and quality of fluorescently labeled primers can be relevant. A simple method to assess the primers on capillary electrophoresis system has been developed, by checking profiles or fluorescence intensity, in comparison with standards (Frasier & White, 2008). This should help reduce variation in amplification among primer batches, and among dyes. Another precaution is to limit the frequency of freeze-thaw cycles that can accelerate the breakdown of the dye attachment to the oligonucleotide, resulting in heterogeneous signals (Butler, 2005a).

In general, for moderate multiplexing (≤ 8 loci), there is no need for extensive optimization if all precautions outlined in Figure 5 are taken. In this respect, the situation has greatly changed compared to a few years ago when primer-to-template ratio, dNTP/$MgCl_2$ balance and PCR buffer concentration had to be carefully optimized and multiple rounds of changes in primer concentration were considered unavoidable (Henegariu *et al.*, 1997; Markoulatos *et al.*, 2002). However, for highly multiplexed sets (>12 SSRs), more advanced strategies might still be necessary. Hill *et al.* (2009) have proposed a method that relies on a core set of co-

amplifying markers to which other primers are added one after another. If difficulties are encountered, the primer causing the problem is identified by successively adding each primer in the multiplex primer mix. However, intensive optimization such as that proposed by Hill *et al*. (2009) must only be considered in exceptional cases.

Sizing precision

Sizing precision is defined as the ability to reproducibly estimate fragment sizes from run to run on a given instrument (Moretti *et al.*, 2001; Greenspoon *et al.*, 2008). It is calculated by averaging the standard deviation of size estimates across alleles at each locus. Imprecise sizing directly translates into genotyping errors, especially when the spacing of alleles is minimal (Ghosh *et al.*, 1997). For alleles 1 base apart, the tolerance level is normally set at a value near 0.2 bp. Precision depends on capillary length and voltage as well as of the detection window and the detection integration time. It can also be affected by temperature fluctuations, polymer and capillary effects (Hartzell *et al.*, 2003; Sgueglia *et al.*, 2003) or by the type of fluorescent dye used (Hahn *et al.*, 2001). Limiting variation in PCR conditions should also help (Ghosh *et al.*, 1997).

"Allelic drift" is the tendency for true allele sizes to differ by a value slightly different from the known repeat length. At di-nucleotide SSRs, for instance, the effective spacing between peaks of observed allele sizes has been shown to vary between 1.8 and 2.2 bp (Amos *et al.*, 2007). Spacing of adjacent alleles decreases with increases in PCR product size, thereby reducing precision (Idury & Cardon, 1997). The precision should however still be sufficient to distinguish reliably one base pair difference for fragments >300 bp (Koumi *et al.*, 2004).

Allele calling and binning

Once large datasets of multiplexed SSR markers have been collected from capillary sequencing machines, the corresponding genotypes need to be read. There are two distinct steps in this process: true allele size calling, i.e. using decimal numbers, and binning, i.e. the conversion of alleles from real-valued DNA fragment sizes into discrete units to which an integer label is assigned (Idury & Cardon, 1997).

The first step of the analysis is allele calling, i.e. identifying peaks that correspond to alleles and measuring the size of the corresponding fragments. Commercial software provided by constructors of capillary electrophoresis systems decrease analysis set-up time through automated correction of common genotyping problems including saturated peaks, excessive

baseline noise, voltage spikes caused by micro-air bubbles or debris in the laser path, and stutter peaks. However, depending on the quality of the markers, allele calling often necessitates additional manual editing. As this step can be labor-intensive and can generate errors, it is important to select well-behaved markers at the outset, as emphasized before (Scandura *et al.*, 2006).

The next step, allelic binning, is critical. In one comparative study, 83% of discrepancies between laboratories in scoring di-nucleotide alleles were due to arbitrary decisions in binning (Weeks *et al.*, 2002). In another study, binning errors accounted for 21% to 40% of all errors (Ewen *et al.*, 2000), confirming the necessity of well-established reading rules. Interestingly, in our survey, most authors (95%) used software with automatic-binning module. We assume that these studies relied on user-friendly automated binning procedure (Table 3) and possibly on manual checks, rather than on direct analysis of raw fragment sizes, hence increasing risks of genotyping errors (Amos *et al.*, 2007).

Since integer labels may not directly reflect the underlying allele sizes, raw allele sizes need to be stored for later reference and comparisons. One efficient and simple procedure is to export raw fragment size data to a spreadsheet and use it to compile cumulative frequency plots of size distributions (Jayashree *et al.*, 2006). New bins for the inferred number of repeats can then be constructed around these distributions, at places where discrete breaks in periodic size classes are evident. In this way, alleles that deviate from the expected periodicity of repeats (i.e. off-ladder microvariants) can be identified. Software has been designed for this step. ALLELOBIN and FLEXIBIN use least-squares minimization procedures and allow for allelic drift (Idury & Cardon, 1997; Amos *et al.*, 2007). TANDEM has been specifically designed for integration into population genetic and genomic workflows and requires no additional reformatting of data files (Matschiner & Salzburger, 2009). MsatAllele is a computer package built on R to visualize and bin the raw microsatellite allele size distributions (Alberto, 2009). It uses files exported from the open-source electropherogram peak-reading program STRand. Genotype files with the resulting binned data can then be exported. In our lab, we developed an Excel macro, inspired from FlexiBin (Amos *et al.*, 2007), Autobin (http://www.pierroton.inra.fr/biogeco/site_pole_agro/telechargement/Binning.html), which automatically analyzes raw data generated with commercial software (Table 3). The number of samples and loci is automatically detected, alleles in raw sizes are sorted and plotted to detect relevant gaps in size (Figure 7), alleles are binned (with

manual checking) and the whole data set is formatted for GENEPOP (Raymond & Rousset, 1995) or STRUCTURE (Pritchard *et al.*, 2000).

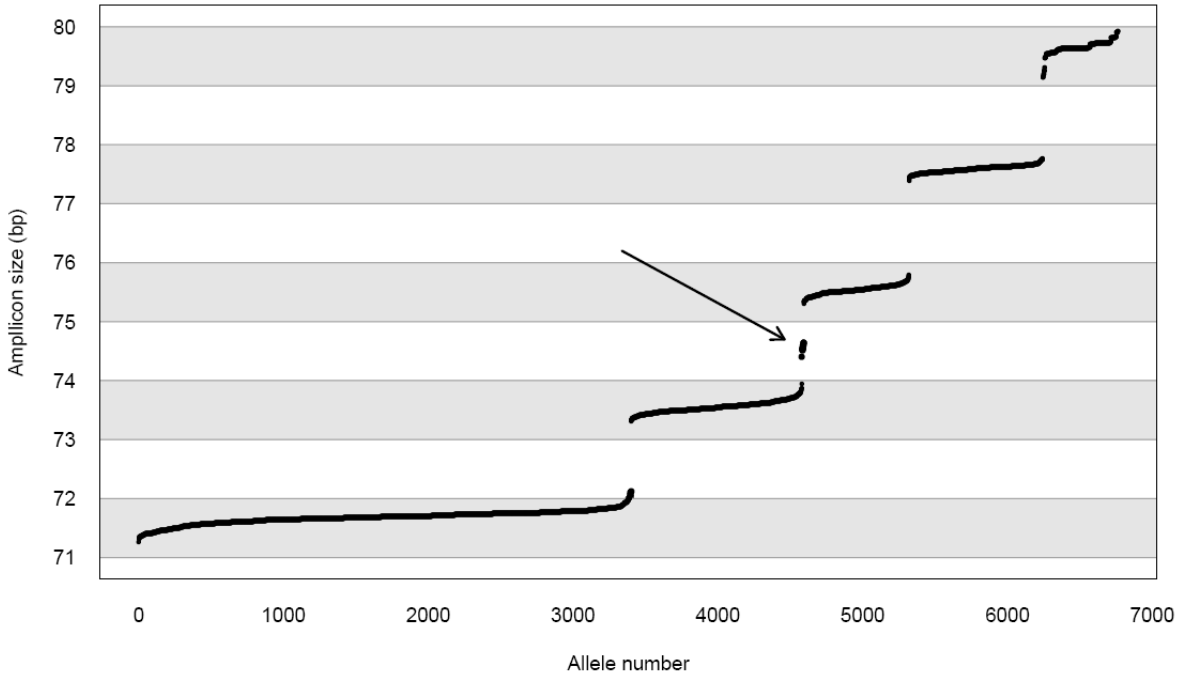


Figure 7: Size distribution of 6762 alleles for one di-nucleotide EST-SSR developed on oaks, achieved with the macro we developed. The arrow indicates the presence of an off-ladder microvariant found in 13 alleles that differs by one base pair from the expected periodicity of 2 bp. Analyses of segregating progenies have confirmed that this variant corresponds to a different allele.

Thousands of datasets that could potentially be expanded as samples become available are regarded as lost because of the effort that would be required to validate congruence of genotypes from old and new data sets (Presson *et al.*, 2008; Morin *et al.*, 2009). To take advantage of past studies, specific software has been designed (ALLELOGRAM and MicroMerge). These two software programs can normalize and bin alleles from multiple data sources using a relatively small set of controls (Table 3). Binning can also be harmonized using reference allelic ladders (e.g. Gill *et al.*, 2001). However, these methods will not be able to correct for poor binning if raw allele sizes have been lost.

Measuring and reporting error rates

Error rates per locus and per individual should be systematically measured and reported in genotyping studies. In our survey, however, genotyping error rates were reported in only 26% of the studies. In genotyping studies relying on multiplexing, measuring error rates is particularly important (Luikart *et al.*, 2008), because information on locus-specific error rates

is necessary to improve multiplex assays. Genotyping error rates can be estimated by counting Mendelian inconsistencies in parent-offspring pairs or by counting mismatches between duplicated genotypes (Bonin *et al.*, 2004; Hoffman & Amos, 2005; Pompanon *et al.*, 2005; DeWoody *et al.*, 2006; Johnson & Haydon, 2007). This second option can be further subdivided in two cases, depending on whether duplicated genotypes include or not a well-characterized control (i.e. concordance checking using standard reference genotypes versus re-genotyping of a random subset of genotypes).

Clearly, none of these approaches allow the identification of all genotyping problems. For instance, in parent-offspring comparisons, not all errors result in Mendelian inconsistencies. Similarly, with duplicated samples, some problems, such as mutations or null alleles, cannot be identified (Ewen *et al.*, 2000). When randomly re-genotyping samples in the absence of reference sample, some errors might remain unnoticed, as when a heterozygous genotype is genotyped twice as a homozygote. Moreover, when the duplicated genotypes differ, the nature of the error can sometimes be difficult to establish. In particular, it might not be possible to distinguish between allelic dropout (failure to amplify one of the two alleles in heterozygotes) and false alleles (caused by polymerase errors, see Broquet & Petit, 2004). This is unfortunate because the two classes of error affect analyses in different ways (Wang, 2004; Hadfield *et al.*, 2006).

Hence, multiple strategies should be used whenever possible, concentrating on pedigree evaluation and re-genotyping with reference samples. Nevertheless, from a practical point of view, re-genotyping to get complete data set in multiplex surveys means that, as a by-product of this process, individuals will be genotyped several times at some of the loci, thereby providing more accurate error rate measurements. Software has been developed to estimate error rates and break them down into different categories (reviewed in Johnson & Haydon, 2007).

Data management

The utility of genotyping techniques is only as good as one's ability to handle the flood of data produced from them. Managing genotyping data can indeed be challenging. In particular, because records for a particular sample might have to be revised over time, the management system must keep track of each DNA sample during the whole process. Genotyping data must be kept as raw data for future work (in the same lab or in another lab)

to avoid laborious normalization work. Database management systems or Laboratory Information Management Systems (LIMS) specialized in genotyping data have been released to meet these demands (Li *et al.*, 2001; Jayashree *et al.*, 2006; van Rossum *et al.*, 2010). Besides serving as workflow managers, these systems also provide visible quality checks and centralization of data, but their use is far from being commonplace.

Software name	Licence	Functionalities	Type of program	Platforms	Reference
Primer detection and design					
AutoDimer	**Free**	Screening for primer-dimer and hairpins	Visual Basic or Web application	Platform independent	(Vallone *et al.*, 2004)
Generunner	**Commercial**	Sequence analysis tool	Unknown	Windows	Hastings Software Inc.
MultiPlx	**Free**	PCR primer compatibility multiplexing	Web application	Linux/Windows/ Solaris	(Kaplinski *et al.*, 2005)
MSATCOMMANDER	**Free**	SSR marker detection and design	Python	Platform independent	(Faircloth, 2008)
NetPrimer	**Free**	Primer design and secondary structure analysis	Java	Mac/Windows	Premier Biosoft Int.
PolySSR	**Free**	SSR marker detection	Web application	Platform independent	(Tang *et al.*, 2008)
Primer3	**Free**	SSR marker design	Web application	Platform independent	(Rozen & Skaletsky, 1999)
QDD	**Free**	SSR marker detection and design	Perl	Linux/Windows	(Meglécz *et al.*, 2010)
SAT	**Free**	SSR analysis tool	Web application	Platform independent	(Dereeper *et al.*, 2007)
STAMP	**Free**	SSR marker design	Extension to the STADEN package	Platform independent	(Kraemer *et al.*, 2009)
Multiplexing					
Multiplex Manager	**Free**	Design and optimization of multiplex PCRs	C++	Linux/Mac/Windows	(Holleley & Geerts, 2009)
Estimation of error rates					
MasterBayes	**Free**	Pedigree reconstruction, analysis and simulation	R package	Mac/Unix/Windows	(Hadfield *et al.*, 2006)
Pedant	**Free**	Estimation of maximum likelihood allelic dropout and false allele errors	Delphi	Windows	(Johnson & Haydon, 2007)
PedManager	**Free**	Inheritance errors and more	Unix	Unix/Windows	(Ewen *et al.*, 2000)
Fragment calling					
GeneMapper	**Commercial**	Genotyping software package	Unknown	Windows	Applied Biosystems
GENOTYPER	**Commercial**	Genotyping software	Unknown	Windows	Applied Biosystems
Peak Scanner	**Free**	Genotyping software	Unknown	Windows	Applied Biosystems
STRand	**Free**	Analysis of DNA fragment length polymorphism	C++ / Visual Basic	Windows	(Toonen *et al.*, 2001)
TrueAllele	**Commercial**	Genotyping software	Matlab	Mac/Unix/Windows	None
Fragment binning and analysis					
ALLELOBIN	**Free**	Automated allele binning	C and Java	Unknown	(Idury & Cardon, 1997)
ALLELOGRAM	**Free**	Allele binning and normalization	Java	Mac/Unix/Windows	(Morin *et al.*, 2009)
Decode-GT	**Free**	Quality measures for allele calling	Unknown	Mac/Unix/Windows	(Palsson *et al.*, 1999)
FLEXIBIN	**Free**	Automated allele binning	Microsoft Visual Basic	Excel macro	(Amos *et al.*, 2007)
MsatAllele	**Free**	Automated allele binning	R package	Mac/Unix/Windows	(Alberto, 2009)
MicroMerge	**Free**	Merging of microsatellite data sets	Unknown	Linux/Windows	(Presson *et al.*, 2008)
TANDEM	**Free**	Automated allele binning	Ruby	Mac/Unix/Windows	(Matschiner & Salzburger, 2009)
AutoBin	**Free**	Automated allele binning	Microsoft Visual Basic	Excel macro	See text
Data Management					
GenoDB	**Free**	Manipulation of dinucleotide SSRs genotype data	Unknown	Unknown	(Li *et al.*, 2001)
SLIMS	**Free**	Sample-based LIMS	Web application	Platform independent	(van Rossum *et al.*, 2010)

Table 3: Non-exhaustive list of software for microsatellites detection and genotyping.

CONCLUSIONS AND PERSPECTIVES

There are many applications in molecular ecology where 10-30 highly polymorphic markers such as SSRs would suffice to provide precise answers (Box 1). During the last years, considerable progresses have been made in SSR development and genotyping, including in associated bioinformatics. However, the efforts remain somewhat disparate and current practices are lagging behind. As a consequence, SSR markers are not used to their full power, as shown by our survey of a sample of the recent literature. Hence, additional efforts to improve SSR isolation, multiplex genotyping and scoring remain critical.

The identification of SSR motifs has long been a bottleneck in studies involving non-model species for which sequence data is not readily available. The use of next-generation sequencing techniques instead of cloning and conventional sequencing to obtain sequence data and identify SSRs in such species is just beginning and appears extremely promising. It provides the optimal conditions for subsequent multiplex development, by detecting many potential SSRs. In fact, the throughput and cost-effectiveness of next-generation sequencing should allow researchers to be more selective in their choice of SSR loci. In particular, sequencing depth should provide sufficient data on sequence variation to focus on conserved regions flanking polymorphic SSR motifs for designing primers, considerably simplifying the whole process of marker testing.

The number of multiplexed markers could be increased, since there is no major limitation in combining up to 30 or 40 SSRs in a single PCR (Gabriel *et al.*, 2009; Hill *et al.*, 2009). Increasing the number of fluorochromes could also help. Multiplexing should not only increase throughput but also accuracy. The latter point might not be immediately obvious. However, designing a good multiplex is demanding, hence forcing researchers to take a number of precautions and to better evaluate candidate loci, which eventually benefits to the whole genotyping process. Better precision could also be achieved with new size standards or improved algorithms (Johansson *et al.*, 2003). Automation, from DNA isolation to capillary electrophoresis, could be developed using appropriate robotics and high-throughput plate formats (384 or 1536 wells). Recently, laboratory-on-a-chip systems relying on microfluidic technology have been tested successfully for DNA amplification (Horsman *et al.*, 2007; Sinville & Soper, 2007; Greenspoon *et al.*, 2008; Bienvenue *et al.*, 2009; Liu & Mathies, 2009). Such systems potentially offer speed, automation, sensitivity and portability (Beyor *et al.*,

2009). Completely different methods amenable to highly parallelized SSR assays might also emerge (e.g. Pettersson *et al.*, 2006; Zajac *et al.*, 2009).

With the outbreak of next-generation sequencing technologies, SSR genotyping could eventually be done via sequencing of amplified fragments. The million reads obtained could make it possible to genotype hundreds of samples at thousands of loci, provided these samples can be identified prior to sequencing (e.g. with short ligated sequence tags). This would result in a drastic reduction of genotyping costs and a substantial improvement of data quality. Indeed, direct access to microsatellite motif sequence (rather than PCR product sizes) would reduce problems of homoplasy in datasets and avoid poor genotyping repeatability among laboratories using different equipments or reagents. However, such processes still need to be set up and must be associated to bioinformatic methods aiming at sorting sequences, correcting for sequencing errors and finally summarizing genotype information.

Box 1: SSRs versus SNPs

To evaluate current trends in genotyping methods, we searched the ISI Web of Knowledge database for papers citing SSRs or SNPs. The former have increased linearly since the early 1990s, whereas the latter have increased exponentially since the late 1990s (Figure 1). Yet, papers citing SSRs still outnumbered those citing SNPs in 2009. Although this should change soon, the continued increase in studies relying on SSRs justifies efforts to improve their effectiveness.

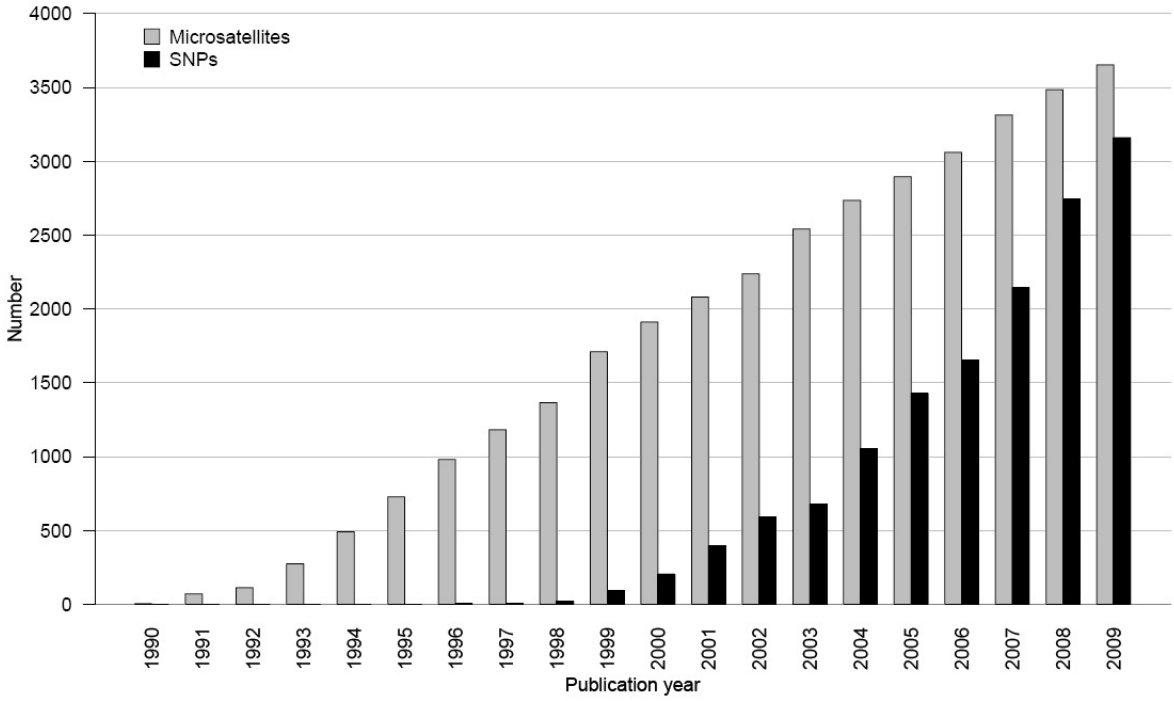


Figure 1 : Evolution of the number of studies relying on SSRs and SNPs since 1990.

Current popularity is not always the best guide to decide which markers to use (Schlötterer, 2004). Instead, data on the relative advantages of each type of marker for various applications should help researchers embarking on new projects in molecular ecology. Following Morin *et al*. (2004), we provide here a brief summary of the relative merits of SSRs and SNPs, focusing successively on the intrinsic differences between the two markers and then on the technical aspects of their analysis.

There are two main differences between SSRs and SNPs. First, SNPs are more numerous than SSRs in the genome of most species. On average, in the human genome, there is one SNP every 100-300 bp (Thorisson *et al.*, 2005), compared to one SSR locus every 2-30 kb (Webster *et al.*, 2002), depending on how SSRs are defined (Kelkar *et al.*, 2010). This can be important for genome-wide association studies but not necessarily for other applications. Second, the mutation rate per generation differs drastically between the two marker types. SSRs have mutation rates ranging from 10^{-3} to 10^{-4} per locus per generation (Ellegren, 2000), compared to about 10^{-9} for SNPs, i.e. several orders of magnitude lower. As a consequence, SNPs are typically diallelic: in humans, less than 0.1% of SNPs are triallelic (Lai, 2001). In contrast, SSR loci generally have high allelic richness, often in excess of 10 alleles. Below, we list the relative merits of SSRs and of SNPs to help researchers decide which type of markers is best suited for their needs.

Advantages of SSRs over SNPs

- SSR loci above a certain number of repeats can be assumed to be polymorphic (Schlötterer, 2004) whereas to identify SNPs, homologous regions must be sequenced from multiple chromosomes.
- SSRs have little ascertainment bias (the bias resulting from the choice of the initial panel of genotypes used to screen for polymorphisms), in contrast to SNPs (e.g. Li *et al.*, 2008).
- The success rate of cross-amplification of SSRs in close relatives is typically higher than for SNPs (up to 50%, Sharma *et al.*, 2007).
- SSR loci are more powerful than SNPs to detect mixtures (Clayton *et al.*, 1998; Gill, 2001).
- SSR accuracy is easy to assess because a larger proportion of errors can be detected in pedigree analyses when there are many alleles per locus; in contrast, for SNPs, which are typically diallelic, many errors will remain undetected when analyzing pedigrees as they will be compatible with Mendelian segregation rules (Palsson *et al.*, 1999).
- SSRs will be more useful for detecting recent population expansions than SNPs, because the accumulation of new mutations requires shorter time periods for rapidly evolving loci than for slowly evolving ones (Morin *et al.*, 2004).

- For many applications, there is not much gain in using more loci after a certain threshold is reached. For instance, low error rates can be achieved in clonal identification using a few highly polymorphic loci. In other cases, using more than a few tens of loci might not be relevant as additional loci become non-independent because of linkage (Santure *et al.*, 2010). In such cases, microsatellites represent a credible alternative. To help researchers decide on the best alternative, we provide indications from the literature on the number of SNPs needed to result in a power equivalent to that of one SSR for different applications (Table 1). The information originates mostly from simulation studies aiming at evaluating the relative power of different markers differing in allelic richness.

Drawbacks of SSRs over SNPs

- The large number of alleles per locus in SSRs implies that for accurate estimation of allelic frequencies, large sample sizes are needed, in contrast to SNPs.
- Spontaneous mutations are more likely to take place at SSRs than at SNPs within a given pedigree, potentially complicating parentage reconstruction (Ellegren, 2000; Phillips *et al.*, 2007; Borsting *et al.*, 2009).
- The high rate of recurrent or backward mutation of SSRs makes them poor indicators of long-term population history (Li *et al.*, 2002; Ellegren, 2004; Morin *et al.*, 2004; Schlötterer, 2004).
- Capillary gel electrophoresis coupled with fluorescence-based detection is the only commonly reported method for the assay of SSRs (Butler *et al.*, 2001; Koumi *et al.*, 2004). In contrast, SNPs are potentially amenable to typing through many techniques, including digital typing methods using chip technology, allowing the development of ultra-high density methods (Syvänen, 2005).
- With SSRs, there is a need to include common controls among studies and across time. In contrast, SNP genotypes are based on the detection of DNA sequence nucleotide differences rather than PCR product size differences, so that genotype data are more easily comparable and portable. In fact, SNP studies can be replicated, performed in parallel across several laboratories, and added to as samples become available without the need to calibrate results at each step in the

process. To date, reduced portability of SSR data across laboratories has resulted in significant data use limitations (e.g. Hoffman *et al.*, 2006).

- PCR amplicons are typically longer for SSRs than for SNPs, so that it is more difficult to study highly degraded DNA samples, such as fecal and other non-invasive samples (Seddon *et al.*, 2005; Morin & McCarthy, 2007), or ancient samples (Sanchez & Endicott, 2006).

In conclusion, the widespread adoption of SSRs lies in the power that they provide to solve biological problems, due in particular to their high allelic richness. In contrast, many disadvantages of SSRs are of a technical nature (Chambers & MacAvoy, 2000). This suggests that SSRs could remain useful in the future if at least some of the technical problems identified are overcome (Glaubitz *et al.*, 2003; Schlötterer, 2004; Ryynänen *et al.*, 2007; Matschiner & Salzburger, 2009). In principle, using blocks of tightly linked SNPs and treating each haplotype as a separate allele could yield genotyping data with properties similar to those obtained with SSR loci (Jones *et al.*, 2009). However, the incidence of missing data will likely be high, whereas compound genotyping errors will quickly increase as multiple PCR reactions are needed to type a single locus.

Application	Relative power of SSRs versus SNPs	Comments	References
Linkage study, individual identification	2-3	Power proportional to heterozygosity *H*: H_{SSR}~2.H_{SNP}	(Kruglyak, 1997; Waits *et al.*, 2001; Seddon *et al.*, 2005)
Parentage analysis	~5	This estimate was obtained using SNPs with minor allele frequency >0.2. Note also that with diallelic SNPs, a heterozygous genotype is a universal donor.	(Glaubitz *et al.*, 2003)
Genetic structure	4-12	SNPs have typically few private alleles as a consequence of the way they are identified, i.e. using a limited panel of genotypes; such private alleles are particularly useful to reconstruct genetic structure.	(Rosenberg *et al.*, 2003; Liu *et al.*, 2005)
Association studies / Linkage disequilibrium	5-20	Expected power of genome-wide LD testing for the detection of a low-frequency disease variant, assuming SNPs have minor allele frequencies >0.2	(Ohashi & Tokunaga, 2003)
Sibling reconstruction	∞	The 4-allele property states that no more than 4 alleles can be found in a full-sib family; this property cannot be used to reconstruct sibships with diallelic SNPs	(Berger-Wolf *et al.*, 2007; Ashley *et al.*, 2009; Wang & Santure, 2009; Jones & Wang, 2010)

Table 1 : Number of SNPs needed to result in a power equivalent to that of one SSR depending on the application.

Box 2: Cost effectiveness of multiplex SSR typing

We have estimated the overall cost of SSR genotyping as a function of the degree of multiplexing, following Renshaw *et al*. (2006). The goal we set was the genotyping of up to 2500 samples at 24 microsatellites. Five strategies were considered: no multiplexing, 2-plex, 4-plex, 8-plex and 12-plex. Cost included consumables (plates, tips) and reagents (Qiagen Multiplex PCR Kit, unlabelled primers, labeled primers, LIZ-600 size standard). Salary costs were based on those of an experienced research assistant in France. We conservatively assumed that in the absence of true-multiplexing, pseudo-multiplexing was used by combining four loci marked with different fluorochromes in one lane.

The results (Figure 2) show that even for a moderate number of samples (100), multiplexing is cost-effective (12-plex is eight times cheaper than simplex PCR). For completeness, this should be balanced with the cost of developing the multiplex. However, most of the work to develop and optimize SSR multiplex is actually represented by phases that are common to all SSR development projects. If primers have been selected with the objective of multiplexing in mind, the extra costs of multiplexing can amount to little more than 2-4 PCR tests for an 8-plex, depending on whether the concentration of some primers has to be optimized or some primers have to be replaced.

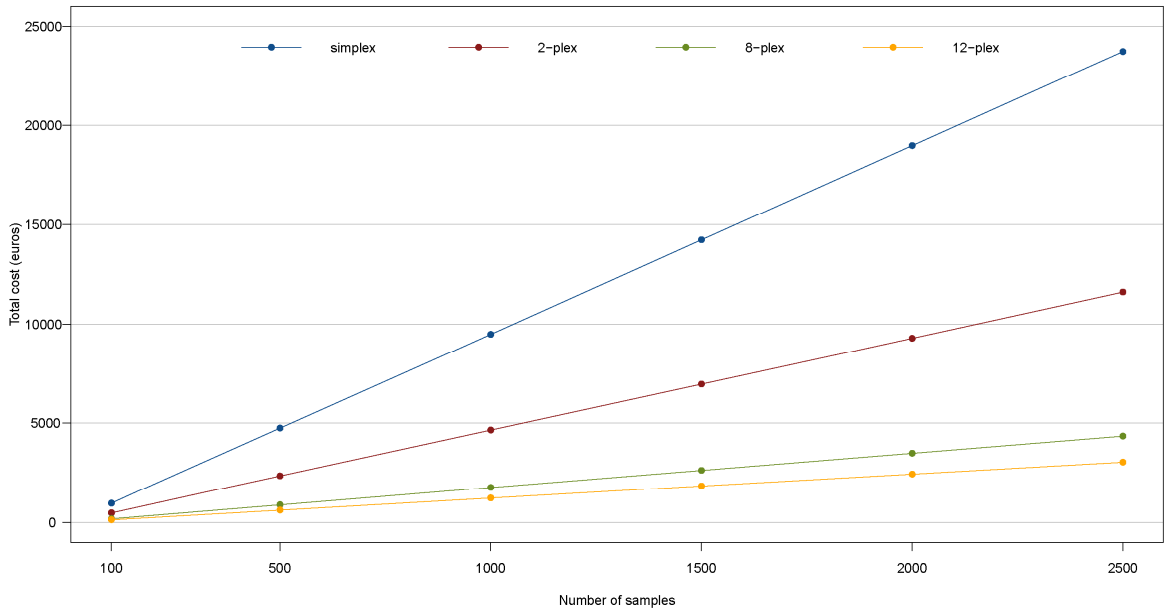


Figure 2: Overall cost for genotyping 24 SSRs, depending on the multiplex strategy and the number of genotyped samples.

Other solutions to decrease costs

The Qiagen Multiplex PCR Kit is the most widely cited commercial kit, with 25% of the papers we surveyed mentioning it. This commercial kit has a high cost per sample, but the final volume can be decreased to 5µl (Lepais & Bacles, 2010) with a final buffer concentration of 0.7× (Qiagen recommends 1×), without compromising reproducibility or specificity (Spathis & Lum, 2008). This reduces the final cost to 0.13€ per sample (compared to 1.88€ with no optimization). Another solution to decrease the costs is to shift to 384 plates as these allow the use of even smaller volumes, down to 2µl (Kenta *et al.*, 2008). Finally, instead of relying on direct fluorescent labeling of primers, it is possible to use universal tailed primers (Oetting *et al.*, 1995), one for each fluorescence detection. Such a method allows the same level of marker multiplexing and accuracy in SSR genotyping attained in regular direct-labeled microsatellite fluorescent detection assays, while significantly reducing the costs (Missiaggia & Grattapaglia, 2006). This procedure is particularly adapted when many SSRs need to be investigated on relatively few samples.

Box 3: Problems arising during SSR amplification

A number of problems can arise during amplification. They can compromise allele calling and binning, resulting in increased error rates or extensive need for manual corrections, and should therefore be identified as early as possible (Figures 3 and 4):

- *Low heterozygote peak height ratios* (Figure 3B). They are caused by mutations in the flanking region, at primers binding sites, resulting in poor amplification of the corresponding allele. Possible solutions to avoid them are similar to those put forth for null alleles below.
- *Stuttering or shadow bands* (Figure 3C). This corresponds to the amplification of PCR products that differ from the original template by multiples of the repeat unit length. This widespread phenomenon complicates the interpretation of electropherograms. Due to a strong bias towards contractions, stutter bands are typically shorter than the original fragment (Shinde *et al.*, 2003). To reduce stuttering, one option is to decrease denaturation temperature to 83°C (Olejniczak & Krzyzosiak, 2006), another is to use new-generation polymerases such as fusion enzymes (Fazekas *et al.*, 2010). However, the best solution is to select loci that present reduced stuttering from the outset (e.g. O'Reilly *et al.*, 2000). Note that M13-tails labeling can result in slight stuttering due to low melting temperature of this primer (53°C), so if primers are first tested in simplex with an M13-tail, some improvements can be expected at the time of multiplexing.
- *Split peaks* (Figure 3D). This is caused by the non-template addition of a nucleotide (generally an adenine) to PCR fragments by the *Taq* polymerase (Clark, 1988; Esselink *et al.*, 2003). When this adenylation is incomplete, it results in double peaks (the original fragment and an additional peak 1 bp longer corresponding to the adenylated fragment), thereby compromising automatic peak recognition, particularly for heterozygote genotypes with nearby alleles. The addition of a guanine base (G), a ''PIG-tail'' (5′-GTTTCTT-3′ or 5′-GTTT-3′), or longer (40 bp) sequences at the 5′ end of the reverse (non-labeled) primer has been shown to promote full adenylation of some fragments during PCR (Brownstein *et al.*, 1996; Binladen *et al.*, 2007; Hill *et al.*, 2009). However, according to our observations, PCR

efficiency can decrease with such tailed primers. This can in some cases be compensated by increasing the number of amplification cycles, as shown for primers with M13 tails (de Arruda *et al.*, 2010). Other suggestions to promote complete adenylation include the reduction of the amount of template DNA, down to 10 ng (Lederer *et al.*, 2000; Butler, 2005b), the decrease of primer concentration, the increase of *Taq* concentration (Fishback *et al.*, 1999), or the use of alternative polymerases (Hu, 1993; Vallone *et al.*, 2008).

- *Null alleles* (Figures 4A and 4B). These are non-amplifying alleles that result in an apparent homozygote when present in heterozygote state and in the lack of amplification when present in homozygote state. In the latter case, they can be confounded with reaction failure (Varshney *et al.*, 2005). Null alleles are produced by mutations in the flanking region, at primer binding sites. When null alleles are present, observed banding patterns represent one of several possible true genotypes. While methods have been developed to mitigate this problem during data analysis (e.g. Wagner *et al.*, 2006; Chapuis & Estoup, 2007), the best approach is to avoid design primers in polymorphic regions, either using prior information on sequence variation (Meglécz *et al.*, 2010) or by checking early on all candidate loci using Mendelian segregation analyses. In our lab, we use 12 or 24 progenies (one mother and seven of her open-pollinated progenies) representing one or two 96-well plates. The use of full-sib families (e.g. the mother, the father and six offspring) would be twice as informative by screening both the mother and the father for the presence of null alleles. If such approaches are not feasible, deviations from Hardy-Weinberg equilibrium proportions can be investigated (van Oosterhout *et al.*, 2004). In the 100 studies that we surveyed, explicit tests of the presence of null alleles were reported in only 40% of the studies.
- *Primer-dimers, artifactual bands* (Figure 4C) *and triallelic patterns* (Figure 4D). These can be caused by the mispriming of primers (Brownie *et al.*, 1997; Hill *et al.*, 2009). Although the artifacts produced could be simply omitted during scoring if they do not interfere with allele calling, they may be a criterion for exclusion or redesign, to facilitate automatic interpretation of electropherograms.

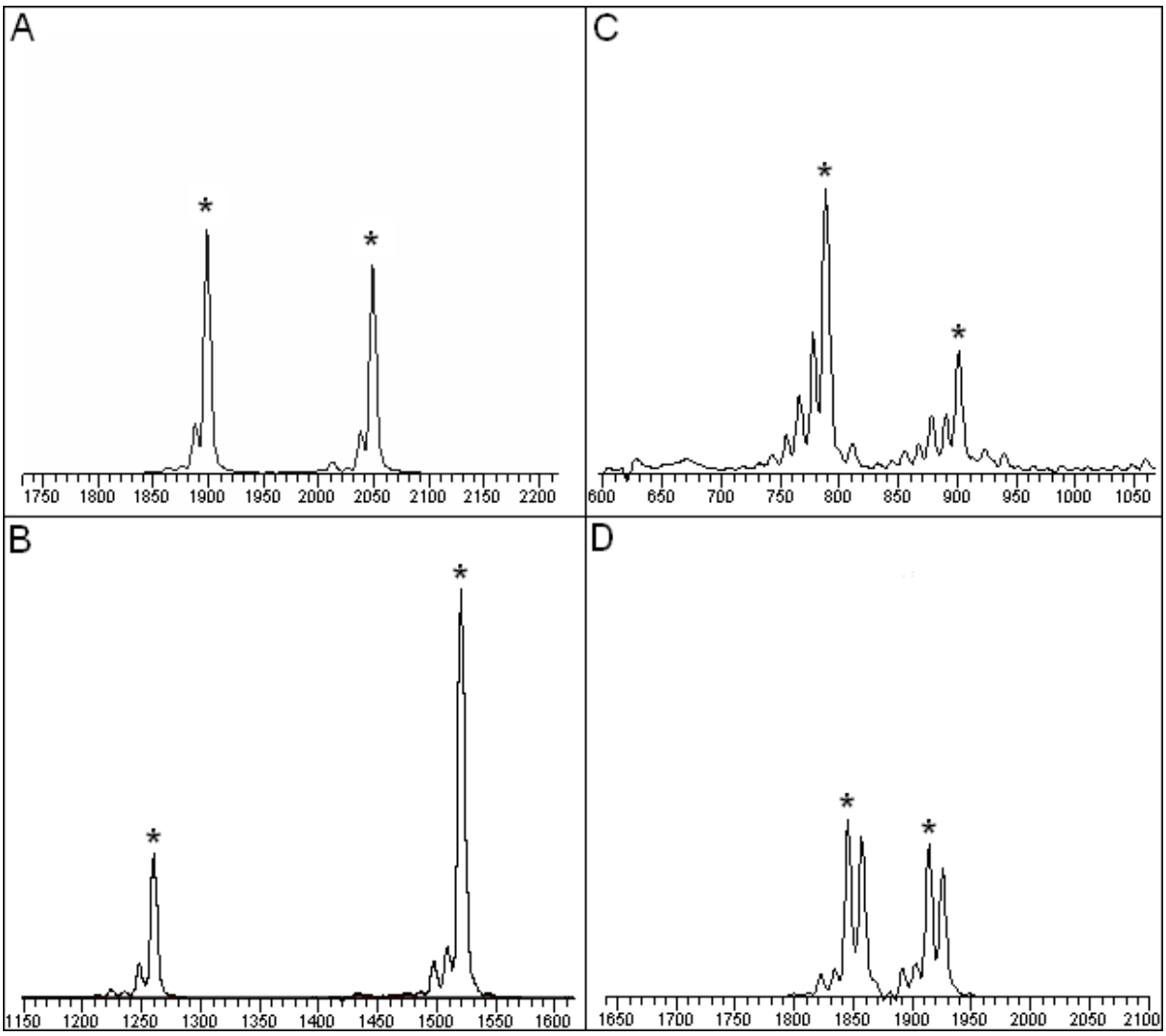


Figure 3: Illustration of SSR profiles generated on capillary sequencer: correct profile (A), low heterozygote peak height ratios (B), excessive stuttering (C), and split peaks (D). Correct alleles are marked with asterisks.

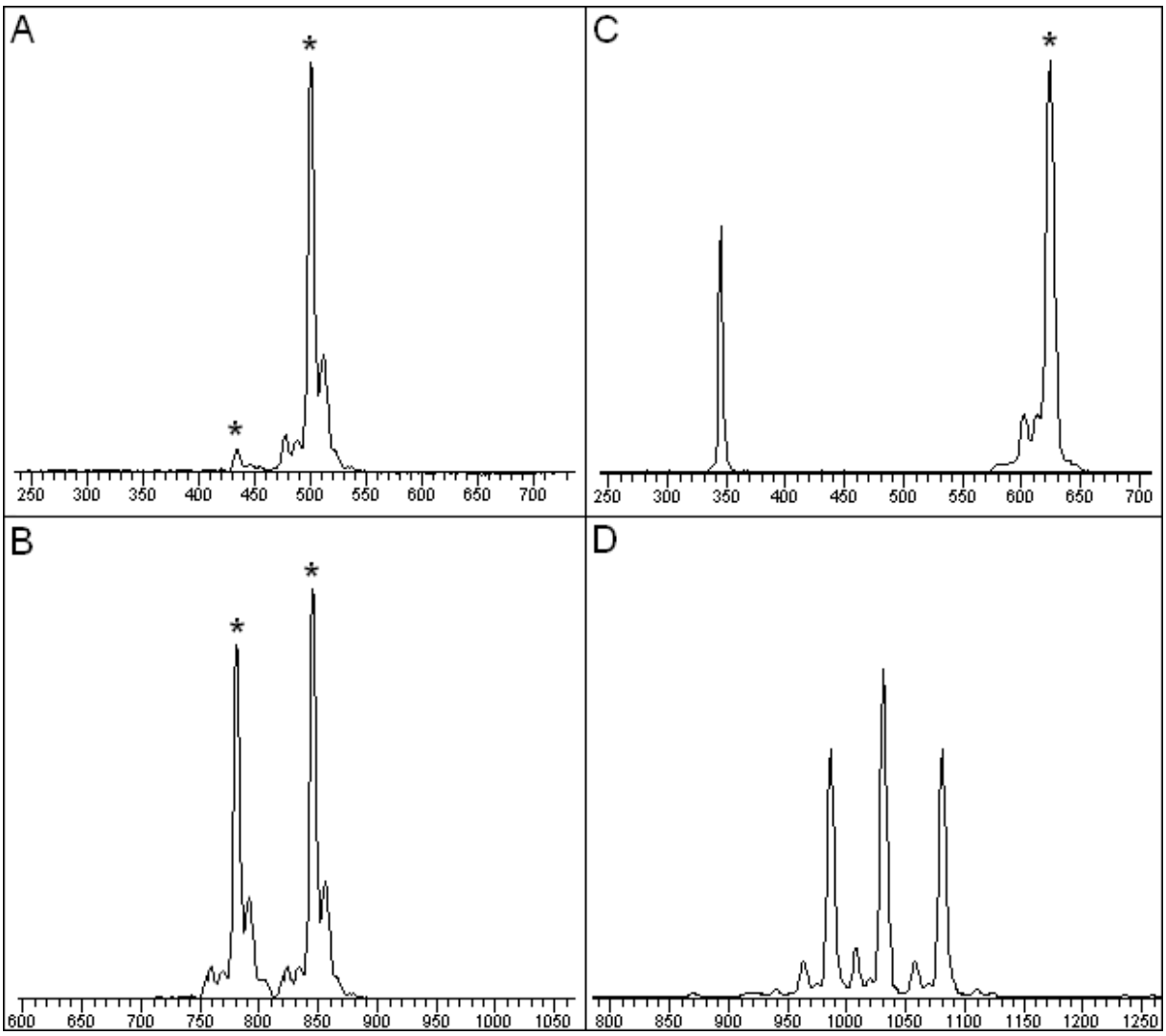


Figure 4: Illustration of SSR profiles generated on capillary sequencer: weak allele before (A) and after (B) successful primer redesign, artifactual band (C) and triallelic pattern (D). Correct alleles are marked with asterisks.

ACKNOWLEDGMENTS

We are especially grateful to Christophe Boury for developing the robotics used in the frame of SSR genotyping and to Sarah Monllor for help with genotyping. We also thank Joëlle Chat, François Hubert and Stephanie Mariette for their useful comments on the paper and Sophie Lefèvre for sharing with us her experience on the development of multiplexed SSRs in beech. The experience on genotyping was gained in our Genome-Transcriptome facility, which is part of the Functional Genomic Center of Bordeaux. We acknowledge financial support from the Aquitaine Region, from the EVOLTREE Network of Excellence supported by the 6th Framework Programme of the European Commission no. 016322 and from the LINKTREE project from the Eranet Biodiversa Programme (ANR-08-BDVA-006).

SUPPORTING INFORMATION

Supporting Information S1: Endnote library of 100 original journal articles relying on SSRs that had been published recently (in 2009-2010) in the journal *Molecular Ecology*.

Supporting Information S2: Endnote library of 15 original journal articles relying on SSR identification using next-generation sequencing techniques, published recently (2009-2010).

REFERENCES

Abbott C, Ebert D, Tabata A, Therriault T (2010) Twelve microsatellite markers in the invasive tunicate, *Didemnum vexillum*, isolated from low genome coverage 454 pyrosequencing reads. *Conservation Genetics Resources* **3**, 79-81.

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques* **46**, 185-192.

Alberto F (2009) MsatAllele_1.0: An R package to visualize the binning of microsatellite alleles. *Journal of Heredity* **100**, 394-397.

Allentoft M, Schuster SC, Holdaway R, *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *BioTechniques* **46**, 195-200.

Amos W, Hoffman JI, Frodsham A, *et al.* (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes* **7**, 10-14.

Anonymous (2002) Multiplex PCR that simply works — the new QIAGEN® Multiplex PCR Kit. *QIAGENews* **5**, 14-16.

Ashley MV, Caballero IC, Chaovalitwongse W, *et al.* (2009) KINALYZER, a computer program for reconstructing sibling groups. *Molecular Ecology Resources* **9**, 1127-1131.

Berger-Wolf TY, Sheikh SI, DasGupta B, *et al.* (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics* **23**, i49-56.

Beyor N, Yi L, Seo TS, Mathies RA (2009) Integrated capture, concentration, polymerase chain reaction, and capillary electrophoretic analysis of pathogens on a chip. *Analytical Chemistry* **81**, 3523-3528.

Bienvenue JM, Legendre LA, Ferrance JP, Landers JP (2009) An integrated microfluidic device for DNA purification and PCR amplification of STR fragments. *Forensic Science International Genetics* **4**, 178-186.

Binladen J, Gilbert MTP, Campos PF, Willerslev E (2007) 5 '-Tailed sequencing primers improve sequencing quality of PCR products. *BioTechniques* **42**, 174-176.

Bonin A, Bellemain E, Eidesen PB, *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261-3273.

Boomer J, Stow A (2010) Rapid isolation of the first set of polymorphic microsatellite loci from the Australian gummy shark, *Mustelus antarcticus* and their utility across divergent shark taxa. *Conservation Genetics Resources* **2**, 393-395.

Borsting C, Rockenbauer E, Morling N (2009) Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Science International Genetics* **4**, 34-42.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* **13**, 3601-3608.

Brownie J, Shawcross S, Theaker J, *et al.* (1997) The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Research* **25**, 3235-3241.

Brownstein MJ, Carpten D, Smith JR (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: Primer modifications that facilitate genotyping. *BioTechniques* **20**, 1004-1010.

Buchan JC, Archie EA, Van Horn RC, Moss CJ, Alberts SC (2005) Locus effects and sources of error in noninvasive genotyping. *Molecular Ecology Notes* **5**, 680-683.

Buggs RJ, Chamala S, Wu W, *et al.* (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology* **19**, 132-146.

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays* **28**, 1040-1050.

Butler JM (2005a) Constructing STR multiplex assays. *Methods in Molecular Biology* **297**, 53-65.

Butler JM (2005b) *Forensic DNA Typing, Second Edition: Biology, Technology, and Genetics of STR Markers* Elsevier Academic Press, London.

Butler JM, Buel E, Crivellente F, McCord BR (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis* **25**, 1397-1412.

Butler JM, Ruitberg CM, Vallone PM (2001) Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. *Fresenius Journal of Analytical Chemistry* **369**, 200-205.

Castoe TA, Poole AW, Gu WJ, *et al.* (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources* **10**, 341-347.

Chamberlain JS, Gibbs RA, Rainer JE, Nguyen PN, Thomas C (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Research* **16**, 11141-11156.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology -- Part B: Biochemistry and Molecular Biology* **126**, 455-476.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* **24**, 621-631.

Chen JW, Uboh CE, Soma LR, *et al.* (2010) Identification of racehorse and sample contamination by novel 24-plex STR system. *Forensic Science International Genetics* **4**, 158-167.

Chistiakov DA, Hellemans B, Volckaert FAM (2006) Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture* **255**, 1-29.

Christians JK, Watt CA (2009) Mononucleotide repeats represent an important source of polymorphic microsatellite markers in *Aspergillus nidulans*. *Molecular Ecology Resources* **9**, 572-578.

Cipriani G, Marrazzo MT, Di Gaspero G, *et al.* (2008) A set of microsatellite markers with long core repeat optimized for grape (*Vitis* spp.) genotyping. *BMC Plant Biology* **8**, 127.

Clark JM (1988) Novel non-templated nucleotide addition-reactions catalyzed by procaryotic and eukaryotic DNA-polymerases. *Nucleic Acids Research* **16**, 9677-9686.

Clayton TM, Whitaker JP, Sparkes R, Gill P (1998) Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International* **91**, 55-70.

Collins HE, Li H, Inda SE, *et al.* (2000) A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Human Genetics* **106**, 218-226.

Cryer N, Butler D, Wilkinson M (2005) High throughput, high resolution selection of polymorphic microsatellite loci for multiplex analysis. *Plant Methods* **1**, 3.

Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered dwarf bulrush (*Typha minima*) using next-generation sequencing technology. *Journal of Heredity* **101**, 789-793.

Dawson DA, Horsburgh GJ, Küpper C, *et al.* (2010) New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility - as demonstrated for birds. *Molecular Ecology Resources* **10**, 475-494.

de Arruda M, Gonçalves E, Schneider M, da Costa da Silva A, Morielle-Versute E (2010) An alternative genotyping method using dye-labeled universal primer to reduce unspecific amplifications. *Molecular Biology Reports* **37**, 2031-2036.

Dereeper A, Argout X, Billot C, Rami JF, Ruiz M (2007) SAT, a flexible and optimized Web application for SSR marker development. *BMC Bioinformatics* **8**, 465.

DeWoody JA, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. *Molecular Ecology Notes* **6**, 951-957.

Dieringer D, Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Research* **13**, 2242-2251.

Dubut V, Grenier R, Meglécz E, *et al.* (2010) Development of 55 novel polymorphic microsatellite loci for the critically endangered *Zingel asper* L. (Actinopterygii: Perciformes: Percidae) and cross-species amplification in five other percids. *European Journal of Wildlife Research*, 10.1007/s10344-10010-10421-x.

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resources* **9**, 673-690.

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**, 746-756.

Edwards MC, Gibbs RA (1994) Multiplex PCR: advantages, development, and applications. *PCR Methods and Applications* **3**, 65-75.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16**, 551-558.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435-445.

Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE (2000) Multiplex PCR: optimization and application in diagnostic virology. *Clinical Microbiology Reviews* **13**, 559-570.

Esselink GD, Smulders MJM, Vosman B (2003) Identification of cut rose (*Rosa hybrida*) and rootstock varieties using robust sequence tagged microsatellite site markers. *Theoretical and Applied Genetics* **106**, 277-286.

Estoup A, Garnery L, Solignac M, Cornuet JM (1995) Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**, 679-695.

Ewen KR, Bahlo M, Treloar SA, *et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics* **67**, 727-736.

Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources* **8**, 92-94.

Fazekas AJ, Steeves R, Newmaster SG (2010) Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques* **48**, 277-281.

Fishback AG, Danzmann RG, Sakamoto T, Ferguson MM (1999) Optimization of semi-automated microsatellite multiplex polymerase chain reaction systems for rainbow trout (*Oncorhynchus mykiss*). *Aquaculture* **172**, 247-254.

Frasier TR, White BN (2008) Increased efficiency of genetic profiling through quantity and quality assessment of fluorescently labeled oligonucleotide primers. *BioTechniques* **44**, 49-52.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current Protocols in Human Genetics* **2**, 1-18.

Galan M, Cosson JF, Aulagnier S, *et al.* (2003) Cross-amplification tests of ungulate primers in roe deer (*Capreolus capreolus*) to develop a multiplex panel of 12 microsatellite loci. *Molecular Ecology Notes* **3**, 142-146.

Ghebranious N, Ivacic L, Mallum J, Dokken C (2005) Detection of ApoE E2, E3 and E4 alleles using MALDI-TOF mass spectrometry and the homogeneous mass-extend technology. *Nucleic Acids Research* **33**, e149.

Ghislain M, Spooner DM, Rodríguez F, *et al.* (2004) Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. *Theoretical and Applied Genetics* **108**, 881-890.

Ghosh S, Karanjawala ZE, Hauser ER, *et al.* (1997) Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group. *Genome Research* **7**, 165-178.

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine* **114**, 204-210.

Gill P, Brenner C, Brinkmann B, *et al.* (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *International Journal of Legal Medicine* **114**, 305-309.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology* **12**, 1039-1047.

Greenspoon SA, Yeung SH, Johnson KR, *et al.* (2008) A forensic laboratory tests the Berkeley microfabricated capillary array electrophoresis device. *Journal of Forensic Sciences* **53**, 828-837.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplex (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources* **in press**.

Gupta PK, Rustgi S, Sharma S, *et al.* (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics* **270**, 315-323.

Gusmão L, Butler JM, Carracedo A, *et al.* (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *International Journal of Legal Medicine* **120**, 191-200.

Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology* **15**, 3715-3730.

Hahn M, Wilhelm J, Pingoud A (2001) Influence of fluorophor dye labels on the migration behavior of polymerase chain reaction - amplified short tandem repeats during denaturing capillary electrophoresis. *Electrophoresis* **22**, 2691-2700.

Hartzell B, Graham K, McCord B (2003) Response of short tandem repeat systems to temperature and sizing methods. *Forensic Science International* **133**, 228-234.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics* **9**, 80.

Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH (1997) Multiplex PCR: critical parameters and step-by-step protocol. *BioTechniques* **23**, 504-511.

Hill CR, Butler JM, Vallone PM (2009) A 26plex autosomal STR assay to aid human identity testing. *Journal of Forensic Sciences* **54**, 1008-1015.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* **14**, 599-612.

Hoffman JI, Matson CW, Amos W, Loughlin TR, Bickham JW (2006) Deep genetic subdivision within a continuously distributed and highly vagile marine mammal, the Steller's sea lion (*Eumetopias jubatus*). *Molecular Ecology* **15**, 2821-2832.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* **46**, 511-517.

Horsman KM, Bienvenue JM, Blasier KR, Landers JP (2007) Forensic DNA analysis on microfluidic devices: A review. *Journal of Forensic Sciences* **52**, 784-799.

Hu G (1993) DNA Polymerase-catalyzed addition of nontemplated extra nucleotides to the 3' of a DNA fragment. *DNA and Cell Biology* **12**, 763-770.

Idury RM, Cardon LR (1997) A simple method for automated allele binning in microsatellite markers. *Genome Research* **7**, 1104-1109.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**, 424-429.

Jayashree B, Reddy PT, Leeladevi Y, *et al.* (2006) Laboratory Information Management Software for genotyping workflows: applications in high throughput crop genotyping. *BMC Bioinformatics* **7**, 383.

Johansson Å, Karlsson P, Gyllensten U (2003) A novel method for automatic genotyping of microsatellite markers based on parametric pattern recognition. *Human Genetics* **113**, 316-324.

Johnson PCD, Haydon DT (2007) Software for quantifying and simulating microsatellite genotyping error. *Bioinformatics and Biology Insights* **2007**, 71-75.

Jones B, Walsh D, Werner L, Fiumera A (2009) Using blocks of linked single nucleotide polymorphisms as highly polymorphic genetic markers for parentage analysis. *Molecular Ecology Resources* **9**, 487-497.

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* **10**, 551-555.

Kaplinski L, Andreson R, Puurand T, Remm M (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* **21**, 1701-1702.

Karaiskou N, Primmer C (2008) PCR multiplexing for maximising genetic analyses with limited DNA samples: an example in the collared flycatcher, *Ficedula albicollis*. *Annales Zoologici Fennici* **45**, 478-482.

Kelkar YD, Strubczewski N, Hile SE, *et al.* (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution* **2**, 620-635.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* **18**, 30-38.

Kenta T, Gratten J, Haigh NS, *et al.* (2008) Multiplex SNP-SCALE: a cost-effective medium-throughput single nucleotide polymorphism genotyping method. *Molecular Ecology Resources* **8**, 1230-1238.

Kim TS, Booth J, Gauch H, *et al.* (2008) Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. *BMC Genomics* **9**, 31.

Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. *BioEssays* **32**, 524-536.

Kirov G, Williams N, Sham P, Craddock N, Owen MJ (2000) Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Research* **10**, 105-115.

Kline MC, Duewer DL, Redman JW, Butler JM (2005) Results from the NIST 2004 DNA quantitation study. *Journal of Forensic Sciences* **50**, 571-578.

Koumi P, Green HE, Hartley S, *et al.* (2004) Evaluation and validation of the ABI 3700, ABI 3100, and the MegaBACE 1000 capillary array electrophoresis instruments for use with short tandem repeat microsatellite typing in a forensic environment. *Electrophoresis* **25**, 2227-2241.

Kraemer L, Beszteri B, Gabler-Schwarz S, *et al.* (2009) STAMP: Extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. *BMC Bioinformatics* **10**, 41.

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* **17**, 21-24.

Lai E (2001) Application of SNP technologies in medicine: lessons learned and future challenges. *Genome Research* **11**, 927-929.

Lederer T, Seidl S, Graham B, Betz P (2000) A new pentaplex PCR system for forensic casework analysis. *International Journal of Legal Medicine* **114**, 87-92.

Lepais O, Bacles C (2010) Comparison of random and SSR-enriched shotgun pyrosequencing for microsatellite discovery and single multiplex PCR optimisation in *Acacia harpophylla* F. Muell. Ex Benth. *Molecular Ecology Resources* **submitted (MER-10-0446)**.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**, 203-221.

Li JL, Deng H, Lai DB, *et al.* (2001) Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. *Genome Research* **11**, 1304-1314.

Li JZ, Absher DM, Tang H, *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**, 2453-2465.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**, S26.

Liu P, Mathies RA (2009) Integrated microfluidic systems for high-performance genetic analysis. *Trends in Biotechnology* **27**, 572-581.

Livingstone D, Freeman B, Tondo CL, *et al.* (2009) Improvement of high-throughput genotype analysis after implementation of a dual-curve Sybr Green I-based quantification and normalization procedure. *HortScience* **44**, 1228-1232.

Luikart G, Zundel S, Rioux D, *et al.* (2008) Low genotyping error rates and noninvasive sampling in bighorn sheep. *Journal of Wildlife Management* **72**, 299-304.

Malausa T, Gilles A, Meglécz E, *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources* **in press**.

Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. *Journal of Clinical Laboratory Analysis* **16**, 47-51.

Martin J-F, Pech N, Meglecz E, *et al.* (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **11**, 560.

Matschiner M, Salzburger W (2009) TANDEM: integrating automated allele binning into genetics and genomics workflows. *Bioinformatics* **25**, 1982-1983.

Meglécz E, Costedoat C, Dubut V, *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404.

Meldgaard M, Morling N (1997) Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations. *Electrophoresis* **18**, 1928-1935.

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science* **12**, 106-117.

Missiaggia A, Grattapaglia D (2006) Plant microsatellite genotyping with 4-color fluorescent detection using multiple-tailed primers. *Genetics and Molecular Research* **5**, 72-78.

Moretti TR, Baumstark AL, Defenbaugh DA, *et al.* (2001) Validation of STR typing by capillary electrophoresis. *Journal of Forensic Sciences* **46**, 661-676.

Morin PA, Luikart G, Wayne RK, the SNPwg (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.

Morin PA, Manaster C, Mesnick SL, Holland R (2009) Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with ALLELOGRAM. *Molecular Ecology Resources* **9**, 1451-1455.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes* **7**, 937-946.

Neff BD, Fu P, Gross MR (2000) Microsatellite multiplexing in fish. *Transactions of the American Fisheries Society* **129**, 584-593.

O'Reilly PT, Canino MF, Bailey KM, Bentzen P (2000) Isolation of twenty low stutter di- and tetranucleotide microsatellites for population analyses of walleye pollock and other gadoids. *Journal of Fish Biology* **56**, 1074-1086.

Oetting WS, Lee HK, Flanders DJ, *et al.* (1995) Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. *Genomics* **30**, 450-458.

Ohashi J, Tokunaga K (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *Journal of Human Genetics* **48**, 487-491.

Olejniczak M, Krzyzosiak WJ (2006) Genotyping of simple sequence repeats factors implicated in shadow band generation revisited. *Electrophoresis* **27**, 3724-3734.

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology* **29**, 294-307.

Palsson B, Palsson F, Perlin M, *et al.* (1999) Using quality measures to facilitate allele calling in high-throughput genotyping. *Genome Research* **9**, 1002-1012.

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180.

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006) EST databases as a source for molecular markers: Lessons from *Helianthus*. *Journal of Heredity* **97**, 381-388.

Petit RJ, Deguilloux M-F, Chat J, *et al.* (2005) Standardizing for microsatellite length in comparisons of genetic diversity. *Molecular Ecology* **14**, 885-890.

Pettersson E, Lindskog M, Lundeberg J, Ahmadian A (2006) Tri-nucleotide threading for parallel amplification of minute amounts of genomic DNA. *Nucleic Acids Research* **34**, e49.

Phillips C, Fang R, Ballard D, *et al.* (2007) Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Science International Genetics* **1**, 180-185.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847-846.

Presson AP, Sobel EM, Pajukanta P, *et al.* (2008) Merging microsatellite data: enhanced methodology and software to combine genotype data for linkage and association analysis. *BMC Bioinformatics* **9**, 317.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rachlin J, Ding C, Cantor C, Kasif S (2005) MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Research* **33**, 544-547.

Rasmussen D, Noor M (2009) What can you do with 0.1x genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). *BMC Genomics* **10**, 382.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2) - Population-genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Renshaw MA, Saillant E, Bradfield SC, Gold JR (2006) Microsatellite multiplex panels for genetic studies of three species of marine fishes: red drum (*Sciaenops ocellatus*), red snapper (*Lutjanus campechanus*), and cobia (*Rachycentron canadum*). *Aquaculture* **253**, 731-735.

Rithidech K, Dunn JJ (2003) Combining multiplex and touchdown PCR for microsatellite analysis. *Methods in Molecular Biology* **226**, 295-300.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* **73**, 1402-1422.

Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**, 365-386.

Ryynänen HJ, Tonteri A, Vasemägi A, Primmer CR (2007) A comparison of biallelic markers and microsatellites for the estimation of population and conservation genetic parameters in Atlantic salmon (*Salmo salar*). *Journal of Heredity* **98**, 692-704.

Saarinen EV, Austin JD (2010) When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa darter (*Etheostoma okaloosae*). *Journal of Heredity* **101**, 784-788.

Sanchez JJ, Endicott P (2006) Developing multiplexed SNP assays with special reference to degraded DNA templates. *Nature Protocols* **1**, 1370-1378.

Santana QC, Coetzee MPA, Steenkamp ET, *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques* **46**, 217-223.

Santure AW, Stapley J, Ball AD, *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology* **19**, 1439-1451.

Scandura M, Capitani C, Iacolina L, Marco A (2006) An empirical approach for reliable microsatellite genotyping of wolf DNA from multiple noninvasive sources. *Conservation Genetics* **7**, 813-823.

Schlötterer C (1998) Genome evolution: Are microsatellites really simple sequences? *Current Biology* **8**, 132-134.

Schlötterer C (2004) The evolution of molecular markers: just a matter of fashion? *Nature Reviews Genetics* **5**, 63-69.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233-234.

Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16-18.

Schwengel DA, Jedlicka AE, Nanthakumar EJ, Weber JL, Levitt RC (1994) Comparison of fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. *Genomics* **22**, 46-54.

Seddon JM, Parker HG, Ostrander EA, Ellegren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology* **14**, 503-511.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.

Sgueglia J, Geiger S, Davis J (2003) Precision studies using the ABI Prism 3100 Genetic Analyzer for forensic DNA analysis. *Analytical and Bioanalytical Chemistry* **376**, 1247-1254.

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. *Trends in Biotechnology* **25**, 490-498.

Shen Z, Qu W, Wang W, *et al.* (2010) MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics* **11**, 143.

Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Research* **31**, 974-980.

Sinville R, Soper SA (2007) High resolution DNA separations using microchip electrophoresis. *Journal of Separation Science* **30**, 1714-1728.

Spathis R, Lum JK (2008) An updated validation of Promega's PowerPlex 16 System: high throughput databasing under reduced PCR volume conditions on Applied Biosystem's 96 capillary 3730xl DNA Analyzer. *Journal of Forensic Sciences* **53**, 1353-1357.

Squirrell J, Hollingsworth PM, Woodhead M, *et al.* (2003) How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology* **12**, 1339-1348.

Sun X, Liu Y, Lutterbaugh J, *et al.* (2006) Detection of mononucleotide repeat sequence alterations in a large background of normal DNA for screening high-frequency microsatellite instability cancers. *Clinical Cancer Research* **12**, 454-459.

Syvänen A-C (2005) Toward genome-wide SNP genotyping. *Nature Genetics* **37**, 5-10.

Tang J, Baldwin SJ, Jacobs JM, *et al.* (2008) Large-scale identification of polymorphic microsatellites using an in silico approach. *BMC Bioinformatics* **9**, 374.

Tautz D, Schlötterer C (1994) Simple sequences. *Current Opinion in Genetics & Development* **4**, 832-837.

Techen N, Arias RS, Glynn NC, *et al.* (2010) Optimized construction of microsatellite-enriched libraries. *Molecular Ecology Resources* **10**, 508-515.

Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. *Genome Research* **15**, 1592-1593.

Toonen RJ, Hughes S (2001) *Increased throughput for fragment analysis on an ABI PRISM 377 automated sequencer using a membrane comb and STRand software* Eaton, Natick, MA, ETATS-UNIS.

Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in Short Tandem Repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine* **107**, 13-20.

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques* **37**, 226-231.

Vallone PM, Hill CR, Butler JM (2008) Demonstration of rapid multiplex PCR amplification involving 16 genetic loci. *Forensic Science International Genetics* **3**, 42-45.

van Asch B, Pinheiro R, Pereira R, *et al.* (2010) A framework for the development of STR genotyping in domestic animal species: Characterization and population study of 12 canine X-chromosome loci. *Electrophoresis* **31**, 303-308.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

van Rossum T, Tripp B, Daley D (2010) SLIMS--a user-friendly sample operations and inventory management system for genotyping labs. *Bioinformatics* **26**, 1808-1810.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23**, 48-55.

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity* **97**, 336-345.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology* **10**, 249-256.

Wallin JM, Holt CL, Lazaruk KD, Nguyen TH, Walsh PS (2002) Constructing universal multiplex PCR systems for comparative genotyping. *Journal of Forensic Sciences* **47**, 52-65.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics* **166**, 1963-1979.

Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**, 1579-1594.

Weber JL (1990) Informativeness of human (dC-dA)n · (dG-dT)n polymorphisms. *Genomics* **7**, 524-530.

Webster MT, Smith NGC, Ellegren H (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 8748-8753.

Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Research* **12**, 430-435.

Zajac P, Öberg C, Ahmadian A (2009) Analysis of Short Tandem Repeats by parallel DNA threading. *PLoS One* **4**, e7823.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology* **11**, 1-16.

CHAPITRE 2

Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus spp*.)

E. Guichoux[1,2,3], L. Lagache[1,2], S. Wagner[1,2,4], P. Léger[1,2] and R.J. Petit[1,2]

[1] INRA, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[2] University of Bordeaux, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[3] Centre de Recherche Pernod Ricard, F-94000 Créteil, France

[4] University of Bonn, Steinmann Institut, D-53115 Bonn, Germany

Published in 2011 in *Molecular Ecology Resources*

INTRODUCTION

Oaks (*Quercus spp.*) are widely distributed across the Northern Hemisphere. They are often dominant forest tree species and play therefore key ecological and economical roles. For instance, in France, they represent 40% of the forests and almost 60% of wood lumber production. The two major temperate European species (*Quercus petraea* and *Q. robur*) have become important models for population genetic and speciation studies (Streiff *et al.*, 1998; Streiff *et al.*, 1999; Muir *et al.*, 2000; Petit *et al.*, 2002; Barreneche *et al.*, 2004; Petit *et al.*, 2004; Scotti-Saintagne *et al.*, 2004; Prida *et al.*, 2007; Lepais *et al.*, 2009; Morin *et al.*, 2010). Studying the evolutionary dynamics of such closely related species requires suitable genetic markers (Vähä & Primmer, 2006). In recent studies, SSRs (Simple Sequence Repeats) have been the markers of choice to study hybridization (Burgarella *et al.*, 2009 voir **Annexe 4**; Viscosi *et al.*, 2009; Ortego & Bonal, 2010; Penaloza-Ramirez *et al.*, 2010) and population genetic structure (Neophytou *et al.*, 2010). At the same time, SNP (Single Nucleotide Polymorphism) genotyping is emerging as a possible alternative, in oaks as in other tree species (Namroud *et al.*, 2008; Eckert *et al.*, 2009; Lascoux & Petit, 2010). Nevertheless, many basic or applied questions in population genetics only require a small number of highly polymorphic markers on large sample numbers. High-density SNP genotyping is not suitable in such cases. Instead, multiplexing SSRs can improve genotyping throughput as well as cost-effectiveness. Multiplexing is the amplification of several markers in a single PCR (Polymerase Chain Reaction) and must be distinguished from pool-plexing, where pooling takes place after PCR. Multiplex PCR is increasingly used (Hayden *et al.*, 2008; Kawalko *et al.*, 2009). However, large multiplexes involving eight or more markers are still uncommon (Hill *et al.*, 2009), due to long development procedures and complex reaction interactions. Since a few years, new tools for multiplex development have appeared, including software for primer design to limit interactions between primers during PCR and for selecting the best combinations of loci (Holleley & Geerts, 2009). Moreover, the generalization of second generation sequencing techniques now allows fast and affordable SSR identification (Abdelkrim *et al.*, 2009; Santana *et al.*, 2009). In oaks, although microsatellites have been available for many years (Dow *et al.*, 1995; Steinkellner *et al.*, 1997; Kampfer *et al.*, 1998), multiplexing efforts were limited, with only two studies reporting multiplexing at no more

than 5 loci (Dzialuk *et al.*, 2005; Lepais *et al.*, 2006). Thus, analyzing large oak populations at multiple markers remains expensive and time-consuming. In the present study, we developed two multiplex kits, a 12-plex of Expressed Sequence Tag-SSRs (eSSRs) and an 8-plex of genomic SSRs (gSSRs), paying particular attention to genotyping accuracy and cost-effectiveness. We describe the whole procedure, with a focus on the binning phase (i.e. the identification of peaks corresponding to the different alleles) by comparing the performance of two genotyping software. Finally, we test the assignment power of both multiplex kits using simulated oak genotypes and study their transferability on congeneric species and on species belonging to other genera within the Fagaceae family.

MATERIAL AND METHODS

Material

Part of the material used is coming from a 5 ha mixed oak stand comprising both *Q. petraea* and *Q. robur* located in the western part of France (Petite Charnie State Forest, Sarthe, latitude: 48.08° N, longitude: 0.17° W). This stand has been intensively studied for many years for gene flow, species differentiation, phenology, and wood characteristics (Bacilieri *et al.*, 1993; Bacilieri *et al.*, 1994; Bacilieri *et al.*, 1995; Streiff *et al.*, 1998; Streiff *et al.*, 1999; Prida *et al.*, 2006; Prida *et al.*, 2007; Lepais *et al.*, 2009). In 2000, 273 adult trees from this stand were grafted in a nursery (Guémené-Penfao, Loire-Atlantique, France). Each genotype was cloned eight times. A total of 898 surviving ramets were sampled (number of ramet per genotype: 1-8, mean: 2.2). In addition, 3780 trees belonging to 51 half-sib families (originating from seeds collected on 28 *Q. robur* and 23 *Q. petraea* adult trees from the Petite Charnie stand) were planted in 1998 and 2001, close to the adult stand. In 2009, we sampled 1257 trees from 35 half-sib families (18 *Q. robur* and 17 *Q. petraea*). For each tree, one leaf or several buds were stored in sealed plastic bags with 10g of silicagel. The taxonomic status of the adult trees had previously been characterized using 19 leaf measures. Trees were classified into three categories, *Q. petraea*, *Q. robur* or intermediate (Kremer *et al.*, 2002). The two multiplex kits were further tested on *Q. pubescens, Q. pyrenaica, Q. alba, Q. rubra, Q. faginea, Q. suber*, *Q. ilex*, *Castanea sativa* and *Fagus sylvatica* (number of samples per species: 5-48), sampled in south west of France in natural populations or in an arboretum (for *Q. alba*, *Q. rubra*, *Q. faginea*, *Q. suber* and *Q. ilex*).

DNA isolation

Five leaf disks (5 mm diameter) or two buds for each tree to standardize the starting quantity of tissue were collected in 96-well plates. DNA was isolated with Invisorb DNA plant HTS 96 kit (Invitek, Germany), following the manufacturer instructions, except for the lysis step (one hour at 65°C). Disruption of plant material was carried out using a Mixer Mill MM300 (Retsch, Germany). In each well of the 96-well plates, a 3mm tungsten bead was added and the plates were frozen in liquid nitrogen for two minutes before a one minute disruption step at 30Hz. DNA quality was estimated on a 1% (w/v) agarose gel stained with GelRed (Biotium, USA). DNA concentration was evaluated on an 8 channel Nanodrop spectrophotometer, and concentration of each sample was adjusted to 10ng/µl on a STARlet 8-channel robot (Hamilton, USA).

Multiplex PCR optimization

Kit-1

Sixty-four eSSRs (Durand *et al.*, 2010) derived from ESTs (Expressed Sequence Tags) were first tested on 24 samples from across the European range (12 *Q. petraea* and 12 *Q. robur* trees). They were analyzed on a 4000L automatic DNA sequencer (LI-COR Biosciences, USA). Criteria for SSR selection were: good amplification quality, no slippage, and high number of alleles (>5). We then determined which specific combination of loci provides the highest species assignment power with the software WHICHLOCI (Banks *et al.*, 2003). A subset of 17 loci was selected for further evaluation.

Kit-2

In the second kit we included highly-validated genomic SSRs (gSSRs) (Dow *et al.*, 1995; Steinkellner *et al.*, 1997; Kampfer *et al.*, 1998), some of which had already been multiplexed (Lepais *et al.*, 2006). We selected 10 loci suitable for species differentiation to develop a second multiplex (8-plex) and to increase taxonomic resolution in combination with *kit-1*.

We first validated all SSRs in simplex using the M13-tail technique (Schuelke, 2000), which allows direct visualization of the PCR product on capillary sequencer. Hence, SSRs presenting low quality profiles, i.e. excessive stuttering, weak alleles, triple bands, unspecific products or heterogeneous profiles (more than 50% of difference in fluorescence intensity

between the two alleles of a heterozygote), were excluded or redesigned from original sequences (Dow *et al.*, 1995; Steinkellner *et al.*, 1997; Kampfer *et al.*, 1998; Durand *et al.*, 2010) using Primer3Plus (Untergasser *et al.*, 2007). To help null-allele detection, 12 families (composed of the female parent and seven offspring) were genotyped at all loci. We also tested microsatellite loci for null alleles, large allele dropout and scoring errors due to stutter peaks with MICRO-CHECKER 2.2.0.3 (Van Oosterhout *et al.*, 2004). Further validations (microsatellites scoring and error rate measurement) were only performed on *kit-1* because gSSRs (*kit-2*) are already highly-validated (Dow *et al.*, 1995; Steinkellner *et al.*, 1997; Kampfer *et al.*, 1998). Once validated in simplex, and prior to multiplexing, primers were examined for possible interactions using a local BLAST. The complementary threshold (the maximum number of AT or CG matches for any two primers within a multiplex reaction) was set to seven (Holleley & Geerts, 2009). The multiplex reactions were then carried out with the Qiagen Multiplex PCR kit (Qiagen, Germany), following the manufacturer instructions, in a 10µl final volume. Final concentration of the Mastermix was also optimized (0.6×), reducing eight times the final cost. Briefly, PCR mix was composed of 3.5µl of sterile water, 3µl of Qiagen Multiplex Buffer (2X), 1µl of primer premix and 2.5µl of DNA (10ng/µl). Concentrations for each primer pair in the primer premix are shown in Table 1. The cycling conditions were: an initial step at 95°C for 15 min; followed by 30 cycles at 94°C for 30 s, 56°C for 1 min and 72°C for 45 s; and a final incubation at 60°C for 10 min. PCR products were separated on 3% agarose gel stained with GelRED (Biotium, USA), diluted 20 times in pure water and run on ABI-3730 (Applied Biosystems, USA), with LIZ600 as internal lane size standard. Similarity between profiles from simplex and multiplex was also checked.

Diversity analyses and assignment power

Allelic richness (A), observed heterozygosity (H_o), F_{IS} and F_{ST} were estimated on 273 adult trees of both species using GENALEX 6 (Peakall & Smouse, 2006). We used simulated data, generated from allele frequencies of purebred individuals with HYBRIDLAB 1.0 (Nielsen *et al.*, 2006), to test the assignment power of the two multiplexes alone and in combination (Burgarella *et al.*, 2009; Lepais *et al.*, 2009). Allele frequencies for *Quercus robur* and *Q. petraea* were first estimated on a subset of 88 purebred samples per species (based on their genotype at 20 SSRs), identified with STRUCTURE 2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2003), with

a burn-in of 50,000 steps followed by 50,000 Markov chain Monte Carlo repetitions. We calculated the average result over 10 runs with *K* (number of groups) set to two, corresponding to the two species, and used a threshold of 0.9 to identify pure individuals from each species. Assignment of simulated genotypes (10,000 purebreds and 10,000 F1 hybrids) relied on the same method, except that we used theoretical intervals of 0-0.25 and 0.75-1 for purebreds and 0.25-0.75 for F1 hybrids (only F1 were generated, not backcrosses, so these thresholds should be optimal to distinguish between parental species and hybrids in the simulations).

Microsatellites scoring (kit-1 only)

Individual genotypes were determined using both Genemapper (Applied Biosystems, USA) and STRand (http://www.vgl.ucdavis.edu/STRand). Alleles were sorted by raw size to detect discrete size variants, with an Excel macro inspired from FLEXIBIN (Amos *et al.*, 2007). The results were used to assign each allele to a bin. We also compared raw sizes between software to test the reproducibility of data obtained with two different algorithms (Advanced Peak Detection Algorithm implemented in Genemapper and Local Southern Algorithm implemented in Strand) on a subset of 490 samples.

Error rate measurement (kit-1 only)

A first error rate was estimated using 80 duplicated samples (6% of the complete dataset) that had been randomly selected, by counting mismatches (Johnson & Haydon, 2007). A second error rate, called "disagreement rate" between human readers, was measured on all 490 samples. Incoherencies were classified as follows: Type A is when one genotype is classified as heterozygous for one reader and as homozygous for the other reader and Type B is when different alleles are selected by both readers. When two different genotypes were obtained for the same sample, we tried to identify a consensus genotype. In a few cases, no consensus genotype could be determined and was considered as missing.

RESULTS

Multiplex PCR optimization

Among the 27 pre-selected SSRs (17 eSSRs and 10 gSSRs), seven were excluded (five with null alleles, one with triple bands and one with low signal once multiplexed). Three primer pairs were re-designed: one locus having a weak allele and two showing overlapping sizes in our first tests. The final profiles obtained for each kit were sharp with homogeneous amplification of the loci (Figures S1 and S2, Supporting Information). Moreover, the analysis of the 35 half-sib families did not reveal a single case of null-allele at any of the 20 SSR markers. Four of the 20 SSR markers, all with di-nucleotide repeat (PIE152, PIE239, PIE258 and PIE271), had one or more off-ladder microvariants (i.e. variants differing from the expected periodicity of two base pairs). These alleles were shown to segregate in progenies and are therefore not amplification artifacts. Interestingly, initial analysis with classical automatic-binning mode (implemented in most commercial software and widely used by many researchers) failed to identify these alleles, resulting in incoherencies when checking for Mendelian segregation (data not shown). With binning based on raw allele size, these alleles are easily identified, increasing the total number of alleles for the corresponding markers. These results confirm the necessity to analyze samples using raw sizes and to bin the alleles afterwards.

SSR properties

We found that gSSRs are more polymorphic than eSSRs (mean allelic richness: 16.9 for gSSRs and 10.3 for eSSRs). This difference is partly due to the presence of SSRs with tri-nucleotide repeats in *kit-1,* as loci with longer repeats are known to be less variable (Kelkar *et al.*, 2008). The loci that best differentiate *Q. robur* from *Q. petraea* are distributed on the two kits (Table 1), with interspecific F_{ST} reaching 0.20 (mean: 0.06, Table 1).

Locus	Primer Sequences (5′ – 3′)	LG	Dye	[C]	Motif	Size (bp)	*A*	H_o	F_{IS}	F_{ST}
PIE020	GCAGAGGCTCTTCTAAATACAGAACT GGGAGGTTTCTGGGAGAGAT	1	FAM	1.00	AG	97-119	11	0.668	-0.002	0.018
PIE223	TAGAAGCCCAACACGGCTAC AGCAAAACACAAACGCACAA	2	FAM	1.00	GGT	197-221	9	0.749	-0.057	0.108
PIE152	TGTACCTCTTTCCTCTCTCTAAAACT GAATTTCTAAACCACTAGCATTGAC	2	FAM	3.75	TA	230-260	15	0.842	-0.024	0.032
PIE242	TGGAGGGAAAAGAACAATGC TTGCAATCCTCCAAATTTAATG	3	VIC	1.00	TA	102-128	12	0.803	0.045	0.038
PIE102	ACCTTCCATGCTCAAAGATG GCTGGTGATACAAGTGTTTGG	11	VIC	0.50	CT	131-161	9	0.722	-0.047	0.008
PIE243	GGGGTCAGTAGGCAAGTCTTC GAGCTGCATATTTTCCTTAGTCAG	10	VIC	0.25	AG	208-222	6	0.151	0.677	0.070
PIE239	TCAACAAATGGCTCAACAGTG CCCATTTGGTAGCAAAGAGTC	NA	PET	0.63	AT	70-83	11	0.590	-0.082	0.159
PIE227	TACCATGATCTGGGAAGCAAC AAGGGCTTGGTTGGGTTAGT	NA	PET	0.38	TGG	156-177	5	0.546	-0.064	0.207
PIE271	CACACTCACCAACCCTACCC GTGCGGTTGTAGACGGAGAT	2	PET	0.50	TC	180-197	10	0.759	0.019	0.021
PIE267	TCCAACCATCAAGGCCATTAC GTGCGAACAGATCCCTTGTC	3	NED	0.25	AG	80-105	10	0.824	-0.038	0.015
PIE258	TTCTCGATCTCAAAACAAAACCA TTTGATTTGTTTAAGGAAAATTGGA	2	NED	0.75	TC	128-159	19	0.880	0.005	0.039
PIE215	TACGAAATGGAGCTGTTGACC TCTCCTTCTCTTCTGCCATGA	12	NED	0.30	GAG	188-206	6	0.553	0.036	0.125
QrZAG7	CAACTTGGTGTTCGGATCAA GTGCATTTCTTTTATAGCATTCAC	2	FAM	0.50	TC	115-153	19	0.874	-0.015	0.025
MsQ13	ACACTCAGACCCACCATTTTTCC TGGCTGCACCTATGGCTCTTAG	6	FAM	0.50	GA	191-221	16	0.785	0.055	0.052
QrZAG112	TTCTTGCTTTGGTGCGCG GTGGTCAGAGACTCGGTAAGTATTC	12	VIC	0.40	GA	85-96	12	0.579	-0.005	0.128
QrZAG20	CCATTAAAAGAAGCAGTATTTTGT GCAACACTCAGCCTATATCTAGAA	1	VIC	0.15	TC	160-200	19	0.874	-0.015	0.025
QpZAG15	CGATTTGATAATGACACTATGG CATCGACTCATTGTTAAGCAC	9	PET	0.50	AG	108-152	14	0.764	-0.026	0.024
* QpZAG110	GGAGGCTTCCTTCAACCTACTT GATCTCTTGTGTGCTGTATTTTT	8	PET	0.50	AG	206-262	16	0.765	0.009	0.024
QrZAG96	CCCAGTCACATCCACTACTGTCC GGTTGGGAAAAGGAGATCAGA	10	NED	0.15	TC	135-194	18	0.628	0.015	0.149
* QrZAG11	CCTTGAACTCGAAGGTGTCC TGGTTGACTAAAGTATGAACTGTTTG	10	NED	0.40	TC	238-267	21	0.828	-0.031	0.075

[1] LG: linkage group (Catherine Bodénès, personal communication), [C]: final concentration in each primer premix (µM), *A*: allelic richness, H_o: observed heterozygosity
NA: Not available
*: redesigned

Table 1: Characteristics of *kit-1* (eSSRs) and *kit-2* (gSSRs), based on 273 samples of *Q. petraea* and *Q. robur* from a mixed oakwood [1]

Assignment power

Results of assignment tests on 20,000 simulated genotypes are shown in Figure 1. The three classes (*Q. robur*, *Q. petraea* and F1 hybrids) are well delimitated, resulting in low assignment error rates, even though *Q. petraea* and *Q. robur* are closely related species. Assignment with all 20 SSRs is much more effective than when using only 8 or 12 loci: the proportion of incorrect assignments is divided by four or five when the two kits are combined, compared to the proportion observed with only one of the two kits (with thresholds of 0.25 and 0.75, see Table 2). Note that the thresholds chosen are considered as optimal. If they had been set to other values, incorrect assignments would have increased for one category (purebreds or F1 hybrids) and decreased for the other one, but the overall error rate would have been increased (Figure S3, Supporting Information).

Kit	Number of markers	Type	*Q. robur*	F1 hybrids	*Q. petraea*	TOTAL
kit-1	12	eSSRs	5.9%	7.5%	5.5%	**6.6%**
kit-2	8	gSSRs	7.5%	6.5%	5.6%	**5.8%**
kit-1 + *kit-2*	20	eSSRs + gSSRs	1.0%	2.0%	0.6%	**1.4%**

Table 2: Incorrect assignment of simulated genotypes with theoretical intervals of 0-0.25 (*Q. robur*), 0.25-0.75 (F1 hybrids) and 0.75-1 (*Q. petraea*), with one and two multiplexes

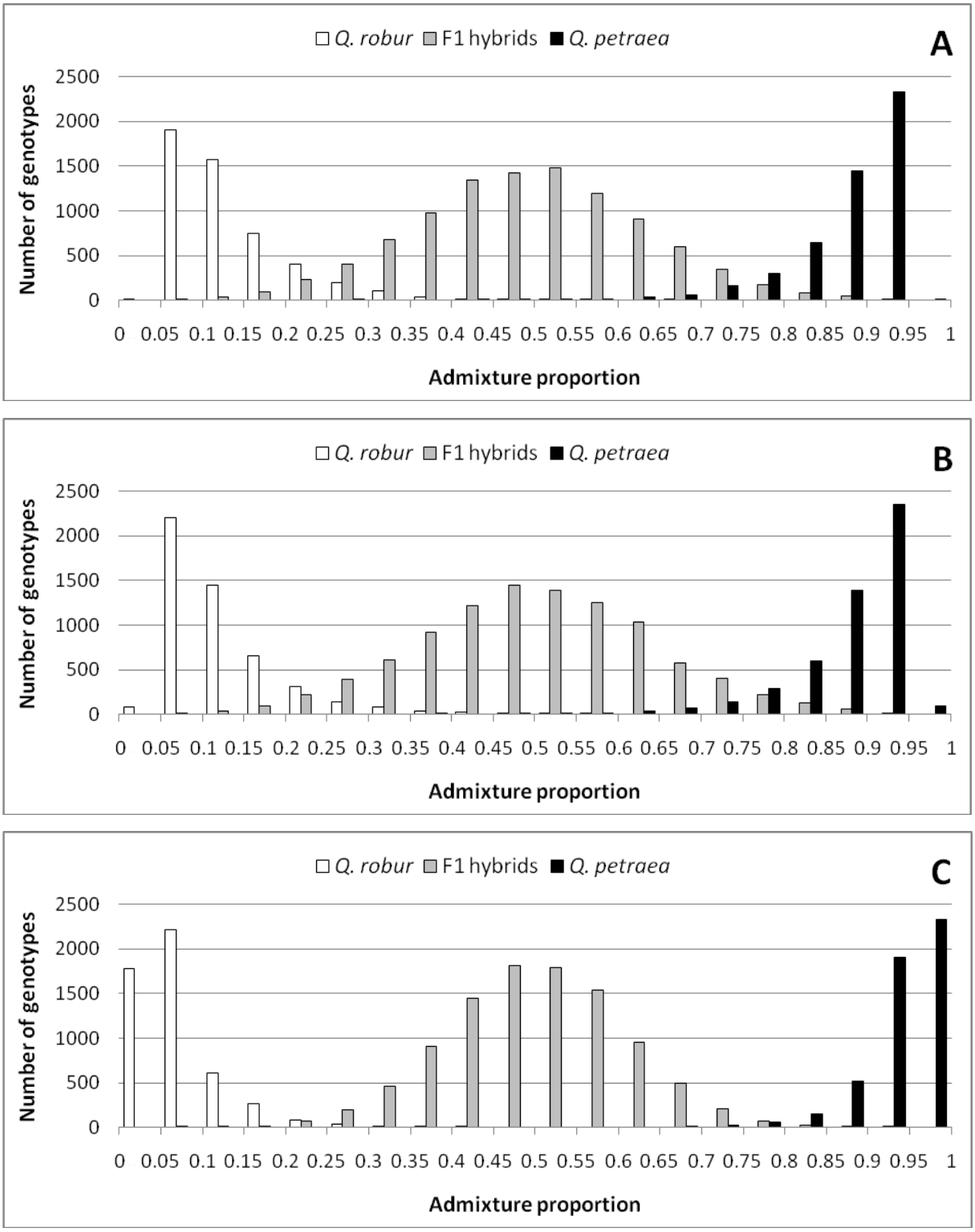


Figure 1: Assignment of 20,000 simulated genotypes (purebred for both parental species and F1 hybrids). A: *kit-1* (12-plex). B: *kit-2* (8-plex). C: *kit-1* + *kit-2* (12-plex + 8-plex)

SSR transferability

All 20 loci amplified in the other oak species tested (*Q. pubescens, Q. pyrenaica, Q. alba, Q. rubra, Q. faginea, Q. suber and Q. ilex*). Our first tests on more distant species showed that all 20 SSRs amplified in *C. sativa*. In *F. sylvatica*, three loci from *kit-1* (PIE020, PIE152 and PIE271) and four from *kit-2* (MsQ13, QpZAG15, QrZAG20 and QrZAG96) failed to amplify with our conditions, even though transferability of gSSRs from *kit-1* has been previously validated in simplex (Barreneche *et al.*, 2004). Depending on the species, we noticed highly heterogeneous profiles and amplification was not successful on all samples, perhaps because of low DNA quality or technical difficulties. The Mendelian segregation analysis and further amplification tests on large populations remain necessary before concluding that these markers can be successfully transferred to these species. Still, it appears that eSSRs (*kit-1*) have a better transferability than gSSRs (*kit-2*), as found in previous studies on other species (Varshney *et al.*, 2005).

Microsatellites scoring and binning (kit-1)

True allele sizes recovered with Genemapper and STRAND were similar (mean deviation: 0.03bp). However, moderate deviation (>0.1bp) was observed between sizes measured with each software in 7.8% of genotypes and large deviation (>0.25bp) was observed in 2.9% of genotypes (maximum deviation: 0.48bp). These deviations are directly induced by the algorithm used to relate internal size marker and allele sizes. This result indicates that even if raw sizes are used for analysis, problems might still occur when samples from different datasets scored with different methods are integrated (Morin *et al.*, 2009).

Error rate measurement (kit-1)

Disagreement rates between both human readers ranged from 0 to 3.6% across all loci (mean 1.1%). Most differences (78%) were due to calling a heterozygous genotype as homozygous by one of the two readers (type A error). Wrong allele calling (type B error) represented only 22% of incoherencies. Type A errors are easily avoidable as they result most of the time in careless mistakes. Type B errors can be decreased by defining clearer reading rules across readers. While corrections involving only 1% of the samples might seem costly in view of the

extra-work involved, it can be critical in studies that are very sensitive to genotyping errors such as parentage analysis (Kalinowski *et al.*, 2007). After establishing consensus genotypes between the two readers, error rates measured by checking the conformity of blindly repeated genotypes ranged from 0% to 1.6%, with a mean of only 0.26% across loci, illustrating the high robustness of markers (Table S4, Supporting Information).

CONCLUSION

Multiplex PCR allows fast, accurate and cost-effective genotyping but requires significant efforts for its development. Primer validation in simplex is the key step of the overall process. If carried out carefully, subsequent multiplexing becomes much easier. Furthermore, if automatic binning seems to save time, genotyping errors appear to be more frequent. As a consequence, we recommend to analyze samples in raw sizes and to bin the data afterwards, which allows accurate analysis of off-ladder microvariants. We believe that these two highly-validated multiplexes will be helpful for future studies on oaks by providing powerful and accurate genotyping tools. In particular, our results confirm the power of microsatellites for hybrid identification. With a larger reference database, assignment rates should be further improved. In addition, with new multiplex SSR or genotyping tools, these markers will be useful in more complex situations involving more than two species or later-generation hybrids. More generally, this development strategy for medium-throughput genotyping assay (presented here from multiplex PCR development to the definition of allele calling rules) could be efficiently transferred to other species.

ACKNOWLEDGEMENTS

We thank Christophe Boury for developing robotic applications used in this project. We thank Alexis Doucousso, François Hubert, Catherine Bodénès, Emilie Chancerel and Jérôme Durand for their assistance during various steps of the project. We thank Nicolas Langlade and two anonymous reviewers for their suggestions that greatly improved the manuscript. Part of the sampling was performed at the State Forest Nursery of Guémené-Penfao with the

assistance of Jean-Pierre Huvelin. Genotyping was performed in the Genome-Transcriptome facility of the Functional Genomic Center of Bordeaux with the help of Franck Salin and Sarah Monllor. Experiments were funded by the Research Center of Pernod Ricard (CRPR) as part of Erwan Guichoux PhD, by the EVOLTREE Network of Excellence supported by the 6th Framework Programme of the European Commission no. 016322 and by the LINKTREE project from the Eranet Biodiversa programme (ANR-08-BDVA-006).

AUTHORS' CONTRIBUTIONS

EG, LL, SW and PL performed the experiments and produced the data. EG analyzed the data and performed the simulations. EG wrote the paper with the help of RJP. All authors have checked and approved the final version of the manuscript.

SUPPORTING INFORMATION

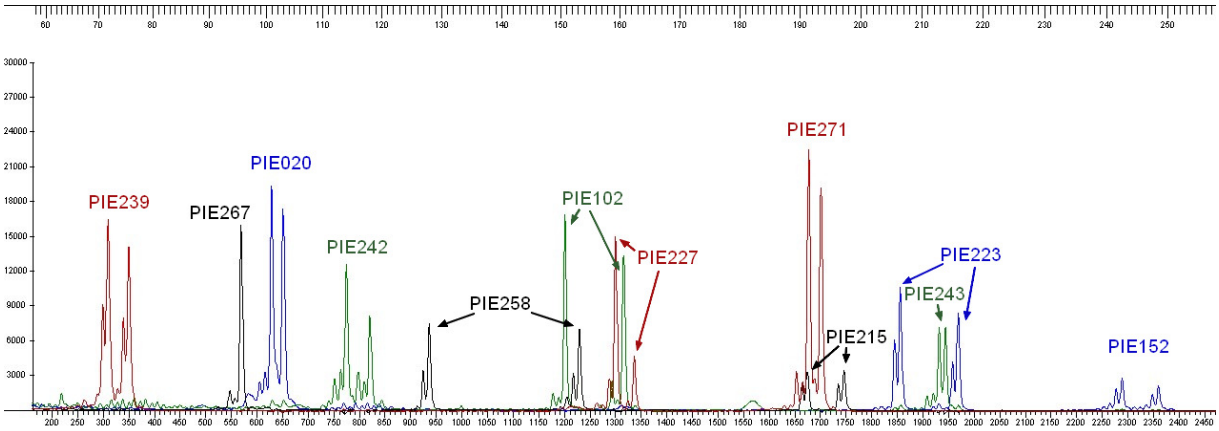


Figure S1: Multiplex profile with *kit-1*

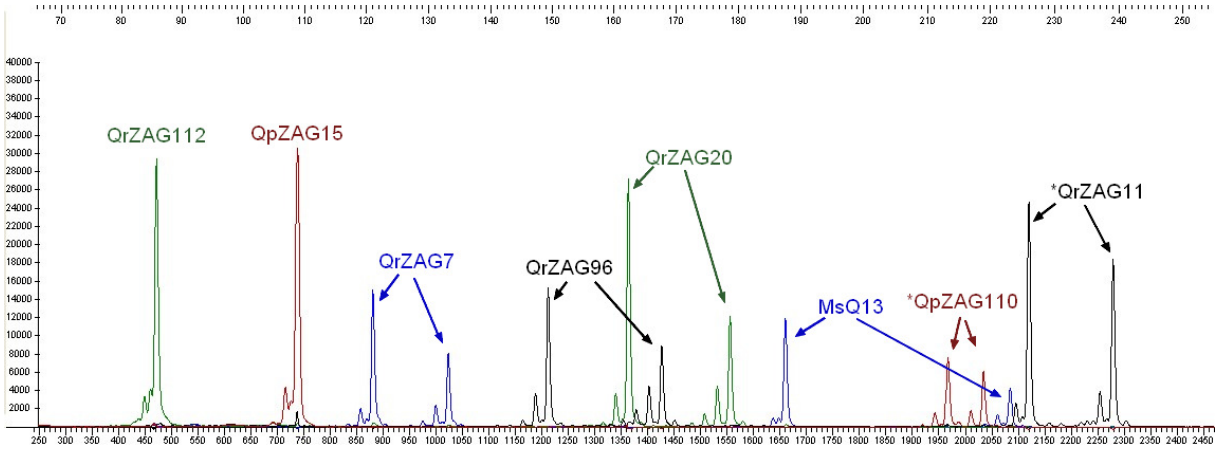


Figure S2: Multiplex profile with *kit-2*

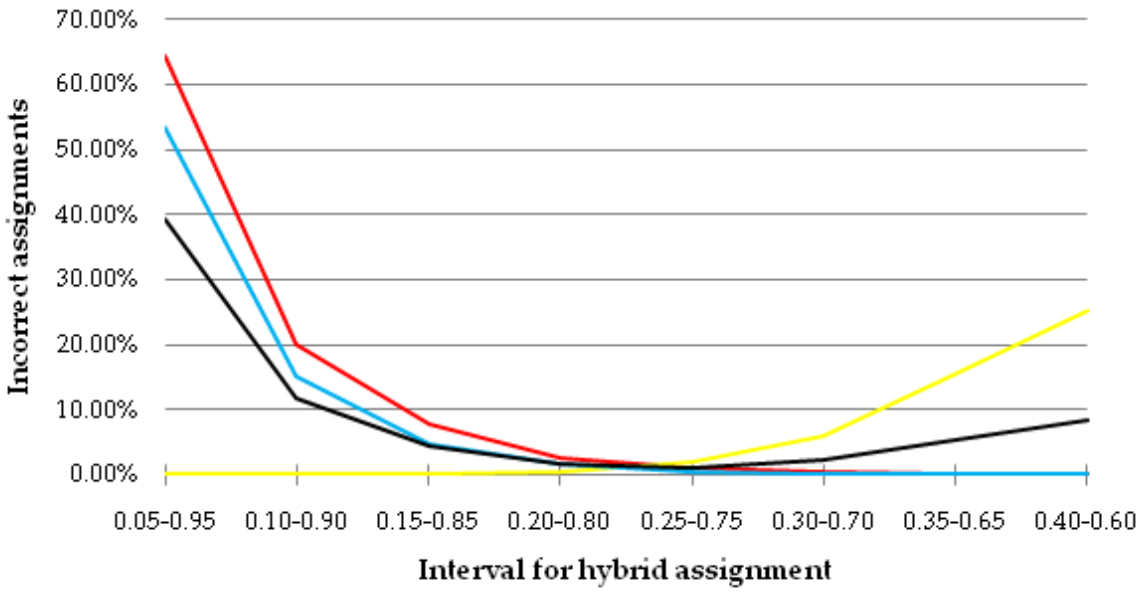


Figure S3: Incorrect assignments for simulated *Q. robur* (red), *Q. petraea* (blue) and F1 hybrids (yellow) with different intervals used for hybrid assignment

Disagreement rate %	PIE020	PIE223	PIE152	PIE242	PIE102	PIE243	PIE239	PIE227	PIE271	PIE267	PIE258	PIE215	Mean
Type A	1.02	4.29	1.02	1.22	0.61	0.82	0.20	0.82	0.41	0.20	0.82	0.00	**0.95**
Type B	0.82	0.41	0.20	0.00	0.41	0.82	0.00	0.00	0.41	0.00	0.20	0.00	**0.27**
TOTAL	1.84	4.69	1.22	1.22	1.02	1.63	0.20	0.82	0.82	0.20	1.02	0.00	**1.22**
Error rate %	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.56	0.00	1.56	0.00	**0.26**

Type A: Heterozygous genotype mistyped as homozygous
Type B: Wrong allele calling

Table S4: Disagreement rate measured on 490 samples with *kit-1*. Error rate was measured on 80 samples (6% of the complete dataset)

PIE002	PIE028	**PIE152**	PIE203	PIE217	PIE235	PIE244	**PIE258**
PIE004	PIE033	PIE193	PIE204	PIE218	PIE236	PIE247	PIE259
PIE013	PIE035	PIE194	PIE208	PIE219	PIE238	PIE249	PIE260
PIE014	PIE036	PIE196	PIE211	**PIE223**	**PIE239**	PIE250	PIE262
PIE020	PIE037	PIE197	PIE212	PIE224	PIE240	PIE252	PIE264
PIE022	PIE039	PIE198	PIE214	**PIE227**	PIE241	PIE253	PIE265
PIE023	PIE041	PIE200	**PIE215**	PIE228	**PIE242**	PIE254	**PIE267**
PIE027	**PIE102**	PIE202	PIE216	PIE233	**PIE243**	PIE257	**PIE271**

Table S5: List of 64 EST-SSRs tested to develop *kit-1*. The 12 selected loci are in red.

REFERENCES

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques* **46**, 185-192.

Amos W, Hoffman JI, Frodsham A, *et al.* (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes* **7**, 10-14.

Bacilieri R, Ducousso A, Kremer A (1995) Genetic, morphological, ecological and phenological differentiation between *Quercus petraea* (Matt) Liebl. and *Quercus robur* L. in a mixed stand of northwest of France. *Silvae Genetica* **44**, 1-10.

Bacilieri R, Labbe T, Kremer A (1994) Intraspecific genetic-structure in a mixed population of *Quercus petraea* (Matt) Liebl. and *Quercus robur* L. *Heredity* **73**, 130-141.

Bacilieri R, Roussel G, Ducousso A (1993) Hybridization and mating system in a mixed stand of sessile and pedunculate oak. *Annals of Forest Science* **50**, 122-127.

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics* **19**, 1436-1438.

Barreneche T, Casasoli M, Russell K, *et al.* (2004) Comparative mapping between *Quercus* and *Castanea* using simple-sequence repeats (SSRs). *Theoretical and Applied Genetics* **108**, 558-566.

Burgarella C, Lorenzo Z, Jabbour-Zahab R, *et al.* (2009) Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity* **102**, 442-452.

Dow BD, Ashley MV, Howe HF (1995) Characterization of highly variable (GA/CT)n microsatellites in the bur oak, *Quercus macrocarpa*. *Theoretical and Applied Genetics* **91**, 137-141.

Durand J, Bodénès C, Chancerel E, *et al.* (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* **11**, 570.

Dzialuk A, Chybicki I, Burczyk J (2005) PCR multiplexing of nuclear microsatellite loci in *Quercus* species. *Plant Molecular Biology Reporter* **23**, 121-128.

Eckert AJ, Pande B, Ersoz ES, *et al.* (2009) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). *Tree Genetics & Genomes* **5**, 225-234.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics* **9**, 80.

Hill CR, Butler JM, Vallone PM (2009) A 26plex autosomal STR assay to aid human identity testing. *Journal of Forensic Sciences* **54**, 1008-1015.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* **46**, 511-517.

Johnson PCD, Haydon DT (2007) Software for quantifying and simulating microsatellite genotyping error. *Bioinformatics and Biology Insights* **2007**, 71-75.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16**, 1099-1106.

Kampfer S, Lexer C, Glössl J, Steinkellner H (1998) Characterization of (GA)n microsatellite loci from *Quercus robur*. *Hereditas* **129**, 183-186.

Kawalko A, Dufkova P, Wojcik JM, Pialek J (2009) Polymerase chain reaction multiplexing of microsatellites and single nucleotide polymorphism markers for quantitative trait loci mapping of wild house mice. *Molecular Ecology Resources* **9**, 140-143.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* **18**, 30-38.

Kremer A, Dupouey J-L, Deans JD, *et al.* (2002) Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Annals of Forest Science* **59**, 777-787.

Lascoux M, Petit RJ (2010) The 'New Wave' in plant demographic inference: more loci and more individuals. *Molecular Ecology* **19**, 1075-1078.

Lepais O, Leger V, Gerber S (2006) Short note: High throughput microsatellite genotyping in oak species. *Silvae Genetica* **55**, 238-240.

Lepais O, Petit RJ, Guichoux E, *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology* **18**, 2228-2242.

Morin PA, Manaster C, Mesnick SL, Holland R (2009) Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with allelogram. *Molecular Ecology Resources* **9**, 1451-1455.

Morin X, Roy J, Sonie L, Chuine I (2010) Changes in leaf phenology of three European oak species in response to experimental climate change. *New Phytologist* **186**, 900-910.

Muir G, Fleming CC, Schlötterer C (2000) Taxonomy: Species status of hybridizing oaks. *Nature* **405**, 1016-1016.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**, 3599-3613.

Neophytou C, Aravanopoulos FA, Fink S, Dounavi A (2010) Detecting interspecific and geographic differentiation patterns in two interfertile oak species (*Quercus petraea* (Matt.) Liebl. and *Q. robur* L.) using small sets of microsatellite markers. *Forest Ecology and Management* **259**, 2026-2035.

Nielsen EE, Bach LA, Kotlicki P (2006) Hybridlab (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes* **6**, 971-973.

Ortego J, Bonal R (2010) Natural hybridisation between kermes (*Quercus coccifera* L.) and holm oaks (*Q. ilex* L.) revealed by microsatellite markers. *Plant Biology* **12**, 234-238.

Peakall R, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Penaloza-Ramirez JM, Gonzalez-Rodriguez A, Mendoza-Cuenca L, *et al.* (2010) Interspecific gene flow in a multispecies oak hybrid zone in the Sierra Tarahumara of Mexico. *Annals of Botany* **105**, 389-399.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164.

Petit RJ, Brewer S, Bordacs S, *et al.* (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* **156**, 49-74.

Prida A, Boulet JC, Ducousso A, Nepveu G, Puech JL (2006) Effect of species and ecological conditions on ellagitannin content in oak wood from an even-aged and mixed stand of *Quercus robur* L. and *Quercus petraea* Liebl. *Annals of Forest Science* **63**, 415-424.

Prida A, Ducousso A, Petit RJ, Nepveu G, Puech JL (2007) Variation in wood volatile compounds in a mixed oak stand: strong species and spatial differentiation in whisky-lactone content. *Annals of Forest Science* **64**, 313-320.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Santana QC, Coetzee MPA, Steenkamp ET, *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques* **46**, 217-223.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233-234.

Scotti-Saintagne C, Mariette S, Porth I, *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea (*Matt.) Liebl.]. *Genetics* **168**, 1615-1626.

Steinkellner H, Fluch S, Turetschek E, *et al.* (1997) Identification and characterization of (GA/CT)n microsatellite loci from *Quercus petraea*. *Plant Molecular Biology* **33**, 1093-1096.

Streiff R, Ducousso A, Lexer C, *et al.* (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Molecular Ecology* **8**, 831-841.

Streiff R, Labbe T, Bacilieri R, *et al.* (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* **7**, 317-328.

Untergasser A, Nijveen H, Rao X, *et al.* (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research* **35**, W71-W74.

Vähä JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology* **15**, 63-72.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23**, 48-55.

Viscosi V, Lepais O, Gerber S, Fortini P (2009) Leaf morphological analyses in four European oak species (*Quercus*) and their hybrids: A comparison of traditional and geometric morphometric methods. *Plant Biosystems* **143**, 564-574.

CHAPITRE 3

DNA-based identification of tree species from wood: application to oak staves

E. Guichoux[1,2,3] and R.J. Petit[1,2]

[1] INRA, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[2] University of Bordeaux, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[3] Centre de Recherche Pernod Ricard, F-94000 Créteil, France

INTRODUCTION

The sensory characteristics of wines are significantly improved during aging in oak barrels (Boidron *et al.*, 1988). When aging in oak barrels, wine undergoes a series of processes that cause important improvements in wine aroma, taste, color, and astringency (Jarauta *et al.*, 2005). Since the Gallo-Roman period, wine maturation has been performed in oak barrels. Recently, alternative methods started to emerge, using either oak planks, staves, chips, cubes, powder or shavings, which can be added to wine held in tanks made of any inert material (Del Alamo Sanza *et al.*, 2004; Young *et al.*, 2010). These techniques allow a better control of wine aging and reduce production cost (Spillman, 1999). In comparison with oak barrels, wood quantities that are necessary for wine maturation are low (only a few grams of oak chips per liter). In this context, precise characterization of wood aromatic properties appears to be particularly relevant. But anticipating the aromatic contribution of a barrel or of some oak chips is difficult in practice.

The majority of winemakers insist on using French oak (*Quercus robur* and *Q. petraea*) for its typical aromatic contribution (Boidron *et al.*, 1988). However, a growing minority uses American white oak (*Q. alba*), for the crafting of wine barrels. Still today, a noticeable proportion of oak woods are primarily selected based on their grain size and geographic origin (Mosedale & Ford, 1996; Feuillat *et al.*, 1998; Ancin *et al.*, 2004; Spillman *et al.*, 2004). Yet, several studies have shown that the major effect explaining aromatic differences among wood batches depends on the botanical species (Doussot *et al.*, 2000; Doussot *et al.*, 2002; Guchu *et al.*, 2006; Prida & Puech, 2006; Prida *et al.*, 2007). The wood of the two French oak species has contrasted aromatic patterns (Feuillat *et al.*, 1997b; Doussot *et al.*, 2000), especially for whisky-lactone. This molecule (β-methyl-γ-octalactone), also known as quercus lactone, provides typical aromas of coconut, resin and celery (Boidron *et al.*, 1988). If *Q. robur* wood has only traces of whisky-lactone, *Q. petraea* wood often present high amount of this compound (Prida *et al.*, 2007). Hence, the ability to sort wood-lots according to the species will allow better anticipation of the aromatic contribution of wood, at least for one of the most aromatic molecule described to date.

Species identification based on wood anatomy is complex and sometimes impossible in the case of taxonomically related species that have similar wood structure (Feuillat *et al.*, 1997a;

Deguilloux *et al.*, 2002; Lens *et al.*, 2005). Chemical analyses can be effective at differentiating species but remain expensive and often suffer from the high level of variability among populations or samples (Deguilloux *et al.*, 2002; Gougeon *et al.*, 2009). Tree species can also be differentiated by near-infrared reflectance spectroscopy (Atkinson *et al.*, 1997; Humphreys *et al.*, 2008) but this technique requires samples with similar preparation conditions and is highly sensitive to ascertainment bias (Russ *et al.*, 2009). Thus, wood identification methods relying on molecular markers have recently been investigated (Eurlings *et al.*, 2010; Finkeldey *et al.*, 2010).

Since two decades, genetic analyses on fresh tissues such as leaves, bud or cambium are commonplace and technological improvements allow the retrieval of DNA from most tree species (Doyle & Doyle, 1990; Lin & Walker, 1997; Csaikl *et al.*, 1998). But the retrieval of DNA from wood itself is more complex because wood properties differ from fresh tissues (Rachmayanti *et al.*, 2009). Once a tree is cut, the quantity and quality of DNA that can be retrieved quickly diminish and the size of fragments that can be amplified decreases after the death of the wood tissue (Bär *et al.*, 1988; Lindahl, 1993; Cano, 1996; Deguilloux *et al.*, 2002). Furthermore, some species, like oaks, have high levels of polyphenols and ellagitannins, which are known to inhibit PCR (Cooper & Poinar, 2000). Amounts of water-soluble ellagitannins increase during wood drying (Mosedale *et al.*, 1998), which could be problematic for genetic analyses targeting aged wood-lots.

Most genetic studies on dry wood have targeted organelle genomes, especially chloroplast DNA (Dumolin-Lapègue *et al.*, 1999; Deguilloux *et al.*, 2002; Deguilloux *et al.*, 2004; Asif & Cannon, 2005; Deguilloux *et al.*, 2006; Elbaum *et al.*, 2006; Rachmayanti *et al.*, 2006; Rachmayanti *et al.*, 2009). In some cases, organelle markers (mitochondrial [mt] or chloroplastic [cp]) are sufficient to identify geographic origins (Deguilloux *et al.*, 2003; Lowe, 2008; Eurlings *et al.*, 2010) or species (Shaw *et al.*, 2005; Group *et al.*, 2009; Duminil *et al.*, 2010). Organelle genomes have two major technical advantages over nuclear genome. The number of organelle genome copies is over 100 times higher than the number of nuclear genome copies (Bendich, 1987), considerably facilitating the amplification of the corresponding DNA sequences (Soltis *et al.*, 1992). They seem also to be more stable following tissue death (Schwarz *et al.*, 2009). However, organelle markers, which are maternally inherited in angiosperms, are more frequently introgressed and hence generally less powerful than

nuclear markers to delimitate species (Currat *et al.*, 2008; Du *et al.*, 2009; Petit & Excoffier, 2009). As a given nuclear DNA sequence is typically represented by only two copies per cell, instead of hundreds or thousands, severe improvements in DNA isolation and amplification protocols are necessary when targeting nuclear markers compared to organelle markers (Rogers & Kaya, 2006; Novaes *et al.*, 2009).

In this study, we illustrate wood-based DNA identification approach using as case study two interfertile oak species, *Q. robur* and *Q. petraea*. Besides their extensive use for wine maturation, these two species have become models in tree population genetic and speciation studies (Streiff *et al.*, 1998; Streiff *et al.*, 1999; Muir *et al.*, 2000; Petit *et al.*, 2002; Petit *et al.*, 2004; Scotti-Saintagne *et al.*, 2004; Prida *et al.*, 2006; Lepais *et al.*, 2009; Lepais & Gerber, 2010). First successful DNA isolation from dry wood was notably achieved on these species in 1999 (Dumolin-Lapègue *et al.*, 1999). However, differentiating these two species using organelle markers proved impossible due to extensive introgression (Muir *et al.*, 2001; Petit *et al.*, 2002; Lepais & Gerber, 2010). In contrast, nuclear markers have been efficiently used for differentiating the species using DNA isolated from fresh material (Bodénès *et al.*, 1997; Samuel, 1999; Bakker *et al.*, 2001; Gömöry *et al.*, 2001; Muir *et al.*, 2001; Coart *et al.*, 2002; Mariette *et al.*, 2002; Scotti-Saintagne *et al.*, 2004; Lepais *et al.*, 2006; Guichoux *et al.*, 2011). The ability to identify oak species from dry wood would have major consequences in various fields. In the first place, it might contribute to improve the control of maturation processes of wine by selecting woods based on their aromatic contribution. But it could also allow *a posteriori* certification of wood-lots when no traceability piece of evidence is present, contributing to combat fraud (Degen & Fladung, 2008; Finkeldey *et al.*, 2008; Tacconi, 2008).

We therefore developed methods of DNA-based identification of oak species from dry wood samples. We paid particular attention to DNA isolation and purification protocols and evaluated benefits of each protocol change with real-time PCR technique. We also adapted published genetic markers to maximize amplification success on dry wood and accurately identify the two species using assignment analysis.

MATERIAL AND METHODS

Plant material

Seven pieces of internal wood (100mm x 50mm x 30mm) were sampled after 18 months of seasoning (wood drying) in barrel staves. Species status (*Q. robur* or *Q. petraea*) had been previously provided by the cooper. Prior to experiments, all samples were sanded on each side and were cleaned with 10% bleach to remove contaminants. Forty-eight fresh leave samples from both species were also used to allow integration of genotypes acquired with specifically developed primers adapted to degraded DNA.

DNA isolation

DNA isolation from dry wood samples is critical, and any contamination from fresh DNA coming from classic lab must be avoided. As a consequence, all DNA isolation steps were carried out in a separate dedicated lab under high-pressure, with severe experimental precautions, following Deguilloux *et al.* (2006). In particular, all surfaces were bleached before experiments and UV light irradiation was performed every day for one hour. Reagents used were never opened before entering the dedicated lab. DNA was isolated from no more than six samples per day and two negative DNA isolation controls were used per experiment. Negative controls were treated in the same way as wood samples. For each sample, DNA was isolated twice in separate experiments and each duplicate was used for further DNA amplification to test the reproducibility of the results. Prior to DNA isolation, external surface of wood was removed over 2mm with a scalpel to avoid external contamination. About 50mg of wood shavings obtained with a scalpel were added into 2ml tubes with two 3mm tungsten beads. Tubes were frozen into liquid nitrogen for 2min before disruption in fine powder using a Mixer Mill MM300 (Retsch, Germany), for 4min at 30Hz. Four DNA isolation protocols were tested: DNeasy Plant Mini Kit (Qiagen, Germany), Nucleospin Plant II (Macherey-Nagel, Germany), Invisorb Spin Plant Mini Kit (Invitek, Germany) and a modified CTAB protocol (Doyle & Doyle, 1990). DNA isolation with commercial kits was achieved following manufacturer instructions except for the lysis step (one hour at 65°C under agitation), additional washings until the elute was clear and a final elution volume of 100µl. CTAB protocol was adapted to degraded DNA. In particular, all reagent volumes were increased to prevent excessive absorption by wood powder.

Proteinase K, Polyvinylpolypyrrolidon (PVPP) and β-mercaptoethanol were added in the lysis buffer (see detailed protocol in Table 1). Additional washings were also performed until elute was clear (three to five washings). Two purification protocols were tested in combination with DNA isolation protocols: OneStep PCR Inhibitor Removal Kit (Zymo Research) and High Pure PCR Product Purification Kit (Roche).

Lysis Buffer: Cetyl TrimethylAmmonium Bromide or CTAB (2%), EthyleneDiamineTetraacetic Acid or EDTA ph8 (0.02M), Tris-HCl ph8 (0.1M), NaCl (1.4M), PVPP (1%), β-mercaptoethanol (0.2%), Proteinase K (0.5mg/ml), distilled water.

Lysis

1- Add 900µl of warm lysis buffer to 50mg of wood powder in a 2ml tube.
2- ortex vigorously and incubate for one hour at 65°C under agitation.

Deproteinization

3- Add 720µl (or 4/5th of buffer volume) of Chloroform/Isoamyl Alcohol 24:1 and vortex vigorously.
4- Centrifuge for 10 min at 13.000rpm.
5- Carefully remove upper phase (do not remove interphase) and transfer to a new 2ml tube.
6- Repeat steps 3 to 5 until no interphase is visible (two to four times).
7- Transfer the clean upper phase in a new 1.5ml tube.

Precipitation

8- Add the same volume of cold isopropanol as upper phase volume retrieved in step 7.
9- Gently shake the tube and store at -20°C for at least one hour.
10- Centrifuge for 10 min at 13.000rpm.
11- Discard the liquid phase.
12- Add 800µl of 70% ethanol and vortex slowly.
13- Centrifuge for 10 min at 13.000rpm.
14- Repeat steps 11 to 13.
15- Add 500µl of absolute ethanol and vortex slowly.
16- Discard the liquid phase and allow complete drying of DNA pellet.

Resuspension

17- Add 100µl of ultrapure water.
18- Incubate at room temperature under agitation for 30 min.
19- Store at 4°C for immediate use or at -20°C for postponed use.

Table 1: modified CTAB protocol for DNA isolation of oak wood samples. Final concentration of each product in lysis buffer is indicated in brackets.

Quantification of dsDNA

DNA was quantified in the classic lab on an Infinate 200 microplate reader (Tecan) using the Quant-iT PicoGreen dsDNA Kit (Invitrogen), which can theoretically detect as little as 25pg/mL of double-stranded DNA (dsDNA). We used an appropriate standard curve with increasing DNA quantities (100pg, 500pg, 1000pg, 2500pg and 5000pg) for more accurate interpretations.

Real-time PCR to optimize DNA isolation

Real-time PCR was achieved with the iQ SYBR Green Supermix (Bio-Rad, USA) in a final volume of 25µl, with UltraPureGold purified primers (Eurogentec, Belgium). PCR mix was composed of 10µl of iQ SYBR Green Supermix (final concentration 1X), 0.3µl of each primer pair (final concentration 120nM), 6.9µl of ultrapure water and 2.5µL of purified DNA diluted ten times. We amplified a chloroplast fragment of 53bp (dt13 from Demesure *et al.* (1995) and a nuclear fragment of 83bp (a-PIE258 derived from Guichoux *et al.* (2011), see Table 2). The cycling conditions were: an initial step at 95°C for 3 min; followed by 55 cycles at 95°C for 45 s and 50°C for 45 s. Fluorescence was measured with a Chromo4 Real-Time PCR Detection System (Bio-Rad, USA) at the end of the annealing period of each cycle to monitor the progress of amplification. After completion, a melting curve was obtained by heating slowly at 1°C/s from 65°C to 95°C with fluorescence acquisition every 1°C. Standard curve was realized from fresh purified DNA previously quantified with the Quant-iT PicoGreen dsDNA Kit. We adapted the standard curve to very low DNA quantities and used increasing quantities of fresh DNA (0.0025pg, 0.025pg, 0.25pg, 2.5pg, 25pg). We used triplicates for each point (standard curve points or wood samples) to evaluate the reproducibility of the results. Absolute quantification of DNA was achieved with Opticon Monitor software (Bio-Rad, USA), by relating the PCR signal to the standard curve. Mean DNA quantity was calculated on triplicates of each sample with a specific protocol combination (DNA isolation and DNA purification). Results were used to compare the efficiency of the different protocols.

Locus	Primer Sequences (5′ – 3′)	Motif	Annealing temperature	Size (bp)	Size shift (bp)
a-PIE215	TGATCATGGCAGAAGAGAAGG TGGCAGGACTCGTGAACC	GAG	56°C	78	125
PIE239	TCAACAAATGGCTCAACAGTG CCCATTTGGTAGCAAAGAGTC	AT	56°C	71	0
a-PIE242	TTGCAATCCTCCAAATTTAATG CAAGGATTAAGATTCAAGATTGTGT	TA	56°C	80	32
a-PIE243	AATCAAATGTCAATTAGAAAGAAAAAG GGCAATGCCTCATCTCTCAC	CT	55°C	99	120
a-PIE258	ACCAAACCAAAACCGAAACC GAGCAAACACAGTTTGGGGTA	CT	56°C	83	66
a-QrZAG7	CGGATTTCGAGACCAGGTTA TGCATTTCTTTTATAGCATTCA	TC	52°C	105	13
a-QrZAG112	GGTGCGCGGGAGAGAAAA GAGACTCGGTAAGTATTCTTATTT	GA	49°C	74	10

Table 2: Characteristics of seven nuclear microsatellites used for species identification of oak wood samples. Amplicon size shifts are estimated between redesigned primers and original primers for the same allele. "a-": redesigned

SSR genotyping

To enhance SSR genotyping success on wood samples, we redesigned primer pairs published by Guichoux *et al.* (2011), except PIE239 (already "degraded-DNA" compatible). Severe criteria were followed: amplicon size below 100bp, primer melting temperature over 55°C and GC-clamp (which promotes specific binding at the 3'-end due to the stronger bonding of G and C bases). Redesigned primers were all labeled with an « a-» prefix. PCR amplifications were performed on a DNA Engine Tetrad 2 Thermal Cycler (Bio-Rad, USA), in a 25µL reaction volume containing 2.5µl Gold Buffer 10X (Applied Biosystems, USA), 3.5 mM M_gCl_2, 0.4 mM dNTP, 0.66 mg/ml BSA, 1 mM each primer, 1.25U AmpliTaq Gold DNA Polymerase and 2.5µl DNA sample diluted ten times. Cycling conditions were: 5 min at 94°C followed by 55 cycles at 94°C for 45 s, 58°C for 45 s and 72°C for 45 s; and a final incubation at 72°C for 10 min. PCR products were separated on 3% agarose gel stained with GelRed (Biotium, USA), diluted 25 times in ultrapure water and run on ABI-3730 (Applied Biosystems, USA), with LIZ600 as internal lane size standard. Size fragment analysis was performed with Genemapper (Applied Biosystems, USA).

PCR inhibitor test

We performed inhibitory tests (Rachmayanti *et al.*, 2006) during amplification of the cpDNA fragment dt13 (Demesure *et al.*, 1995). We used DNA extract from fresh material combined with an increasing proportion of DNA extract from wood (0%, 10%, 25%, 50%, 75% and 90%). Amplification success was estimated by real-time PCR.

Species identification of wood samples

We used assignment methods based on Bayesian clustering approaches to confirm the species of wood samples. Admixture proportion of each sample was estimated using STRUCTURE v.2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2003), with a burn-in of 50,000 steps followed by 50,000 Markov chain Monte Carlo repetitions. We calculated the average result over 10 runs with *K* (number of clusters) set to two, corresponding to the two species. Samples were classified as purebreds if their admixture proportion was over 0.875 for one of the two clusters (Guichoux, 2011b, submitted). All fresh samples (24 *Q. robur* and 24 *Q. petraea*) have been previously assigned to one species using 20 SSRs (Guichoux *et al.*, 2011) and 273 samples from both species. Here, we repeated this analysis with seven loci (PIE215, PIE239, PIE242, PIE243, PIE258, QrZAG7 and QrZAG112) to test the assignment stability of the 48 fresh samples using a subset of markers. To allow comparison of genotypes obtained from wood samples generated with the redesigned primer, we genotyped the 48 fresh samples with the redesigned primers and estimated size shifts for each locus. This way, we could integrate the genotypes generated with the redesigned primers in the same original genetic database used for assignment analysis (Figure 1).

RESULTS AND DISCUSSION

Quantification of dsDNA

Final concentration of DNA extracts from wood ranged from 20pg/µl to 157pg/µL, with a high heterogeneity between replicates of the same sample (two different DNA isolation experiments of the same wood sample with the same protocol). PCR amplification control with cpDNA fragment dt13 was positive on agarose gel, regardless of the DNA isolation protocol used. As a consequence, we could not discriminate among these protocols with this approach. DNA quantification techniques relying on intercalating agent (PicoGreen,

ethidium bromide or SYBR Green I) are very sensitive but quantify all dsDNA isolated from the wood sample, not only oak DNA. Moreover, even very small fragments of degraded DNA not amenable to amplification using our primers are quantified. This will lead to an over-estimation of available DNA quantity, making this information difficult to use for the optimization of DNA isolation protocols.

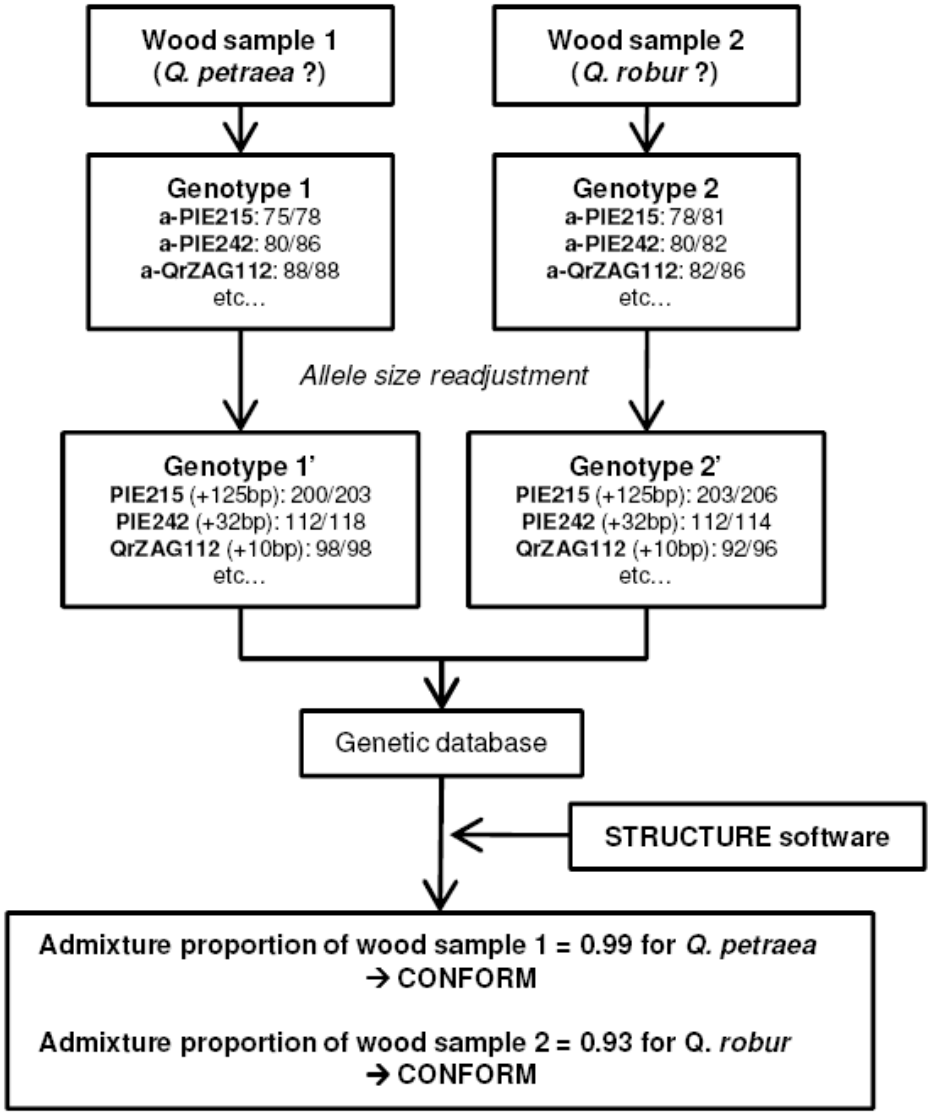


Figure 1: Species conformity methodology with the example of two samples from both species genotyped at seven nuclear SSRs.

Real-time PCR to optimize DNA isolation

Negative controls (both in DNA isolation and real-time PCR amplification) remained negatives, confirming the absence of contamination despite the high number of PCR cycles used (55). DNA isolation protocols based on Nucleospin Plant or Invisorb Spin Plant Mini Kit were successful on a very low number of samples, whatever the purification protocol used. Thus, they were excluded upstream. Real-time PCR results with nuclear microsatellite a-PIE258 gave unexploitable results. PCR efficiency was low and heterogeneous and melting curve analysis revealed the amplification of unspecific products. DNA quantification was also highly heterogeneous between triplicates (data not shown). As a consequence, DNA quantification was not performed for PCR amplification of nuclear fragments. Real-time PCR results on cpDNA fragment dt13 showed high repeatability between triplicates and high PCR efficiency (>85%). For the same sample, mean cpDNA quantities were compared between different combination of DNA isolation and purification protocols to detect significant improvements (Figure 2). The two DNA isolation protocols (DNeasy Plant Mini Kit and modified CTAB) gave similar results with dt13 (Figure 3).

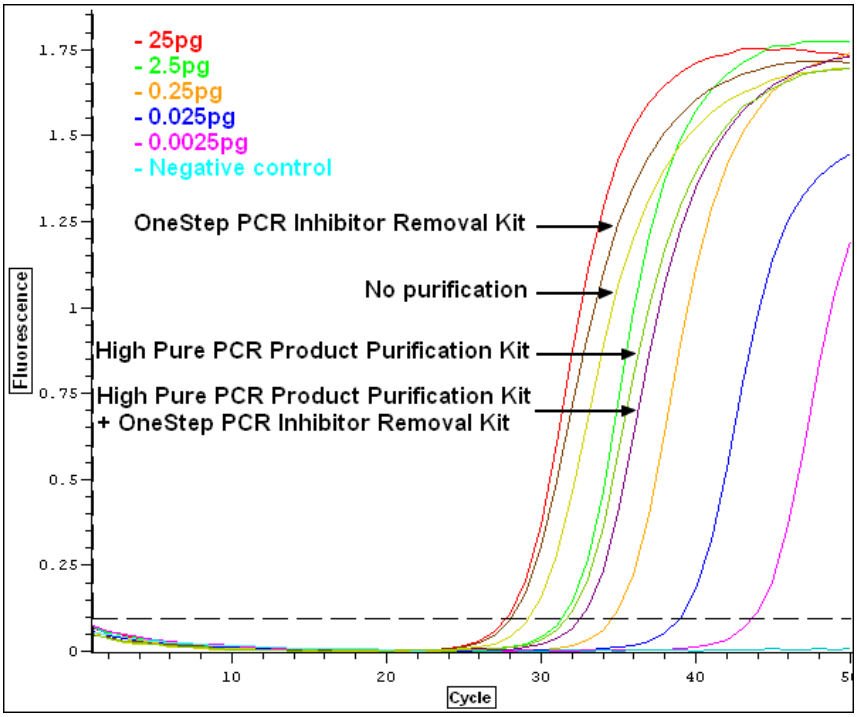


Figure 2: Fluorescence intensity kinetic during 50 cycles of real-time PCR on chloroplastic fragment dt13. DNA of the same *Q. petraea* sample (PR1) has been isolated with modified CTAB protocol and the purification treatments are indicated by black arrows. Only one kinetic per triplicate is showed to improve visualization.

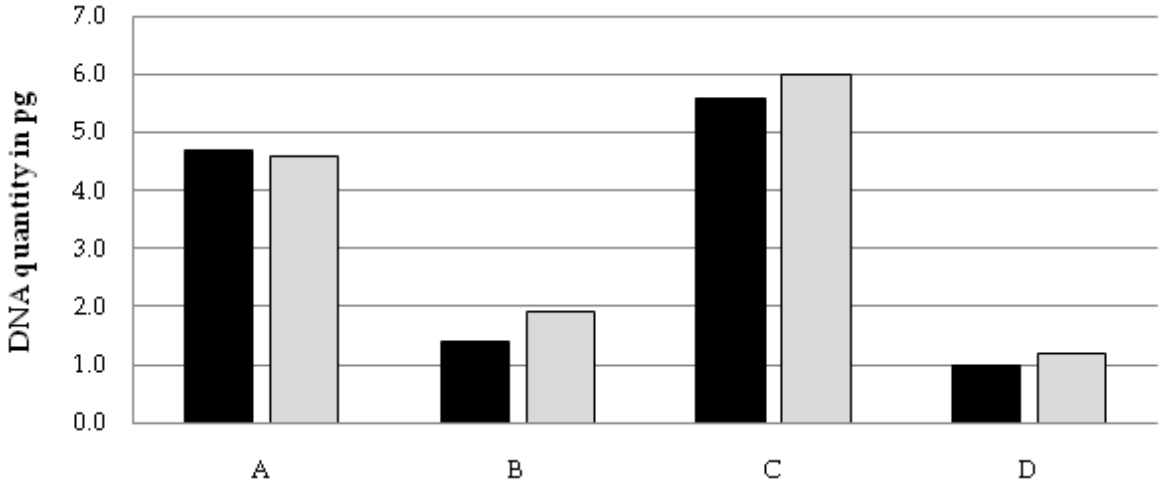


Figure 3: Mean quantities of cpDNA (in pg) estimated on seven wood extracts using different protocols for DNA isolation and purification. cpDNA quantities were estimated by real-time PCR amplification on chloroplastic fragment dt13 (53bp). Black: DNeasy Plant Mini Kit (Qiagen). Grey: modified CTAB protocol. A: no purification. B: High Pure PCR Product Purification Kit (Roche). C: OneStep PCR Inhibitor Removal Kit (Zymo Research). D: both purification kits combined. Each wood extract was independently analyzed two times with one specific protocol combination.

Interestingly, efficiency was slightly better for all *Q. petraea* samples (data not shown). This may be due to differences in inhibitory content between the two species, *Q. robur* having higher levels of ellagitannins (Feuillat *et al.*, 1997b). On the other hand, purification protocols gave different results: High Pure PCR Product Purification Kit led to a significant decrease of mean DNA quantity (-65%) whereas OneStep PCR Inhibitor Removal Kit improved the DNA quantity (+20%). The combination of both purification protocols led to a stronger decrease (-77%). Purification kits relying on column matrices eliminate small fragments (generally <100pb) and as a consequence final DNA concentration is lower. The fact that OneStep PCR Inhibitor Removal Kit improved the quantity of usable DNA for PCR suggests that inhibitors that limit DNA amplification were successfully removed. Purification with OneStep PCR Inhibitor Removal Kit offset the loss of some DNA molecules inherent in purification on columns.

Real-time PCR results also confirmed that the addition of β-mercaptoethanol and PVPP in the lysis buffer, as proposed by Rachmayanti *et al.* (2006) or Mitchell *et al.* (2005) was not sufficient to yield conclusive results on our samples when amplifying nuclear loci, and that DNA purification is crucial for this purpose. But our results also confirmed that DNA purification kits, which eliminate short fragments, are unsuitable for degraded DNA

whenever short amplicons (between 60 and 100pb) are targeted. On the other hand, DNA purification is crucial for amplification of low copy number DNA. Hence, purification methods such as OneStep PCR Inhibitor Removal Kit, which limit DNA fragment loss while removing efficiently PCR inhibitors, are particularly well-suited to degraded DNA.

SSR genotyping

Even if the two DNA isolation protocols gave similar results with real-time PCR using cpDNA primers, several loci did not amplify in classic PCR with the DNeasy Plant Mini Kit. With modified CTAB protocol combined with OneStep PCR Inhibitor Removal Kit, all loci successfully amplified. Hence, real-time PCR results on cpDNA fragments do not give perfect predictions for nuclear fragments but they are nevertheless relevant indicators: DNA isolation protocols that perform poorly on cpDNA fragments will most likely be inefficient on nuclear DNA fragments. Similarly, the most efficient DNA isolation protocol on cpDNA fragments will most likely be also the most efficient on nuclear DNA fragments. Despite the use of optimized protocols (modified CTAB in combination with OneStep PCR Inhibitor Removal Kit), preliminary tests on wood samples with original primers were inconclusive, confirming the necessity to design dedicated primers for amplification of degraded DNA. With an appropriate primer redesign, we successfully genotyped all wood samples at seven nuclear microsatellites except one locus (a-PIE215) that failed to amplify on all *Q. robur* samples. Although one locus had very low melting temperature (49°C for a-QrZAG112), missing data ratio was low (on average <15%, Table 3). We noticed discrepancies between some duplicates, resulting most of the time in the absence of one allele, a phenomenon called allelic dropout that has been described previously (Soulsbury *et al.*, 2007; Tvedebrink *et al.*, 2009). We also detected sporadic profiles with unexpected peaks (artifacts or extra allele). By repeating these ambiguous profiles three to five times, a consensus genotype could be retrieved in most cases.

PCR inhibitor test

With the optimal combination of protocols (modified CTAB and OneStep PCR Inhibitor Removal Kit), mean apparent quantity of DNA (as estimated through real-time PCR) only decreased by 5% between no DNA extract from wood and 90% of DNA extract from wood in the total DNA mixture. Even with a very high proportion of DNA extract from wood (90%),

PCR efficiency remained high (85%) and amplification was complete. These results underline the benefit of dedicated methods designed to remove inhibitors such as polyphenolic compounds or ellagitannins found in wood. In all cases, real-time PCR approaches can supplant classic inhibitor tests by directly quantifying the impact of inhibitors on PCR efficiency.

Species identification of wood samples

Using seven nuclear microsatellites and a threshold value of 0.875 for the targeted purebred cluster, admixture proportion of the 273 samples from the reference genetic database was still accurate. In particular, all 48 pure samples used for allele size synchronization between datasets remained strongly assigned to one species (>0.875 for the targeted cluster). Genotyping of the 48 fresh samples with redesigned primers at seven nuclear microsatellites allowed the determination of size shift between original and redesigned primers (Table 2). Thus, multilocus genotypes of wood samples were transformed to allow their integration into the reference genetic database (Figure 1).

Assignment results for the seven wood samples were consistent with the species announced by the cooper (Table 3). In all cases, assignment scores were high (>0.875) except for sample PR7 (assignment value of 0.850), maybe due to missing data at one locus (re-QrZAG112) that has high species discriminatory power (Scotti-Saintagne *et al.*, 2004).

Sample	Species	Missing data (%)	Assignment values		Conformity
			Q. petraea cluster	*Q. robur* cluster	
PR1	*Q. petraea*	0	**0.995**	0.005	+++
PR2	*Q. petraea*	29	**0.993**	0.007	+++
PR3	*Q. petraea*	14	**0.986**	0.014	+++
PR4	*Q. petraea*	0	**0.995**	0.005	+++
PR5	*Q. robur*	14	0.070	**0.930**	+++
PR6	*Q. robur*	14	0.010	**0.990**	+++
PR7	*Q. robur*	29	0.146	**0.855**	++

+++: highly conform (>0.875)
++: conform (>0.75)

Table 3: Genetic analysis of 18 months-old wood samples (four *Q. petraea* and three *Q. robur*) with seven nuclear microsatellites. Assignment values are calculated over ten runs of STRUCTURE.

CONCLUSION AND PERSPECTIVES

In this study, we proved that DNA-based certification of oak species from dry wood samples is possible provided some precautions are taken. First, contamination must be avoided, so all experiments prior to DNA amplification have to be done in a dedicated clean lab. Second, primers must be chosen on their ability to amplify short fragments, given that wood contains only largely degraded DNA. Third, we showed that DNA isolation protocols as well as purification methods can to some extent be optimized using real-time PCR analysis of cpDNA fragments. This approach can indirectly help with the improvement of nuclear DNA amplification techniques. DNA isolation protocols that perform poorly in real-time PCR on cpDNA fragments should be excluded upstream. Fourth, reference genetic database used for assignment analysis should be as complete as possible to allow accurate estimation of admixture level of wood samples when they will be integrated. Following this methodology, we successfully confirmed the species of wood samples dried for 18 months.

However, some limits were identified in this study, and these will have to be taken into account in the future. Identification of optimal DNA isolation protocols through real-time PCR only allowed the detection of major differences (King *et al.*, 2009; Schwarz *et al.*, 2009). Heterogeneity detected between replicates highlighted the limit of this approach and only relative DNA quantification of wood extracts could be achieved. Whereas real-time PCR triplicates proved highly repeatable, DNA isolation duplicates showed large heterogeneity. Despite precautions, standardization of wood sample preparation remains difficult and could lead to different amplification success rates. Thus, high number of replicates (same samples extracted many times) should be used to make the most of real-time PCR approach for protocol optimization.

Microsatellites genotyping on degraded DNA underlined the limits of multiallelic markers for species identification on wood samples. Despite optimized DNA isolation and purification protocols and the use of efficient primers, we repeatedly found ambiguous profiles (with triple bands or showing allelic dropout), which obliged us to perform up to five repetitions to obtain reliable genotypes. Assignment results may have been different if such precautions had not been taken, potentially compromising assay quality.

Given the technical limits for isolation of longer DNA fragments inherent to dry wood, di-allelic markers such as Single Nucleotide Polymorphisms (SNPs) appear promising (Asari *et al.*, 2009; Ogden *et al.*, 2009). Genotyping errors are more limited with only two alleles and amplicons can be further shortened, down to a minimum of 45-50bp. Hence, amplification success should be higher than with microsatellites. On the other hand, assignment methods with SNPs will require more markers (Glover *et al.*, 2010; Haasl & Payseur, 2011) unless more differentiated loci can be identified. The generalization of next-generation sequencing on non-model species will soon allow the detection of such informative markers. In view of our efforts to optimize DNA amplification and of the many SNP genotyping techniques currently available (High-Resolution Melting analysis, allele specific PCR, Derived Cleaved Amplified Polymorphic Sequences), we are optimistic that accurate DNA-based identification of species from dry wood samples might be feasible at acceptable cost in the near future.

ACKNOWLEDGMENTS

We thank Frédéric Lagane for preparing wood samples and Grégoire Le Provost for help with real-time PCR development on dry wood. Wood samples were provided by the Research Center of Pernod Ricard with the contribution of Petitrenaud sawmill (Dirol, France). Genotyping was performed in the Genome-Transcriptome facility of the Functional Genomic Center of Bordeaux. Experiments were funded by the Research Center of Pernod Ricard (CRPR) as part of Erwan Guichoux PhD, by the LINKTREE project from the Eranet Biodiversa program (ANR-08-BDVA-006) and by EVOLTREE Network of Excellence supported by the 6th Framework Programme of the European Commission no. 016322.

REFERENCES

Ancin C, Garde T, Torrea D, Jimenez N (2004) Extraction of volatile compounds in model wine from different oak woods: effect of SO2. *Food Research International* 37, 375-383.

Asari M, Watanabe S, Matsubara K, Shiono H, Shimizu K (2009) Single nucleotide polymorphism genotyping by mini-primer allele-specific amplification with universal reporter primers for identification of degraded DNA. *Analytical Biochemistry* 386, 85-90.

Asif MJ, Cannon CH (2005) DNA extraction from processed wood: A case study for the identification of an endangered timber species (*Gonystylus bancanus*). *Plant Molecular Biology Reporter* 23, 185-192.

Atkinson MD, Jervis AP, Sangha RS (1997) Discrimination between *Betula pendula*, *Betula pubescens*, and their hybrids using near-infrared reflectance spectroscopy. *Canadian Journal of Forest Research* 27, 1896-1900.

Bakker EG, Van Dam BC, Van Eck HJ, Jacobsen E (2001) The description of clones of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. with microsatellites and AFLP in an ancient woodland. *Plant Biology* 3, 616-621.

Bär W, Kratzer A, Mächler M, Schmid W (1988) Postmortem stability of DNA. *Forensic Science International* 39, 59-70.

Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *Bioessays* 6, 279-282.

Bodénès C, Joandet S, Laigret F, Kremer A (1997) Detection of genomic regions differentiating two closely related oak species *Quercus petraea* (Matt) Liebl and *Quercus robur* L. *Heredity* 78, 433-444.

Boidron JN, Chatonnet P, Pons M (1988) Influence du bois sur certaines substances odorantes des vins. *Connaiss. Vigne Vin* 22, 275-294.

Cano RJ (1996) Analysing ancient DNA. *Endeavour* 20, 162-167.

Coart E, Lamote V, De Loose M, *et al.* (2002) AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl] in Flemish populations. *Theoretical and Applied Genetics* 105, 431-439.

Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at all. *Science* 289, 1139.

Csaikl UM, Bastian H, Brettschneider R, *et al.* (1998) Comparative analysis of different DNA extraction protocols: A fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. *Plant Molecular Biology Reporter* 16, 69-86.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: massive introgression by local genes. *Evolution* 62, 1908-1920.

Degen B, Fladung M (2008) Use of DNA-markers for tracing illegal logging. In: *Proceedings of the International Workshop Fingerprinting Methods for the Identification of Timber Origins* (ed. Degen B), pp. 6-14.

Deguilloux MF, Bertel L, Celant A, *et al.* (2006) Genetic analysis of archaeological wood remains: first results and prospects. *Journal of Archaeological Science* 33, 1216-1227.

Deguilloux MF, Pemonge MH, Bertel L, Kremer A, Petit RJ (2003) Checking the geographical origin of oak wood: molecular and statistical tools. *Molecular Ecology* 12, 1629-1636.

Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proceedings of the Royal Society B-Biological Sciences* 269, 1039-1046.

Deguilloux MF, Pemonge MH, Petit RJ (2004) DNA-based control of oak wood geographic origin in the context of the cooperage industry. *Annals of Forest Science* 61, 97-104.

Del Alamo Sanza M, Fernandez Escudero JA, De Castro Torio R (2004) Changes in phenolic compounds and colour Parameters of red wine aged with oak chips and in oak barrels. *Food Science and Technology International* 10, 233-241.

Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic noncoding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology* 4, 129-131.

Doussot F, De Jeso B, Quideau S, Pardon P (2002) Extractives content in cooperage oak wood during natural seasoning and toasting; Influence of tree species, geographic location, and single-tree effects. *Journal of Agricultural and Food Chemistry* 50, 5955-5961.

Doussot F, Pardon P, Dedier J, De Jeso B (2000) Individual, species and geographic origin influence on cooperage oak extractible content (*Quercus robur* L. and *Quercus petraea* Liebl.). *Analusis* 28, 960-965.

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus* 12, 13-15.

Du FK, Petit RJ, Liu JQ (2009) More introgression with less gene flow: chloroplast vs. mitochondrial DNA in the *Picea asperata* complex in China, and comparison with other Conifers. *Molecular Ecology* 18, 1396-1407.

Duminil J, Heuertz M, Doucet JL, *et al.* (2010) CpDNA-based species identification and phylogeography: application to African tropical tree species. *Molecular Ecology* 19, 5469-5483.

Dumolin-Lapègue S, Pemonge MH, Gielly L, Taberlet P, Petit RJ (1999) Amplification of oak DNA from ancient and modern wood. *Molecular Ecology* 8, 2137-2140.

Elbaum R, Melamed-Bessudo C, Boaretto E, *et al.* (2006) Ancient olive DNA in pits: preservation, amplification and sequence analysis. *Journal of Archaeological Science* 33, 77-88.

Eurlings MCM, van Beek HH, Gravendeel B (2010) Polymorphic microsatellites for forensic identification of agarwood (*Aquilaria crassna*). *Forensic Science International* 197, 30-34.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567-1587.

Feuillat F, Dupouey JL, Sciama D, Keller R (1997a) A new attempt at discrimination between *Quercus petraea* and *Quercus robur* based on wood anatomy. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 27, 343-351.

Feuillat F, Keller R, Huber F (1998) Grain and quality of cooperage oak (*Quercus robur* L. and *Q. petraea* Liebl.): myth or reality? *Revue des Oenologues et des Techniques Vitivinicoles et Oenologiques*, 11-15.

Feuillat F, Moio L, Guichard E, *et al.* (1997b) Variation in the concentration of ellagitannins and cis- and trans-beta-methyl-gamma-octalactone extracted from oak wood (*Quercus robur* L., *Quercus petraea* Liebl.) under model wine cask conditions. *American Journal of Enology and Viticulture* 48, 509-515.

Finkeldey R, Leinemann L, Gailing O (2010) Molecular genetic tools to infer the origin of forest plants and wood. *Applied Microbiology and Biotechnology* 85, 1251-1258.

Finkeldey R, Rachmayanti Y, Nuroniah H, *et al.* (2008) Identification of the timber origin of tropical species by molecular genetic markers - the case of dipterocarps. In: *Proceedings of the International Workshop Fingerprinting Methods for the Identification of Timber Origins* (ed. Degen B), pp. 20-27.

Glover KA, Hansen MM, Lien S, *et al.* (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics* 11, 2.

Gömöry D, Yakovlev I, Zhelev P, Jedinakova J, Paule L (2001) Genetic differentiation of oak populations within the *Quercus robur/Quercus petraea* complex in Central and Eastern Europe. *Heredity* 86, 557-563.

Gougeon RD, Lucio M, Frommberger M, *et al.* (2009) The chemodiversity of wines can reveal a metabologeography expression of cooperage oak wood. *Proceedings of the National Academy of Sciences of the United States of America* 106, 9174-9179.

Group CPW, Hollingsworth PM, Forrest LL, *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106, 12794-12797.

Guchu E, Diaz-Maroto MC, Diaz-Maroto IJ, Vila-Lameiro P, Perez-Coello MS (2006) Influence of the species and geographical location on volatile composition of Spanish oak wood (*Quercus petraea* Liebl. and *Quercus robur* L.). *Journal of Agricultural and Food Chemistry* 54, 3062-3066.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplex (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources* in press.

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* 106, 158-171.

Humphreys JR, O'Reilly-Wapstra JM, Harbard JL, *et al.* (2008) Discrimination between seedlings of *Eucalyptus globulus*, *E. nitens* and their F-1 hybrid using near-infrared reflectance spectroscopy and foliar oil content. *Silvae Genetica* 57, 262-269.

Jarauta I, Cacho J, Ferreira V (2005) Concurrent phenomena contributing to the formation of the aroma of wine during aging in oak wood: An analytical study. *Journal of Agricultural and Food Chemistry* 53, 4166-4177.

King CE, Debruyne R, Kuch M, Schwarz C, Poinar HN (2009) A quantitative approach to detect and overcome PCR inhibition in ancient DNA extracts. *Biotechniques* 47, 941-949.

Lens F, Jansen S, Caris P, Serlet L, Smets E (2005) Comparative wood anatomy of the primuloid clade (*Ericales* s.I.). *Systematic Botany* 30, 163-183.

Lepais O, Gerber S (2010) Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution* 65, 156-170.

Lepais O, Léger V, Gerber S (2006) Short note: High throughput microsatellite genotyping in oak species. *Silvae Genetica* 55, 238-240.

Lepais O, Petit RJ, Guichoux E, *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology* 18, 2228-2242.

Lin H, Walker MA (1997) Extracting DNA from cambium tissue for analysis of grape rootstocks. *Hortscience* 32, 1264-1266.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362, 709-715.

Lowe A (2008) Can we use DNA to identify the geographic origin of tropical timber? In: *Proceedings of the International Workshop Fingerprinting Methods for the Identification of Timber Origins* (ed. Degen B), pp. 15-19.

Mariette S, Cottrell J, Csaikl UM, *et al.* (2002) Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Quercus petraea* Liebl. and *Quercus robur* L. stands. *Silvae Genetica* 51, 72-79.

Mitchell, Alyson E, Hong, *et al.* (2005) A comparison of polyvinylpolypyrrolidone (PVPP), silica xerogel and a polyvinylpyrrolidone (PVP)-silica co-product for their ability to remove polyphenols from beer. *Journal of the Institute of Brewing* 111, 6.

Mosedale JR, Feuillat F, Baumes R, Dupouey JL, Puech JL (1998) Variability of wood extractives among *Quercus robur* and *Quercus petraea* trees from mixed stands and their relation to wood anatomy and leaf morphology. *Canadian Journal of Forest Research* 28, 994-1006.

Mosedale JR, Ford A (1996) Variation of the flavour and extractives of European oak wood from two French forests. *Journal of the Science of Food and Agriculture* 70, 273-287.

Muir G, Fleming CC, Schlötterer C (2000) Taxonomy: Species status of hybridizing oaks. *Nature* 405, 1016-1016.

Muir G, Fleming CC, Schlötterer C (2001) Three divergent rDNA clusters predate the species divergence in *Quercus petraea* (matt.) liebl. and *Quercus robur* L. *Molecular Biology and Evolution* 18, 112-119.

Novaes RML, Rodrigues JG, Lovato MB (2009) An efficient protocol for tissue sampling and DNA isolation from the stem bark of Leguminosae trees. *Genetics and Molecular Research* 8, 86-96.

Ogden R, McGough HN, Cowan RS, *et al.* (2009) SNP-based method for the genetic identification of ramin *Gonystylus spp*. timber and products: applied research meeting CITES enforcement needs. *Endangered Species Research* 9, 255-261.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist* 161, 151-164.

Petit RJ, Brewer S, Bordacs S, *et al.* (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* 156, 49-74.

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends in Ecology & Evolution* 24, 386-393.

Prida A, Boulet JC, Ducousso A, Nepveu G, Puech JL (2006) Effect of species and ecological conditions on ellagitannin content in oak wood from an even-aged and mixed stand of *Quercus robur* L. and *Quercus petraea* Liebl. *Annals of Forest Science* 63, 415-424.

Prida A, Ducousso A, Petit RJ, Nepveu G, Puech JL (2007) Variation in wood volatile compounds in a mixed oak stand: strong species and spatial differentiation in whisky-lactone content. *Annals of Forest Science* 64, 313-320.

Prida A, Puech JL (2006) Influence of geographical origin and botanical species on the content of extractives in American, French, and East European oak woods. *Journal of Agricultural and Food Chemistry* 54, 8115-8126.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.

Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2006) Extraction, amplification and characterization of wood DNA from *Dipterocarpaceae*. *Plant Molecular Biology Reporter* 24, 45-55.

Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2009) DNA from processed and unprocessed wood: Factors influencing the isolation success. *Forensic Science International-Genetics* 3, 185-192.

Rogers SO, Kaya Z (2006) DNA from ancient cedar wood from King Midas' tomb, Turkey, and Al-Aksa Mosque, Israel. *Silvae Genetica* 55, 54-62.

Russ A, Fiserova M, Gigac J (2009) Preliminary study of wood sepcies identification by NIR spectroscopy. *Wood Research* 54, 23-32.

Samuel R (1999) Identification of hybrids between *Quercus petraea* and *Q. robur* (*Fagaceae*): results obtained with RAPD markers confirm allozyme studies based on the Got-2 locus. *Plant Systematics and Evolution* 217, 137-146.

Schwarz C, Debruyne R, Kuch M, *et al.* (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research* 37, 3215-3229.

Scotti-Saintagne C, Mariette S, Porth I, *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea (*Matt.) Liebl.]. *Genetics* 168, 1615-1626.

Shaw J, Lickey EB, Beck JT, *et al.* (2005) The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92, 142-166.

Soltis PS, Soltis DE, Smiley CJ (1992) An rbcL sequence from a Miocene *Taxodium* (bald cypress). *Proceedings of the National Academy of Sciences of the United States of America* 89, 449-451.

Soulsbury C, Iossa G, Edwards K, Baker P, Harris S (2007) Allelic dropout from a high-quality DNA source. *Conservation Genetics* 8, 733-738.

Spillman PJ (1999) Wine quality biases inherent in comparisons of oak chip and barrel systems. *Australian and New Zealand Wine Industry Journal* 14, 25-33.

Spillman PJ, Sefton MA, Gawel R (2004) The effect of oak wood source, location of seasoning and coopering on the composition of volatile compounds in oak-matured wines. *Australian Journal of Grape and Wine Research* 10, 216-226.

Streiff R, Ducousso A, Lexer C, *et al.* (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Molecular Ecology* 8, 831-841.

Streiff R, Labbe T, Bacilieri R, *et al.* (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* 7, 317-328.

Tacconi L (2008) *Illegal logging: law enforcement, livelihoods and the timber trade*, Luca Tacconi edn. Earthscan.

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N (2009) Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International-Genetics* 3, 222-226.

Young OA, Kaushal M, Robertson JD, Burns H, Nunns SJ (2010) Use of species other than oak to flavor wine: an exploratory survey. *Journal of Food Science* 75, S490-S498.

CHAPITRE 4

Genes under selection provide unique insights on oak trees demography

E. Guichoux[1,2,3], P. Garnier-Géré[1,2], L. Lagache[1,2], C . Boury[1,2] and R.J. Petit[1,2]

[1] INRA, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[2] University of Bordeaux, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France

[3] Centre de Recherche Pernod Ricard, F-94000 Créteil, France

INTRODUCTION

When reconstructing population history, a large number of presumably neutral markers are typically used for making demographic inferences. Those showing the footprints of selection are generally discarded from further analyses. The rationale for excluding these loci is simple. In population genetics, neutral theory is used to infer demographic processes from genetic data. Locus-specific effects caused by selection could potentially complicate the demographic inference process. For instance, according to Luikart *et al.* (2003), genes under selection could "bias estimates of parameters such as gene flow, population size and structure, and therefore should not be used". Similarly, for Beaumont (2005) "recovering the historical demography of populations through an analysis of genetic variation [is] a project that demands the absence of selection", whereas for Helyar *et al.* (2011) "loci under selection leads to violation of assumptions for most neutral population genetic models and may cause erroneous inference about population demographic parameters, such as rates of genetic drift and migration between individual demes".

In an interesting move away from these established ideas, researchers have recently suggested to preferentially focus on genes under selection to reconstruct genetic structure in high gene flow species (Nielsen *et al.*, 2007; O'Malley *et al.*, 2007; Westgaard & Fevolden, 2007; Gebremedhin *et al.*, 2009; Nielsen *et al.*, 2009; Andre *et al.*, 2010). As pointed out by André *et al*. (2010), when selection pressure is temporally stable, genetic markers influenced by selection can help detect population structure even when neutral loci have not diverged substantially. While promising, these studies have typically relied on one or a few genes under selection and did not attempt to reconstruct actual patterns of gene flow. Yet, using genes under selection to study gene flow would not be without ground. It has long been known that some forms of selection reduce or enhance effective gene flow (Bengtsson, 1985). This might be of interest not only to delimitate populations adapted to different environmental conditions but also to reconstruct specific demographic processes including characteristics of gene flow itself.

One circumstance that could make it relevant to use selected genes rather than neutral ones is when interspecific gene flow occurs at high rates between species coexisting in sympatry. Such genetic exchanges could confound the reconstruction of intraspecific demographic

processes, including gene flow and genetic drift. Focusing on selected loci with low interspecific gene flow should improve the visualization of intraspecific processes whenever interspecific gene flow is of a magnitude comparable or higher than gene flow among conspecific populations (Neophytou *et al.*, 2010; Raduski *et al.*, 2010).

Another circumstance where there could be an advantage in using markers under selection rather than neutral ones is when gene flow is too high to accurately estimate patterns of gene exchanges. The difficulty to study demographic history in "high gene flow" species has been emphasized previously (Waples, 1998). When gene flow is very high, differences in allelic frequencies between populations are very slight and become hard to measure, with errors assuming relatively more importance. Under such circumstances, focusing on markers that are exchanged at reduced rates between populations, due e.g. to their association with genes under local selection, could be of interest to evaluate relative differences in levels of gene flow.

In fact, genetic exchanges, both within and between species, are often asymmetric, with some populations acting as sources of migrants and others as sinks (Sweigart & Willis, 2003; Curtu *et al.*, 2007; Palstra *et al.*, 2007; Addison & Pogson, 2009; Gomes *et al.*, 2009; Hertwig *et al.*, 2009; Boratynski *et al.*, 2010). Such an asymmetry can be caused by many processes, including differences in abundance (Lepais *et al.*, 2009) or in population dynamics (Currat *et al.*, 2008). To measure this asymmetry, genetic markers can be used; typically, the sink population will be more genetically variable while also harboring more unique alleles (Ritland, 1989). However, if gene flow is high, the signal will be low, even in cases where gene flow is nearly exclusively unidirectional. We predict that in such cases, markers under divergent selection will outperform neutral markers, because the signature of asymmetric gene flow will be easier to detect with markers experiencing reduced effective gene flow.

Thousands of Single Nucleotide Polymorphisms (SNPs) can now be isolated in non-model species at low cost, offering multiple possibilities to decide which loci will be used for downstream analysis (Garvin *et al.*, 2010; Helyar *et al.*, 2011). Divergent selection is a form of selection that acts in contrasting directions in different populations (Rundle & Nosil, 2005). As not all parts of the genome will be experiencing selection to a similar extent, heterogeneous differentiation across the genome will result (Wu & Ting, 2004). So called "outlier" loci show a genetic differentiation that exceeds the upper level of genetic

divergence expected under neutrality (Kelley *et al.*, 2006; Nosil *et al.*, 2009). Detection of outlier loci can be achieved by studying linkage disequilibrium between markers (for model species with exhaustive genomic resources) or by looking at differences in levels of genetic variation and levels of genetic divergence between samples (Helyar *et al.*, 2011). For this latter approach, all recent methods (Beaumont & Nichols, 1996; Beaumont, 2005; Foll & Gaggiotti, 2006; Excoffier *et al.*, 2009, Narum & Hess, 2011) are derived from the original Lewontin–Krakauer test, which compares single locus estimates of F_{ST} to an expected neutral distribution of F_{ST} (Lewontin & Krakauer, 1973). While such outliers should be comparatively rare, genome-wide surveys will provide many candidates, even in non-model species. Hence, they could be used in combination for the purpose of reconstructing "average" aspects of population demography and history, just as presumably neutral markers have been used in the past.

In this study, we focus on two interfertile white oak species, *Quercus robur* and *Q. petraea*, which have been used as models to discuss hybridization, species concepts and speciation ever since Darwin mentioned them in his chapter 2 of the *Origins* (Stebbins, 1950; Burger, 1975; van Valen, 1976; Grant, 1981; Coyne & Allen Orr, 2004). The two oak species are widely distributed over Europe and have overlapping ranges (*Q. petraea* being largely included within the distribution range of *Q. robur*). They frequently occur together in mixed stand and frequently hybridize (Petit *et al.*, 2004; Jensen *et al.*, 2009; Lepais & Gerber, 2010). However, despite recurrent interspecific gene flow, *Q. robur* and *Q. petraea* remain ecologically and morphologically differentiated (Dering & Lewandowski, 2007; Lepais & Gerber, 2010).

Accurately delimiting these two widespread and abundant oak species has been a long term goal of botanists and geneticists (Bodénès *et al.*, 1997; Muir *et al.*, 2000; Coart *et al.*, 2002; Scotti-Saintagne *et al.*, 2004; Kelleher *et al.*, 2005; Guichoux *et al.*, 2011). Without reference samples, this can be tricky. If markers likely under selection are being used for this purpose, due to their greater discriminatory power, it is important to check that an unbiased classification is obtained that does not overly depends on the inclusion of a particular locus under strong divergent selection.

Once species have been accurately delimitated, other analyses become possible. Following the last ice age, the more pioneer oak species, *Q. robur*, is thought to have recolonized first, to be later followed by the more shade-tolerant *Q. petraea* (Petit *et al.*, 2004). Given this

demographic asymmetry, one might predict that *Q. petraea*, the later invader, should be more introgressed than the pioneer *Q. robur* (Currat *et al.*, 2008). Indeed, *Q. petraea*, during its invasion, will be initially at low densities. It will tend to mate mostly with the already established species, *Q. robur*, due to a lack of conspecific mates, resulting in increased rates of hybridization (Hubbs, 1955). Those genes from *Q. robur* that leak into in the genome of *Q. petraea* will then be amplified by the logistic demographic growth of this invading species, resulting in large scale introgression. In contrast, little introgression is expected towards the resident *Q. robur*, as this species is already at carrying capacity (Currat *et al.*, 2008). Hence, asymmetric gene flow is expected to have taken place between these two oak species. Unfortunately, this asymmetry might be hard to document if interspecific gene flow is high anyway, which appears to be the case (Bacilieri *et al.*, 1993; Bacilieri *et al.*, 1994; Streiff *et al.*, 1998; Streiff *et al.*, 1999; Lepais *et al.*, 2009; Chybicki & Burczyk, 2010).

In addition, the more pioneer *Q. robur* species has been shown to better disperse its pollen than the more shade-tolerant and competitive *Q. petraea* (Jensen *et al.*, 2009, Lagache *et al.* unpublished results). This result makes sense in view of the more diffuse distribution at the edge of forests and lighter pollen of *Q. robur* compared to *Q. petraea* (Rushton, 1976; Petit *et al.*, 2004). Hence, greater differentiation is expected among *Q. petraea* populations than among *Q. robur* populations. However, this prediction could be compromised by high rates of interspecific gene flow, which have been reported to be of comparable magnitude than rates of gene flow among populations (Neophytou *et al.*, 2010).

To accurately delimitate species and test both predictions regarding direction of introgression and levels of intraspecific genetic structure, we relied on outlier markers identified in multilocus scans and compared the results with those obtained with presumably neutral markers. We predict that oak genomic regions experiencing divergent selection between species should outperform neutral markers for the purposes of (i) delimitating species, (ii) identifying the primary direction of interspecific gene flow and (iii) measuring intraspecific gene flow. We genotyped 855 oak samples from six mixed forests at 262 SNPs. Half of the SNPs used had been selected for their ability to differentiate the two species, using appropriate criteria (Jost, 2008; Gerlach *et al.*, 2010; Meirmans & Hedrick, 2010). The second half contains genes that are typically much less differentiated between species. Using two complementary approaches based on assignment methods and genotype

likelihoods, we demonstrate that genes under selection (outlier loci) have outstanding power to delimitate the two oak species and provide unique insights on intra- and interspecific gene flow, whereas genes lacking such a signature (putatively neutral loci) provide little or no resolution. These results contradict the received knowledge that only neutral markers should be used to reconstruct demographic processes and indicate that even gene flow studies can benefit from the use of markers under selection.

MATERIAL & METHODS

Material

We sampled 855 oak trees in six populations in northern France (Petite Charnie, Vitrimont, Charmes, Lure, Cuve, Mondon, see geographic allocation and sampling sizes in Supporting Information S1). All populations are mixed stands of *Q. robur* and *Q. petraea*. One stand (Petit Charnie) includes adult trees (278) and their offspring (380 samples in 51 half-sib families) from both species (Guichoux *et al.*, 2011). This population has been intensively studied for many years for gene flow, species differentiation, phenology, and wood characteristics (Bacilieri *et al.*, 1993; Bacilieri *et al.*, 1994; Bacilieri *et al.*, 1995; Streiff *et al.*, 1998; Streiff *et al.*, 1999; Prida *et al.*, 2006; Prida *et al.*, 2007; Lepais *et al.*, 2009). All samples were identified in the field as purebreds or putative hybrids using morphological criteria. Leaves or buds were sampled and stored immediately at -20°C or in silica gel.

DNA isolation

DNA was isolated from leaves or buds with Invisorb DNA plant HTS 96 kit (Invitek, Berlin, Germany), following the manufacturer instructions, except for the lysis step (one hour at 65°C). DNA quality was estimated on a 1% (w/v) agarose gel and DNA concentration was evaluated on an Infinate 200 microplate reader (Tecan, Männedorf, Switzerland) using the Quant-it dsDNA Broad-Range Assay Kit (Invitrogen, Carlsbad, USA). Concentration of each sample was adjusted to 50ng/µL on a STARlet 8-channel robot (Hamilton, Reno, USA).

SNP selection

1000 DNA fragments were resequenced from candidate genes potentially involved in adaptive differentiation (i.e. linked to drought stress tolerance, hypoxia, reproduction, phenology, host-pathogens interactions), in a panel of 24 genotypes from both species (*Q.*

robur and *Q. petraea*) sampled across their natural distribution. The sequence data production strategy included the development of bioinformatics tools adapted to a non-model species, building on 100,000 assembled Sanger ESTs, and followed different steps of data quality testing for optimizing the overall success rate (Garnier-Géré *et al.*, unpublished data). A total of 12,469 SNPs were detected among assembled ESTs with a Perl script, *snp2illumina* (Lepoittevin *et al.*, 2010), which automatically detects SNPs in FASTA sequences and makes them compatible with Illumina Assay Design Tool software (Illumina Inc., San Diego, USA). We selected a subset of 384 polymorphic SNPs for the present study. Criteria included amplification success (>2/3 in each species) and Illumina score (> 0.6). Among the 384 SNPs, 200 were selected on their ability to differentiate the two species. The 184 others were selected among putative genes involved in drought stress.

SNP genotyping

SNP genotyping was achieved with the 384-plex GoldenGate assay (Illumina Inc., San Diego, USA), which is based on the VeraCode technology. We followed the manufacturer's instructions, using 250 ng of DNA as starting quantity. Three negative controls were added in each batch of five 96-well plates. Analysis (i.e., clustering) was realized with BeadStudio software (Illumina Inc.) according to Lepoittevin *et al.* (2010), except that we did not automatically discarded compressed clusters (i.e. when the two homozygous clusters are closer to each other than expected) and SNPs lacking one homozygote cluster. We used well-established progenies from Petite Charnie population to validate *a posteriori* all SNPs (Guichoux *et al.*, 2011). Monomorphic loci and loci in total LD were discarded from subsequent analyses.

Assignment methods for accurate species delimitation

Bayesian clustering approaches implemented STRUCTURE 2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2003) were used to classify individual genotypes on the basis of data from 262 validated SNPs. A burn-in of 50,000 steps followed by 50,000 Markov chain Monte Carlo repetitions was used. We then calculated the average assignment score over 10 runs with *K* (number of groups) set to two, corresponding to the two species. The use of appropriate threshold values to classify samples into appropriate categories (purebreds, hybrids and backcrosses) is critical because inappropriate thresholds might result in wrong estimation of

the proportion of purebreds or introgressed samples (Vähä & Primmer, 2006). Typically, thresholds used to assign pure samples are 0.1-0.9 (Vähä & Primmer, 2006). We used slightly different thresholds, based on a biological rather than on an empirical approach. Considering putative purebreds with admixture levels of 0 (for species A) and 1 (for species B), and putative F1 hybrids with admixture level of 0.5, we expect backcrosses to have admixture levels distributed around 0.25 for species A and 0.75 for species B. Assuming that the sample consists of a mixture of purebreds, hybrids and backcrosses, the optimal thresholds for purebreds should be respectively 0-0.125 and 0.875-1. To confirm this choice, we simulated 5000 genotypes using HYBRIDLAB 1.0 (Nielsen *et al.*, 2006) and compared their assignment scores with expectations. Using allelic frequencies of 200 purebred samples from each species delimitated using the above thresholds, we generated 1000 genotypes from each of the following category: purebreds (2), F1 and backcrosses (2). Assignment level for each sample estimated as before using STRUCTURE was compared to theoretical expectations and false assignments were counted. The five categories (*Q. petraea*: 0-0.125, backcrosses with *Q. petraea*: 0.125-0.375, F1 hybrids: 0.375-0.625, backcrosses with *Q. robur*: 0.625-0.875, *Q. robur*: 0.875-1) were used to establish a gold standard based on 262 validated SNPs for further comparison with only a subset of the loci. To validate this gold reference, we performed new admixture analyses on two independent subsets of 131 SNPS, randomly selected among the 262 SNPs. We compared the assignment values for each sample obtained with the two analyses. We also measured power, accuracy and performance as described in (Vähä & Primmer, 2006). Comparison with microsatellites relied on an existing dataset obtained by screening the same genotypes with 12 EST-SSRs that had been selected for their ability to distinguish these two species (Guichoux *et al.*, 2011). We also carried out an analysis using different subsets of SNPs to test their ability to correctly assign pure samples. Increasing numbers of loci (2, 4, 8, 16, 32, 64, 128 and 256) ranked by decreasing interspecific D_{EST} were used. The analyses were performed with STRUCTURE 2.3.3, using the same conditions as previously described.

Diversity analyses

Allelic frequencies, genotypic frequencies, expected heterozygosity (H_e) and F_{IS} were estimated for each of the following group: *Q. robur*, *Q. petraea* and intermediates, as defined using assignment methods and appropriate thresholds. Intra- and interspecific F_{ST} were then

estimated by considering only purebred individuals. To obtain diversity-independent parameters (Gerlach *et al.*, 2010), we also calculated intra- and interspecific D_{EST} for all loci as follows:

$$D_{EST} = \frac{H_T - H_S}{1 - H_S} \times \frac{n}{n-1} \quad \text{(Jost, 2008)}$$

where H_T is the heterozygosity of the pooled populations, H_S is the mean heterozygosity of the individual populations and n is the number of populations.

Detection of outlier loci

Loci showing very high levels of genetic differentiation between species ("outlier loci") were detected with the "detecting loci under selection" module implemented in ARLEQUIN 3.5.1.2 (Excoffier *et al.*, 2009). This module inspired from Beaumont & Nichols (1996) infers heterozygosity between populations (H_{BP}) from the average heterozygosity within populations (H_{WP}). We simulated 20.000 genotypes with the number of demes set to two from a subset of loci with mean F_{ST}=0.03, as including loci under selection in the initial F_{ST} estimate may generate a bias in the simulated distribution (Helyar *et al.*, 2011). This F_{ST} value corresponds to the overall mean F_{ST} between the two species evaluated by Scotti-Saintagne *et al.* (2004) on 389 markers (isozymes, AFLPs, SCARs, microsatellites, and SNPs). To avoid diversity-dependence of F_{ST} values, D_{EST} was used to detect these outlier loci, and was plotted against heterozygosity. We used the 95th percentile as proposed by Beaumont & Nichols (1996) to evaluate the lower D_{EST} limit for loci under selection. All other diversity analyses were performed using GENALEX 6.4 (Peakall & Smouse, 2006) and XLSTAT (Addinsoft, France).

Comparison of genotype likelihoods between species

Following Paetkau *et al.* (1995) and Waser & Strobeck (1998), we plotted genotype likelihoods for *Q. robur*, *Q. petraea* and intermediates. Likelihood of one genotype is estimated as "the square of the observed allele frequency for homozygotes or twice the product of the two allele frequencies for heterozygotes and likelihoods for each locus were multiplied together, under an assumption of independence between loci, to yield an overall likelihood" (Paetkau *et al.*, 2004). This simple method allows direct visualization of genetic differentiation between two or more clusters. We used three different subsets of loci (12 EST-SSRs, 262 SNPs, outlier

loci with D_{EST}>0.2) to estimate genotype likelihoods. For each cluster, we calculated the coordinates of the barycenter and the associated D_{LR} (mean distance of individuals from the diagonal center line, as proposed by Paetkau *et al.* (2004)). All genotype likelihood analyses were performed with GENALEX 6.4 (Peakall & Smouse, 2006), using the leave-one-out option for allelic frequencies estimation. To facilitate interpretation, likelihoods were log-transformed before plotting, highest values (near zero) indicating the most likely population (Peakall & Smouse, 2006). To validate this approach, we simulated 5000 genotypes, following the observed proportion of each class (purebreds, backcrosses and F1 hybrids) estimated using our gold reference.

RESULTS

SNP genotyping

After all validations steps, 262 out of 384 SNPs were retained (68.3%). Cluster compressions occurred in 18% of SNPs and 6.3% had one homozygous genotype lacking (Supporting Information S2). Previously validated mother-offspring relationships allowed the validation of the majority of ambiguous profiles (i.e. compressed clusters). In contrast, 12 SNPs (1.5%) were excluded based on incompatibility with well-established relationships (parent pair analysis), warning against automatic validations. With these methods, we also detected single errors for nine other loci. Considering that these errors may be due to point mutations, we included these SNPs in the analysis. If severe precautions for SNP analysis had been taken on our dataset, as proposed in recent studies (Close *et al.*, 2009; Lepoittevin *et al.*, 2010), we would have discarded many valid loci, unnecessarily reducing success rate (down to an estimated 50%).

Species assignment

Assignment results based on 262 SNPs highlighted a low proportion of intermediate samples (8% of F1 hybrids and backcrosses), about twice lower than the estimate based on 12 EST-SSRs (15%). The estimated proportion of F1 hybrids was not significantly different with the two datasets (3.2% versus 3.6%). This proportion is much lower than previously described between these two species (Lepais *et al.*, 2009), most probably due to the number and the nature of the loci used for assignment analysis. Assignment stability was very high for

purebreds (97.4% of correspondence between the two subsets of 131 SNPs, see Figure 1). Assignment of intermediate samples was more variable between the two subsets of loci (correspondence of 77.3% for F1 hybrids and 59.6% for backcrosses), underlying the fact that assignment errors remain present for these categories. Assignment values were also very stable between 12 EST-SSRs and 262 SNPs (95.2% of correspondence for purebreds, data not shown). When using a small number of highly-differentiated SNPs (those having the highest interspecific D_{EST}), efficiency, accuracy and performance were very high for both species. However, performance for *Q. robur* was always better, regardless of the number of highly-differentiated SNPs used. Hence, *Q. robur* samples require less SNPs than *Q. petraea* to be efficiently assigned (Figure 2).

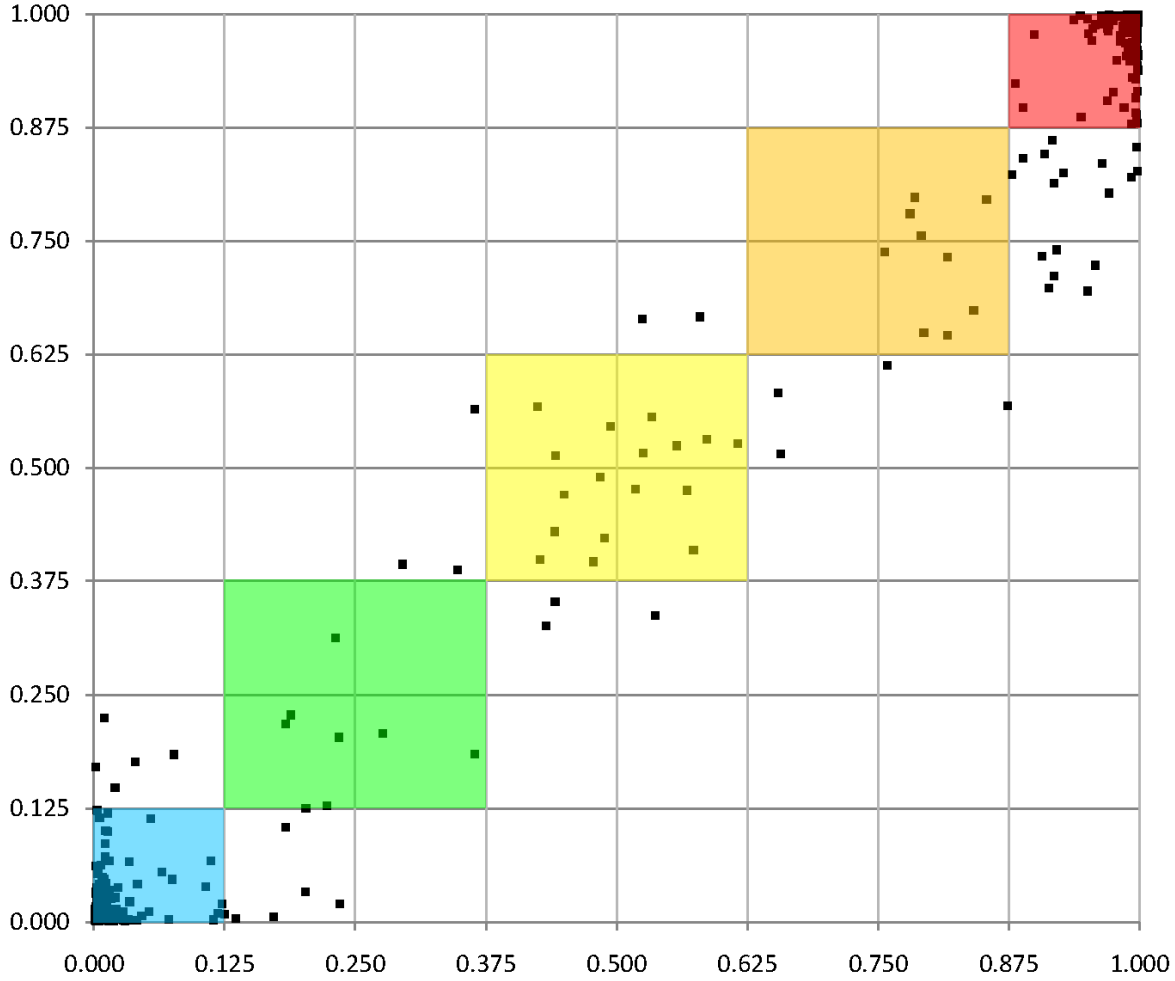


Figure 1: assignment correspondence between two subsets of 131 SNPs (randomly chosen among all 262 SNPs) for all 855 samples. Assignment values for subset A are on x-axis, assignment values for subset B are on y-axis. Points outside colored areas represent divergences in assignments between the two subsets. Red: *Q. robur*. Blue: *Q. petraea.* Yellow: F1 hybrids. Orange: backcrosses with *Q. robur*. Green: backcrosses with *Q. petraea*.

Assignment value for each simulated genotype was compared with its expected value (*Q. petraea*: 0-0.125, backcrosses with *Q. petraea*: 0.125-0.375, F1 hybrids: 0.375-0.625, backcrosses with *Q. robur*: 0.625-0.875, *Q. robur*: 0.875-1). The five classes were clearly separated, with only few false assignments (1% for each backcross class and 0.3% for F1 hybrids). The choice of these thresholds minimize the number of false assignments: different thresholds would have decreased false assignments but at the expense of an increase error rate in the adjacent category (data not shown; see also suppl. fig S3 in Guichoux *et al.* (2011)). Assignment values for simulated purebreds were asymmetric as *Q. robur* genotypes were more strongly assigned than *Q. petraea* genotypes (Supporting Information S3).

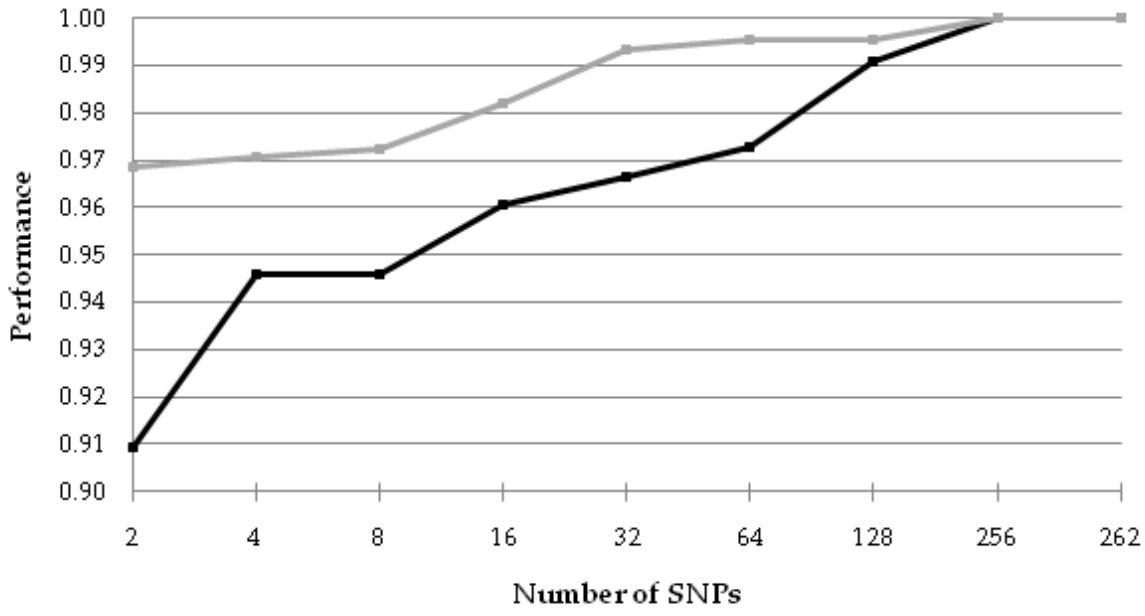


Figure 2: assignment performance (in comparison with the gold reference) for *Q. robur* (grey line) and *Q. petraea* (black line) with increasing number of SNPs with the highest D_{EST}

Genetic differentiation among populations

A posteriori assignment of multilocus genotypes to categories (purebreds, hybrids and backcrosses) makes it possible to compare genetic diversity between these categories. Mean expected heterozygosity (H_e) across loci was significantly higher for intermediates (0.268) than for either purebreds (0.220 and 0.216, p<0.001). Mean F_{IS} across loci were very close to zero and did not show differences between *Q. robur* and *Q. petraea* (both F_{IS}~0, p=0.7, see

Table 1), but F_{IS} for intermediates was lower than F_{IS} for purebreds (-0.069 vs. -0.002, p<0.001, see Table 1). Mean interspecific F_{ST} across loci was high (0.139, see Table 2), with some SNPs showing very high values (up to 0.85). D_{EST} was also high (mean=0.14, maximum=0.92, see Supporting Information S4). Mean intraspecific D_{EST} across loci was slightly higher for *Q. petraea* than for *Q. robur* (0.052 vs 0.041, p=0.001 and see Table 1).

Class	N	F_{IS}	H_e	Intraspecific D_{EST}	Interspecific D_{EST}
Quercus robur	451	-0.004	0.220	0.041	0.141
Quercus petraea	336	0.000	0.216	0.052	
Intermediates	68	-0.069	0.268	-	-

Table 1: Sample size (N), F_{IS} and H_e for the three classes (*Q. robur*, *Q. petraea* and intermediates), based on 262 SNPs. D_{EST} are provided for purebreds only.

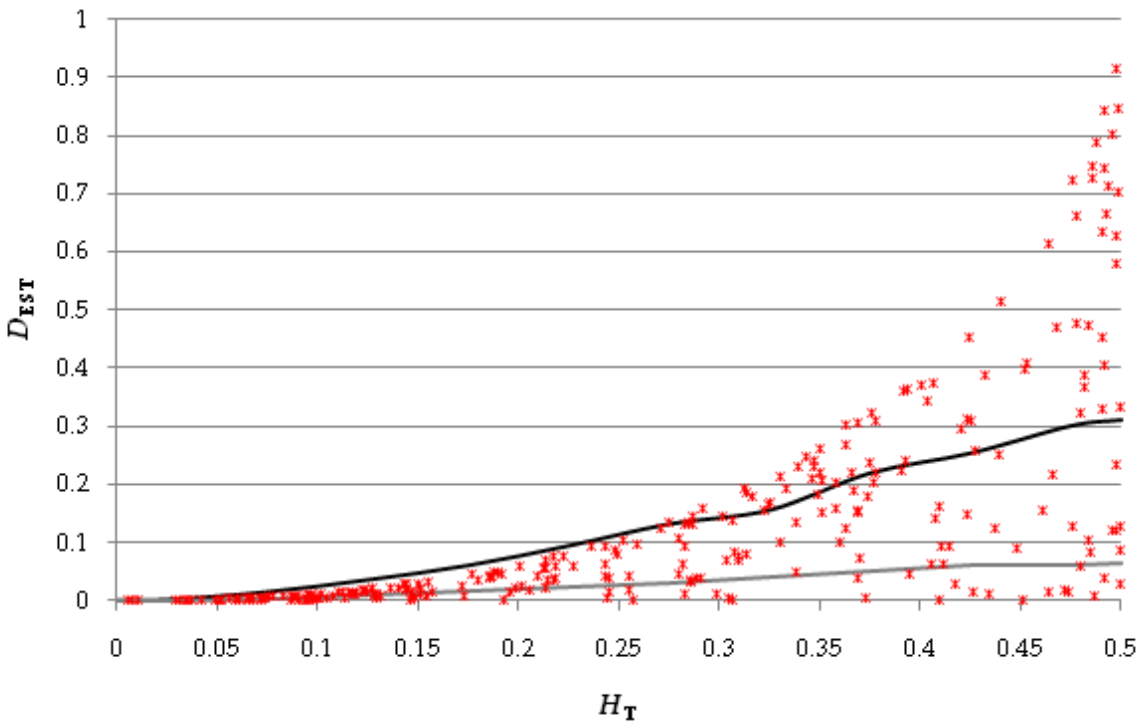


Figure 3: D_{EST} estimated from 262 SNPs with ARLEQUIN 3.5.1.2 (Excoffier *et al.*, 2009), plotted against total heterozygosity (H_T) to detect outlier loci. Black line: 95th quantile distribution. Grey line: 50th quantile distribution.

Outlier loci were detected (i.e. plotted outside the neutral envelop illustrated by the 95th quantile distribution) but only for H_T>0.25 and D_{EST}>0.13 (Figure 3). The proportion of outlier loci was 25.3% of all SNPs. To be more conservative, we decided to declare "true outliers" only loci with D_{EST}>0.2 (representing 23.4% of all SNPs). When focusing on SNPs with low or moderate differentiation (interspecific D_{EST}<0.2), intraspecific D_{EST} for both species were similar (Table 2). For outlier loci (interspecific D_{EST}>0.2), intraspecific D_{EST} for *Q. petraea* was twice as large as intraspecific D_{EST} for *Q. robur*, the difference being highly significant (Table 2).

Loci class	**N**	**Interspecific D_{EST}**	**Intraspecific D_{EST}**		**p-value**
			Q. robur	***Q. petraea***	
D_{EST} <0.2	197	0.049	0.043	0.046	0.890
D_{EST} >0.2	65	0.418	0.036	0.072	<0.001***
All loci	262	0.141	0.041	0.052	0.015*

Table 2: Loci number (N), interspecific D_{EST}, intraspecific D_{EST} for *Q. robur* and *Q. petraea* (and associated p-values), with different classes of loci: "neutral loci" with interspecific D_{EST} <0.2, "outlier loci" with interspecific D_{EST} >0.2 and all loci.

Genotype likelihoods and asymmetric introgression

An interesting visualization of genetic differentiation can be achieved by plotting genotype likelihoods for purebreds and intermediates. Log-likelihoods based on 5000 simulated genotypes and 262 SNPs (with appropriate proportions of each class) revealed clear separation of all five categories except for a few *Q. petraea* backcrosses (Figure 4a), confirming the great ability of our loci to differentiate all categories. With 12 EST-SSRs, intermediate samples could not be distinguished from purebreds, and even purebreds from both species were not fully separated (Figure 4b). In addition, there was no evidence of asymmetry between species, as D_{LR} for intermediates was null, and mean log-likelihoods for purebreds of each species were similar. With 262 SNPs, intermediates were correctly differentiated from purebreds (Figure 4c). D_{LR} for intermediates was slightly negative, meaning that intermediate genotypes were genetically closer to *Q. petraea* purebreds than to *Q. robur*. However, log-likelihoods for both purebreds were similar, with no apparent

asymmetry. In contrast, at those loci having the highest D_{EST}, a strong asymmetry was revealed (Figure 4d). Intermediates were still well delimited, but the associated D_{LR} showed that they were much closer to *Q. petraea* than to *Q. robur* (despite the higher proportion of *Q. robur* in our sample set). The comparison of each purebred class also revealed a strong asymmetry: *Q. petraea* samples with the highest log-likelihoods had lower values than *Q. robur* samples with the lowest log-likelihoods (for the targeted species). With all SNPs, log-likelihoods for the non-targeted species differed between species.

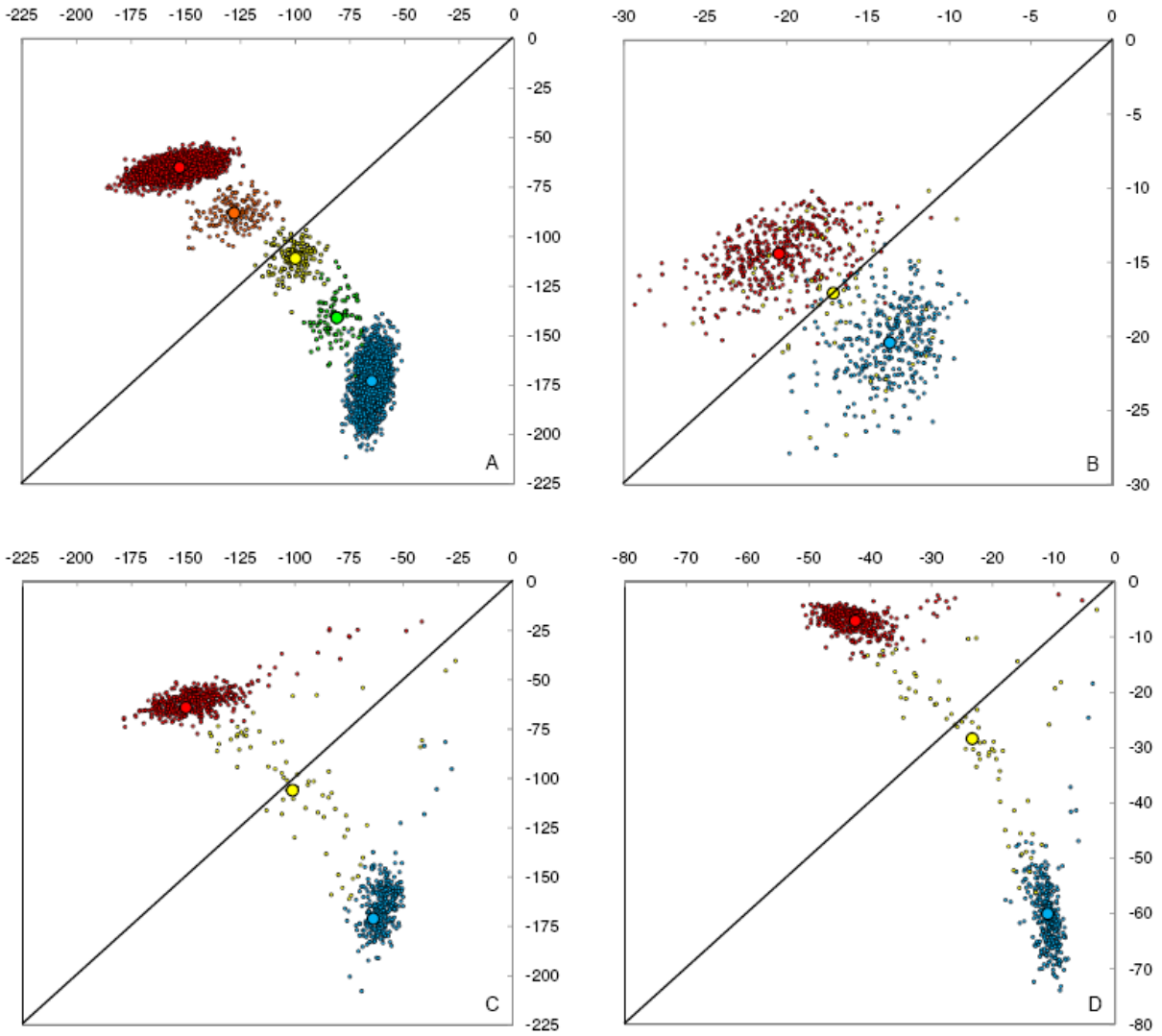


Figure 4: Log likelihood for *Q. robur, Q. petraea,* F1 hybrids and backcrosses with different subsets of markers.
x-axis : Log likelihood for *Q. petraea*. y-axis: Log likelihood for *Q. robur*
A: 5000 simulated genotypes (262 SNPs). B: 12 EST-SSRs. C: 262 SNPs. D: SNPs with D_{EST}>0.2.
Barycenters for each category are marked as larger circles. Red: *Q. robur*. Blue: *Q. petraea.* Yellow: F1 hybrids. Orange: backcrosses with *Q. robur*. Green: backcrosses with *Q. petraea*.

DISCUSSION

Advances in sequencing and in associated bioinformatics make it possible to generate genomic resources and investigate non-model species at an unprecedented rate (Ekblom & Galindo, 2011; Neale & Kremer, 2011; Rice *et al.*, 2011). In particular, direct analyses of sequence variation including single nucleotide polymorphisms (SNPs) are becoming standard, supplanting other markers such as microsatellites for a large range of studies. The ability to isolate thousands of loci at moderate cost raises the question of which loci should be used for reconstructing population structure and demographic history.

Until recently, loci showing non-neutral behavior (for example loci with very high level of differentiation between two species) were systematically excluded to avoid bias in interpretation of demographic processes (Luikart *et al.*, 2003; Beaumont, 2005). Within the past five years, alternative points of view started to emerge, as outlier loci have proved informative to reconstruct population genetic structure in high gene flow species (Nielsen *et al.*, 2007; O'Malley *et al.*, 2007; Westgaard & Fevolden, 2007; Nielsen *et al.*, 2009; Andre *et al.*, 2010). In this study, we deliberately enriched our dataset with highly-differentiated loci between the two species. We confirmed the relevance of outlier loci for species delimitation but also to reconstruct some aspects of gene flow within and between species. In particular, we showed that some population demographic processes could only be unraveled with these loci, a somewhat heretic proposal.

Species delimitation

SNPs used in this study were in part chosen for their ability to efficiently differentiate *Q. robur* from *Q. petraea.* The loci identified correspond to genome components that are likely under divergent selection, as confirmed by the genome scan test, which revealed a high proportion of outliers (23.4%). The results showed that highly differentiated SNPs allow unbiased, efficient and accurate species delimitation, even with a moderate number of loci (Cornuet *et al.*, 1999; Manel *et al.*, 2005). Our assignment results based on 262 SNPs, including a high proportion of highly-differentiated loci, largely outperform those from previous studies of the same species, based on much smaller sets of loci (Muir *et al.*, 2000; Jensen *et al.*, 2009; Lepais *et al.*, 2009). Validation of assignment performances requires the use of independent samples (Waples, 2010). In our case, the use of two independent subsets of 131

SNPs confirmed the repeatability of our results. However, detection of intermediates categories required more loci than purebred detection. In particular, backcrosses identification remains sensitive, as highlighted by discrepancies observed in assignment performance with the two subsets of 131 SNPs (ranging from 44% to 92%). Similarly, most of the assignment differences between SNPs and EST-SSRs involved backcrosses. Results obtained with increasing numbers of SNPs with high D_{EST} confirmed that backcrosses and, to a lesser extent, F1 hybrids, require more loci for correct assignment (data not shown). As a consequence, hybrid proportion observed with 262 SNPs should be more accurate than that found with SSRs. This suggests that proportions of backcrosses and F1 hybrids are prone to overestimation unless particular precautions are taken, as previously noticed (Vähä & Primmer, 2006). It also appears that the limited diversity of SNPs compared to SSRs (Rosenberg *et al.*, 2003) can be compensated for by selecting loci with appropriate criteria (Liu *et al.*, 2005). In fact, using only the two loci with the highest D_{EST} (mean=0.88), assignment performance reached 94% for both species, whereas as many as 49 loci with lowest D_{EST} (mean=0.002) were necessary to reach this value.

Directional interspecific gene flow

Assignment results based on 262 SNPs highlighted a genetic asymmetry between the two species. Whatever the loci used for clustering, *Q. robur* genotypes were more easily assigned than *Q. petraea* ones. Yet, the relative proportion of backcrosses is the same for the two species and assignment performance for these categories was equal. Results obtained on simulated samples (with appropriate proportion of each category) further confirmed that *Q. petraea* samples are globally less strictly assigned than *Q. robur* purebreds. Genotype likelihoods also revealed a strong genetic asymmetry between *Q. robur* and *Q. petraea*, with likelihoods for *Q. petraea* trees being typically lower than those for *Q. robur*. This fits with our expectations for the introgression dynamics between these two species: asymmetric introgression towards *Q. petraea* should increase diversity and decrease assignments of samples from this species. Interestingly, the evidence of asymmetry is much clearer when a subset of loci with the highest differentiation is used. Hence, focusing on loci under strong divergent selection allowed a better detection of patterns of interspecific gene exchanges expected from demographic differences.

Measuring intraspecific gene flow

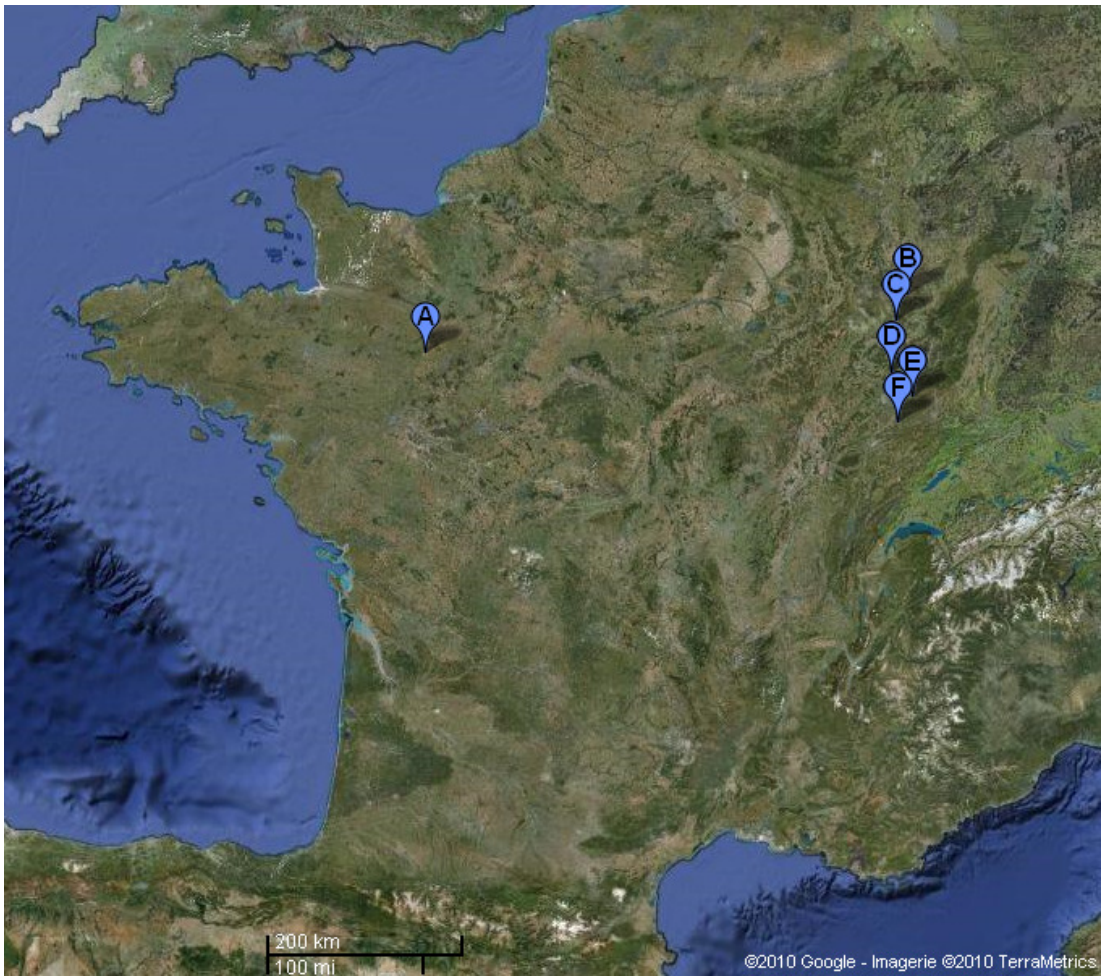
We also showed that targeting outlier loci facilitates the detection of intraspecific processes. For presumably neutral loci (D_{EST}<0.2), genetic differentiation between species was of the same magnitude than genetic differentiation among populations of the same species flow. As a consequence, gene flow among populations of the same species could affect patterns of interspecific gene flow. In contrast, at outlier loci (i.e. genes for which interspecific gene flow is considerably reduced), intraspecific differentiation for *Q. petraea* is twice as large as for *Q. robur*. These results fit with those expected from life history characteristics as well as from paternity studies, which point to reduced pollen gene flow in *Q. petraea* compared to *Q. robur* (Petit *et al.*, 2004; Jensen *et al.*, 2009). Hence, only estimates of genetic differentiation based on outlier loci match with biological expectations: loci that are less prone to interspecific gene flow appear more appropriate to reconstruct intraspecific demographic processes. Interestingly, only results using corrected differentiation indexes (such as D_{EST}) allowed the visualization of these differences between *Q. robur* and *Q. petraea* (Meirmans & Hedrick). If the classical F_{ST} index had been used, such differences between species would not have been detected, even with outliers (Supporting Information S5). This suggests that for comparative purposes, appropriate differentiation measures such as D_{EST} or $F_{ST'}$ (Meirmans & Hedrick, 2010) should be used, as these estimators do not a priori depend on heterozygosity, in contrast to F_{ST}.

Perspectives

Our study showed that loci under strong divergent selection are highly efficient for refined species delimitation. Such delimitation is unbiased. Hence, concerns that using a few outlier loci will reveal only locus-specific effects are not warranted. They also demonstrate that such markers can provide new insights on demographic events, including on gene flow. Of course, measuring average levels of neutral gene flow require the use of loci that are not overwhelmingly affected by selection processes. Yet, precious indications on patterns of gene flow (e.g. on the predominant direction of gene flow between two populations or species) could benefit from the choice of sets of loci under strong divergent selection. Similarly, in cases such as those of the two oak species we study, reconstructing patterns of intraspecific gene movements could benefit from the use of loci experiencing less interspecific gene flow

as a consequence of their association with genomic regions under divergent selection between species.

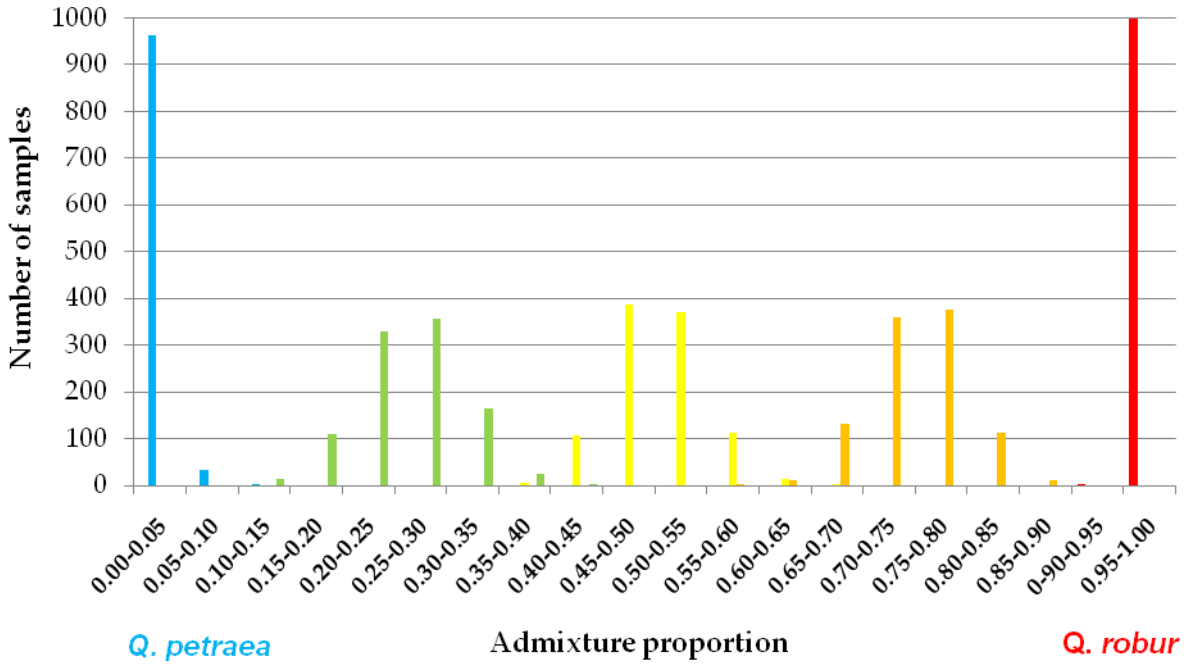
SUPPORTING INFORMATION



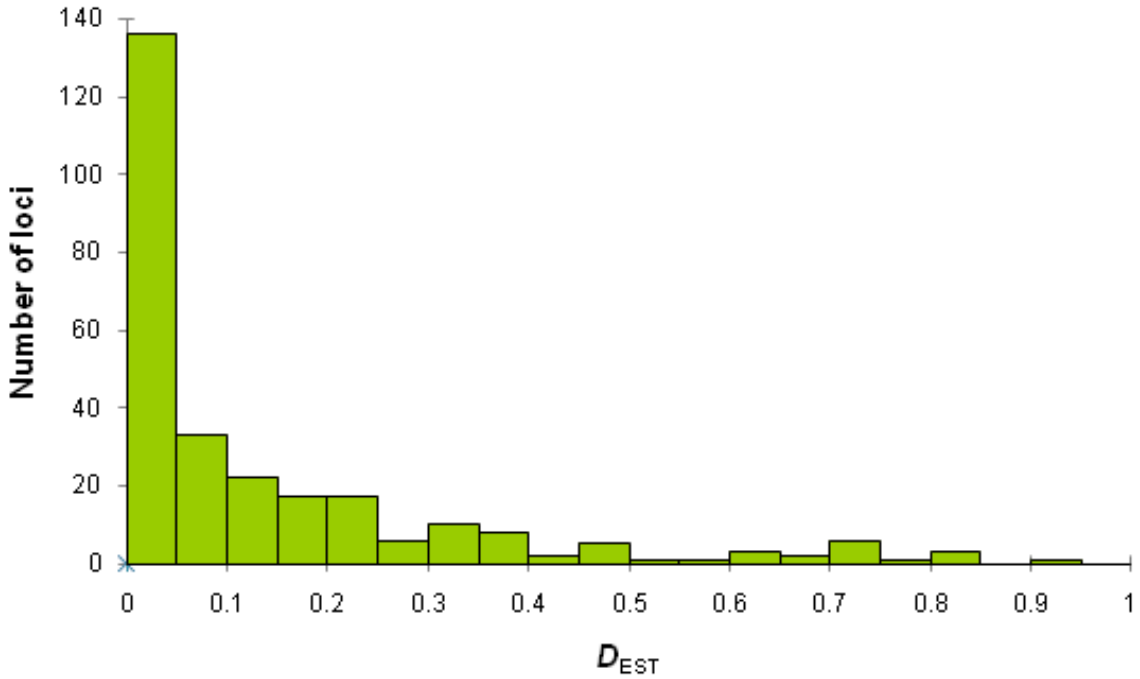
Supporting Information S1: Sampling sites. A: Petite Charnie (659 samples). B: Vitrimont (37 samples). C: Charmes (42 samples). D: Cuve (36 samples). E: Lure (42 samples). F: Mondon (39 samples).

Category	% of SNPs
No problem	48.2%
Cluster compression	18%
More than 3 clusters	13%
No cluster	8.1%
One homozygous is missing	6.3%
Monomorphic	3.9%
Total LD	1.5%
No amplification	1%

Supporting Information S2: Characteristics of the 384 SNPs genotyped



Supporting Information S3: Assignment of 1000 simulated genotypes for each category with 262 SNPs. Red: *Q. robur*. Blue: *Q. petraea.* Yellow: F1 hybrids. Orange: backcrosses with *Q. robur*. Green: backcrosses with *Q. petraea*.



Supporting Information S4: Distribution of D_{EST} (calculated between purebreds over 262 SNPs)

Loci class	N	Interspecific F_{ST}	Intraspecific F_{ST}		p-value
			Q. robur	*Q. petraea*	
F_{ST} <0.2	199	0.069	0.024	0.029	0.001
F_{ST} >0.2	63	0.351	0.022	0.034	0.009
All loci	262	0.137	0.023	0.030	<0.001***

Supporting Information S5: Loci number (N), interspecific F_{ST}, intraspecific F_{ST} for *Q. robur* and *Q. petraea* (and associated p-values), with different classes of loci: "neutral loci" with interspecific F_{ST} <0.2, "outlier loci" with interspecific F_{ST} >0.2 and all loci.

REFERENCES

Addison JA, Pogson GH (2009) Multiple gene genealogies reveal asymmetrical hybridization and introgression among strongylocentrotid sea urchins. *Molecular Ecology* **18**, 1239-1251.

Andre C, Larsson LC, Laikre L, *et al.* (2010) Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity* **106**, 270-280.

Bacilieri R, Ducousso A, Kremer A (1995) Genetic, morphological, ecological and phenological differentiation between *Quercus petraea* (Matt) Liebl. and *Quercus robur* L. in a mixed stand of northwest of France. *Silvae Genetica* **44**, 1-10.

Bacilieri R, Labbé T, Kremer A (1994) Intraspecific genetic-structure in a mixed population of *Quercus petraea* (Matt) Liebl. and *Quercus robur* L. *Heredity* **73**, 130-141.

Bacilieri R, Roussel G, Ducousso A (1993) Hybridization and mating system in a mixed stand of sessile and pedunculate oak. *Annals of Forest Science* **50**, 122-127.

Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? *Trends in Ecology & Evolution* **20**, 435-440.

Beaumont MA, Nichols RA (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.

Bengtsson BO (1985) The flow of genes through a genetic barrier. In: *Evolution. Essays in honour of John Maynard Smith.*, pp. 31-42. Cambridge University Press, Cambridge UK.

Bodénès C, Joandet S, Laigret F, Kremer A (1997) Detection of genomic regions differentiating two closely related oak species *Quercus petraea* (Matt) Liebl and *Quercus robur* L. *Heredity* **78**, 433-444.

Boratynski A, Marcysiak K, Lewandowska A, Jasinska A, Iszkulo G (2010) Interrelations among con-generic and co-occurring tree species: asymmetric hybridization and the high success of *Quercus petraea* (Matt.) Liebl. regeneration in mixed *Q. petraea*/*Q. robur* stands. *Polish Journal of Ecology* **58**, 273-283.

Burger WC (1975) The species concept in *Quercus*. *Taxon* **24**, 45-50.

Chybicki IJ, Burczyk J (2010) Realized gene flow within mixed stands of *Quercus robur* L. and *Q. petraea* (Matt.) L. revealed at the stage of naturally established seedling. *Molecular Ecology* **19**, 2137-2151.

Close T, Bhat P, Lonardi S, *et al.* (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**, 582.

Coart E, Lamote V, De Loose M, *et al.* (2002) AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl] in Flemish populations. *Theoretical and Applied Genetics* **105**, 431-439.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989-2000.

Coyne JA, Allen Orr H (2004) *Speciation* Sinauer Associates Inc.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: massive introgression by local genes. *Evolution* **62**, 1908-1920.

Curtu AL, Gailing O, Finkeldey R (2007) Evidence for hybridization and introgression within a species-rich oak (*Quercus spp.*) community. *BMC Evolutionary Biology* **7**.

Dering M, Lewandowski A (2007) Unexpected disproportion observed in species composition between oak mixed stands and their progeny populations. *Annals of Forest Science* **64**, 413-417.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of Populations. *Genetics* **174**, 875-891.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources* **10**, 915-934.

Gebremedhin B, Ficetola GF, Naderi S, *et al.* (2009) Frontiers in identifying conservation units: from neutral markers to adaptive genetic variation. *Animal Conservation* **12**, 107-109.

Gerlach G, Jueterbock A, Kraemer P, Deppermann J, Harmand P (2010) Calculations of population differentiation based on GST and D: forget GST but not all of statistics! *Molecular Ecology* **19**, 3845-3852.

Gomes B, Sousa CA, Novo MT, *et al.* (2009) Asymmetric introgression between sympatric molestus and pipiens forms of *Culex pipiens* (*Diptera*: *Culicidae*) in the Comporta region, Portugal. *BMC Evolutionary Biology* **9**.

Grant V (1981) *Plant speciation* Columbia University Press edition.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplex (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources* **in press**.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 1-14.

Hertwig ST, Schweizer M, Stepanow S, *et al.* (2009) Regionally high rates of hybridization and introgression in German wildcat populations (*Felis silvestris*, Carnivora, Felidae). *Journal of Zoological Systematics and Evolutionary Research* **47**, 283-297.

Hubbs CL (1955) Hybridization between fish species in nature. *Systematic Zoology* **4**, 1-20.

Jensen J, Larsen A, Nielsen LR, Cottrell J (2009) Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark. *Annals of Forest Science* **66**.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology* **17**, 4015-4026.

Kelleher CT, Hodkinson TR, Douglas GC, Kelly DL (2005) Species distinction in Irish populations of *Quercus petraea* and *Q. robur*: Morphological versus molecular analyses. *Annals of Botany* **96**, 1237-1246.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16**, 980-989.

Lepais O, Gerber S (2010) Reproductive patterns shape introgession dynamics and species succession within the european white oak species complex. *Evolution*.

Lepais O, Petit RJ, Guichoux E, *et al.* (2009) Species relative abundance and direction of introgression in oaks. *Molecular Ecology* **18**, 2228-2242.

Lepoittevin C, Frigerio J-M, Garnier-Géré P, *et al.* (2010) *In Vitro vs In Silico* detected SNPs for the development of a genotyping array: what can we learn from a non-model species? *PLoS ONE* **5**, e11034.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175-195.

Liu NJ, Chen L, Wang S, Oh CG, Zhao HY (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**, S26.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions techniques with appropriate. *Trends in Ecology & Evolution* **20**, 136-142.

Meirmans PG, Hedrick PW (2010) Assessing population structure: FST and related measures. *Molecular Ecology Resources* **11**, 5-18.

Muir G, Fleming CC, Schlötterer C (2000) Taxonomy: Species status of hybridizing oaks. *Nature* **405**, 1016-1016.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, in press.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**, 111-122.

Neophytou C, Aravanopoulos FA, Fink S, Dounavi A (2010) Detecting interspecific and geographic differentiation patterns in two interfertile oak species (*Quercus petraea* (Matt.) Liebl. and *Q. robur* L.) using small sets of microsatellite markers. *Forest Ecology and Management* **259**, 2026-2035.

Nielsen, Einar E, Mackenzie, *et al.* (2007) Historical analysis of Pan I in Atlantic cod (*Gadus morhua*) : temporal stability of allele frequencies in the southeastern part of the species distribution. *Canadian journal of fisheries and aquatic sciences* **64**, 1448-1455.

Nielsen EE, Bach LA, Kotlicki P (2006) Hybridlab (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes* **6**, 971-973.

Nielsen EE, Hemmer-Hansen J, Poulsen NA, *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology* **9**, 11.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375-402.

O'Malley KG, Camara MD, Banks MA (2007) Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **16**, 4930-4941.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347-354.

Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology* **13**, 55-65.

Palstra FP, O'Connell MF, Ruzzante DE (2007) Population structure and gene flow reversals in Atlantic salmon (*Salmo salar*) over contemporary and long-term temporal scales: effects of population size and life history. *Molecular Ecology* **16**, 4504-4522.

Peakall R, Smouse PE (2006) Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164.

Prida A, Boulet JC, Ducousso A, Nepveu G, Puech JL (2006) Effect of species and ecological conditions on ellagitannin content in oak wood from an even-aged and mixed stand of *Quercus robur* L. and *Quercus petraea* Liebl. *Annals of Forest Science* **63**, 415-424.

Prida A, Ducousso A, Petit RJ, Nepveu G, Puech JL (2007) Variation in wood volatile compounds in a mixed oak stand: strong species and spatial differentiation in whisky-lactone content. *Annals of Forest Science* **64**, 313-320.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Raduski AR, Rieseberg LH, Strasburg JL (2010) Effective population size, gene flow, and species status in a narrow endemic sunflower, *Helianthus neglectus*, compared to its widespread sister species, *H. petiolaris*. *International Journal of Molecular Sciences* **11**, 492-506.

Rice AM, Rudh A, Ellegren H, Qvarnstrom A (2011) A guide to the genomics of ecological speciation in natural animal populations. *Ecology Letters* **14**, 9-18.

Ritland K (1989) Genetic differentiation, diversity, and inbreeding in the mountain monkeyflower (*Mimulus caespitosus*) of the washington cascades. *Canadian Journal of Botany* **67**, 2017-2024.

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* **73**, 1402-1422.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters* **8**, 336-352.

Rushton BS (1976) Pollen grain size in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. *Watsonia* **11**, 137-140.

Scotti-Saintagne C, Mariette S, Porth I, *et al.* (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics* **168**, 1615-1626.

Stebbins GL (1950) *Variation and Evolution in Plants* Columbia University Press, New-York.

Streiff R, Ducousso A, Lexer C, *et al.* (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Molecular Ecology* **8**, 831-841.

Streiff R, Labbe T, Bacilieri R, *et al.* (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* **7**, 317-328.

Sweigart AL, Willis JH (2003) Patterns of nucleotide diversity in two species of *Mimulus* are affected by mating system and asymmetric introgression. *Evolution* **57**, 2490-2506.

Vähä J-P, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology* **15**, 63-72.

van Valen L (1976) Ecological species, multispecies, and oaks. *Taxon* **25**, 233-239.

Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity* **89**, 438-450.

Waples RS (2010) High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology* **19**, 2599-2601.

Waser PM, Strobeck C (1998) Genetic signatures of interpopulation dispersal. *Trends in Ecology & Evolution* **13**, 43-44.

Westgaard JI, Fevolden SE (2007) Atlantic cod (*Gadus morhua* L.) in inner and outer coastal zones of northern norway display divergent genetic signature at non-neutral loci. *Fisheries Research* **85**, 306-315.

Wu CI, Ting CT (2004) Genes and speciation. *Nature Reviews Genetics* **5**, 114-122.

CONCLUSIONS ET PERSPECTIVES

L'objectif principal de ce travail de thèse était de développer des outils moléculaires efficaces qui permettent d'identifier avec précision l'espèce (*Q. robur* ou *Q. petraea*) à partir d'un échantillon de bois, plus ou moins sec (du merrain à la douelle). En effet, cette caractérisation permettrait d'anticiper en grande partie le potentiel aromatique d'un bois destiné à la maturation des vins et spiritueux. L'utilisation de marqueurs moléculaires pour différencier finement deux espèces si proches génétiquement était en soi un challenge. Le transfert de ces marqueurs sur bois, avec toutes les contraintes inhérentes au travail sur de l'ADN dégradé, s'est également avéré complexe.

De l'importance du choix des marqueurs génétiques

Il est rapidement apparu qu'un des points clés de ce travail d'identification d'espèce serait le choix des marqueurs retenus. Les marqueurs microsatellites sont depuis de nombreuses années les marqueurs génétiques les plus utilisés dans une grande variété de travaux (étude de la structure des populations ou des espèces, étude de l'hybridation, études des flux de gènes, …). Le développement, au cours des 20 dernières années, d'un grand nombre de marqueurs microsatellites génomiques pour le genre *Quercus* (Dow *et al.*, 1995; Steinkellner *et al.*, 1997; Kampfer *et al.*, 1998) a naturellement orienté le choix initial vers ce type de marqueurs. De plus, la constitution de plusieurs banques de ADNc au sein de notre laboratoire a permis d'isoler plus de 250 microsatellites issus d'EST (Durand *et al.*, 2010), offrant ainsi un large choix de marqueurs pour différencier les deux espèces de chênes (Banks *et al.*, 2003). Comme nous l'avons vu dans le **Chapitre 1**, les microsatellites présentent un grand nombre d'avantages par rapport aux SNPs. De part leur nature multi-allélique, ils sont plus informatifs que les SNPs, généralement di-alléliques. Par conséquent, moins de microsatellites que de SNPs seront nécessaires pour correctement identifier l'espèce de chêne, diminuant d'autant le coût des analyses. Dans l'optique d'une application industrielle, ce point est particulièrement important. Les microsatellites, en particulier ceux issus d'ESTs (Varshney *et al.*, 2005), sont également plus facilement transférables entre espèces proches. Cependant, identifier avec précision l'espèce à l'aide de ces marqueurs reste délicat. Premièrement, comme rappelé dans les **Chapitres 1 et 2**, le génotypage de ces marqueurs microsatellites, bien plus que celui des SNPs, peut présenter des taux d'erreur importants (Bonin *et al.*, 2004; Hoffman & Amos, 2005; Pompanon *et al.*, 2005), avec des conséquences parfois significatives sur les résultats d'affectation (Luikart *et al.*, 2008).

Deuxièmement, en l'absence de microsatellites diagnostiques entre ces deux espèces, l'identification passe nécessairement par le développement d'une base de données génétiques de référence qui permet dans un second temps de tester de nouveaux échantillons (en particulier les échantillons de bois). Cette base de données est d'autant plus efficace qu'elle est conséquente et qu'elle capte un maximum de la diversité de ces deux espèces (à l'échelle de l'aire de distribution géographique). L'intégration dans ce travail de thèse de nombreuses populations de chênes français a très certainement permis d'augmenter la précision de nos méthodes d'identification basée sur l'affectation génétique. Il n'est cependant pas acquis qu'un échantillon d'une population très éloignée soit aussi précisément affecté avec ces outils et cette base de données. Pour développer des bases de données microsatellites à une échelle suffisante pour permettre une affectation efficace (plusieurs centaines d'individus génotypés à quelques dizaine de loci), le multiplexage apparait comme la méthode idéale. Pourtant, lors d'une méta-analyse présentée dans le **Chapitre 1**, j'ai constaté que seulement 42% des études récentes basées sur les microsatellites utilisaient cette technique. Ce faible taux de pénétration semble être lié à des a priori sur le temps et le coût de développement de ces multiplex. Or, comme je l'ai démontré dans les **Chapitres 1 et 2**, ce qui est le plus limitant dans cette méthode est la validation des marqueurs en simplex, inhérente à tout développement de nouveaux marqueurs microsatellites. Avec une méthodologie adaptée et grâce aux outils récemment dédiés au multiplexage (Holleley & Geerts, 2009; Shen *et al.*, 2010), développer un multiplex de marqueurs microsatellites est possible à des coûts abordables, et ce même sur des espèces non-modèles. Cependant, certaines espèces demeurent réfractaires à de telles techniques, principalement à cause de la nature de leur génome parfois pauvre en motifs microsatellites (Parchman *et al.*, 2010). Un récent développement au sein de notre équipe d'un multiplex sur mélèze (*Larix spp.)* à partir de données de séquences 454 (Roche) a confirmé que dans le cas d'espèce pauvres en motifs microsatellites ou présentant des motifs peu variables, la mise au point d'un multiplex devient beaucoup plus difficile (Stefanie Wagner, communication personnelle).

Les contraintes liées à l'amplification d'ADN nucléaire à partir de bois

Les marqueurs microsatellites, s'ils sont correctement choisis et couplés à des méthodes d'analyses haut-débit, sont donc de puissants outils pour identifier les espèces. Mais bien que parfaitement validés sur matériel végétal frais, le transfert sur bois de ces

marqueurs s'est avéré difficile. Les principales études génétiques sur bois de chêne concernaient jusqu'à ce jour l'origine géographique des échantillons (Dumolin-Lapègue *et al.*, 1999; Deguilloux *et al.*, 2002; Deguilloux *et al.*, 2003; Deguilloux *et al.*, 2004; Deguilloux *et al.*, 2006). Ces études ciblaient uniquement le génome chloroplastique, suffisamment informatif pour confirmer l'origine géographique (Petit *et al.*, 2002). L'obligation de travailler sur le génome nucléaire pour identifier l'espèce d'un échantillon de bois a considérablement compliqué les choses, comme détaillé dans le **Chapitre 3**. Le génome nucléaire, présent en seulement deux copies par cellule (contre plusieurs centaines de copies pour le génome chloroplastique) est d'autant plus difficile à amplifier qu'il se fragmente rapidement avec le temps (Bär *et al.*, 1988; Lindahl, 1993; Cano, 1996). J'ai confirmé dans le **Chapitre 3** qu'une des étapes clés du succès de ces analyses se situait au niveau de l'extraction de l'ADN. Comme théoriquement une seule copie d'ADN suffit à réaliser une amplification par PCR, la quantité d'ADN extrait sur bois est assez peu limitante. Cependant, plus l'ADN ciblé est en faible quantité, plus le risque d'amplifier de l'ADN contaminant est important. D'autre part, les inhibiteurs de PCR doivent être éliminés pour obtenir de l'ADN ayant une qualité suffisante pour permettre une amplification efficace et complète du fragment d'ADN ciblé. Les protocoles le plus efficaces pour purifier l'ADN sont aussi ceux qui ont les plus faibles rendements en ADN. Il a donc fallut trouver un équilibre entre purification et rendement. L'utilisation de la PCR en temps réel sur ADN dégradé est assez courante (Poinar *et al.*, 2003; Gugerli *et al.*, 2005) mais elle sert généralement à observer les altérations de l'ADN sur des échantillons à authentifier plus ou moins anciens (Hofreiter *et al.*, 2001; Alonso *et al.*, 2004; Schwarz *et al.*, 2009), ou à détecter des contaminations (Pruvost & Geigl, 2004). A contrario, cette technique est encore peu utilisée pour optimiser directement les protocoles d'extraction (Vural, 2009). Or j'ai pu démontrer dans le **Chapitre 3** que la PCR en temps réel sur génome chloroplastique est une technique indirecte efficace pour optimiser les protocoles d'extraction en vue d'analyses génétiques sur génome nucléaire. Le deuxième point clé de l'identification d'espèce d'un échantillon de bois se situe au niveau des marqueurs moléculaires utilisés. Sur les 20 marqueurs microsatellites ayant servi à développer la base de données génétiques de référence (**Chapitre 2**), seuls sept marqueurs ont pu être transférés sur bois. L'ADN dégradé pose de nombreux problèmes lors de l'analyse de motifs répétés : amplification préférentielle d'un allèle (« *allelic drop-out* ») et bandes parasites ponctuelles

(Soulsbury *et al.*, 2007; Tvedebrink *et al.*, 2009) sont fréquents. L'obtention d'un génotype complet pour un échantillon de bois a nécessité parfois plusieurs essais successifs (augmentant par la même occasion le coût) mais a permis cependant de confirmer avec précision l'espèce annoncée d'échantillons tests fournis par le Centre de Recherche Pernod Ricard. Dans ces conditions, il apparait quand même difficile de transférer cette technique basée sur les marqueurs microsatellites pour une application dans un contexte industriel.

Les microsatellites supplantés par les SNPs ?

Même si le nombre d'études utilisant les microsatellites continue de croître, les marqueurs moléculaires les plus utilisés sont dorénavant les SNPs. Ces marqueurs ont comme principal avantage d'être très fréquents dans le génome, en moyenne toutes les 50 pb chez *Quercus* (Pauline Garnier-Géré, communication personnelle). Avec les techniques de séquençage nouvelle génération (voir **Chapitre 1**), il est possible d'isoler des milliers de SNPs sur des espèces non-modèles à des coûts modérés (Ekblom & Galindo, 2010; Helyar *et al.*, 2011; Neale & Kremer, 2011). Quel que soit le type de marqueurs moléculaires, disposer de nombreux marqueurs permet d'être plus exigeant dans la sélection, en prenant en compte des critères techniques (qualité de l'amplification) et biologiques (marqueurs les plus différenciés entre espèces). Disposer de près de 13000 SNPs validés pour développer des méthodes d'identification d'espèce m'a permis de détecter les meilleurs loci pour différencier les deux espèces de chêne. A titre d'exemple, le SNP le plus différencié entre *Q. robur et Q. petraea* (Stress_WZ0AQRAQ4YD18FM1_02_477) a un F_{ST} de 0.85 alors que le microsatellite le plus différencié entre ces même espèces (PIE227) a un F_{ST} de « seulement » 0.21. Dans la mesure où ils peuvent être détectés plus facilement en très grand nombre, les SNPs apparaissent donc comme des marqueurs plus efficaces que les microsatellites pour identifier les espèces de chênes. Aucun marqueur microsatellite n'est diagnostique pour ces deux espèces alors que plusieurs SNPs le sont quasiment au niveau génotypique (Figure 1).

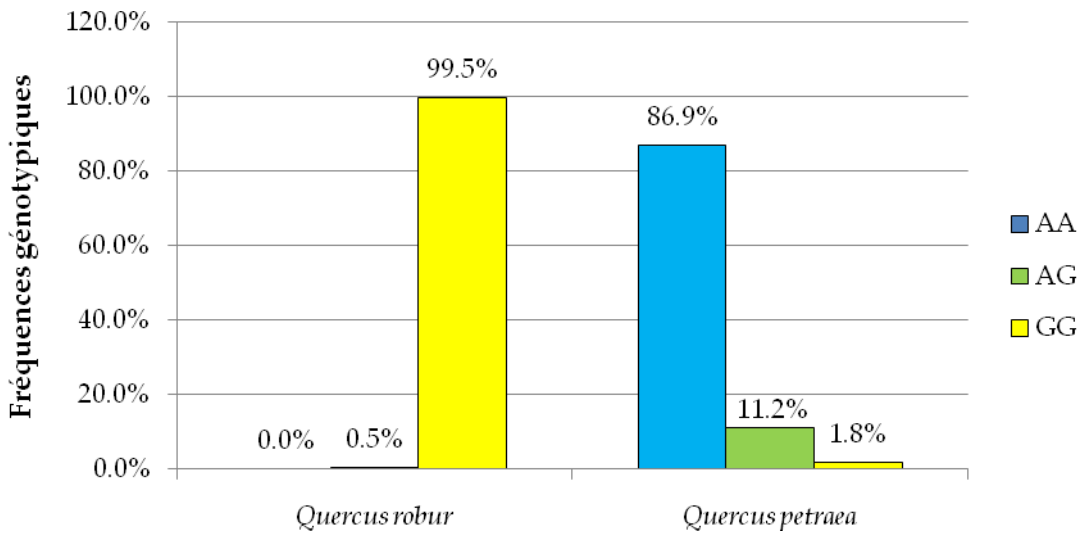


Figure 1 : Fréquences génotypiques chez *Q. robur* et *Q. petraea* au SNP « Stress_WZ0AQRAQ4YD18FM1_02_477 », sur 787 individus des deux espèces. Les deux espèces ont été délimitées par des méthodes d'affectation sur la base des génotypes multilocus des individus (262 SNPs).

Les méthodes d'identification d'espèces ou de populations basées sur l'affectation de génotypes multilocus (Pritchard *et al.*, 2000; Falush *et al.*, 2003) nécessitent généralement plus de SNPs que de microsatellites, car des marqueurs multi-alléliques sont plus informatifs que des marqueurs di-alléliques (Paschou *et al.*, 2007; Glover *et al.*, 2010; Haasl & Payseur, 2011). Mais si seuls les meilleurs marqueurs sont utilisés, alors les SNPs peuvent s'avérer plus efficaces que les microsatellites pour ce type d'approches (Liu *et al.*, 2005). Dans le cas présent, avec très peu de marqueurs SNPs très différenciés entre espèces, les résultats d'affectation génétique sont rapidement proches de l'optimum (94% des individus purs de chaque espèce sont bien affectés avec les deux meilleurs SNPs, voir **Chapitre 4**).

D'un point de vue technique, les SNPs sont plus adaptés à l'ADN dégradé que les microsatellites. Les fragments amplifiés peuvent être réduits au maximum (jusqu'à 45pb contre au minimum 70 à 80pb pour les microsatellites) et les erreurs associées au génotypage sont réduites car il n'y a généralement que deux allèles possibles. Autre avantage, de très nombreuses techniques de génotypage sont disponibles pour ce type de marqueurs, généralement haut-débit et souvent beaucoup moins coûteuses que les analyses classiquement utilisées pour les microsatellites, basées sur des amorces fluorescentes spécifiques. Des techniques utilisant la spectrométrie de masse (Sequenom) ou l'analyse des courbes de fusion (HRM - *High Melting Resolution*) semblent particulièrement adaptées à un

contexte industriel, pour lequel il est nécessaire de disposer de méthodes robustes et économiques. Les premiers tests de génotypage SNP sur des échantillons de bois (18 mois de séchage) réalisés à la fin de cette thèse se sont d'ailleurs révélés concluants. Il reste cependant à valider ces méthodes d'identification d'espèce sur des échantillons plus récalcitrants utilisés par la filière. Ainsi, de nouvelles améliorations techniques devront être poursuivies pour analyser des douelles ou des copeaux préalablement chauffés, l'ADN se dégradant très vite sous l'effet de la chaleur (Threadgold & Brown, 2003; Bonnet *et al.*, 2009). Pour ces applications difficiles, les technologies de séquençage nouvelle génération de fragments très courts, qui sont déjà très largement répandues dans le domaine de l'ADN ancien (Knapp & Hofreiter, 2010), semblent particulièrement adaptées.

Perspectives appliquées pour la filière bois

Les SNPs présentent de nombreux avantages pour le genre d'applications pratiques envisagées ici. Ils sont abondants, peu coûteux à isoler, relativement faciles à génotyper, ils peuvent être très puissants pour différencier des espèces génétiquement proches. Il apparait donc fort probable que les analyses génétiques sur bois pour identifier les espèces ou l'origine géographique se fassent désormais à l'aide de marqueurs SNPs, les limites liées à l'extraction d'ADN étant indépendantes du type de marqueur utilisé. De plus, d'un point de vue très appliqué pour la filière bois, seuls des marqueurs di-alléliques comme les SNPs permettent de développer des méthodes innovantes pour tester la conformité d'un lot de bois à une espèce (ou à une origine géographique). Je me suis inspiré des tests diagnostiques utilisés en médecine pour évaluer l'efficacité des traitements médicaux pour mettre au point des tests diagnostiques d'espèce pour les marqueurs SNPs (**Annexe 5**). Pour chaque marqueur, à l'aide des fréquences alléliques ou génotypiques au sein des deux espèces, il est possible de quantifier les erreurs associées à chaque test (déclarer conforme un échantillon qui ne l'est pas – déclarer non-conforme un échantillon qui l'est). Cet effort de caractérisation des erreurs associées à un test diagnostique s'intègre bien dans une optique appliquée, pour les industriels de la filière mais également pour les gestionnaires forestiers (ONF par exemple). De tels outils peuvent servir de contrôle a posteriori des lots de bois, mais peuvent aussi être utilisés comme outil de certification a priori. Enfin, au-delà de la conformité avec une espèce ou une origine géographique pour le seul genre *Quercus*, cette méthodologie, tout comme les améliorations techniques liées à l'extraction d'ADN sur bois, pourra s'appliquer à

de nombreuses autres espèces. Le seul point limitant est la nécessité de développer une base de données génétiques la plus complète possible, même si cette contrainte devrait être dépassée grâce à la généralisation des technologies de séquençage nouvelle génération. On peut donc imaginer que de tels outils de contrôle permettront de limiter la fraude et le commerce illégal de bois.

A court terme, la caractérisation des espèces sur des lots de bois se fera donc vraisemblablement à l'aide de marqueurs SNPs. Ainsi, les professionnels en aval de la filière (tonneliers, vignerons et œnologues) disposeront d'outils fiables pour anticiper une partie du potentiel aromatique de leur bois. A moyen terme et avec les avancées constantes dans le domaine de la génomique, la caractérisation du potentiel aromatique d'un bois pourra être encore améliorée, en ciblant par exemple des gènes directement impliqués dans le contrôle de l'expression des molécules aromatiques majeures (whisky-lactone, vanilline, eugénol, ...).

L'apport significatif des marqueurs sous sélection

Au-delà de leur capacité à différencier finement ces deux espèces de chênes, les SNPs m'ont également permis de mieux caractériser les flux de gènes chez *Q. robur* et *Q. petraea* (**Chapitre 4**). A ce jour, aucune étude n'a mis en évidence l'intérêt des marqueurs soumis à sélection pour étudier les flux de gènes. Par contre, l'intérêt de ces gènes selectionnés pour mettre en évidence la structure génétique commence à être bien établi (Nielsen *et al.*, 2007; O'Malley *et al.*, 2007; Westgaard & Fevolden, 2007; Gebremedhin *et al.*, 2009; Nielsen *et al.*, 2009; Andre *et al.*, 2010). Le fait, comme ce fut mon cas, de disposer d'un très grand nombre de marqueurs, chose difficilement envisageable avant l'apparition des technologies de séquençage nouvelle génération, permet d'étudier séparément les loci supposés neutres et les loci « outliers » fortement différenciés entre espèces, tout en conservant des effectifs suffisants pour disposer d'une puissance intéressante (**Chapitre 4**). Mes résultats indiquent que la proportion de loci « outliers » est très importante (23%). Ces travaux démontrent que les SNPs soumis à sélection permettent d'étudier des processus démographiques complexes, bien mieux que des marqueurs présumés neutres, et permettent de valider certains modèles d'évolution (Petit *et al.*, 2004). Ce résultat était complètement inattendu. Les flux de gènes étudiés ici se situent aux niveaux inter- et intraspécifique, mais il est probable que les avancées dans le domaine de la génomique des populations permettront à court terme d'étudier des processus démographiques plus fins, entre populations proches d'une même

espèce, ou entre individus d'une même population qui présentent une variabilité pour certains caractères (études d'association génome entier). Grâce aux méthodes de séquençage nouvelle génération qui permettent d'isoler des milliers de SNPs sur des espèces non-modèles, il est vraisemblable que ce genre d'approches basée sur les « outliers » se développe dans les années à venir sur d'autres modèles biologiques (jusqu'ici seules certaines espèces de poissons, et désormais les chênes, avaient été étudiées dans cette optique). Cela transformera en profondeur les approches de génétique des populations et en démo-génétique, permettant une meilleure compréhension des processus d'évolution qui affectent les génomes.

REFERENCES

Alonso A, Martín P, Albarrán C, *et al.* (2004) Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies. *Forensic Science International* **139**, 141-149.

Andre C, Larsson LC, Laikre L, *et al.* (2010) Detecting population structure in a high gene-flow species, Atlantic herring (*Clupea harengus*): direct, simultaneous evaluation of neutral vs putatively selected loci. *Heredity* **106**, 270-280.

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics* **19**, 1436-1438.

Bär W, Kratzer A, Mächler M, Schmid W (1988) Postmortem stability of DNA. *Forensic Science International* **39**, 59-70.

Bonin A, Bellemain E, Eidesen PB, *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261-3273.

Bonnet J, Colotte M, Coudy D, *et al.* (2009) Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Research* **38**, 1531-1546.

Cano RJ (1996) Analysing ancient DNA. *Endeavour* **20**, 162-167.

Deguilloux MF, Bertel L, Celant A, *et al.* (2006) Genetic analysis of archaeological wood remains: first results and prospects. *Journal of Archaeological Science* **33**, 1216-1227.

Deguilloux MF, Pemonge MH, Bertel L, Kremer A, Petit RJ (2003) Checking the geographical origin of oak wood: molecular and statistical tools. *Molecular Ecology* **12**, 1629-1636.

Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proceedings of the Royal Society B-Biological Sciences* **269**, 1039-1046.

Deguilloux MF, Pemonge MH, Petit RJ (2004) DNA-based control of oak wood geographic origin in the context of the cooperage industry. *Annals of Forest Science* **61**, 97-104.

Dow BD, Ashley MV, Howe HF (1995) Characterization of highly variable (GA/CT)n microsatellites in the bur oak, *Quercus macrocarpa*. *Theoretical and Applied Genetics* **91**, 137-141.

Dumolin-Lapègue S, Pemonge MH, Gielly L, Taberlet P, Petit RJ (1999) Amplification of oak DNA from ancient and modern wood. *Molecular Ecology* **8**, 2137-2140.

Durand J, Bodénès C, Chancerel E, *et al.* (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* **11**, 570.

Ekblom R, Galindo J (2010) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **1**.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Gebremedhin B, Ficetola GF, Naderi S, *et al.* (2009) Frontiers in identifying conservation units: from neutral markers to adaptive genetic variation. *Animal Conservation* **12**, 107-109.

Glover KA, Hansen MM, Lien S, *et al.* (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics* **11**, 2.

Gugerli F, Parducci L, Petit RJ (2005) Ancient plant DNA: review and prospects. *New Phytologist* **166**, 409-418.

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* **106**, 158-171.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, no-no.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* **14**, 599-612.

Hofreiter M, Jaenicke V, Serre D, Haeseler Av, Paabo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* **29**, 4793-4799.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* **46**, 511-517.

Kampfer S, Lexer C, Glössl J, Steinkellner H (1998) Characterization of (GA)n microsatellite loci from *Quercus robur*. *Hereditas* **129**, 183-186.

Knapp M, Hofreiter M (2010) Next generation sequencing of ancient DNA: Requirements, strategies and perspectives. *Genes* **1**, 227-243.

Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715.

Liu NJ, Chen L, Wang S, Oh CG, Zhao HY (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**, S26.

Luikart G, Zundel S, Rioux D, *et al.* (2008) Low genotyping error rates and noninvasive sampling in bighorn sheep. *Journal of Wildlife Management* **72**, 299-304.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* **12**, 111-122.

Nielsen, Einar E, Mackenzie, *et al.* (2007) Historical analysis of Pan I in Atlantic cod (*Gadus morhua*) : temporal stability of allele frequencies in the southeastern part of the species distribution. *Canadian journal of fisheries and aquatic sciences* **64**, 1448-1455.

Nielsen EE, Hemmer-Hansen J, Poulsen NA, *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology* **9**, 11.

O'Malley KG, Camara MD, Banks MA (2007) Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology* **16**, 4930-4941.

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180.

Paschou P, Ziv E, Burchard EG, *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet* **3**, 1672-1686.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2004) Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164.

Petit RJ, Brewer S, Bordacs S, *et al.* (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *Forest Ecology and Management* **156**, 49-74.

Poinar H, Kuch M, McDonald G, Martin P, Pääbo S (2003) Nuclear gene sequences from a late Pleistocene sloth coprolite. *Current biology* **13**, 1150-1152.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847-846.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pruvost M, Geigl E-M (2004) Real-time quantitative PCR to assess the authenticity of ancient DNA amplification. *Journal of Archaeological Science* **31**, 1191-1197.

Schwarz C, Debruyne R, Kuch M, *et al.* (2009) New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research* **37**, 3215-3229.

Shen Z, Qu W, Wang W, *et al.* (2010) MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics* **11**, 143.

Soulsbury C, Iossa G, Edwards K, Baker P, Harris S (2007) Allelic dropout from a high-quality DNA source. *Conservation Genetics* **8**, 733-738.

Steinkellner H, Fluch S, Turetschek E, *et al.* (1997) Identification and characterization of (GA/CT)n microsatellite loci from *Quercus petraea*. *Plant Molecular Biology* **33**, 1093-1096.

Threadgold J, Brown TA (2003) Degradation of DNA in artificially charred wheat seeds. *Journal of Archaeological Science* **30**, 1067-1076.

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N (2009) Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International-Genetics* **3**, 222-226.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23**, 48-55.

Vural HC (2009) Quantification and presence of human ancient DNA in burial place remains of Turkey using real time polymerase chain reaction. *African Journal of Biotechnology* **8**, 5163-5168.

Westgaard JI, Fevolden SE (2007) Atlantic cod (*Gadus morhua* L.) in inner and outer coastal zones of northern norway display divergent genetic signature at non-neutral loci. *Fisheries Research* **85**, 306-315.

ANNEXES

ANNEXE 1

 doi: 10.1111/j.1365-294X.2009.04137.x

Species relative abundance and direction of introgression in oaks

O. LEPAIS,*†‡ R. J. PETIT,*† E. GUICHOUX,*† J. E. LAVABRE,*†§ F. ALBERTO,*† A. KREMER*† and S. GERBER*†

*INRA, UMR 1202 BIOGECO, 69 route d'Arcachon, F-33612 Cestas cedex, France, †Université de Bordeaux, UMR 1202 BIOGECO, 69 route d'Arcachon, F-33612 Cestas cedex, France

Abstract

Successful hybridisation and subsequent introgression lead to the transfer of genetic material across species boundaries. In this process, species relative abundance can play a significant role. If one species is less abundant than the other, its females will receive many heterospecific gametes, increasing mate-recognition errors and thus hybridisation rate. Moreover, first-generation hybrids will also more likely mate with the more abundant species, leading to asymmetric introgression. These predictions have important fundamental consequences, especially during biological invasions or when a rare species threatened by extinction is surrounded by individuals from a related species. However, experimental tests in nature of the importance of the relative abundance of each species on hybridisation dynamics remain scarce. We assess here the impact of species relative abundance on hybridisation dynamics among four species from the European white oak species complex. A total of 2107 oak trees were genotyped at 10 microsatellite markers and Bayesian clustering methods were used to identify reference trees of each species. We then used these reference trees to simulate purebred and hybrid genotypes to determine optimal threshold for genetic assignment. With this approach, we found widespread evidence of hybridisation between all studied oak species, with high occurrence of hybrids, varying from 11% to 31% according to stand and sampling strategies. This finding suggests that hybridisation is a common phenomenon that plays a significant role in evolution of this oak species complex. In addition, we demonstrate a strong impact of species abundance on both hybridisation rate and introgression directionality.

Keywords: frequency-dependent process, genetic assignment, hybridisation, microsatellites, *Quercus*, species delimitation

Received 10 November 2008; revision received 16 January 2009; accepted 26 January 2009

Introduction

Interspecific mating associates heterogeneous genomes, giving rise to new allelic combinations (Rieseberg & Carney 1998). When hybridisation is successful, first-generation hybrids may mate with parental species, producing back-crossed individuals. This leads to gene introgression with transfer of genetic material across species boundaries (Anderson 1949; Martinsen *et al*. 2001; Kim *et al*. 2008). Hybridisation and introgression imply some contact between species so that mating can occur. It has long been argued that local species abundance will impact hybridisation dynamics (Hubbs 1955; Mayr 1963). The rationale is that in species where females exert male choice through prezygotic isolation, hybridisation rate will increase when species relative abundances become sharply unbalanced, because the females belonging to the rare species then receive too many heterospecific gametes and are more likely to make mate-recognition errors (Wirtz 1999; Chan *et al*. 2006). Such a mechanism, sometimes called Hubbs' principle, has been hypothesised in animals (reviewed by Rhymer & Simberloff 1996; Wirtz 1999) and in plants (reviewed by Rieseberg 1997). Differences

Correspondence: Gerber Sophie, UMR BIOGECO, Genetic Team, 69 route d'Arcachon, F-33612 Cestas, France. Fax: +33(0)557122881; E-mail: sophie@pierroton.inra.fr

‡Present address: School of Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, Scotland, UK.

§Present address: Estacion Biologica de Donana, CSIC, Integrative Ecology Group, Pabellon del Peru, Avda. M. Luisa S/N, E-41013 Sevilla, Spain.

© 2009 Blackwell Publishing Ltd

in species proportion could have consequences beyond the first hybrid generation. This is because first generation hybrids (F_1) will also be more likely to mate with the more abundant species, producing backcrossed individuals that will be more similar to the common species (Anderson & Hubricht 1938; Rieseberg 1997). The validity of Hubbs' prediction is interesting to check because it has important practical and fundamental consequences. For instance, if the minority species is represented by only few individuals that produce a high proportion of hybrids, the species might become locally extinct, by pollen swamping and dilution of the genome of the rare species, although its genes will persist at least temporarily in hybrid individuals (Levin *et al*. 1996; Rhymer & Simberloff 1996). Another situation where species proportion can be highly unbalanced is when a colonising species spreads in an area already occupied by a related species. In this case, the invading species is initially rare, and matings with the local species are likely. Genetic material of the local species incorporated into the invading species can then reach high frequency as the invading population experiences rapid demographic growth, resulting in asymmetric introgression of neutral genes (Currat *et al*. 2008). Clearly, species relative abundance can have important consequences on hybridisation dynamics, affecting both hybridisation rates and the direction of introgression. Although some researchers have acknowledged the fact that species proportion can play an important role in introgression dynamics, only few have experimentally demonstrated its reality in nature (e.g. Buggs 2007; but see Burgess *et al*. 2005; Prentis *et al*. 2007; Field *et al*. 2008; Zhou *et al*. 2008). Additional empirical surveys addressing this issue with different organisms are therefore needed.

Hybridisation has been intensively studied in the genus *Quercus* (Arnold 2006). In particular, hybridisation and introgression are suspected to play a role in postglacial recolonisation of Europe by oaks (Petit *et al*. 2003). Detailed studies of mating system of the two species involved (*Quercus robur* and *Quercus petraea*) in controlled crosses (Steinhoff 1993; Steinhoff 1998; Kleinschmit & Kleinschmit 2000) or in natural populations (Bacilieri *et al*. 1996; Streiff *et al*. 1999) have shown that prezygotic and postzygotic barriers exist, but few studies have focused on the consequences of species abundance on hybridisation dynamics within this species complex. In one recent study, hybridisation rate between two oak species (*Q. petraea* and *Q. pyrenaica*) seemed unrelated to species relative abundance, but the number of investigated stands was limited (Valbuena-Carabaña *et al*. 2007). While oak species are only weakly genetically differentiated, they present important morphological and ecological differences. In forests where several oak species are found in sympatry, species are often clustered according to their ecological requirements (Bacilieri *et al*. 1995). Thus, relative proportions of oak species are expected to vary between stands as a result of local ecological conditions as well as stand history (including forest management). These species represent therefore a good model to test the hypothesis that species proportion affects hybridisation and introgression.

In this study, we adopted a blind (i.e. no a priori classification) approach (Duminil *et al*. 2006) to assign oaks to species and identify hybrids using microsatellite markers and Bayesian clustering methods. We analysed several populations from the four most common species of the European white oak complex in France. We first applied a clustering analysis to all trees studied and then used the results to identify reference trees of each species. These were used to generate artificial genotypes of known ancestry (pure species, hybrids and backcrosses) to determine objective and optimal thresholds for genetic assignment. We analysed several populations and stands with different species composition. This allowed us to test whether relative species abundance influences hybridisation dynamics in this species complex. The specific aims of this paper are (i) identifying hybrid individuals, (ii) estimating the pattern of hybridisation across species and populations, and (iii) testing the effect of parental species proportions on hybridisation rate and introgression.

Materials and methods

Species description

Four oak species were included in this study: *Quercus robur* L. (pedunculate oak), *Q. petraea* (Matt.) Liebl. (sessile oak), *Q. pubescens* Willd. (pubescent or downy oak) and *Q. pyrenaica* Willd. (Pyrenean or rebollo oak). *Quercus robur* and *Q. petraea* are widely distributed in Europe. *Quercus pyrenaica* is found along the Atlantic coast from Morocco and northwestern Spain to western France. *Quercus pubescens* is localised around the Mediterranean Basin with a northern latitudinal limit up to 50 degrees. Distribution range and local species presence are governed by climatic and edaphic factors (Rameau *et al*. 1989). In brief, *Q. pubescens* grows on limestones and in thermophilous stations, whereas *Q. pyrenaica* prefers sandy acidic soils. *Q. robur* is found on rich and deep soils and can support flooding, unlike the other oak species, while *Q. petraea* is found on poorer and dryer soils. Whereas the other three oak species are postpioneer species capable of colonising open land, *Q. petraea* is a late-successional species that grows in stable and well-established forest environment. Thus in the Aurignac region, composed of small forests and woodlands (see below), *Q. petraea* is found in the centre of the stands (Gonzalez *et al*. 2008). The species are traditionally identified during the growing season by examining leaf morphology. *Quercus robur* leaves have short petioles, several secondary veins and their basal parts are typically lobated (Kremer *et al*. 2002). *Quercus petraea* leaves have a longer petiole, no secondary veins and a regular leaf shape. *Quercus*

© 2009 Blackwell Publishing Ltd

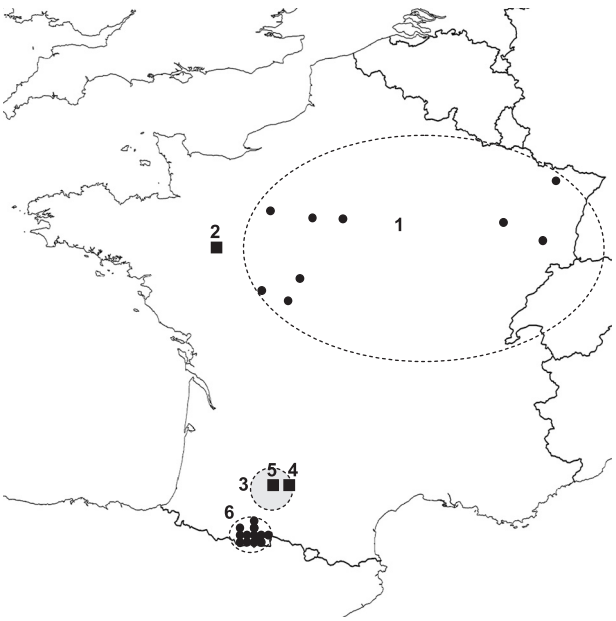


Fig. 1 Location map of the intensively studied stands (squares) and the other sampled populations (encircled) in France (see Table S1 for more details). 1, ONF (National Forest Office) populations; 2, Petite Charnie stand; 3, Aurignac region; 4, Briouant stand; 5, Paguères stand; and 6, Pyrenean populations.

pubescens is similar to *Q. petraea* but the leaves have a higher number of lobes and the abaxial part is densely hairy (Dupouey & Badeau 1993; Curtu *et al*. 2007). *Quercus pyrenaica* leaves are hairy on both sides and have a particular leaf shape with numerous lobes and deep sinuses.

Sampling strategy

A total of 2107 oak trees belonging to the species complex described above were sampled in 53 populations in France (Fig. 1, Table S1, Supporting information). This material had been sampled in the frame of several studies with different objectives, so the sampling strategies are contrasted. The large size of the combined data set should help improve assignment tests (Pritchard *et al*. 2000; Waples & Gaggiotti 2006). In three regions, 10–79 individuals were collected from many populations in France: in the south (Aurignac and Pyrenean stands) and in the north (ONF stands), representing a total of 889 individuals in 50 populations (see Table S1 for more details). In the other areas, stands were more intensively sampled with two stands exhaustively collected, regardless of leaf morphology (Petite Charnie and Briouant) and a third one regularly sampled along a grid (Paguères).

ONF populations consisted in high forests composed mostly of *Q. robur* and *Q. petraea*. Oaks showing typical species morphology were sampled whenever possible. Pyrenean populations were sampled in two valleys at an altitude ranging form 100 to 1600 m. Only *petraea*-like individuals were collected in this study. The Petite Charnie stand has been intensively studied for a long time (Bacilieri *et al*. 1995; Streiff *et al*. 1998; Streiff *et al*. 1999) and only *Q. robur* and *Q. petraea* have been described in this stand, which is part of a continuous high forest. In Aurignac, oak trees showing typical morphology of all three locally abundant oak species (*Q. robur*, *Q. petraea* and *Q. pubescens*) were collected. We sampled one to three individuals by stand (2.5 on average) in 29 forest fragments located within a radius of 30 km around Paguères stand. Briouant and Paguères are two coppice stands localised with Aurignac populations in the long term Ecological Research (LTER-Europe) site 'Vallées et Coteaux de Gascogne'. Paguères includes *Q. robur*, *Q. pubescens* and few *Q. petraea* oaks whereas in Briouant *Q. pyrenaica* is the most frequent species, followed by *Q. robur*, *Q. pubescens* and only few *Q. petraea*.

Two leaves per tree were sampled and kept at 4 °C until stored at –80 °C in the laboratory or immediately dried in silica gel and kept at room temperature. Global positioning system coordinates and morphological species identification using the morphological criteria described above were recorded for each collected tree. Moreover, a detailed morphological analysis was available for the trees from the Petite Charnie (Bacilieri *et al*. 1995) and Briouant (Viscosi *et al*. 2009). Either a discriminant function based on two morphological characters (Kremer *et al*. 2002) was used to distinguish *Q. robur* and *Q. petraea* in the ONF stands or 10 morphological characters were measured to perform a morphological analysis in the case of Pyrenean populations (E. Guichoux, unpublished data and F. Alberto, unpublished data, respectively). When species status was uncertain, oaks were recorded as undetermined species.

Genetic analyses

DNA isolation was performed with a cetyltrimethyl ammonium bromide (CTAB) protocol as previously described (Lepais *et al*. 2006) except for the ONF populations for which the QIAGEN DNeasy Plant Mini Kit was used following the manufacturer's instructions. Ten microsatellite loci selected for their relatively high degree of genetic differentiation between species (Scotti-Saintagne *et al*. 2004; P. G. Goicoechea, unpublished data) were analysed using a multiplex protocol (Lepais *et al*. 2006). Briefly, two polymerase chain reaction were carried out with an MJ Research DNA Engine Tetrad2 thermocycler to amplify the 10 microsatellites: QpZAG110 (Steinkellner *et al*. 1997), QrZAG11, QrZAG112, QrZAG39, QrZAG96, QrZAG7, QrZAG87, QrZAG65, QrZAG5, QrZAG20 (Kampfer *et al*. 1998). Amplified fragments were analysed with an Amersham MegaBace1000 capillary sequencer and individual genotypes were determined with the Fragment Profiler software version 1.2 using the same parameters for all populations.

© 2009 Blackwell Publishing Ltd

Table 1 Number of simulated individuals (rows) assigned to the different species or hybrid classes (columns) and computed efficiency, accuracy and global performance of the assignment method (at the bottom). Correct assignments are highlighted in bold

Simulated/assigned	Rob	Pet	Pyr	Pub	Hyb RobPet	Hyb RobPyr	Hyb RobPub	Hyb PetPyr	Hyb PetPub	Hyb PyrPub	Total
Rob	**996**				3	1					1000
Pet		**988**			7			4	1		1000
Pyr			**992**			3		2		3	1000
Pub				**972**			4		14	10	1000
F_1_RobPet		3			**27**						30
bc_RobPet	19	1			**38**	2					60
bc_PetRob		20			**38**			2			60
F_1_RobPyr			1			**28**				1	30
bc_RobPyr	15				2	**40**	3				60
bc_PyrRob	1		16			**41**		1		1	60
F_1_RobPub				1			**29**				30
bc_RobPub	17				4	1	**38**				60
bc_PubRob				11	1		**44**		1	3	60
F_1_PetPyr			1					**29**			30
bc_PetPyr		16			2	1		**38**	3		60
bc_PyrPet			23			2		**32**		3	60
F_1_PetPub							1		**29**		30
bc_PetPub		19	1		1			4	**35**		60
bc_PubPet				29			2		**27**	2	60
F_1_PyrPub								1	2	**27**	30
bc_PyrPub			16			1		4	1	**38**	60
bc_PubPyr			1	25				1	2	**31**	60
Total	1048	1047	1051	1038	123	120	121	118	115	119	4900
Efficiency (percentage)	99.6	99.8	99.2	97.2	68.7	72.7	74.0	66.0	60.7	64.0	
Accuracy (percentage)	95.0	94.4	94.4	93.6	83.7	90.8	91.7	83.9	79.1	80.7	
Performance (percentage)	94.7	93.2	93.6	91.0	57.5	66.0	67.9	55.4	48.0	51.6	

Hyb, hybrids; F_1, first generation hybrids; bc, backcrosses; Rob, *Q. robur*; Pet, *Q. petraea*; Pub, *Q. pubescens*; Pyr, *Q. pyrenaica*.

Admixture analyses

Bayesian clustering of the genetic data was performed using Structure version 2.1 (Pritchard *et al*. 2000; Falush *et al*. 2003). To determine the optimal number of groups (K), we ran Structure with K varying from 1 to 10, with 10 runs for each K value, to find the K value with the highest posterior probabilities. We also used the ΔK statistics to evaluate the change in likelihood (Evanno *et al*. 2005). Our parameters were 50 000 burn-in periods and 100 000 Markov chain Monte Carlo repetitions after burn-in with admixture and correlated allele models without any prior information. For the most likely number of clusters ($K = 4$), we calculated the average result over 10 runs to get the final admixture analysis.

Hybrid simulation and genetic assignment

For each of the four species, we selected at random 65 individuals that had high probabilities (admixture coefficient, $Q > 0.90$) to belong to each of the four corresponding clusters identified in the admixture analysis. This allowed us to estimate allelic frequencies of the four species. We then simulated pure species and hybrid genotypes using these allele frequencies and the R statistic software (R Development Core Team 2005). We simulated 1000 genotypes for each species, 30 F_1 hybrids and 60 backcrosses for all combinations of possible crosses between each pair of species. The number of simulated hybrids is somewhat arbitrary but reflects the expected hybrid percentage observed in real populations (see Results section). We analysed these simulated data set with the Structure software, with $K = 4$ and the same parameters as before, to test the performance of the software to distinguish between pure species and hybrids, and to determine thresholds to assign individuals to these categories to reach a high correct classification rate. We then assigned individuals with the determined threshold (see Results section) and computed efficiency (the proportion of correctly assigned individual), accuracy (the proportion of true hybrids or purebreds assigned in each hybrid or purebred classes) and overall performance (the product of efficiency and accuracy) of the assignment procedure (Vaha & Primmer 2006).

Distance-based analyses

Using the individual tree assignment results, we computed Cavalli-Sforza and Edwards genetic distances (DS;

© 2009 Blackwell Publishing Ltd

Cavalli-Sforza & Edwards 1967) between each pair of species or hybrid classes in each population (provided there were a minimum of 10 individuals) with the Populations software (Langella 1999). The resulting distance matrix was used to build an unrooted neighbour-joining tree using the R package APE (Analysis of Phylogenetics and Evolution; Paradis *et al.* 2004).

Hybridisation characteristics and direction of introgression

We further analysed the three intensively sampled stands (Briouant, Petite Charnie and Paguères) to characterise introgression between species. We first performed global analyses to check if there was a difference in the contribution of each species to hybridisation. For $K = 4$, each individual is characterised by a vector of four admixture coefficients. In each stand, we defined two groups of individuals: purebred (whatever their species) and hybrids. We then computed the average of each of the four individual admixture coefficient within groups, resulting in a vector of four averaged admixture coefficients for purebred and a vector of averaged admixture coefficients for hybrids. These two vectors characterised the global genetic composition of purebreds and hybrids in each stand. The null expectation was that each species would contribute to the hybrid gene pool in proportion to its abundance in the stand; that is, the global genetic composition of purebreds should be the same as the global genetic composition of hybrids. To test this hypothesis, we compared the differences between averaged admixture coefficients in purebred and in hybrids using a Student *t*-test. We then investigated the effect of species abundance on differences in genetic composition between hybrids and pure categories. We computed the difference between hybrids and purebreds of each averaged admixture coefficients, considered as an estimate of hybrid excess. This measure of hybrid excess was correlated to the corresponding species relative abundance and tested with a linear model using the R package effects (Fox 2003) to estimate the confidence interval of the linear regression.

We then performed a detailed analysis to test for an effect of parental species relative abundance on introgression directionality. In each stand, we grouped hybrid individuals in one of the six plausible hybrid classes (each characterised by their two parental species). We first computed the average admixture coefficient of each hybrid class in each stand. The genetic composition of each hybrid class is characterised by a vector of four averaged admixture coefficients, among them, the two corresponding to the parental species have a high value while the other have a very low value. We then computed parental species relative abundance for each hybrid class (ratio between the number of oaks of the most abundant parental species and the total number of oaks of the two parental species) in each stand

Table 2 Number (and percentage) of pure species and hybrid oaks as assigned by the Structure software in the different studied stands and populations

Population	Sampling strategy	*N*	Rob	Pet	Pub	Pyr	Hyb RobPet	Hyb RobPub	Hyb RobPyr	Hyb PetPub	Hyb PetPyr	Hyb PubPyr	Total species	Total hybrids
Briouant	Exhaustive, stand	807	240 (29.7%)	3 (0.4%)	83 (10.3%)	235 (29.1%)	28 (3.5%)	35 (4.3%)	48 (5.9%)	7 (0.9%)	40 (5.0%)	88 (10.9%)	561 (69.5%)	246 (30.5%)
Petite Charnie	Exhaustive, stand	262	128 (48.9%)	84 (32.1%)	–	–	15 (5.7%)	11 (4.2%)	12 (4.6%)	7 (2.7%)	5 (1.9%)	–	212 (80.9%)	50 (19.1%)
Paguères	Partial grid-based, stand	149	87 (58.4%)	1 (0.7%)	28 (18.8%)	–	12 (8.1%)	16 (10.7%)	–	2 (1.3%)	–	3 (2.0%)	116 (77.9%)	33 (22.1%)
Aurignac	Partial, 29 populations	75	24 (32.0%)	14 (18.7%)	29 (38.7%)	–	4 (5.3%)	1 (1.3%)	1 (1.3%)	1 (1.3%)	–	1 (1.3%)	67 (89.3%)	8 (10.7%)
Pyrenees	Partial, 12 populations	288	1 (0.3%)	223 (77.4%)	4 (1.4%)	–	31 (10.8%)	3 (1.0%)	1 (0.3%)	9 (3.1%)	15 (5.2%)	1 (0.3%)	228 (79.2%)	60 (20.8%)
ONF	Partial, 9 populations	526	117 (22.2%)	321 (61.0%)	2 (0.4%)	–	24 (4.6%)	11 (2.1%)	4 (0.8%)	26 (4.9%)	21 (4.0%)	–	440 (83.7%)	86 (16.3%)
Total		2107	597 (28.3%)	646 (30.7%)	146 (6.9%)	235 (11.2%)	114 (5.4%)	77 (3.6%)	66 (3.1%)	52 (2.5%)	81 (3.8%)	93 (4.4%)	1624 (77.1%)	483 (22.9%)

N indicates the number of sampled oaks; Hyb, hybrids; Rob, *Q. robur*; Pet, *Q. petraea*; Pub, *Q. pubescens*; Pyr, *Q. pyrenaica* and all hybrid classes between these species by pairs.

© 2009 Blackwell Publishing Ltd

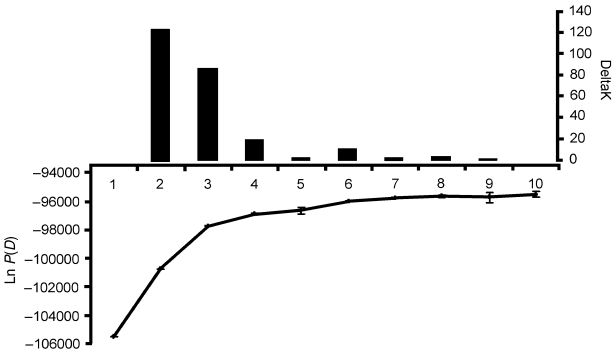


Fig. 2 Estimated number of populations (K) derived from the Structure clustering analyses. Mean and standard deviation probabilities of the data over 10 replicated runs (below) and ΔK (above) are plotted as a function of the number of clusters (K from 1 to 10).

and plotted it against the averaged admixture coefficient of hybrid class that corresponds to the most abundant parental species. If hybridisation is strictly bidirectional or restricted to the first generation (F_1 only), one would expect that the hybrids have an average admixture coefficient value of 0.5. However, if hybridisation is not restricted to the first generation and hybrids themselves can reproduce freely with their parental species, one would expect relative parental species abundance to affect hybrid genetic composition, that is, hybrids would be genetically more similar to the more frequent parental species.

Results

Admixture analysis

The likelihood of the partition of the data increased sharply from $K = 1$ to $K = 3$ and then increased only slightly from $K = 3$ to $K = 6$, where it reached a plateau (Fig. 2). The statistics ΔK indicates that $K = 2$ corresponds to the optimal number of groups, but the statistics also gives some support for $K = 3$ or even for $K = 4$ or $K = 6$. We thus report admixture results for $K = 2$, $K = 3$, $K = 4$ and $K = 6$ to compare them (Fig. 3). For $K = 2$, one cluster corresponds to *Quercus robur* (green) and the second to the three remaining morphological species. When adding a third cluster ($K = 3$), *Quercus petraea* is grouped into a specific cluster (yellow) while *Quercus pubescens* and *Q. pyrenaica* are grouped together in the third cluster (pink). For $K = 4$, we get different solutions depending on the run. In seven out of the 10 runs, each species is grouped in one cluster ($K = 4$, Fig. 3: *Q. robur* in the green cluster, *Q. petraea* in the yellow, *Q. pubescens* in the blue and *Q. pyrenaica* in the violet). The other solutions for $K = 4$ (not shown) group *Q. pubescens* and *Q. pyrenaica* in the same cluster while partitioning *Q. robur* and *Q. petraea* in three clusters. Finally, for $K = 6$, only one solution was found: *Q. pubescens* and *Q. pyrenaica* were distinguished as before but *Q. robur* and *Q. petraea* occupied two clusters each. This substructure in *Q. petraea* and *Q. robur* follows a north–south trend with one intraspecific cluster (dark green for *Q. robur* and brown for *Q. petraea*) more frequent in the northern populations while the other (light green for *Q. robur* and orange for *Q. petraea*) is more frequent among southern populations. The genetic distances between the intraspecific clusters are 10-fold smaller than the distances between clusters corresponding to different species, giving strength to the $K = 4$ clustering solution (Fig. S1, Supporting information).

Performance of assignment methods

Distribution of admixture coefficients (Q) of simulated individuals (Fig. S2, Supporting information) shows that a threshold value of 0.90 allows separating pure species from hybrids (including F_1 and backcrosses) with the lowest misclassification rate. We thus classified each individual with $Q > 0.90$ as pure species and $Q < 0.90$ as hybrids. However, individuals with $Q < 0.90$ for one cluster but $Q < 0.10$ for each of the three remaining clusters (2.1% of simulated individuals) were supposed to have the majority of their

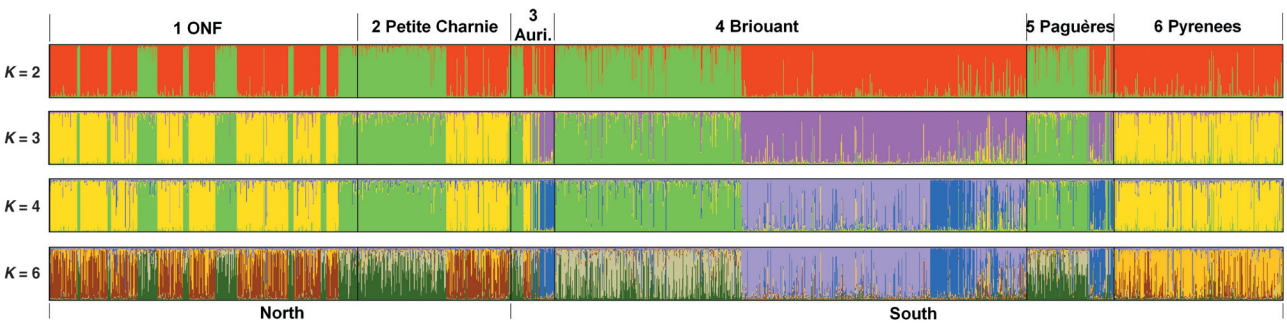


Fig. 3 Structure clustering results obtained for 2, 3, 4 and 6 clusters (K). Each individual is represented by a thin vertical line partitioned into K coloured segments proportional to its membership in the corresponding genetic cluster. Black lines separate individuals from different populations as indicated at the top, classified according to their latitude, indicated at the bottom. Within populations, individuals are grouped according to their species morphological aspect as determined in the forest (information not used in the clustering analysis).

© 2009 Blackwell Publishing Ltd

genome from one species without any significant influence from other species, and they were thus also classified as pure species (changing this rule did not affect the main conclusions of this work, results not shown). For hybrids, we considered that the two species with the highest assignment probability correspond to the hybrid parental species, whatever the probabilities of the third cluster (i.e. the existence of tri-hybrid individuals was ruled out). Note however, that among assigned hybrids 4.9% show a significant contribution ($Q > 0.10$) from a third cluster. Nevertheless, this assignment strategy provides high efficiency and accuracy (Table 1). The overall performance of the method varies from 94.7% to 91.0% depending on the species. Only 0.4% of pure simulated *Q. robur* individuals are wrongly assigned to a hybrid class but the proportion reaches 2.8% for pure *Q. pubescens*. The overall performance is lower for hybrid identification. The majority of simulated F_1 hybrids are correctly assigned to their hybrid class but simulated first-generation backcrosses often fall into the corresponding pure species category (Table 1, Fig. S2). This results in a decrease in the accuracy of pure species identification and in the efficiency of hybrid assignment, as 32% of these backcrosses are wrongly assigned to a pure species class. However, these wrongly assigned individuals are always classified into their parental species class (the species to which the hybrid is backcrossed). Moreover, 2.7% of F_1 and 6.9% of backcrosses are assigned to another hybrid class. Overall, this strategy should result in a conservative approach to hybrid identification (high accuracy at the expense of a decreased efficiency).

Hybridisation between oak species across populations

We assigned all individuals from natural populations using the method indicated above. Among the 1624 trees assigned to pure species, 226 (14%) showed signs of slight introgression (less than 0.90 probability to belong to their own species but less than 0.10 probability to belong to any other species). Among the 483 assigned hybrids (23%), 96 (20%) have a probability, higher than 0.10, to belong to a third species. Those individuals that escape the strict 0.90 threshold rule are far more numerous than in the case of simulated individuals (2.1% and 4.9% in simulated genotypes, respectively, as described above). This result indicates that in real populations, interspecific crosses may be more complex than the ones modelled in simulations. First, the existence of third-generation or later-generation hybrids could explain the high percentage of slightly introgressed trees in nature. Second, hybridisation involving more than two species seems to happen in natural populations.

Overall, we detected a high occurrence of hybrids in all studied populations (Table 2). The percentage of hybrids was higher in the intensively studied stands (Briouant, Petite Charnie and Paguères), ranging from 19.1% to 30.5% (23.9% on average) compared with 10.7% to 20.8% (15.9% on average) in populations where we sampled a limited number of individuals per stand (Aurignac, Pyrenees and ONF) (Table 2).

We identified hybrids between all pairs of species investigated, in particular in Briouant where the four species co-occur (Table 2). Additionally, we detected a number of hybrids involving a species present in the population and another species not identified during field work. This finding is particularly remarkable in the well characterised Petite Charnie stand where only pedunculate and sessile oaks had been described but where hybrids involving *Q. pubescens* and *Q. pyrenaica* were detected using molecular markers (Table 2). A similar finding was made in populations from the Pyrenees and in the ONF stands where hybrids with *Q. pyrenaica* (not known in these areas) were observed. To test if these results can be explained by assignment error, we used the results from the simulated data set (Table 1). Among 2000 simulated pure *Q. robur* and *Q. petraea* trees, six individuals were wrongly assigned to *Q. pubescens* or *Q. pyrenaica* hybrids (0.3%). Out of 150 simulated *Q. robur* × *Q. petraea* hybrids, we wrongly assigned four trees considered to represent *Q. pubescens* or *Q. pyrenaica* hybrids (3%). Assuming that we only have *Q. robur* and *Q. petraea* species and their hybrids in Petite Charnie, we expect to falsely assign less than one individual from the 212 pure species trees to *Q. pubescens* or *Q. pyrenaica* hybrids and less than 1.5 tree from the 50 hybrids to *Q. pubescens* or *Q. pyrenaica* hybrids (Table 2). Thus in total, if the Petite Charnie stand was only composed by *Q. robur* and *Q. petraea* and their hybrids, we would expect less than three erroneous assignments to *Q. pubescens* or *Q. pyrenaica* hybrids. By contrast, we identified 35 hybrid types involving these species (Table 2), a figure that cannot be explained by assignment errors alone.

Analyses of genetic distances between groups confirmed species and hybrid identification. Pure species oaks identified in each population group together in the same common node (Fig. 4). Furthermore, hybrids involving the same pair of species, whatever their geographical origin, share a common node or are localised in the same part of the tree. This is clearly the case for *Q. robur* × *Q. petraea*, *Q. robur* × *Q. pubescens* and *Q. robur* × *Q. pyrenaica* hybrids (Fig. 4).

Genetic composition of species and hybrids

We computed the average of each of the four admixture coefficients for the two categories (pure species and hybrids). In the three intensively studied stands, the overall genetic composition differed between pure species and hybrids (Fig. 5). The fact that the genetic composition of the pure species category differs from that of the hybrid category indicates that the four species are not involved proportionally in the formation of hybrids and backcrosses. In Petite Charnie, *Q. robur* and *Q. petraea* genes seem to be equally

© 2009 Blackwell Publishing Ltd

Fig. 4 Phylogenetic neighbour-joining tree based on Cavalli-Sforza and Edwards genetic distances (Cavalli-Sforza & Edwards 1967) between pure species and hybrids as assigned by the Structure software in the different populations. Only groups with more than 10 individuals were used to build the tree, the scale line represents a genetic distance of 0.05. Large branches represent pure oak species with colours corresponding to Fig. 3 at $K = 4$. Thinner branches illustrate hybrid groups with each colour corresponding to a specific hybrid type. Labels at the tip of the branches indicate the corresponding species or hybrid type (Rob, *Quercus robur*; Pet, *Q. petraea*; Pub, *Q. pubescens*; Pyr, *Q. pyrenaica*; and hyb, hybrid) and populations' names are given in the subscript (Bri, Briouant stand; PC, Petite Charnie stand; Pag, Paguères stand; Auri, Aurignac populations; ONF, ONF populations; Pyr, Pyrenean populations).

represented in species and hybrid trees but *Q. pubescens* and *Q. pyrenaica* genes are significantly overrepresented among hybrids ($P < 0.001$ and $P < 0.01$, respectively). In Briouant, *Q. robur* genes are far less present in the hybrid category than in the pure species category ($P < 0.001$) whereas *Q. petraea* and *Q. pubescens* genes are significantly more frequent among the hybrid category ($P < 0.001$ and $P < 0.001$, respectively). In Paguères, we also found that *Q. robur* genes are under-represented among hybrid trees ($P < 0.001$), whereas *Q. petraea* and *Q. pyrenaica* genes are over-represented among hybrids ($P < 0.001$ and $P < 0.001$, respectively).

Species frequency-dependent hybridisation and introgression

Differences in genetic composition between hybrids and pure-bred individuals suggest that genes of the more abundant species are under-represented in hybrids (Fig. 5). To formally test this hypothesis, we have plotted the species relative abundance in each stand against the difference in its genetic composition in hybrids vs. purebreds (Fig. 6). There is a clear negative relationship (Fig. 6, $R^2 = 0.83$, $F_{1,10} = 52.86$, $P < 0.001$). This result comforts our observation that abundant species are proportionally less involved in hybridisation

© 2009 Blackwell Publishing Ltd

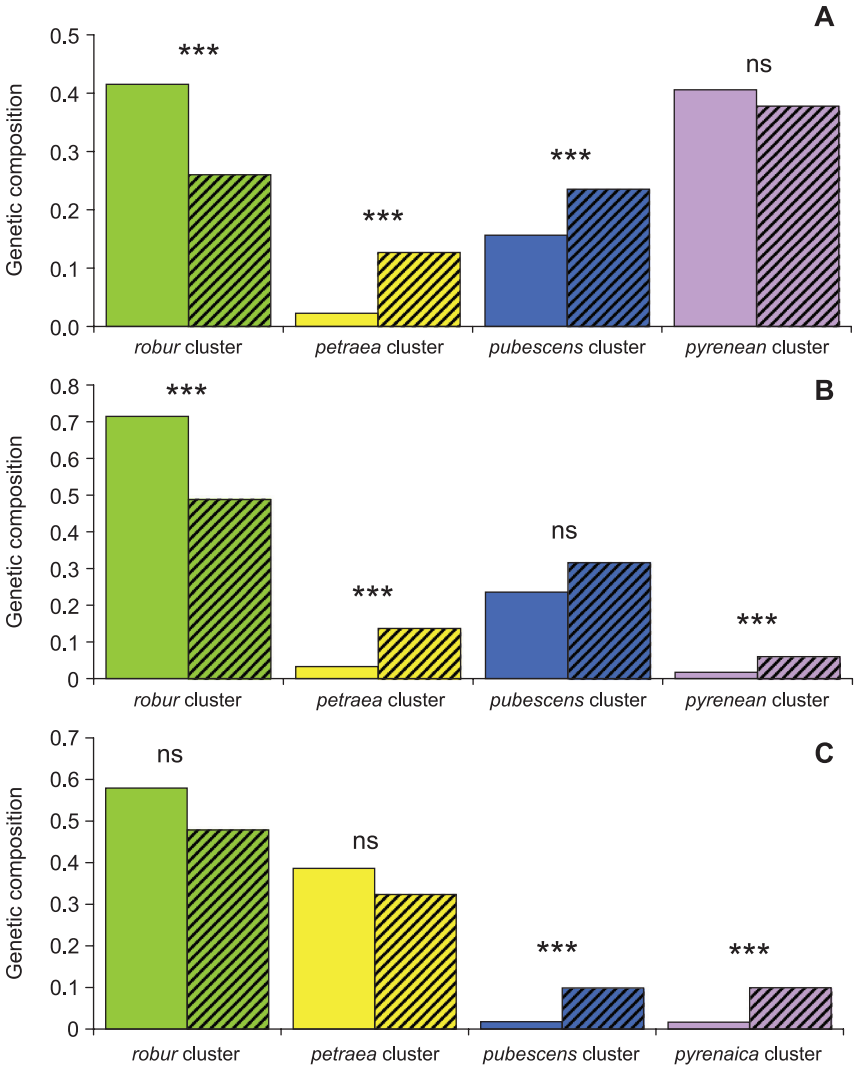


Fig. 5 Comparisons of genetic composition (averaged admixture coefficients from each of the four clusters) for pure species (plain colours) and hybrids (dashed colours) in Briouant (A), Paguères (B) and Petite Charnie (C) stands. Differences were tested with a Student's *t*-test (***: $P < 0.001$, NS: not significant).

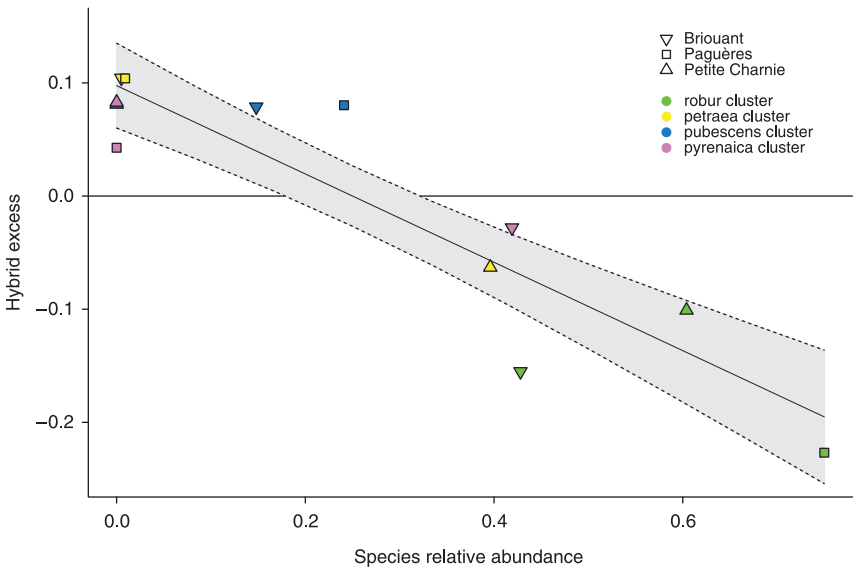


Fig. 6 Change in admixture coefficient between hybrids and purebreds as a function of the corresponding species relative abundance in the stand. The continuous black line indicates no difference between averaged admixture coefficients for hybrids and purebreds. A positive value indicates over-representation of the corresponding cluster in hybrid individuals whereas a negative value indicates over-representation of the corresponding cluster in purebred oaks. Dashed lines and grey shading indicate the confidence interval of the linear regression (large black line; $R^2 = 0.83$, $F_{1,10} = 52.86$, $P < 0.001$). The shapes of the symbols represent the different stands (down-pointing triangle, Briouant; square, Paguères; up-pointing triangle, Petite Charnie) and colours represent clusters (green, *Q. robur* cluster; yellow, *Q. petraea* cluster; blue, *Q. pubescens* cluster; and purple, *Q. pyrenaica* cluster).

© 2009 Blackwell Publishing Ltd

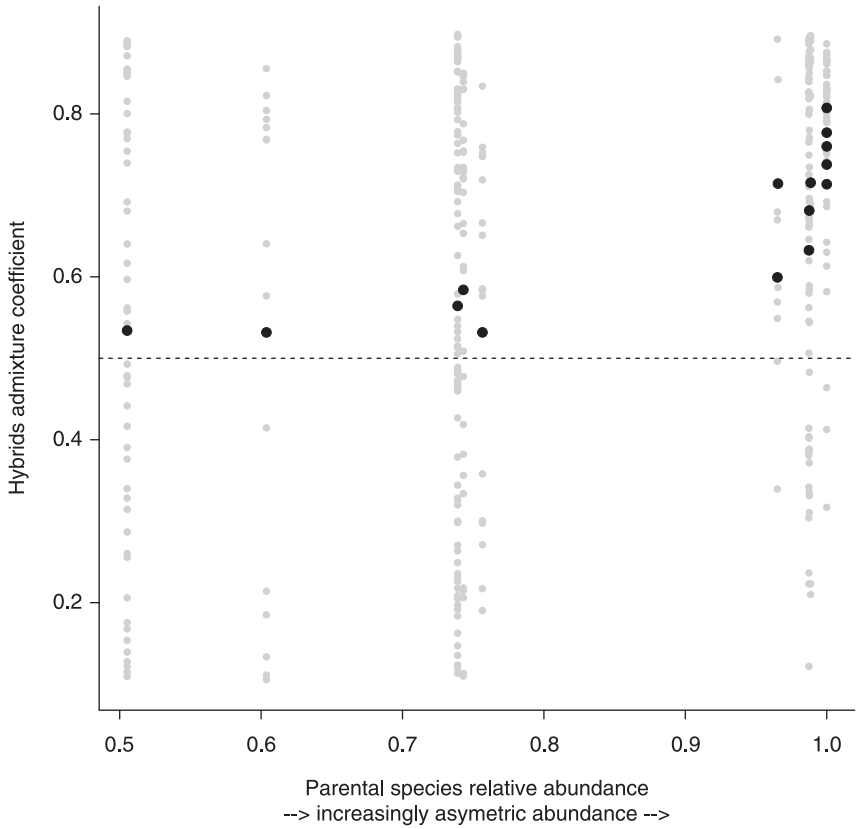


Fig. 7 Effect of parental species relative abundance on hybrid admixture coefficients. Small grey points represent admixture coefficient of each hybrid individuals whereas large black points represent the averaged admixture coefficient for each hybrid class in each stand. For each hybrid class, we used the admixture coefficient corresponding to the most abundant parental species. The horizontal dashed line gives the expected admixture coefficient if introgression was not directional.

than minority species. We then performed a detailed analysis of genetic composition of hybrid classes by using admixture coefficients (Fig. 7). Hybrid individuals admixture coefficients have a large distribution, indicating that hybridisation is not restricted to the first generation (i.e. numerous hybrids had admixture coefficient between 0.65 and 0.9, values that are unlikely for F_1 hybrids, see Fig. S2). Moreover, the averaged admixture coefficient of hybrid classes showed that some classes have an intermediate admixture value, pointing to balanced bidirectional introgression, whereas others hybrid classes have a genetic composition closer to one of the parental species (Fig. 7), indicating directional introgression. Hence, bidirectional introgression seems to take place when parental species are equally represented, whereas directional introgression appears to predominate when parental species differ greatly in abundance (Fig. 7).

Discussion

Our work has addressed the effect of species relative abundance on natural hybridisation and introgression. There are surprisingly few such studies in natural populations. We showed that relative species abundance affects both hybridisation rates and introgression directionality. Previous studies have reported hybridisation patterns between pairs of oak species (Muir *et al*. 2000; Muir & Schlötterer 2005; Valbuena-Carabaña *et al*. 2005, 2007; Gugerli *et al*. 2007) or have studied more species but in one restricted area (Curtu *et al*. 2007). Our extended analyses of 2107 oaks belonging to four species and several populations provide new insights into hybridisation and introgression dynamics within the European white oak species complex. Such large sample sizes should provide accurate estimates of allelic frequencies in the different oak species for use in species delineation and hybrid identification (Waples & Gaggiotti 2006). Using genetic clustering and simulations, we assigned the species or hybrid origin of each sampled oak. We found that hybrids (*sensu lato*: including introgressed individuals) are common in all studied populations, supporting previous claims that hybridisation is ongoing among these oak species (Gugerli *et al*. 2007). Moreover, intensive sampling in three stands allowed us to demonstrate the importance of stand species composition in hybridisation patterns and introgression dynamics.

From clustering to assignment analysis

In the clustering analyses, we found stable results for $K = 6$, highlighting not only differences between species but also a geographical structure within *Quercus robur* and *Quercus petraea*. Such a result might be due to a geographical gradient in allele frequencies, as demonstrated for allozyme data in

© 2009 Blackwell Publishing Ltd

Q. petraea (Zanetto & Kremer 1995; Kremer & Zanetto 1997; Le Corre *et al*. 1998). Using more loci on a wider sampling area covering the distribution range of the species could improve the understanding of these subspecific genetic patterns. In any case, it is clear that intraspecific differences are subsidiary to species differences, and thus intraspecific variation does not compromise species identification. The leaf morphology of a subset of the individuals had been previously analysed (Viscosi *et al*. 2009), showing a clear concordance between genetic cluster and morphological features in these oak species.

We then tested the performance of species assignment and hybrid identification using data-based simulations. Our results show that classes of pure and admixed individuals detected with Structure had been reconstructed with good accuracy and efficiency. However, our 10 microsatellites were not able to differentiate first from second-generation hybrids, an objective that has been shown to require more than 48 loci in cases of low genetic divergence, such as the one observed in these oaks (Vaha & Primmer 2006). Note that our estimates of hybrid abundance are conservative since the threshold we selected ($Q = 0.90$) to distinguish pure species from hybrids should slightly underestimate hybrid proportions and minimise assignment error rate among hybrid classes. Altogether, the results indicate that assignment methods, if used with caution, can be efficient to delimitate species across broad geographical ranges, without prior morphological information, as already shown by Duminil *et al*. (2006). They further indicate that assignments are still relevant when more than two species are present and when an intraspecific geographical structure is detected.

Widespread occurrence of hybrids in the European white oak species complex

Our genetic assignment analysis also confirms that sympatric species from the European white oaks complex do hybridise. Overall hybrid frequencies differ among areas (11–30%, Table 2) with more hybrids detected in intensively sampled stands (19–30%) than in less intensively sampled populations (11–21%). Sampling a small proportion of individuals in a stand can lead to an underestimation of hybridisation if oaks with typical leaf morphology are preferentially sampled. In a detailed multivariate analysis of leaf morphology, hybrid individuals were on average morphologically intermediate between parental species (Viscosi *et al*. 2009). Hence, some (but not all) hybrid oaks could be characterised by an intermediate leaf morphology and intentionally (or not) avoided during sampling (Lexer *et al*. 2006). Estimated hybridisation rates based on non-exhaustive sampling should thus be taken with caution.

The hybrid frequencies found in our populations are comparable with, although slightly higher than, previously found in other studies using comparable approaches. An analysis of three stands in Spain comprising *Q. petraea* and *Q. pyrenaica* detected between 6% and 22% of hybrids depending on the stand (Valbuena-Carabaña *et al*. 2007). Likewise, genetic assignment in a four-oak-species stand in Romania detected between 2% and 16% hybrids depending on the species pairs (Curtu *et al*. 2007). These estimates suggest that hybridisation is not a rare event in oaks and that it is a contemporary process. We were able to identify hybrids between all species pairs studied, indicating that no strict reproductive barriers exist. However, the frequency of the different hybrid classes varies among stands, suggesting that local conditions can affect the outcome of hybridisation. The simultaneous analysis of forests located far apart, with material from all four species included as reference, allowed us to detect hybridisation between species pairs in situations where one of the parental species is locally absent. In the Petite Charnie stand, for instance, only *Q. robur* and *Q. petraea* oaks have been described so far (Bacilieri *et al*. 1995; Streiff *et al*. 1998; Streiff *et al*. 1999) but we identified 13% of *Q. pubescens* and *Q. pyrenaica* hybrid types in this stand (Table 2, Fig. 4), compared with only 6% of *Q. robur* × *Q. petraea* hybrids. This finding highlights the importance of including all species potentially connected by gene flow when studying hybridisation with genetic assignment methods. A separate analysis of the Petite Charnie stand, for example, would have resulted in the detection of only two clusters without any chance to identify *Q. pubescens* and *Q. pyrenaica* hybrids.

The presence of hybrids in the absence of one parental species has also been demonstrated in American red oaks (Dodd & Afzal-Rafii 2004), pinyon pines (Lanner & Phillips 1992) and *Aesculus* tree species (DePamphilis & Wyatt 1989; Thomas *et al*. 2008). Two hypotheses can explain such observations: hybridisation by long-distance pollen dispersal or past local extinction of one of the two parental species (Buggs 2007; Thomas *et al*. 2008). Massive deforestation during the last 3000 years by human exploitation and land clearing for agriculture render difficult to estimate original species distribution ranges and thus the possibility of local extinction of *Q. pubescens* and *Q. pyrenaica* to explain the occurrence of their hybrids. Occasional long-distance hybridisation is not unlikely in these highly outcrossing wind pollinated species. The nearest *Q. pubescens* or *Q. pyrenaica* populations are localised some tens of kilometres from Petite Charnie. Because *Q. pubescens* and *Q. pyrenaica* are more drought tolerant and thermophilous than *Q. robur* and *Q. petraea*, dispersal by long-distance pollen hybridisation could be a mechanism to speed up their northern migration facing climate warming.

Frequency-dependent hybridisation and introgression

Species relative abundance is one of the factors that can affect hybridisation pattern and introgression dynamics

© 2009 Blackwell Publishing Ltd

(Anderson & Hubricht 1938; Nason *et al*. 1992; Burgess *et al*. 2005). Our detailed analysis of three stands differing in species composition allowed us to estimate the relative species abundance and its impact on the outcome of hybridisation.

Hybridisation rate

We found a deficit of hybrids involving locally dominant species (e.g. *Q. robur* and *Q. pyrenaica* in Briouant and *Q. robur* and *Q. petraea* in Petite Charnie), whereas less frequent or rare species tend to be over-represented among hybrids (Figs 5 and 6). Several hypotheses could account for this observation. First, dominant species are expected to be well adapted to local environmental conditions; their hybrids may therefore have a lower competitive ability. Limited hybrid formation between dominant species in a stand would then be caused by differential selection between hybrid and parental species. Second, if these hybrids were selected against, the strength of reproductive barriers between dominant species could increase as a result of reinforcement (Dobzhansky 1937; Butlin 1987). This would lead to a higher reproductive isolation and a lower hybridisation rate between dominant species, compared with species that came more recently in contact, for which reinforcement would not have time to develop. Comparative analyses of open-pollinated progenies with contrasted species abundance situations would be useful to test the hypothesis of reinforcement. Third, rare species could be over-represented among hybrids because of their difficulty to mate with other rare conspecific partners. Such minority species should receive abundant heterospecific pollen, which would increase hybridisation rate (Rieseberg & Gerber 1995). Relative species abundance and underlying causal factors such as local environment and forest management could have a major influence on hybridisation rate. However, this prediction should be tested by manipulating the proportion of pollen from several species received by female flowers using controlled crosses experiments.

Direction of introgression

As we were unable to differentiate F_1 from backcrosses using direct genetic assignment, we computed the mean admixture coefficients of the different hybrid classes in each stand to get some insight into the genetic composition of hybrid individuals compared to their parental species. A mean admixture coefficient of 0.5 would imply that only first-generation hybrids exist or that each parental species mates in the same proportion with hybrids, producing a balanced number of each type of backcrosses. On the contrary, if the backcrosses were biased towards one of the parental species, we should observe a mean cluster value between 0.5 and 0.9 because a majority of the hybrids would be closer to the successfully backcrossing species. Clearly, the observed distribution of individual admixture coefficients in hybrids indicates that backcrosses are more numerous than F_1, as the majority of hybrids showed admixture coefficient between 0.65 and 0.90 (Fig. 7). These results show that hybridisation is not restricted to the formation of F_1 but instead involves further generations of backcrosses between pure species and F_1 hybrids.

Our results show that the direction of introgression strongly depends on the relative frequency of the parental species in the studied stands (Fig. 7). Knowledge of mating system of oak hybrids are lacking, with the exception of one study using controlled crosses on a fertile *Q. robur* × *Q. petraea* hybrid (Olrik & Kjaer 2007). In our study, we found that the direction of the backcrosses was predominantly towards the more numerous species. Additional analyses of hybrid reproductive behaviour would greatly improve our understanding of the hybridisation dynamic in this species complex. However, it is already clear that interspecific gene flow is a widespread and ongoing process among oak species. Since the species remain morphologically and ecologically distinct (Kremer *et al*. 2002; Petit *et al*. 2003), this observation indicates that collective evolution (*sensu* Morjan & Rieseberg 2004) takes place within these species in the face of extensive interspecific gene flow. It would be interesting now to study if collective evolution can simultaneously take place higher in the hierarchy, within groups of closely related species, as first suggested by Pernès (1984). The European white oaks would seem to be good candidates to test this idea, in view of the high rate of interspecific gene flow they experience. In any case, our results indicate that the rate of exchange between species belonging to the same species complex should not be viewed as a fixed parameter but as a variable one that depends on several factors such as the local composition of the community.

Acknowledgements

We thank Jean-Marc Louvet, Jérôme Willm, Maya Gonzalez and Alain Cabanettes for sampling assistance and sharing their field knowledge. We are grateful to Patrick Léger, Valerie Léger, Pierre-Yves Dumolin and Franck Salin for technical assistance. O.L. is grateful to Martin Lascoux for his invitation at the Evolutionary Biology Centre of Uppsala University. We thank Richard Abbott, Alex Buerkle and three anonymous reviewers for their suggestions that greatly improved the manuscript. Genotyping presented in this publication was performed at the Genotyping and Sequencing facility of Bordeaux (grants from the Conseil Régional d'Aquitaine n°20030304002FA, n°20040305003FA and from the European Union, FEDER n°2003227). Experiments were funded by the Interregional Project Aquitaine/Midi-Pyrénées: 'Évolution de la biodiversité des forêts sous l'effet des changements globaux (changements d'usage et changements climatiques)', by the French Research Agency (ANR) through the QDIV project: 'Quantification of the effects of global changes on plant diversity' (n°ANR-05-BDIV-009-01), and

© 2009 Blackwell Publishing Ltd

by the European Union supported project (QLRT-1999-30690) OAKFLOW 'Intra- and interspecific gene flow in oaks as mechanisms promoting genetic diversity and adaptive potential', as well as by the Office National des Forêts ('Traçabilité géographique et identification taxonomique du bois de chêne des forêts domaniales françaises').

References

Anderson E (1949) *Introgressive Hybridization*. Wiley & Sons, New York.

Anderson E, Hubricht L (1938) Hybridization in Tradescantia. III. The evidence for introgressive hybridization. *American Journal of Botany*, **25**, 396–402.

Arnold ML (2006) *Evolution through Genetic Exchange*. Oxford University Press, USA & Oxford, UK.

Bacilieri R, Ducousso A, Kremer A (1995) Genetic, morphological, ecological and phenological differentiation between *Quercus petraea* (Matt) Liebl and *Quercus robur* L. in a mixed stand of Northwest of France. *Silvae Genetica*, **44**, 1–10.

Bacilieri R, Ducousso A, Petit RJ, Kremer A (1996) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution*, **50**, 900–908.

Buggs RJA (2007) Empirical study of hybrid zone movement. *Heredity*, **99**, 301–312.

Burgess KS, Morgan M, Deverno L, Husband BC (2005) Asymmetrical introgression between two *Morus* species (*M. alba*, *M. rubra*) that differ in abundance. *Molecular Ecology*, **14**, 3471–3483.

Butlin R (1987) Speciation by reinforcement. *Trends in Ecology & Evolution*, **2**, 8–13.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, **19**, 233–257.

Chan C, Ballantyne KN, Aikman H *et al*. (2006) Genetic analysis of interspecific hybridisation in the world's only Forbes' parakeet (*Cyanoramphus forbesi*) natural population. *Conservation Genetics*, **7**, 493–506.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: massive introgression by local genes. *Evolution*, **62**, 1908–1920.

Curtu AL, Gailing O, Finkeldey R (2007) Evidence for hybridization and introgression whithin a species-rich oak (*Quercus* spp.) community. *BMC Evolutionary Biology*, **7**, 218.

DePamphilis CW, Wyatt R (1989) Hybridization and introgression in buckeyes (*Aesculus*, Hippocastanaceae): a review of the evidence and a hypothesis to explain long-distance gene flow. *Systematic Botany*, **14**, 593–611.

Dobzhansky T (1937) *Genetics and the Origin of Species*. Columbia University Press, New York.

Dodd RS, Afzal-Rafii Z (2004) Selection and dispersal in a multispecies oak hybrid zone. *Evolution*, **58**, 261–269.

Duminil J, Caron H, Scotti I, Cazal SO, Petit RJ (2006) Blind population genetics survey of tropical rainforest trees. *Molecular Ecology*, **15**, 3505–3513.

Dupouey JL, Badeau V (1993) Morphological variability of oaks (*Quercus robur* L., *Quercus petraea* (Matt) Liebl, *Quercus pubescens* Willd) in northern France: preliminary results. *Annales des Sciences Forestières*, **50**, 35s–40s.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Field DL, Ayre DJ, Whelan RJ, Young AG (2008) Relative frequency of sympatric species influences rates of interspecific hybridization, seed production and seedling performance in the uncommon *Eucalyptus aggregata*. *Journal of Ecology*, **96**, 1198–1210.

Fox J (2003) Effect display in R for generalised linear models. *Journal of Statistical Software*, **8**, 1–27.

Gonzalez M, Deconchat M, Balent G, Cabanettes A (2008) Diversity of woody plant seedling banks under closed canopy in fragmented coppice forests. *Annals of Forest Science*, **65**, 511.

Gugerli F, Walser JC, Dounavi K, Holderegger R, Finkeldey R (2007) Coincidence of small-scale spatial discontinuities in leaf morphology and nuclear microsatellite variation of *Quercus petraea* and *Q. robur* in a mixed forest. *Annals of Botany*, **99**, 713–722.

Hubbs CL (1955) Hybridization between fish in nature. *Systematic Zoology*, **4**, 1–20.

Kampfer S, Lexer C, Glossl J, Steinkellner H (1998) Characterization of $(GA)_n$ microsatellite loci from *Quercus robur*. *Hereditas*, **129**, 183–186.

Kim M, Cui M-L, Cubas P *et al*. (2008) Regulatory genes control a key morphological and ecological trait transferred between species. *Science*, **322**, 1116–1119.

Kleinschmit J, Kleinschmit JGR (2000) *Quercus robur — Quercus petraea*: a critical review of the species concept. *Glasnik Za Smske Pokuse*, **37**, 441–452.

Kremer A, Zanetto A (1997) Geographical structure of gene diversity in *Quercus petraea* (Matt.) Liebl. II. Multilocus patterns of variation. *Heredity*, **78**, 476–489.

Kremer A, Dupouey JL, Deans JD *et al*. (2002) Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Annals of Forest Science*, **59**, 777–787.

Langella O (1999) *Populations*, Version 1.2.28. Available from URL: http://www.pge.cnrs-gif.fr/bioinfo/populations/index.php.

Lanner RM, Phillips AM III (1992) Natural hybridization and introgression of pinyon pines in northwestern Arizona. *International Journal of Plant Sciences*, **153**, 250–257.

Le Corre V, Roussel G, Zanetto A, Kremer A (1998) Geographical structure of gene diversity in *Quercus petraea* (Matt.) Liebl. III. Patterns of variation identified by geostatistical analyses. *Heredity*, **80**, 464–473.

Lepais O, Leger V, Gerber S (2006) Short note: high throughput microsatellite genotyping in oak species. *Silvae Genetica*, **55**, 238–240.

Levin DA, Francisco-Ortega J, Jansen RK (1996) Hybridization and the extinction of rare plant species. *Conservation Biology*, **10**, 10–16.

Lexer C, Kremer A, Petit RJ (2006) Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Molecular Ecology*, **15**, 2007–2012.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.

Mayr E (1963) *Animal Species and Evolution*. Harvard University Press, Cambridge, Massachusetts.

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, **13**, 1341–1356.

Muir G, Schlötterer C (2005) Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci

© 2009 Blackwell Publishing Ltd

differentiating two hybridizing oaks (*Quercus* spp.). *Molecular Ecology*, **14**, 549–561.

Muir G, Fleming CC, Schlötterer C (2000) Species status of hybridizing oaks. *Nature*, **405**, 1016–1016.

Nason JD, Ellstrand NC, Arnold ML (1992) Patterns of hybridization and introgression in populations of oaks, manzanitas and irises. *American Journal of Botany*, **79**, 101–111.

Olrik DC, Kjaer ED (2007) The reproductive success of a *Quercus petraea* × *Q. robur* F_1-hybrid in back-crossing situations. *Annals of Forest Science*, **64**, 37–45.

Paradis E, Claude J, Strimmer K (2004) APE. Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pernès J (1984) *Gestion des ressources génétiques. Tome 2: Manuel*. Agence de Coopération Culturelle et Technique, Paris, France.

Petit RJ, Bodenes C, Ducousso A, Roussel G, Kremer A (2003) Hybridization as a mechanism of invasion in oaks. *New Phytologist*, **161**, 151–164.

Prentis PJ, White EM, Radford IJ, Lowe AJ, Clarke AR (2007) Can hybridization cause local extinction: a case for demographic swamping of the Australian native *Senecio pinnatifolius* by the invasive *Senecio madagascariensis*? *New Phytologist*, **176**, 902–912.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2005) R: a languaage and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from URL: http://www.R-project.org, edn.

Rameau JC, Mansion D, Dumé G (1989) *Flore forestière française: guide écologique illustré, 1: Plaines et collines*. Institut pour le Développement Forestier, Paris, France.

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Rieseberg LH (1997) Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, **28**, 359–389.

Rieseberg LH, Carney SE (1998) Plant hybridization. *New Phytologist*, **140**, 599–624.

Rieseberg LH, Gerber D (1995) Hybridization in the Catalina Island Mountain Mahogany (*Cercocarpus traskiae*). *RAPD Evidence. Conservation Biology*, **9**, 199–203.

Scotti-Saintagne C, Mariette S, Porth I *et al*. (2004) Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. Petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615–1626.

Steinhoff S (1993) Results of species hybridization with *Quercus robur* L. & *Quercus petraea* (Matt.) Liebl. *Annales Des Sciences Forestieres*, **50**, 137s–143s.

Steinhoff S (1998) Controlled crosses between pendunculate and sessile oak: results and conclusion. *Allgemeine Forst und Jagdzeitung*, **169**, 163–168.

Steinkellner H, Fluch S, Turetschek E *et al*. (1997) Identification and characterization of $(GA/CT)_n$-microsatellite loci from *Quercus petraea*. *Plant Molecular Biology*, **33**, 1093–1096.

Streiff R, Labbe T, Bacilieri R *et al*. (1998) Within-population genetic structure in *Quercus robur* L. & *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology*, **7**, 317–328.

Streiff R, Ducousso A, Lexer C *et al*. (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. Petraea* (Matt.) Liebl. *Molecular Ecology*, **8**, 831–841.

Thomas DT, Ahedor AR, Williams CF *et al*. (2008) Genetic analysis of a broad hybrid zone in *Aesculus* (Sapindaceae): is there evidence of long-distance pollen dispersal? *International Journal of Plant Sciences*, **169**, 647–657.

Vaha JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.

Valbuena-Carabaña M, González-Martínez SC, Sork VL *et al*. (2005) Gene flow and hybridisation in a mixed oak forest (*Quercus pyrenaica* Willd. and *Quercus petraea* (Matts.) Liebl.) in central Spain. *Heredity*, **95**, 457–465.

Valbuena-Carabaña M, González-Martínez SC, Hardy OJ, Gil L (2007) Fine-scale spatial genetic structure in mixed oak stands with different levels of hybridization. *Molecular Ecology*, **16**, 1207–1219.

Viscosi V, Lepais O, Gerber S, Fortini P (2009) Leaf morphological analyses in four European oak species (*Quercus*) and their hybrids: a comparison of traditional and geometric morphometric methods. *Plant Biosystems* in press.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Wirtz P (1999) Mother species-father species: unidirectional hybridization in animals with female choice. *Animal Behaviour*, **58**, 1–12.

Zanetto A, Kremer A (1995) Geographical structure of gene diversity in *Quercus. petraea* (Matt.) Liebl. I. Monolocus patterns of variation. *Heredity*, **75**, 506–517.

Zhou R, Gong X, Boufford D, Wu CI, Shi S (2008) Testing a hypothesis of unidirectional hybridization in plants: observations on Sonneratia, Bruguiera and Ligularia. *BMC Evolutionary Biology*, **8**, 149.

This article is a part of O.L.'s PhD thesis focusing on hybridization dynamics between European white oak species. O.L. has a wide interest in application of molecular markers for studying the ecology, evolution and history of species. R.J.P. is a population geneticist with broad interest in evolution, phylogeography and mating system of trees. E.G. is a PhD student working on the characteristics of oak species used by the barrel industry. J.L. collaborated with O.L. during her Master; she is currently doing a PhD on the spatial and temporal variability of the mutualistic interaction between *Taxus baccata* L. and its frugivores' community. F.A. is a PhD student working on the adaptation of *Quercus petraea* (Matt.) Liebl. along an altitudinal gradient in the Pyrenean Mountains. A.K. has long standing interests in the evolution of temperate and tropical forest trees with particular emphasis on population differentiation at various levels where diversity is expressed (from genes to phenotypes). S.G. is a geneticist interested in population genetics and gene flow studies in forest trees, she supervised O.L.'s thesis.

© 2009 Blackwell Publishing Ltd

Supporting information

Additional supporting information may be found in the online version of this article:

Fig. S1 Neighbour-joining tree illustrating the net nucleotide genetic distances, as computed by the STRUCTURE software, between clusters at K = 6.

Fig. S2 Admixture coefficients distribution for simulated individuals: (A) pure species, (B) first generation hybrids (F1), (C) second generation hybrids (backcrosses) and (D) averaged distribution of pure species, first and second generation hybrids.

Table S1 Details of the sampled populations.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2009 Blackwell Publishing Ltd

ANNEXE 2

Analyses sensorielles sur copeaux

Ces tests s'inscrivent dans le cadre de ma thèse sur la caractérisation du bois de chêne utilisé pour la maturation des vins et alcools. L'objectif est ici de tester votre capacité à détecter des différences sensorielles entre des échantillons sous forme de copeaux de bois. Les échantillons sont issus des deux espèces de chêne (sessile et pédonculé), qui présentent des propriétés aromatiques plus ou moins contrastées.

Les tests que vous allez effectuer sont des tests triangulaires, c'est-à-dire que seul un échantillon est différent des deux autres. C'est celui-ci que vous devrez tenter de reconnaitre en sentant les copeaux. Les différences sont qualitatives (odeurs différentes) ou quantitative (odeur plus ou moins forte).

Chaque test prend moins d'une minute et il y a 30 tests au total. Il n'y a pas d'ordre particulier pour effectuer les tests.

PRECAUTIONS : Pour chaque échantillon, il suffit de soulever le couvercle, de sentir les copeaux et de refermer le couvercle. Puis de passer à l'échantillon suivant. Pensez à ne soulever qu'un couvercle à la fois pour ne pas les inverser. Cochez ensuite sur le questionnaire quel échantillon est différent des deux autres. Puis passez au test suivant. Si vous avez un doute, les verres et les couvercles sont identifiés par le numéro du test (1 à 60) et l'échantillon (A, B ou C).

N'hésitez pas à me contacter pour toute question (poste 28.27 ou erwan.guichoux@pierroton.inra.fr)

Un grand merci pour votre participation.

Nom :

Prénom :

Date :

Heure :

Numéro du test	Quel échantillon est différent des deux autres?		
1	☐A	☐ B	☐ C
3	☐A	☐ B	☐ C
5	☐A	☐ B	☐ C
7	☐A	☐ B	☐ C
9	☐A	☐ B	☐ C
11	☐A	☐ B	☐ C
13	☐A	☐ B	☐ C
15	☐A	☐ B	☐ C
17	☐A	☐ B	☐ C
19	☐A	☐ B	☐ C
21	☐A	☐ B	☐ C
23	☐A	☐ B	☐ C
25	☐A	☐ B	☐ C
27	☐A	☐ B	☐ C
29	☐A	☐ B	☐ C
31	☐A	☐ B	☐ C
33	☐A	☐ B	☐ C
35	☐A	☐ B	☐ C
37	☐A	☐ B	☐ C
39	☐A	☐ B	☐ C
41	☐A	☐ B	☐ C
43	☐A	☐ B	☐ C
45	☐A	☐ B	☐ C
47	☐A	☐ B	☐ C
49	☐A	☐ B	☐ C
51	☐A	☐ B	☐ C
53	☐A	☐ B	☐ C
55	☐A	☐ B	☐ C
57	☐A	☐ B	☐ C
59	☐A	☐ B	☐ C

Commentaires :

Numéro du test	Quel échantillon est différent des deux autres?		
2	☐A	☐ B	☐ C
4	☐A	☐ B	☐ C
6	☐A	☐ B	☐ C
8	☐A	☐ B	☐ C
10	☐A	☐ B	☐ C
12	☐A	☐ B	☐ C
14	☐A	☐ B	☐ C
16	☐A	☐ B	☐ C
18	☐A	☐ B	☐ C
20	☐A	☐ B	☐ C
22	☐A	☐ B	☐ C
24	☐A	☐ B	☐ C
26	☐A	☐ B	☐ C
28	☐A	☐ B	☐ C
30	☐A	☐ B	☐ C
32	☐A	☐ B	☐ C
34	☐A	☐ B	☐ C
36	☐A	☐ B	☐ C
38	☐A	☐ B	☐ C
40	☐A	☐ B	☐ C
42	☐A	☐ B	☐ C
44	☐A	☐ B	☐ C
46	☐A	☐ B	☐ C
48	☐A	☐ B	☐ C
50	☐A	☐ B	☐ C
52	☐A	☐ B	☐ C
54	☐A	☐ B	☐ C
56	☐A	☐ B	☐ C
58	☐A	☐ B	☐ C
60	☐A	☐ B	☐ C

Commentaires :

ANNEXE 3

Durand et al. BMC Genomics 2010, 11:570
http://www.biomedcentral.com/1471-2164/11/570

BMC Genomics

RESEARCH ARTICLE Open Access

A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study

Jérôme Durand[1,2], Catherine Bodénès[1,2], Emilie Chancerel[1,2], Jean-Marc Frigerio[1,2], Giovanni Vendramin[3], Federico Sebastiani[3], Anna Buonamici[3], Oliver Gailing[4,5], Hans-Peter Koelewijn[6], Fiorella Villani[7], Claudia Mattioni[7], Marcello Cherubini[7], Pablo G Goicoechea[8], Ana Herrán[8], Ziortza Ikaran[8], Cyril Cabané[9], Saneyoshi Ueno[1,2,10], Florian Alberto[1,2], Pierre-Yves Dumoulin[1,2], Erwan Guichoux[1,2], Antoine de Daruvar[9], Antoine Kremer[1,2], Christophe Plomion[1,2*]

Abstract

Background: Expressed Sequence Tags (ESTs) are a source of simple sequence repeats (SSRs) that can be used to develop molecular markers for genetic studies. The availability of ESTs for *Quercus robur* and *Quercus petraea* provided a unique opportunity to develop microsatellite markers to accelerate research aimed at studying adaptation of these long-lived species to their environment. As a first step toward the construction of a SSR-based linkage map of oak for quantitative trait locus (QTL) mapping, we describe the mining and survey of EST-SSRs as well as a fast and cost-effective approach (bin mapping) to assign these markers to an approximate map position. We also compared the level of polymorphism between genomic and EST-derived SSRs and address the transferability of EST-SSRs in *Castanea sativa* (chestnut).

Results: A catalogue of 103,000 Sanger ESTs was assembled into 28,024 unigenes from which 18.6% presented one or more SSR motifs. More than 42% of these SSRs corresponded to trinucleotides. Primer pairs were designed for 748 putative unigenes. Overall 37.7% (283) were found to amplify a single polymorphic locus in a reference full-sib pedigree of *Quercus robur*. The usefulness of these loci for establishing a genetic map was assessed using a bin mapping approach. Bin maps were constructed for the male and female parental tree for which framework linkage maps based on AFLP markers were available. The bin set consisting of 14 highly informative offspring selected based on the number and position of crossover sites. The female and male maps comprised 44 and 37 bins, with an average bin length of 16.5 cM and 20.99 cM, respectively. A total of 256 EST-SSRs were assigned to bins and their map position was further validated by linkage mapping. EST-SSRs were found to be less polymorphic than genomic SSRs, but their transferability rate to chestnut, a phylogenetically related species to oak, was higher.

Conclusion: We have generated a bin map for oak comprising 256 EST-SSRs. This resource constitutes a first step toward the establishment of a gene-based map for this genus that will facilitate the dissection of QTLs affecting complex traits of ecological importance.

Background

Catalogues of Expressed Sequence Tags (ESTs) are developed from cDNA libraries to obtain expressional sequence information in contrasting environmental conditions or across developmental stages. When available, they also offer an inexpensive source of gene-based DNA markers, in particular SSRs [1]. Such collections of ESTs were produced in several plants providing a unique opportunity for searching SSR motifs and further develop the corresponding microsatellite markers [2]. Alternative and promising strategies to develop SSR markers from genome shotgun sequencing have recently emerged with the development of new generation sequencing technologies [3]. However, because ESTs correspond to coding DNA, the flanking sequences of EST-SSRs are located in well-conserved regions across phylogenetically related species, making them markers

* Correspondence: plomion@pierroton.inra.fr
[1]INRA, UMR1202 BIOGECO, F-33610 Cestas, France
Full list of author information is available at the end of the article

BioMed Central
© 2010 Durand et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of choice for comparative mapping and relevant functional and positional candidate genes to study their colocation with quantitative trait loci (QTLs).

The construction of a high resolution genetic map populated with SSRs requires considerable efforts, including the development of several hundreds of markers (depending on the number of linkage groups) and the genotyping of a large number of plants to ensure that most of the markers are correctly ordered, i.e. with a high LOD support for local ordering. Alternatively, bin-mapping or selective mapping [4] offers a less accurate but faster and cost-effective approach to locate many markers on an already existing framework map. This mapping strategy consists of genotyping a subset of highly informative offspring (the bin set) that are selected based on the number and position of crossover sites. In brief, the optimal bin set of a given size presents the maximum number of breaking points evenly spaced throughout the map, ideally resulting in a number of bins that is close to the number of framework marker intervals. This approach has been used successfully in peach [5], melon [6], strawberry [7] and apple [8,9]. Here, we use this approach for the first time in a forest tree species: oak.

Oaks represent a major component of the northern hemisphere forest. In particular, pedunculate *(Quercus robur L.)* oak is widely spread throughout Europe, from Spain to Russia (Ural mountains). This species is associated with important environmental (carbon sequestration, water cycle, reservoir of biodiversity...) and economic (carpentry, furniture, cabinet making, veneer, cask industry, fuel wood, hunting and fungus gathering) services. It has been used for years to study the genetic architecture of forest tree adaptation through common garden experiments [10,11], where natural populations growing in their native environments have been transplanted in a common environment, and QTL mapping studies [12-16], as well as to decipher the molecular mechanisms underlying adaptive traits such as bud phenology [17], water-use efficiency [18] and response to root hypoxia [15].

Different types of molecular markers were developed in *Q. robur* for linkage mapping to study the genetic architecture of adaptive traits. The different versions of the map included hundreds of random amplified polymorphic DNA (RAPD) markers [19], amplified fragment length polymorphisms (AFLP) [12] markers, and a set of 56 simple sequence repeats obtained from enriched genomic libraries (gSSRs) [20]. Because of their highly polymorphic nature and high degree of transferability across species, SSRs proved to be very useful markers to align different maps of *Q. robur* as well as to initiate a comparative mapping analysis with *Castanea sativa* (chestnut), another important Fagaceae species [20,21]. Despite combining interesting features (typically codominant and multiallelic, high polymorphism information content, evenly distributed throughout the genome, and high reproducibility) too few SSRs have been yet made available in oak to advance to more detailed genetic studies. The high cost associated with their development from enriched genomic libraries [22] and the lack of sequences for the genus *Quercus* genus probably contributed to the delay of the construction of a large battery of SSRs.

In this context, the main objectives of this study were: i/ to screen the oak ESTs for SSR motifs (*i.e.* type, frequency, and distribution of SSR motifs), ii/ to develop a set of EST-SSR markers and compile the data in a dedicated database, iii/ to compare their polymorphism information content with gSSR, iv/ to test the transferability of these markers in chesnut and v/ to map as much SSR loci as possible on two parental framework linkage maps of *Q. robur* using a bin-mapping approach. This study constitutes the first step toward the establishment of a consensus linkage map for oak based on SSRs segregating in several mapping populations.

Results

SSR mining and EST-SSRs frequency

SSRs were searched among the 28,024 unigene elements obtained from the assembly of 103,000 ESTs into 13,477 contigs and 14,547 singletons, using STACKpack™. The search was performed for di- (with a repeat count n ≥ 5 repeat units), tri- (n ≥ 4), tetra- (n ≥ 3), penta- (n ≥ 3) and hexa- (n ≥ 3) nucleotides, using the mreps software [23]. A total of 3,893 unigene elements contained at least one SSRs, resulting into 5,218 microsatellites, ie. a SSR frequency of 18.6%, taking into account multiple occurrences of SSRs in some unigene elements. As expected, the most frequent type of microsatellites corresponded to trimeric SSRs (2,212 unigene elements, i.e. 42% of the detected SSRs). This was followed by dimeric (1,713, 34%) and hexameric (574, 11%) SSRs. The abundance of tetrameric and pentameric SSRs was lower, representing only 8% and 5% of the microsatellites, respectively. The size of the SSR string varied from 10 bp (5 repeats for dinucleotide motifs) to 132 bp (66 repeats for an AG SSR) and the average number of repeats were 8.8 for dimeric (see additional file 1- table S1 for the distribution), 5 for trimeric (48.8% with 4 repeats), 3.5 for tetrameric (65.6% with 6 repeats), 3.2 for pentameric (81.2% with 3 repeats), and 3.4 for hexameric (72.5% with 3 repeats) SSRs. Among the dimeric SSRs, AG was found as the most common motif (70%), followed by AT (19%), AC (10.5%) and CG (0.1%). Similarly, for trimeric SSRs, the most common motifs were AAG (28%), ACC (14%) and AAC (12.4%). For the three other classes, the most common SSR types corresponded to AAAN (for tetrameric SSRs),

AAAAN (for pentameric SSRs), and AAAAAN (for hexameric SSRs). All these SSRs were made available in additional file 1 - table S1, which compiles information such as number of repeats, size of the motif, annotation *etc*.

Distribution of EST-SSRs

For 86% of the 5,218 SSRs, ESTscan [24] succeeded in estimating whether SSRs were located in non-coding (untranslated) (41.8%, including 21.5% di-, 8.5% tri- 2.8% hexa-SSRs) vs. coding (translated) (43.3%, including 2.2% di-, 31.3% tri- 7.5% hexa-SSRs) regions of each EST. The occurrence of each category in coding and non-coding regions is shown in Figure 1a. Overall, 67.3% and 32.7% of the non-coding SSRs were located at 5'- and 3'-UTR, respectively. Using FrameDP, 83% of the 5,218 SSRs was estimated in at least one predicted peptide (Figure 1b). As ESTScan, FrameDP prediction showed that smaller numbers of SSRs were located in non-coding (37.4%, including 14.6% di-, 11.1% tri- 3.7% hexa-SSRs) compared to coding regions (47.9%, including 11.4% di-, 27.5% tri- and 6.2% hexa-SSRs). Overall, 53.8% and 46.2% of the non-coding SSRs were located at the 5'- and 3'- UTRs, respectively. The most remarkable result obtained by FrameDP was the increased ratio of SSRs predicted in coding regions (from 43.3% to 47.9%), that can be attributed to a higher frequency among dinucleotide motifs compared to ESTscan.

Marker development

Of the 5,218 SSRs motifs identified, we designed primer pairs for 748 SSRs (additional file 2 - table S1), including 348 di-, 320 tri-, 2 tetra-, 1 penta-, and 77 hexa-nucleotide SSRs. Locus ID, forward and reverse primer sequences, type of motif and length, amplification and polymorphism in the tested full-sib pedigree have been reported in additional file 3 - table S1. A total of 568 primer pairs (75.8%) amplified a PCR product, among which 283 (154 di-, 107 tri-, 1 tetra-, 1 penta- and 20 hexa-nucleotide SSRs) were found to amplify a single polymorphic locus, i.e. 37.7% of the total number of tested primers. It was also found that the level of polymorphism depended on the type of motif (Figure 2). These loci segregated in the testcross configuration, i.e. 1:1 ratio (65 loci in the male and 77 loci in the female parent), or in the intercross configuration, i.e. 1:1:1:1 ratio (135 loci in both parents) or 1:2:1 ratio (6 loci in both parents). Markers segregating 1:1:1:1 were recoded in the 1:1 ratio in the male and female parents.

Transferability of EST-SSRs

A subset of oak EST-SSRs were also tested for their transferability in chestnut (*Castanea sativa*) another important Fagaceae species. A total of 100 dinucleotide EST-SSRs were tested for their amplification on two DNA specimen (additional file 4 - table S1), from which 63% amplified a single PCR product, a figure that is significantly higher than that obtained for the transferability of dinucleotide genomic SSRs from oak to chestnut, i.e. 47% in [20]. In addition, electronic PCR was carried out against unigene elements for *Quercus mongolica* (Qm) [25] and *Castanopsis sieboldii* (Cs) [26]. There were 52 oak primer pairs that amplified Qm with no mismatch and product size similar to that for European oaks. Six primer pairs amplified two different Qm sequences. For Cs, there were 18 primer pairs that can amplify Cs with no mismatch. One primer pair

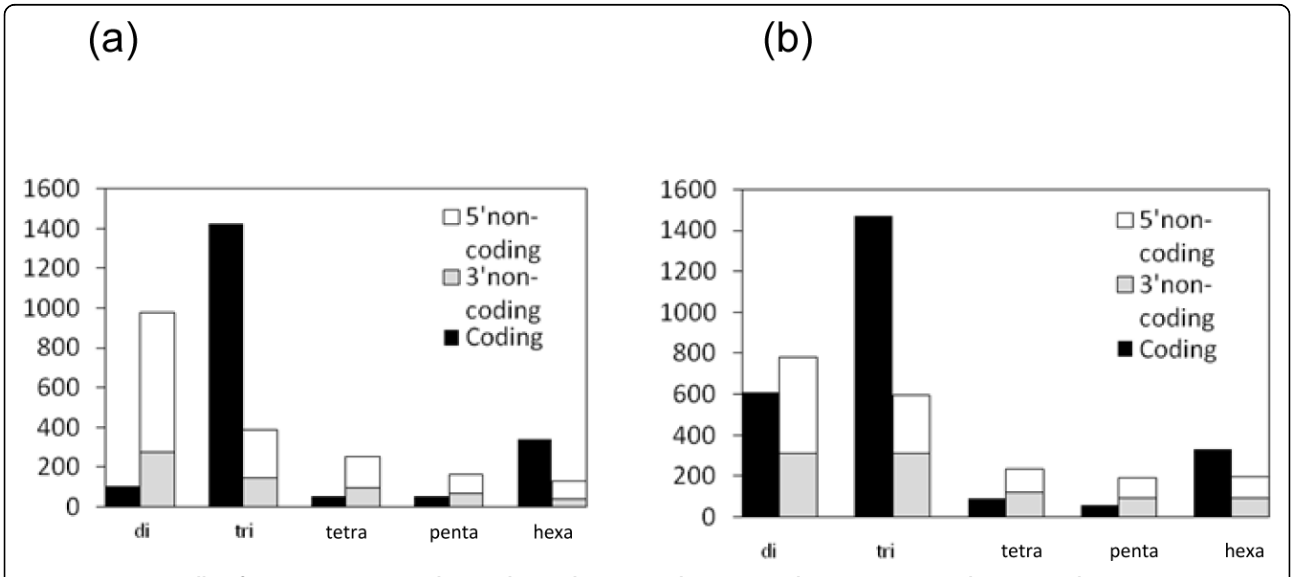


Figure 1 Microsatellite frequency among coding and 5' and 3' non-coding regions by ESTScan (a) and FrameDP (b).

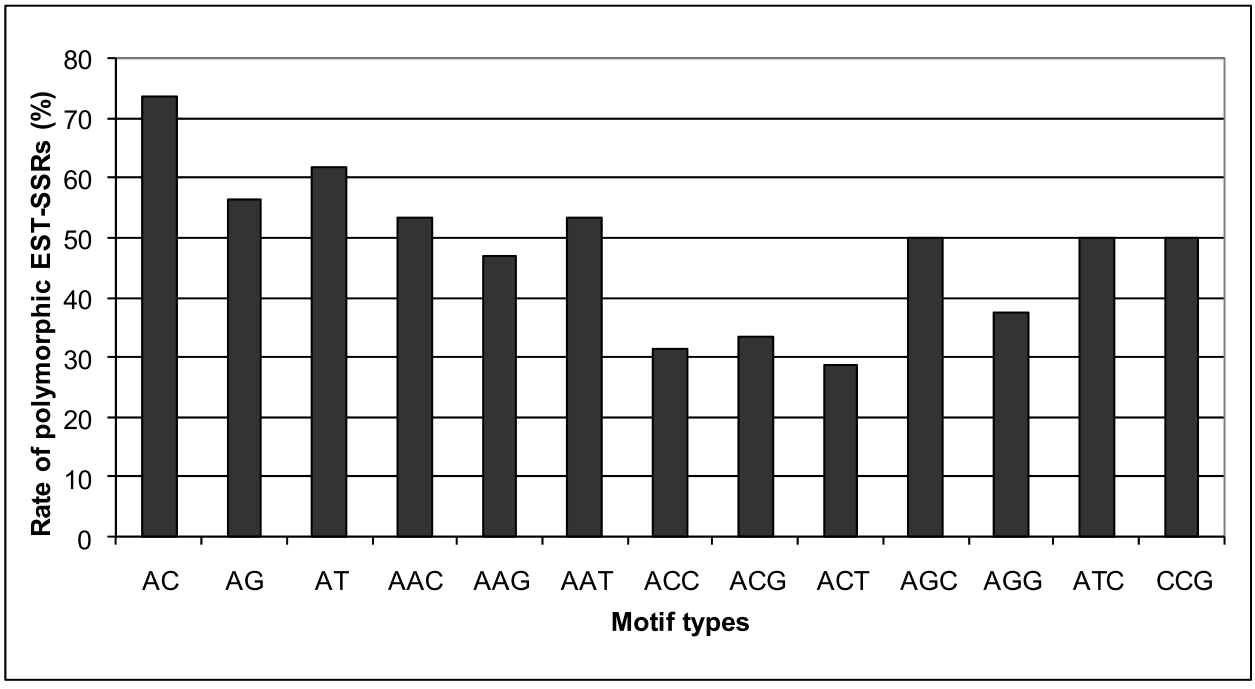


Figure 2 Rate of polymorphism for different types of di- and tri- SSRs.

amplified two different Cs sequences. Seven primer pairs produced ePCR products for both Qm and Cs. Three primer pairs in the present study targeted three unigene elements for which SSR markers were already developed for Qm.

Comparison between genomic and EST-derived SSRs

A total of 16 dinucleotide genomic SSRs from Alberto *et al*. [27] and 16 dinucleotide EST-SSRs (from this study) were genotyped on the same set of 288 *Q. petraea* genotypes described in [27]. The comparison (taking into account heterogeneous sample size using the rarefaction methods from El Mousadik and Petit, [28] of genetic diversity (He) and allelic richness (A) showed that gSSRs were more polymorphic (He = 0.82 A = 4.34) than EST-SSRs (He = 0.77 and A = 3.78). Other diversity statistics as the size range of the SSR motifs and the number of alleles confirmed the lower level of polymorphism of EST-SSRs compared to gSSRs. The size of the SSR motif was on average 46.75 bp for gSSRs and 26.25 bp for EST-SSRs. The total number of alleles present in the tested population, regardless of their frequency was 21.06 vs. for gSSRs and 12.25 bp for EST-SSRs

Bin mapping

The two parental maps established by Saintagne *et al*. [12] using Mapmaker 2.0 [29] were first reconstructed (Figure 2) using Joinmap v4.0 [30] based on the same 128 framework markers and 278 progenies. The female map was covered by 38 AFLPs, 6 RAPDs and 28 gSSRs resulting in 63 marker intervals spanning 728.8 cM. The male map was divided by 60 marker intervals and comprised 43 AFLPs, 4 RAPDs and 23 gSSRs for a total map length of 776.9 cM. Each linkage map consisted in 12 linkage groups that corresponded to the number of haploid chromosomes in oak. Compared to the map previously constructed using Mapmaker, very few differences were noticed, consisting mainly in few inversions (ZQR5a and E-AAC/M-CAC-202/3 on LG8F, E-AAG/M-CTA-150/5 and E-AAC/M-CTT-120 on LG4M) and three unlinked markers (E-AAG/M-CTT-168 on LG10F, and E-AAG/M-CTT-363 on LG10M and P-CCA/M-ATA-335 on LG12M). The total map lengths were however quite different (929 vs. 728.8 cM for the female map and 890 vs. 776.9 cM for the male map, using Mapmaker and Joinmap, respectively). Similar results have been reported elsewhere (e.g. [31] and [32]) and is attributed to the method used by the software to calculate Kosambi genetic distances.

Using the bin set of 14 offsprings, the framework maps were divided into 44 and 37 bins resulting in an average bin length of 16.5 cM and 20.9 cM for the female and male map, respectively. Double crossing-overs were taken into account to define the bin set in order to minimize the effect of possible genotyping errors. The longest bins identified spanned 38.1 cM (bin 10.2) for the female and 79.9 cM (bin 5.1) for the male map. On average, there were 1.88 and 1.80 different

genotypic points between contiguous bins in the female and male maps. Therefore, more genotypic combinations might exist to fit within intermediate positions.

A total of 283 polymorphic EST-SSRs were genotyped on the bin set and the parents of the full- sib pedigree (Figures 3, 4). Overall 256 markers were assigned by graphical genotyping (i.e. graphical representation of genotypic information for individual genotypes as defined by Young and Tanksley [33]) to their respective bin. The remaining 27 markers corresponded either to markers segregating 1:2:1 (6 loci) or presented ambiguous bin positions (21 loci) and were therefore left out from the analysis. On the female map, 198 markers were assigned to bins, giving an average of 4.5 markers per bins ranging from 0 (bin 5.1, 6.3, 9.5) to 18 (bin 2.6). On the male map, 185 markers were assigned to bins, giving an average of 5

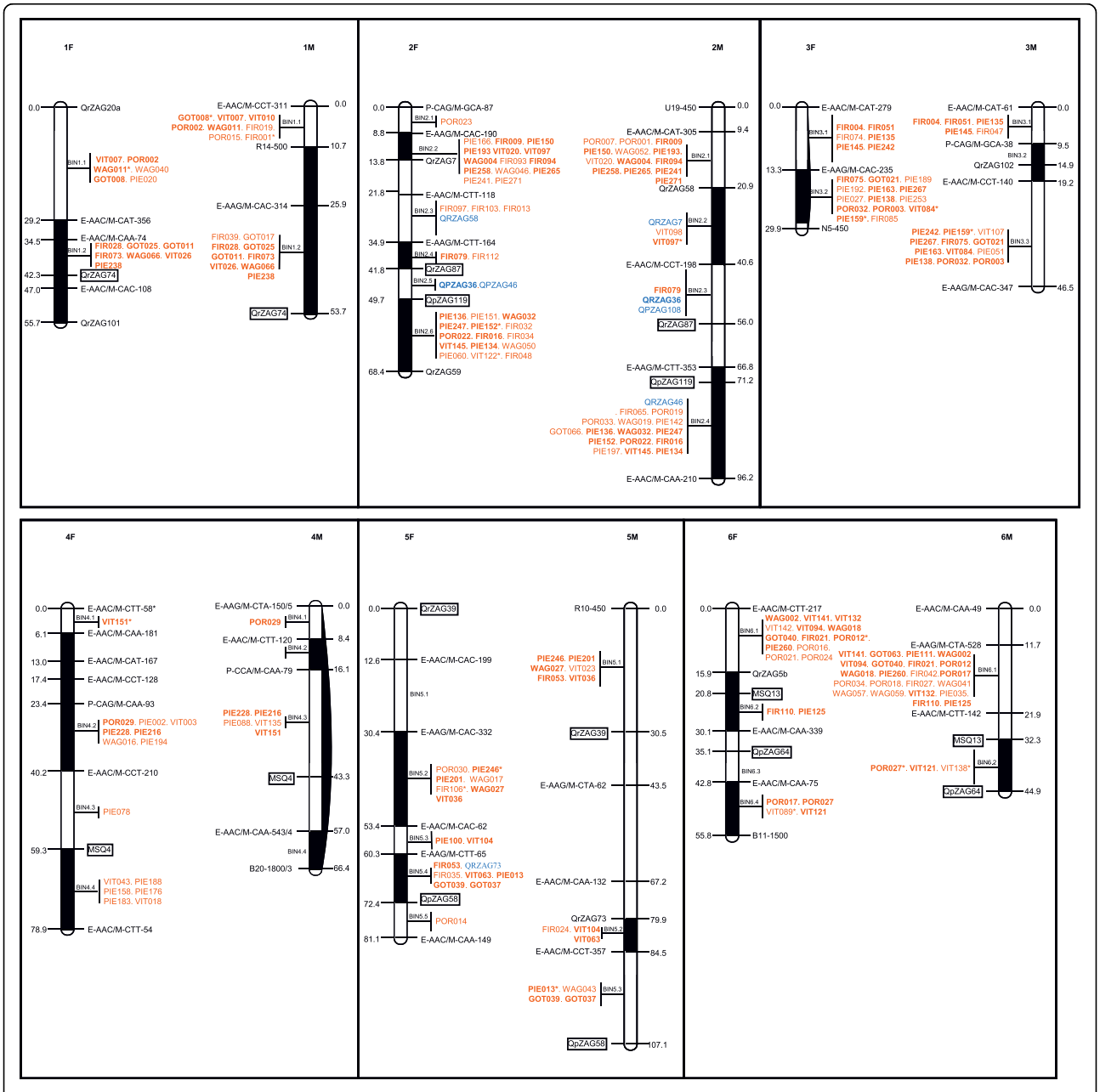


Figure 3 Bin position of EST-SSRs for linkage groups 1 to 6. In black: framework markers (AFLP, RAPD, gSSR), in red: EST-SSRs, in blue: gSSRs, squared: fully informative gSSR framework markers. An asterisk indicates SSRs with ambiguous position. Bold type indicates fully informative EST-SSRs and gSSRs. F: female map, M: male map.

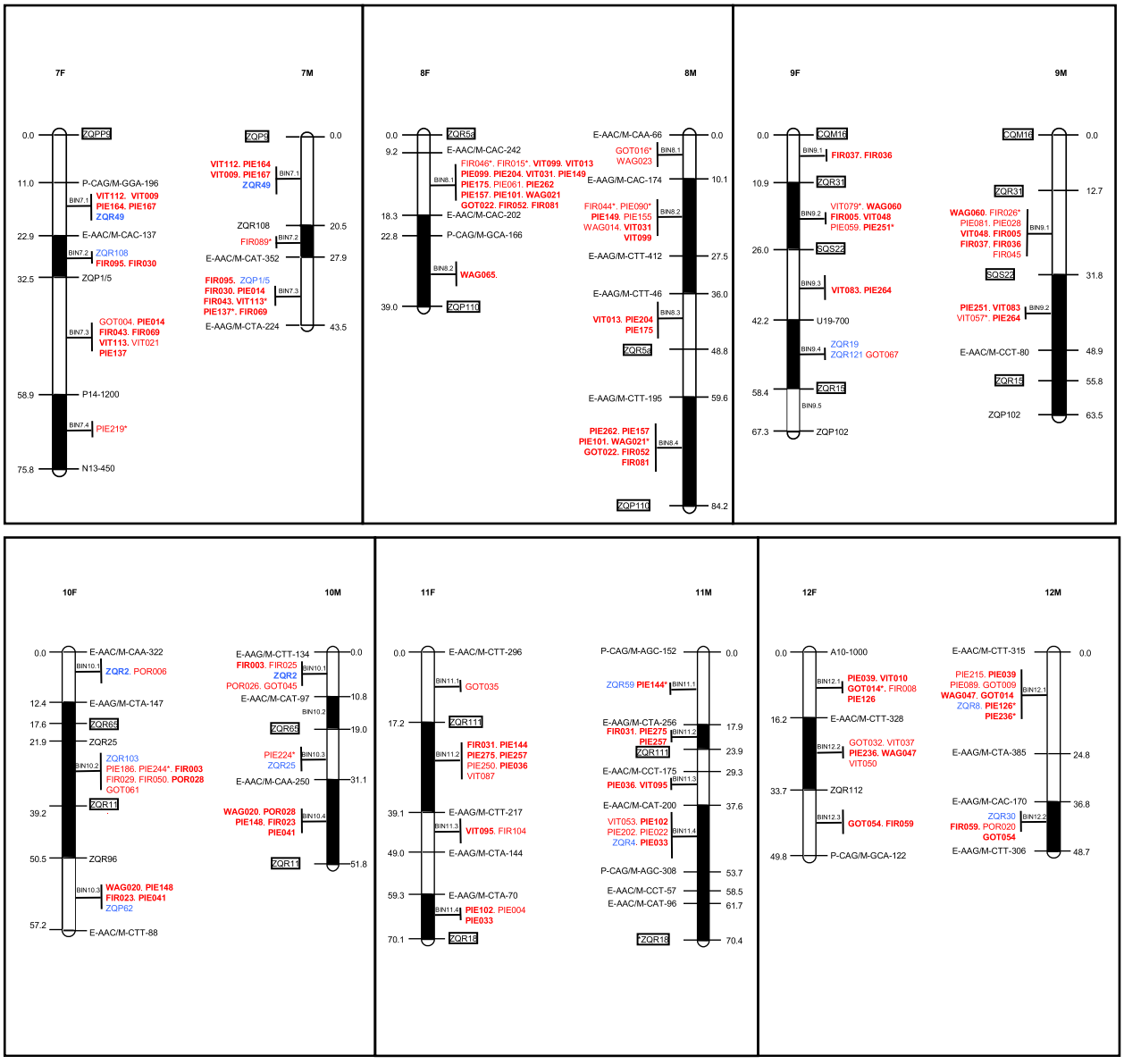


Figure 4 Bin position of EST-SSRs for linkage groups 7 to 12. In black: framework markers (AFLP, RAPD, gSSR), in red: EST-SSRs, in blue: gSSRs, squared: fully informative gSSR framework markers. An asterisk indicates SSRs with ambiguous position. Bold type indicates fully informative EST-SSRs and gSSRs. F: female map, M: male map.

markers per bin ranging from 0 (bin 3.2, 4.2, 4.4, 10.2) to 22 (bin 6.1). Overall, EST-SSRs were evenly distributed across the linkage groups. More precisely, respectively 69 and 78 markers for the female and the male map presented exactly the same genotypic information as bin framework markers, i.e. these markers were positioned at the same location as the markers used for the definition of bins. The others, 104 and 86 markers in the female and in the male map, respectively, were positioned in the bins, presenting a genotype that was compatible with an intermediate bin between two successive bin markers. This is likely the result of large average bin size defined over low marker density framework maps. Only 25 and 21 markers in the female and male maps were involved in one or more double crossing-overs, respectively. Their genotypes were double checked, confirming this observation. These markers were visually assigned to their most probable bins.

Validation of bin assignment

To test the efficiency of bin mapping, we first compare the known map location of 19 accessory gSSRs (blue

type in Figures 3, 4) from the map constructed by Barreneche et al. [20], to their bin positions inferred from the graphical genotyping of 14 F1s. In all cases, both approaches agreed (additional file 5 - table S1), i.e. markers were located either on the same bin (18 markers of class A according to the categories presented in the methods section) or an adjacent bin (1 marker of class B: ZQR49). An *a posteriori* validation was also performed for 146 EST-SSRs (on the female map 47 markers corresponding exactly to bin markers and 54 markers characterized with ambiguous position, on the male map 47 markers corresponding exactly to bin markers and 47 markers characterized with ambiguous position) genotyped on 46 progenies. On the female map, 77 markers showed identical positions between bin assignment and map location (class A), 12 were located in an adjacent bin (class B), 1 was mapped on the same linkage group (class C), and 11 presented a LOD score for linkage < 2 (class D). Overall, the bin assignment was validated for 89% of the markers (class A+B). For the male map, 72, 11, 0 and 11 markers were of class A, B, C and D, respectively, corresponding to a validation rate of 88%. A slightly higher validation rate was obtained for another set of 65 EST-SSRs (53 inter-cross, 7 female and 5 male test-cross markers) genotyped on 92 offsprings, i.e. 98.3% on the female map (53 A, 6 B and 1 D markers), 94.8% on the male map (51 A, 2 B, 2C and 3 D markers).

Macro-synteny and colinearity

About the conservation of macro-synteny between the male and female maps, it should be noticed that all the 129 inter-cross markers (indicated in bold in Figures 3, 4) were found on homologous linkage groups. A conserved macro-colinearity was also verified based on the 55 inter-cross markers (21 gSSRs and 34 EST-SSRs) genotyped on the extended set of 92 progenies. These markers presented the same order on both maps as illustrated in additional file 6 - figure S1, but with one exception on LG9. Given the number of comparisons, 2 occurrences with different orders were expected by chance alone at a 5% type I error rate. This investigation also provided the opportunity to test whether the male and female gametes presented different levels of recombination. Based on 33 intervals flanked by the same adjacent markers in the male and female maps, no statistical difference was found using a t-test for paired comparisons (data not shown).

Discussion

Frequency, distribution and polymorphism of the oak EST-SSRs

EST-derived SSRs have been searched for many years in plant, animal and microbial species. Despite a lower rate of polymorphisms compared to genomic SSRs (confirmed in the present study), EST-SSRs offer a number of advantages over genomic SSRs [2]: (i) their development requires no investment in *de novo* sequencing; (ii) they detect variation in the expressed portion of the genome; (iii) the conservation of primer sites makes them readily transferable across closely related species as illustrated here between oak and chestnut; and (iv) in most cases they can be exploited for population genetic analysis [1].

The number of SSRs detected in ESTs largely depends on the size of the EST catalogue, the algorithm [34] and criteria (type of repeat motif and minimum number of repeat units) used to detect SSR-containing sequences. It is therefore difficult to conclude about the percentage of genes harbouring SSR motifs. This is apparent from several studies: (i) in *Oryza sativa* 40.4% [35] and 50% [36] of EST-SSRs were detected using different software and criteria; (ii) Kumpatla and Mukhopadhyay [37] analysed 1.5 million ESTs derived from 55 dicotyledonous species and found that 2.6 to 16.8% of ESTs contained at least one SSR; and (iii) because the level of polymorphism is positively correlated with the length of the repeats region (see next paragraph), some authors have chosen to use more stringent criteria (i.e. increase the minimum number of repeat units in the detection phase) to increase the probability to find polymorphic SSR markers.

The availability of several genome sequences in angiosperms makes it possible to more accurately estimate the proportion of gene models harbouring SSRs in transcribed and UTR regions. In poplar for example, about 6,000 SSRs were found in coding regions and UTRs [38]. Therefore, taking into account the 45,000 putative protein-coding genes [39], 13.4% of the genes would present a SSR. In *Arabidopsis thaliana*, 44% of the 27,158 putative genes contain one or more SSRs [40], but this figure also includes non transcribed regions.

In oak we found that 18.6% of the unigenes presented at least one SSR motif. In two other Fagaceae species, *Quercus mongolica* [25] and *Castanopsis sieboldii* [26] and it was found that 11.8% and 12.8% of the putative unigenes presented microsatellite motifs (from di- to tetra-nucleotide repeats). Taking into account only di-, tri- and tetra-nucleotide repeats, these figures are very similar to our finding (13.4%), although the detection parameters were different (9 for di-, 6 for tri-, 5 for tetra-nucleotides). Also in terms of the abundance of motif types, our study agrees to that of Ueno *et al.* [25,26] and other studies performed in dicotyledonous species (reviewed by Kumpatla and Mukhopadhyay [37]), i.e. AG and AAG were the most abundant di- and trimeric SSRs, respectively. The

extremely low number of SSR motifs containing C and G (2 CGs out of 1,713 dimeric SSRs and 103 CCGs out of 2,212 trimeric SSRs) could be attributed to the composition of dicot genes being less rich in G+C compared to monocots due to codon usage bias [41] and to the intrinsic negative correlation between GC content and slippage rate [42].

As expected, the most frequent SSR class corresponded to trinucleotides (42%). This suggests that many of the detected EST-SSRs are in protein-coding regions because changes in trinucleotide repeat number will not cause frame shifts unlike changes in other types of motifs [43]. Indeed, the analysis of the distribution of the EST-SSRs clearly showed that this type of SSR was frequently found (ranging from 27.5% to 31.3% based on FrameDP or ESTscan analysis, respectively) in coding regions in contrast to other SSRs. As for dimeric SSRs, the second most abundant type, our results confirm what has been obtained in other studies, i.e. they were mostly located in non-coding regions, despite a noticeable difference obtained between FrameDP (14.6%) and ESTscan (21.5%). Overall, it should also be noticed that most of the EST-SSRs found in non-coding region were located in the 5' UTR (ranging from 53.8% to 67.3% based on FrameDP or ESTscan analysis, respectively). Higher density of SSR in the 5' UTR was also found in rice [44]. This result could be attributed to either a technical bias (ESTs being mainly generated from their 5'-ends) or a biological feature of plant genes as discussed by Grover et al. [44] and Fujimori et al. [45]. These authors found that rice and Arabidopsis genes presented a higher rate of SSRs in the 5' flanking regions of the genes and interpreted this finding as a regulatory role in gene expression.

To further explore the accuracy of FrameDP and ESTscan results, we carried out a complementary analysis using poplar full length cDNAs for which structural annotations were available [46]. The result of this analysis is provided as supplemental data (additional file 7 - figure S1). By comparing the SSR location based on true structural annotations it was clearly shown that ESTscan performed better than FrameDP, the later over-estimating the presence of dinucleotide motifs in coding regions as was found with the oak data. In agreement with the data reported in rice and Arabidopsis, it was also found that SSRs were more frequent in the 5'UTR of poplar genes (additional file 7 - figure S1).

A total of 748 primer pairs were designed and tested on a set of 4 genotypes, among which 568 (75.8%) yielded amplicons. The failure for 24.2% of the primers to generate an amplicon can be explained: i/ by the presence of large intronic regions preventing genomic DNA to be amplified, ii/ the presence of SNPs/INDEL variation in the priming site of the tested genotypes, preventing the hybridization between the primers and the target DNA, iii/ by the fact that a single PCR program was used without further optimisation, iv/ because the M13 tail (that was added to each forward primer) may interfer with appropriate PCR amplification [47], and v/ because primers could have been designed for chimeric unigene elements. A large proportion (285 out of 568, i.e. 50%) of the successful primer pairs were either monomorphic (163 EST-SSRs) or produced multi-banding patterns or yielded faint amplification (122 EST-SSRs), thereby preventing the development of single copy SSRs. This study reveals that polymorphic SSRs (283 loci) tended to have a higher number of repeats (based on the EST data), ie. 10.58 for di, 7.27 for tri- and 3.4 for hexa-SSRs, compared to monomorphic ones (163 loci), i.e. 9.80 for di-, 6.29 for tri-, and 3.20 for hexa-SSRs. The effect of repeat number and motif on the polymorphism was surveyed using logistic regression model by the R software v. 2.6.2 (R Development Core Team 2008), and the effect of repeat number was highly significant (estimate of correlation coefficient for repeat number = 0.237 and $P < 0.001$). This result agrees with the significant positive correlation that was found between SSR length and polymorphism rate in plants and animals [48].

In oak, polymorphic markers were not evenly distributed among repeat classes, amounted to 58.7%, 44.3% and 36% for di- tri- and hexa- repeats, respectively. These figures confirm the higher level of polymorphism of dinucleotide repeats among plants [49-51]. The lower level of polymorphism for tri- and hexa- SSRs is mainly related to their location in translated sequences compared to dimeric SSRs that were preferentially distributed in UTRs. These observations suggest that natural selection limit both the number and polymorphism rate of SSRs in translated regions of the genes. Moreover, a closer examination among perfect di-and tri- oak SSRs showed that the level of polymorphism (Figure 2) depended on the type of motif. In particular, SSR markers with dinucleotide AC were the most polymorphic loci. These considerations should be taken into account for the development of additional polymorphic SSRs in oak that are conserved among the Fagaceae species, comparative genomics being our ultimate goal. In that respect, we showed that oak dinucleotide EST-SSRs were highly transferable to European chestnut.

Bin mapping

Linkage mapping is a time consuming process that requires large size recombinant populations (from which progenies are randomly chosen) to locate polymorphic markers onto a genetic map. Other methods that do not rely on meiotic recombination have also been developed to assign any genes to chromosomal locations, such as

the use of aneuploid and deletion stocks in polyploids or radiation hybrid panels. One important advantage of these methods is that any sequence of interest is readily placed on a radiation hybrid or deletion map. In contrast, only polymorphic markers can be mapped on a genetic map. However, such approaches have been limited to a handful of plant species, including wheat [52,53]. Alternatively, a computational method was developed [4] to optimize the construction of high-density linkage maps using a reduced sample of selected offsprings presenting complementary recombinational events throughout the genome. A prerequisite to such selective/bin mapping approach is the availability of a high-confidence framework map. The first bin mapping approach was recently implemented in peach [5]. Using only 6 F_2 progenies, their F_1 hybrid parent and one of the grand-parental lines, these authors successfully assigned 264 SSRs to 67 bins of the peach map. The bin mapping strategy was also used in melon (121 SSRs/14 plants [6]; 200 SNP-based markers/14 plants [54]), apple (31 SSRs/14 plants [8]) and strawberry (103 SSRs/8 plants [7]).

A bin mapping approach was developed for the first time in a forest tree species to increase the density of SSR markers in the oak linkage map and provide orthologous anchor markers for comparative mapping within the Fagaceae. The selection of the bin set combined the use of Mappop software and visual inspection of the data. It resulted in the selection of 14 plants, which was considered as a suitable size, as a set of 16 samples (14 F1s and both parents) fits in standard 96-well PCR plates. With this subset, 44 (for the female map) and 37 (for the male map) bins were obtained. As expected based on the number of different genotypic points between adjacent bins, about half of the markers presented a genotype that was compatible with a putative bin between two contiguous bins. To investigate the accuracy of the bin mapping approach, a large number of EST-SSRs was genotyped on an extended set of genotypes (46 or 92 F1s). Most markers assigned to bins or putative bins were placed in the expected position, validating the bin mapping strategy for oak, despite the low number of bins compared to similar studies [5,6]. At this stage, it is difficult to propose a general guideline for further bin mapping studies, but some general recommendations can be made: i/ Number of individuals to be included in the bin set: it largely depends on the population and marker types. For instance, there are more genotypic informations in F2s as compared to F1s for codominant markers (3 *vs*. 2 genotypic classes, respectively). Therefore, less individuals will be needed to define the bins with F2 genotypes. It also depends on technical constraints, 14 individuals emerging as a magic number in the few bin mapping studies published so far in plants, since 16 samples, corresponding to 14 offsprings and two parental lines, fits well in a single raw of a 384-well microtiter plate!, ii/ Number of bins: it obviously depends on the number of linkage groups and on the number of individuals included in the bin set (i.e. the more individuals, the more number of bins).

Conclusion

In the present study we used an EST catalog produced for *Quercus petraea* and *Q. robur*, to mine and develop EST-derived SSRs. We observed a relatively high abundance of single sequence repeats in the oak transcriptome, 18.6% of the unigene elements harboring at least one SSR. Despite being less polymorphic than gSSRs, their many advantages make them markers of choice for genetic analyses. In particular, these functional markers directly sample variations in genes, which enhance their value for analyzing the genetic basis of forest tree adaptation through the use of non neutral, so called "functional" markers in genetic diversity analysis, QTL and association mapping studies as well as comparative genomics.

The present study contributed 283 gene-derived microsatellite markers, 255 of which were efficiently assigned to a bin position using 14 informative individuals. The development and distribution of this reference set of highly recombinant genotypes to the "European oak mapping community" has been instrumental for the development and mapping of this new set of high quality markers that also proved to be useful in a related species (chestnut).

Methods

Plant material and DNA extraction

The bin set and the verification panel were selected from the *Quercus robur* full-sib family (3PxA4) described by Saintagne *et al*. [12] The population that was used to compare the level of polymorphism between genomic SSRs and EST-SSRs is described by Alberto *et al*. [27]. DNA was extracted from leaves using DNeasy plant mini kit (Qiagen, Hilden, Germany).

EST-SSRs detection

SSR motifs (5, 4, 3, 3, and 3 repeats at least for di-, tri-, tetra-, penta- and hexa-nucleotides, respectively) were searched within the first version of the oak unigene set established from the assembly of 103,000 ESTs (available at EMBL). These ESTs were derived from about 20 cDNA libraries constructed from mRNA extracted from 4 tissues (bud, leaf, xylem and root) collected on *Q. robur* and *Q. petreae* genotypes. The main objective to generate such a large number of ESTs was to catalogue as many as possible non-redundant genes (unigene set) of oak. These ESTs were assembled to avoid redundancy in SSR detection using the transcript reconstruction

system stackPACK™ [55] from the SAMBI Institute. This pipeline uses the following programs: Cross_Match [56] to clean up the sequences, d2_cluster [57] to perform a loose first stage clustering, PHRAP [58] to assemble these clusters into contigs and finally CRAW [59] to generate the longest consensi.

SSRs motifs were searched using mreps (v. 2.5) [23]. In a comparative study in *Pinus pinaster* (G. Le Provost, unpublished) mreps was found to be more stringent compared to SSRIT [60] and Sputnik v1.22 (http://abajian.net/sputnik/). Once detected, SSRs located 35 nucleotides from either end of each unigene element were discarded to keep enough sequence information for primer design. In addition, those SSRs that were immediately adjacent to each other (separated by less than 30 nucleotides) were merged into a single SSR. The output of mreps was converted into a standard csv file corresponding to the SSR database structure put in place in the frame of the Evoltree project. Specific information for each SSR included the unigene element ID and the annotation, the repeat motif, its length and position (additional file 3 - table S1, also available through the *Quercus* portal (https://w3.pierroton.inra.fr:8443/QuercusPortal/Home.jsf).

ESTscan [24] and FrameDP [61] were used to estimate the location of a coding region within unigenes. By combining the output from mreps, the location of EST-SSR (either coding or noncoding regions) was estimated. Microsatellites, for which no results were returned by each software or location was covered across both coding and non-coding regions, were discarded. Because there are no annotated full-length genes available for oak yet, we used *Arabidopsis thaliana* sequences as a training set for the analysis performed by ESTScan. The resulting matrix was used for peptide prediction of oak unigenes. For the analysis using FrameDP, no specific training set is required.

SSR genotyping

Primer pairs were designed for 748 unigene elements (including 348 di-, 320 tri-, 2 tetra-, 1 penta-, 77 hexa-nucleotides) using Primer3 [62]. A M13 tail (TGT AAA ACG ACG GCC AGT) [63] was added to the 5'-end of the forward primer to facilitate exchange of primers between the partners of the network that used different capillary electrophoresis systems: i.e. ABI3730 (Applied Biosystems, Carlsbad, CA, USA), Licor 4300 (Licor, Lincoln, NB, USA), Megabace (GE Healthcare, Buckinghamshire, UK). PCR reactions were performed in a final volume of 10 μL containing: 1× PCR-buffer [10 mM Tris-HCl, 50 mM KCl 1.5 mM $MgCl_2$, pH 8.3 at 25°C] (BioLabs, Ipswich, England), 100 μM of dNTPs, 0.045 μM of forward primers, 0.165 μM of reverse primer (5 μM), 0.165 μM of M13 primer, 0.25 U of Taq polymerase (BioLabs) and 6 ng of plant DNA. The cycling conditions were as described by Shuelke *et al* [60]: i.e., a first denaturation at 94°C during 4 minutes, 35 cycles at three temperatures, 94°C for 30 s, 56°C for 45 s, and 72°C for 45 s. Additionally 9 cycles were run at 94°C for 30 s, 53°C for 45 s, and 72°C for 45 s and a final extension at 72°C for 10 minutes and a cooling at 10°C. Data generated were analysed using the GeneScan 3.7 and Genotyper 3.7 softwares for ABI, 4300 DNA analyser software for Licor and Fragment Analyser version 1.2 for MegaBace sequencing machine.

Nomenclature of the markers

EST-SSR marker ID consisted of: three letters to identify the lab where they were developed i.e, PIE for those designed in Pierroton (INRA, France) followed by a serial number. Genomic markers were designated according to the restriction enzymes and the primer combination used, and their amplification size. RAPD markers were named as follows: the letter and the first digit refers to the identification of the OPERON primers [64] and the last digits correspond to the molecular weight of the polymorphic bands.

Bin mapping strategy

A total of 748 primer-pairs were tested for amplification and polymorphism on both parental trees and two progenies. Given the relatively high number of putative markers, a bin mapping approach was followed (summarized in additional file 8 - figure S1) with the main objective of minimizing the number of trees to be genotyped, while assigning the markers to their most probable map location. From the initial dataset (278 F1s × 953 markers) a double screen was first applied, consisting of selecting individuals with < 50% missing data and markers with a LOD support for local ordering ≥3 (i.e. framework markers according to Saintagne *et al.* [12]), resulting in a total of 66 individuals and 128 testcross (1:1 segregation) and intercross (1:1:1:1 segregation recoded as 1:1 in each parent) markers. Male and female framework maps were then generated under the two-way pseudo-testcross mapping strategy [65] using the regression mapping algorithm of Joinmap v4.0 [30]. These two datasets were used to select a smaller number of highly recombinant progenies as follows: i/ a first set of 46 plants was selected based on maximizing the number of breakpoints along the 24 linkage groups (12 in the male and 12 in the female maps), using the Mappop software [4,66], and ii/ a final subset of 14 F1s (the bin set: #109, #110, #116, #121, #127, #128, #131, #151, #162, #165, #166, #172, #176, #196) was retained by visual inspection, combining three additional criteria: i) selection of individuals with missing data < 10% and presenting a minimum of duplicated bins; ii)

optimisation of both female and male map coverage with the smallest bin size as possible, and iii) minimization of double crossing-over between adjacent framework markers. The bin set (and the parental lines) were finally genotyped for all "mappable" markers segregating in testcross (1:1 ratio), intercross (1:2:1) and outcross (1:1:1:1 ratio) configurations. The EST-SSRs were assigned to their most probable bin by matching their genotypic profile to that of the framework markers. Bins were coded by a two-digit number, the first corresponding to the linkage group ID (1 to 12) and the second to their numerical order.

Validation of bin assignment

To further test the efficiency of the bin mapping approach, we compared the bin location (obtained as described above) with the map location of SSRs. The map position was estimated on an extended set of genotypes using the two-point test for linkage implemented in Joinmap. An *a priori* validation was first carried out based on 19 genomic SSRs (indicated in blue in Figure 2) that were already genotyped and mapped by Barreneche *et al.* [20]. An *a posteriori* validation was also performed for 146 and 65 non-overlapping EST-SSRs that were genotyped on 46 and 92 progenies, respectively. Markers presenting a LOD score for linkage > 2 (for 46 F1s) or 3 (for 92 F1s) were classified into three categories: class A for markers for which the nearest framework marker (FM) was included in the bin, class B for markers for which the nearest FM was found in an adjacent bin, and class C for markers for which the nearest FM was located in a more distant bin or else in another linkage group. Markers presenting a LOD score for linkage below these thresholds were classified as D marker.

Genetic diversity analysis

Genetic diversity statistics (gene diversity He [67]) and allelic richness (A) were estimated for 16 genomic and 16 EST-derived SSRs using the program Fstat 2.9.3.2 [68]. Allelic richness (A) was calculated using the rarefaction method developed by El Mousadik and Petit [28].

Additional material

Additional file 1: Table S1. Occurrence of non-redundant SSRs in the oak unigene, according to the SSR motif and number of repeats.

Additional file 2: Table S1. Characteristics of the *Quercus* EST-SSRs.

Additional file 3: Table S1. SSR database.

Additional file 4: Table S1. Transferability of dinucleotide EST-SSRs from oak to chesnut.

Additional file 5: Table S1. Segregation, bin and map position of *Quercus* gSSRs and EST-SSRs.

Additional file 6: Figure S1. A macrosynteny map for oak based on 55 intercross SSRs. In black: framework markers (AFLP, RAPD), in red: EST-SSRs, in blue: gSSRs. Bold types indicate fully informative SSRs. Female linkage groups on the left (F), male linkage group on the right (M).

Additional file 7: Figure S1 Location of EST-SSRs based on FrameDP (a), ESTscan (b) and structural annotation (c) for a set of 4,664 poplar genes. Methods. 1. 4,664 full-length cDNA sequences of poplar, downloaded from Genbank. 2. SSRs searched using mreps program with default parameters. 3. Coding sequences estimated by FrameDP and ESTScan. A matrix based on Arabidopsis CDS was used for ESTScan. 4. SSR location (coding or non-coding) inferred by combining FrameDP and mreps results (Figure S1a) and ESTScan and mreps results (Figure S1b). SSR locations were also determined using mreps results and structural annotation for the corresponding cDNA (Figure S1c). Results. Figure S1a: SSR location based on the estimation by FrameDP. Figure S1b: SSR location based on the estimation by ESTScan. Figure S1c: SSR location based on structural annotation.

Additional file 8: Figure S1. Schematic representation of the bin mapping strategy.

Acknowledgements
The study has been carried out with financial support from the European Commission under the FP6 program (FP6-2004-GLOBAL-3, Network of Excellence EVOLTREE "Evolution of Trees as drivers of Terrestrial Biodiversity", N°016322). JD was supported by doctoral fellowships from EVOLTREE. The authors thank two anonymous referees for their thorough review and highly valuable comments and suggestions, which significantly contributed to improving the quality of the paper.

Author details
[1]INRA, UMR1202 BIOGECO, F-33610 Cestas, France. [2]Université de Bordeaux, UMR1202 BIOGECO, F-33610 Cestas, France. [3]Plant Genetics Institute, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino (FI), Italy. [4]Forest Genetics and Forest Tree Breeding Büsgen Institute Faculty of Forest Sciences and Forest Ecology Göttingen University, Büsgenweg 2, Göttingen, 37077, Germany. [5]School 07 Forest Resources and Environmental Science, Michigan Technological University, Houghton 49931, Michigan, USA. [6]ALTERRA - Wageningen UR, PO Box 47, Wageningen, 6700 AA, The Netherlands. [7]CNR Istituto di Biologia Agroambientale e Forestale, Porano (TR), 05010, Italy. [8]NEIKER, Dpto Biotecnologia, Vitoria-Gasteiz, 01080, Spain. [9]CBiB - Université Victor Segalen Bordeaux 2 146, rue Léo Saignat, 33076 Bordeaux, France. [10]Forestry and Forest Products Research Institute, Department of Forest Genetics, Tree Genetics Laboratory, 1 Matsunosato, Tsukuba, Ibaraki, 305-8687, Japan.

Authors' contributions
This article is a part of JD's PhD thesis supervised by CB and CP. The idea of the study was developed by CB, AK and CP. AK coordinated the Evoltree project than funded this research. CB coordinated the present study. Marker development was carried out within the Evoltree network by JD, EC, GV, AB, FS, CM, MC, OG, HPK, FV, CM, MC, PGG, AH, ZI. The writing of the manuscript was performed by JD and CP. JD conducted the bin mapping approach and the verification steps. SU performed the ESTscan and Frame DP analysis. The bioinformatics was performed by JMF (EST assembly) and CC (SSR search and databasing) in covariation with JD. AdD supervised the work of CC. FA, PYD, EG, CB and AK performed the diversity analysis. GGV, FS and AB performed the transferability analysis. All the authors read and approved the final version of the manuscript.
Accession numbers for *Quercus robur* and *Quercus petraea* ESTs can be obtained by searching the EMBL database with keyword for organism name "quercus".

Received: 7 December 2009 Accepted: 15 October 2010
Published: 15 October 2010

References

1. Ellis JR, Burke JM: **EST-SSRs as a resource for population genetic analyses.** *Heredity* 2007, **99(2)**:125-132.

2. Varshney RK, Graner A, Sorrells ME: **Genic microsatellite markers in plants: features and applications.** *Trends Biotechnol* 2005, **23(1)**:48-55.
3. Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P: **Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek).** *BMC Plant Biology* 2009, **9(1)**:137-149.
4. Vision TJ, Brown DG, Shmoys DB, Durrett RT, Tanksley SD: **Selective mapping: A strategy for optimizing the construction of high-density linkage maps.** *Genetics* 2000, **155(1)**:407-420.
5. Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG, Arus P: **Mapping with a few plants: Using selective mapping for microsatellite saturation of the *Prunus* reference map.** *Genetics* 2005, **171(3)**:1305-1309.
6. Fernandez-Silva I, Eduardo I, Blanca J, Esteras C, Pico B, Nuez F, Arus P, Garcia-Mas J, Monforte AJ: **Bin mapping of genomic and EST-derived SSRs in melon (*Cucumis melo* L.).** *Theoretical and Applied Genetics* 2008, **118(1)**:139-150.
7. Sargent DJ, Cipriani G, Vilanova S, Gil-Ariza D, Arus P, Simpson DW, Tobutt KR, Monfort A: **The development of a bin mapping population and the selective mapping of 103 markers in the diploid *Fragaria* reference map.** *Genome* 2008, **51(2)**:120-127.
8. Celton JM, Tustin DS, Chagne D, Gardiner SE: **Construction of a dense genetic linkage map for apple rootstocks using SSRs developed from Malus ESTs and Pyrus genomic sequences.** *Tree Genetics and Genomes* 2009, **5(1)**:93-107.
9. Han Y, Chagné D, Gasic K, Rikkerink EHA, Beever JE, Gardiner SE, Korban SS: **BAC-end sequence-based SNPs and Bin mapping for rapid integration of physical and genetic maps in apple.** *Genomics* 2009, **93(3)**:282-288.
10. Ducousso A, Guyon JP, Kremer A: **Latitudinal and altitudinal variation of bud burst in western populations of sessile oak (*Quercus petraea* (Matt) Liebl).** *Annales Sciences Forestières* 1996, **53**:775-782.
11. Jensen JS, Hansen JK: **Geographical variation in phenology of *Quercus petraea* (Matt.) Liebl and *Quercus robur* L. oak grown in a greenhouse.** *Scandinavian Journal of Forest Research* 2008, **23(2)**:179-188.
12. Saintagne C, Bodénès C, Barreneche T, Pot D, Plomion C, Kremer A: **Distribution of genomic regions differentiating oak species assessed by QTL detection.** *Heredity* 2004, **92(1)**:20-30.
13. Scotti-Saintagne C, Bodénès C, Barreneche T, Bertocchi E, Plomion C, Kremer A: **Detection of quantitative trait loci controlling bud burst and height growth in *Quercus robur* L.** *Theoretical and Applied Genetics* 2004, **109(8)**:1648-1659.
14. Brendel O, Le Thiec D, Scotti-Saintagne C, Bodénès C, Kremer A, Guehl JM: **Quantitative trait loci controlling water use efficiency and related traits in *Quercus robur* L.** *Tree Genet Genomes* 2008, **4(2)**:263-278.
15. Parelle J, Zapater M, Scotti-Saintagne C, Kremer A, Jolivet Y, Dreyer E, Brendel O: **Quantitative trait loci of tolerance to waterlogging in a European oak (*Quercus robur* L.): physiological relevance and temporal effect patterns.** *Plant, Cell and Environment* 2007, **30(4)**:422-434.
16. Derory J, Scotti-Saintagne C, Bertocchi E, Le Dantec L, Graignic N, Jauffres A, Casasoli M, Chancerel E, Bodénès C, Alberto F, Kremer A: **Contrasting correlations between diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks.** *Heredity* 2010, **104**:438-448.
17. Derory J, Leger P, Garcia V, Schaeffer J, Hauser MT, Salin F, Luschnig C, Plomion C, Glossl J, Kremer A: **Transcriptome analysis of bud burst in sessile oak (*Quercus petraea*).** *New Phytol* 2006, **170(4)**:723-738.
18. Roussel M, Dreyer E, Montpied P, Le-Provost G, Guehl JM, Brendel O: **The diversity of 13C isotope discrimination in a *Quercus robur* full-sib family is associated with differences in intrinsic water use efficiency, transpiration efficiency, and stomatal conductance.** *Journal of Experimental Botany* 2009, **60(8)**:2419-2431.
19. Barreneche T, Bodénès C, Lexer C, Trontin JF, Fluch S, Streiff R, Plomion C, Roussel G, Steinkellner H, Burg K, *et al*: **A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 55 rDNA markers.** *Theoretical and Applied Genetics* 1998, **97(7)**:1090-1103.
20. Barreneche T, Casasoli M, Russell K, Akkak A, Meddour H, Plomion C, Villani F, Kremer A: **Comparative mapping between *Quercus* and *Castanea* using simple-sequence repeats (SSRs).** *Theoretical and Applied Genetics* 2004, **108(3)**:558-566.
21. Kremer A, Casasoli M, Barreneche TT, Bodénès C, Sisco P, Kubisiak T, Scalfi M, Leonardi S, Bakker EG, Buiteveld J, Romero-Severson J, Arumuganathan K, Derory J, Scotti-Saintagne C, Roussel G, Bertocchi ME, Lexer C, Porth I, Hebard F, Clark C, Carlson J, Plomion C, Koelewijn H, Villani F: **Fagaceae trees.** In *Genome Mapping & Molecular Breeding. Forest Trees.* Edited by: Kole CR. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo; 2007:**5**:161-187.
22. Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Wang C, Lu K, Shi B: **A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in Gossypium.** *Genetics* 2007, **176(1)**:527.
23. Kolpakov R, Bana G, Kucherov G: **mreps: efficient and flexible detection of tandem repeats in DNA.** *Nucleic Acids Res* 2003, **31(13)**:3672-3678.
24. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999, 138-48.
25. Ueno S, Tsumura Y: **Development of ten microsatellite markers for *Quercus mongolica* var. crispula by database mining.** *Conserv Genet* 2008, **9(4)**:1083-1085.
26. Ueno S, Aoki K, Tsumura Y: **Generation of Expressed Sequence Tags and development of microsatellite markers for *Castanopsis sieboldii* var. sieboldii (*Fagaceae*).** *Annals of Forest Science* 2009, **66(5)**:509-509.
27. Alberto F, Niort J, Derory J, Lepais O, Vitalis R, Galop D, Kremer A: **Population differenciation of sessile oak at the altitudinal front of migration in the French Pyrenees.** *Mol Ecol* .
28. El Mousadik A, Petit RJ: **High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L) Skeels] endemic to Morocco.** *Molecular Ecology* 1996, **5**:547-555.
29. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L: **MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations.** *Genomics* 1987, **1(2)**:174-181.
30. Van Ooijen JW: **JoinMap® 4.** *Software for the calculation of genetic linkage maps in experimental populations* 2006.
31. Chagné D, Lalanne C, Madur D, Kumar S, Frigerio JM, Krier C, Decroocq S, Savouré A, Bou-Dagher-Kharrat M, Bertocchi E: **A high density genetic map of maritime pine based on AFLPs.** *Annals of Forest Science* 2002, **59**:627-636.
32. Qi X, Stam P, Lindhout P: **Comparison and integration of four barley genetic maps.** *Genome* 1996, **39**:379-394.
33. Young ND, Tanksley SD: **Restriction fragment length polymorphism maps and the concept of graphical genotypes.** *Theoretical and Applied Genetics* 1989, **77(1)**:95-101.
34. Leclercq S, Rivals E, Jarne P: **Detecting microsatellites within genomes: significant variation among algorithms.** *BMC Bioinformatics* 2007, **8(1)**:125-143.
35. Parida SK, Anand Raj Kumar K, Dalal V, Singh NK, Mohapatra T: **Unigene derived microsatellite markers for the cereal genomes.** *Theoretical and applied genetics* 2006, **112(5)**:808-817.
36. La Rota M, Kantety RV, Yu JK, Sorrells ME: **Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley.** *BMC Genomics* 2005, **6(1)**:23-34.
37. Kumpatla SP, Mukhopadhyay S: **Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species.** *Genome* 2005, **48(6)**:985-998.
38. Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR, Tuskan GA: **Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome.** *New Phytol* 2009, **181(2)**:498-503.
39. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313(5793)**:1596-1604.
40. Sharopova N: **Plant simple sequence repeats: distribution, variation, and effects on gene expression.** *Genome* 2008, **51(2)**:79-90.
41. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nature Genetics* 2002, **30(2)**:194-200.
42. Schlotterer C, Tautz D: **Slippage synthesis of simple sequence DNA.** *Nucleic Acids Res* 1992, **20(2)**:211-215.

43. Metzgar D, Bytof J, Wills C: **Selection against frameshift mutations limits microsatellite expansion in coding DNA.** *Genome Research* 2000, **10(1)**:72-80.
44. Grover A, Aishwarya V, Sharma PC: **Biased distribution of microsatellite motifs in the rice genome.** *Molecular Genetics and Genomics* 2007, **277(5)**:469-480.
45. Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S: **A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription.** *FEBS Letters* 2003, **554(1)**:17-22.
46. Ralph SG, Chun HJ, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJ, Marra MA, Bohlmann J: **Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding.** *BMC Genomics* 2008, **9(1)**:57-75.
47. Zhou Y, Bui T, Auckland LD, Williams CG: **Direct fluorescent primers are superior to M13-tailed primers for *Pinus taeda* microsatellites.** *Biotechniques* 2002, **32(1)**:46-52.
48. Brandström M, Ellegren H: **Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias.** *Genome Research* 2008, **18(6)**:881-887.
49. Morgante M, Olivieri AM: **PCR-amplified microsatellites as markers in plant genetics.** *The Plant Journal* 1993, **3(1)**:175-182.
50. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S: **Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential.** *Genome Research* 2001, **11(8)**:1441-1452.
51. Kantety RV, La Rota M, Matthews DE, Sorrells ME: **Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat.** *Plant Molecular Biology* 2002, **48(5)**:501-510.
52. Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, Bhamidimarri S, Gonzalez-Hernandez JL, Maan SS, Kianian SF: **High-resolution radiation hydrid map of wheat chromosome 1D.** *Genetics* 2006, **173(2)**:1089-1999.
53. Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A: **A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat.** *Genetics* 2004, **168(2)**:701-712.
54. Deleu W, Esteras C, Roig C, Gonzalez-To M, Fernandez-Silva I, Gonzalez-Ibeas D, Blanca J, Aranda MA, Arus P, Nuez F: **A set of EST-SNPs for map saturation and cultivar identification in melon.** *BMC Plant Biology* 2009, **9(1)**:90-99.
55. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA: **A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base.** *Genome Research* 1999, **9(11)**:1143-1155.
56. Green P: **SWAT/Crossmatch/PHRAP package.** University of Washington 1999 [http://www.phrap.org].
57. Burke J, Davison D, Hide W: **d2_cluster: a validated method for clustering EST and full-length cDNA sequences.** *Genome Research* 1999, **9(11)**:1135-1142.
58. Green P: **Documentation for phrap.** *Genome Center* University of Washington 1996.
59. Chou A, Burke J: **CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences.** *Bioinformatics* 1999, **15(5)**:376-381.
60. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S: **Computational and Experimental Analysis of Microsatellites in Rice (*Oryza sativa* L.): Frequency, Length Variation, Transposon Associations, and Genetic Marker Potential.** *Genome Research* 2001, **11**:1441-52.
61. Gouzy J, Carrere S, Schiex T: **FrameDP: sensitive peptide detection on noisy matured sequences.** *Bioinformatics* 2009, **25(5)**:670-671.
62. **Primer 3.0.** [http://frodo.wi.mit.edu/primer3/].
63. Schuelke M: **An economic method for the fluorescent labeling of PCR fragments.** *Nature Biotechnology* 2000, **18(2)**:233-234.
64. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers.** *Nucleic Acids Res* 1990, **18(22)**:6531-6535.
65. Grattapaglia D, Sederoff R: **Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers.** *Genetics* 1994, **137(4)**:1121-1137.
66. **Mappop.** [http://www.bio.unc.edu/faculty/vision/lab/mappop/].
67. Nei M: **Molecular Evolutionary Genetics.** Columbia University Press: New York 1987.
68. Goudet J: **FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3).** 2001 [http://www2.unil.ch/popgen/softwares/fstat.htm], Updated from Goudet (1995).

doi:10.1186/1471-2164-11-570
Cite this article as: Durand *et al.*: A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 2010 **11**:570.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

ANNEXE 4

Heredity (2009), 1–11
© 2009 Macmillan Publishers Limited All rights reserved 0018-067X/09 $32.00
www.nature.com/hdy

ORIGINAL ARTICLE

Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*)

C Burgarella[1,6], Z Lorenzo[1,2], R Jabbour-Zahab[3], R Lumaret[3], E Guichoux[4,5], RJ Petit[4,5], Á Soto[1,2] and L Gil[1,2]

[1]*G I Genética y Fisiología Forestal, Universidad Politécnica de Madrid, Madrid, Spain;* [2]*Unidad Mixta de Genética y Ecofisiología Forestal INIA-UPM, Universidad Politécnica de Madrid, Madrid, Spain;* [3]*Département de Biologie des populations, Unité Mixte de Recherche (UPR) 5175, Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, Montpellier Cedex, France;* [4]*INRA, UMR1202 BIOGECO, Cestas, France and* [5]*Université de Bordeaux, UMR1202 BIOGECO, Cestas, France*

Powerful and accurate detection of first-generation (F1) hybrids and backcrosses in nature is needed to achieve a better understanding of the function and dynamics of introgression. To document the frequency of ongoing interspecific gene exchange between two Mediterranean evergreen oaks, the cork oak (*Quercus suber*) and the holm oak (*Q. ilex*), we analyzed 1487 individuals originating from across the range of the two species using eight microsatellite loci and two Bayesian clustering approaches (implemented in the programs STRUCTURE and NEWHYBRIDS). Simulated data were used to assess the differences between the two clustering methods and to back up the choice of the threshold value for the posterior probability to discriminate admixed from pure individuals. We found that the use of STRUCTURE resulted in the highest power to detect hybrids, whereas NEWHYBRIDS provided the highest accuracy. Irrespective of the approach, the two species were clearly distinguished as independent genetic entities without any prior information. In contrast with previous reports, we found no evidence for unidirectional introgression. The overall hybridization rate was very low (<2% of introgressed individuals). Only two individuals were identified as F1 hybrids and five as early backcrosses. This work shows that the combined application of the two complementary Bayesian approaches and their systematic validation with simulations, fit for the case at hand, helps gain resolution in the identification of admixed individuals.
Heredity advance online publication, 25 February 2009; doi:10.1038/hdy.2009.8

Keywords: introgressive hybridization; clustering analysis; simulation; *Quercus suber*; *Q. ilex*; microsatellites

Introduction

Natural hybridization and introgression are widespread phenomena in plants, with important evolutionary implications (Rieseberg and Carney, 1998). The movement of genes across species boundaries can promote the appearance of new lineages (Seehausen, 2004), adaptive solutions (Rieseberg *et al*., 2003) or colonization abilities (Potts and Reid, 1988; Petit *et al*., 2004). Measuring the frequency of hybrids and describing their geographic distribution should help focus measures directed to conservation or breeding programs (Burgess *et al*., 2005; Kothera *et al*., 2007). Different types of molecular markers can inform on different spatial and temporal scales of the hybridization–introgression dynamics. Chloroplast and mtDNA have been used to describe past episodes of introgression (Palmé *et al*., 2004; Heuertz *et al*., 2006) whereas nuclear loci have been useful to infer contemporary rates of interspecific gene exchange (Lexer *et al*., 2005; Fernández-Manjarrés *et al*., 2006). However, identifying hybrid individuals in nature using molecular markers still represents an important challenge. Availability of hypervariable codominant markers (for example, microsatellites) and powerful statistical procedures (that is, Bayesian clustering methods, which do not rely on *a priori* morphological classification) has facilitated the detection of first-generation (F1) hybrids and backcrosses. However, the choice of the method that will provide the best resolution needs to be established for a given situation.

Oaks represent good models for such studies. Interspecific hybridization is the most frequently invoked mechanism to account for the existence of plants morphologically and ecologically intermediate between extant oak species (Jensen *et al*., 1993; Howard *et al*., 1997; González-Rodríguez *et al*., 2004) and to interpret the extensive local sharing of organelle and nuclear genes between species (Whittemore and Schaal, 1991; Howard *et al*., 1997; Petit *et al*., 1997; Dumolin-Lapègue *et al*., 1999). However, in some cases, interspecific gene exchanges have been detected with molecular markers in the absence of obvious morphologically intermediate forms (Whittemore and Schaal, 1991; Dodd and Afzal-Rafii, 2004). Moreover, the possibility that shared alleles represent ancestral segregating polymorphisms rather than the outcome of hybridization has been suggested (Muir and Schlötterer, 2005; but see Lexer

Correspondence: Dr Á Soto, GI Genética y Fisiología Forestal, ETSI Montes, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, Madrid 28040, Spain.
E-mail: alvaro.soto.deviana@upm.es
[6]*Current address: Departamento de Sistemas y Recursos Forestales, Instituto Nacional para la Investigación Agraria y Alimentaria (CIFOR-INIA), Madrid, Spain.*
Received 15 July 2008; revised 23 December 2008; accepted 7 January 2009

et al., 2006). Environmental variation, disturbance as well as the degree of contact between species can affect the frequency and the spatial distribution of hybrids in natural oak populations (Nason, 1992; Rushton, 1993; Howard *et al*., 1997; Dumolin-Lapègue *et al*., 1999; Dodd and Afzal-Rafii, 2004; Tovar-Sanchez and Oyama, 2004; Curtu *et al*., 2007; Valbuena-Carabaña *et al*., 2007). Although hybridization between some oak species, such as the closely related species *Quercus robur* and *Q. petraea*, has been analyzed extensively for nuclear, chloroplast and mitochondrial variation, our understanding of the underlying processes is still unclear.

In this study we focus on two distantly related oak species, *Q. suber* (cork oak) and *Q. ilex* (holm oak), which have partially overlapping geographic distributions in the western part of the Mediterranean basin. The two evergreen species have a major ecological function in many Mediterranean woody ecosystems and constitute key elements of seminatural systems of high economical and social importance (for example, cork extraction and silvopastoral uses; Plieninger *et al*., 2003; Martín Vicente and Fernández Alés, 2006). Cork oaks and holm oaks are easily discriminated by a few morphological traits, including bark (that is, cork layer is found exclusively in *Q. suber*), leaf and fruit features (Amaral Franco, 1990). Some concerns exist about the effect of hybridization on cork quality and on breeding programs of *Q. suber* (Oliveira *et al*., 2007). Within the section *Cerris* (subgenus *Quercus*), *Q. suber* and *Q. ilex* belong to different clades (groups *Cerris* and *Ilex*, respectively), which are thought to have diverged during the middle Tertiary (Manos *et al*., 2001). Despite their deep phylogenetic divergence, clearly supported by internal transcribed spacer, amplified fragment length polymorphisms and isozyme variation (Manos *et al*., 1999; Toumi and Lumaret, 2001; Bellarosa *et al*., 2005; López de Heredia *et al*., 2007b), hybridization has been inferred on the basis of morphological and molecular markers (Elena-Rosselló *et al*., 1992; Toumi and Lumaret, 1998; Lumaret *et al*., 2002; Oliveira *et al*., 2003; Bellarosa *et al*., 2005). Furthermore, extensive surveys of chloroplast DNA diversity of both species and of other relatives (such as *Q. coccifera*) across the whole distribution range have demonstrated widespread cytoplasmic introgression, mainly localized along a northeast-southwest line, from French Catalonia and eastern Iberia to Morocco (reviewed in Lumaret *et al*., 2005). Interspecific exchanges seem to be limited to introgression of *Q. ilex* cpDNA and mtDNA into *Q. suber*, with only very few cases of *Q. suber* cpDNA introgressing into *Q. ilex* (Belahbib *et al*., 2001; Lumaret *et al*., 2002; Jiménez *et al*., 2004; Staudt *et al*., 2004). Because organelle DNA is maternally inherited in *Quercus* (Dumolin *et al*., 1995), this asymmetry implies that *Q. ilex* has acted predominantly as the maternal species in interspecific crosses. Boavida *et al*. (2001) provided experimental support for this hypothesis by showing that F1 hybrids are more easily produced when *Q. suber* is the pollen donor. In addition, unidirectional mating can be favored by phenology (*Q. ilex* flowers earlier) combined with protandry (that is, male flowers appear earlier than female flowers; Varela and Valdiviesso, 1996).

To date, no data are available on mating preferences in later hybrid generations, as hybrid individuals with known pedigree remain extremely rare in oaks. In such a context, identifying F1 hybrids and backcrosses would be important, particularly when the proportion of hybrid individuals is low and when they are morphologically cryptic (as seems to be the case for *Q. suber* and *Q. ilex*; Lumaret *et al*., 2002; Staudt *et al*., 2004). We present here a broad-scale survey of molecular variation across the overlapping range of *Q. suber* and *Q. ilex* to explore the extent and pattern of nuclear introgressive hybridization, using a panel of eight highly discriminating microsatellite loci. Our specific aims are (1) to assess the effectiveness of two Bayesian clustering approaches to distinguish hybrid individuals without knowledge of their pedigree and (2) to document the frequency of contemporary interspecific gene exchange in natural populations of cork and holm oaks, and hence evaluate previously proposed hybridization scenarios. For these purposes, we use admixture analysis of multilocus microsatellite genotypes from a range-wide sample of sympatric and allopatric populations of the two species. Furthermore, we simulate hybrid genotypes to assess the performance and the limits of the procedure used to detect hybrid individuals and to distinguish among hybrid classes.

Materials and methods

Sampling strategy

We sampled 597 *Q. suber* and 515 *Q. ilex* from 13 populations across the distribution range of cork oak and the overlapping range of holm oak (Figure 1). Five mixed woods were more intensively sampled (775 individuals). Two of them (Castilla-La Mancha and Sicily) include part of the individuals used in Soto *et al*. (2007) and Burgarella *et al*. (2007). In the mixed population of Minorca, the sample includes all existing cork oaks on the island (67 individuals). As additional reference, another set of 375 cork oaks have been included, sampled from an international provenance trial established in 1998 in the frame of the *Q. suber* network from the European Programme for the Conservation of Forest Genetic Resources (EUFORGEN), which covered the complete distribution range of the species (35 provenances). Reference codes, geographic allocations and sampling sizes are given in Table 1. Individuals were tentatively assigned to each species according to their morphology.

Microsatellite typing

Individuals were genotyped at eight microsatellite loci: MSQ4, MSQ13 (Dow *et al*., 1995), QpZAG9, QpZAG15, QpZAG36, QpZAG46 (Steinkellner *et al*., 1997), QrZAG11 and QrZAG20 (Kampfer *et al*., 1998). A detailed description of the protocols has been published elsewhere (Soto *et al*., 2003, 2007). At MSQ13, 25% of *Q. ilex* genotypes had three or four alleles, possibly due to gene duplication in this species. On the contrary, *Q. suber* showed a normal banding pattern. MSQ13 is a highly informative locus, because allele sizes do not overlap between the two species (Soto *et al*., 2003). To include this locus in the following analyses, we pooled the alleles typical of *Q. ilex*. To identify them, we defined the pure genotype pool of each species with the other seven loci, performing a preliminary clustering analysis with STRUCTURE (same settings described below).

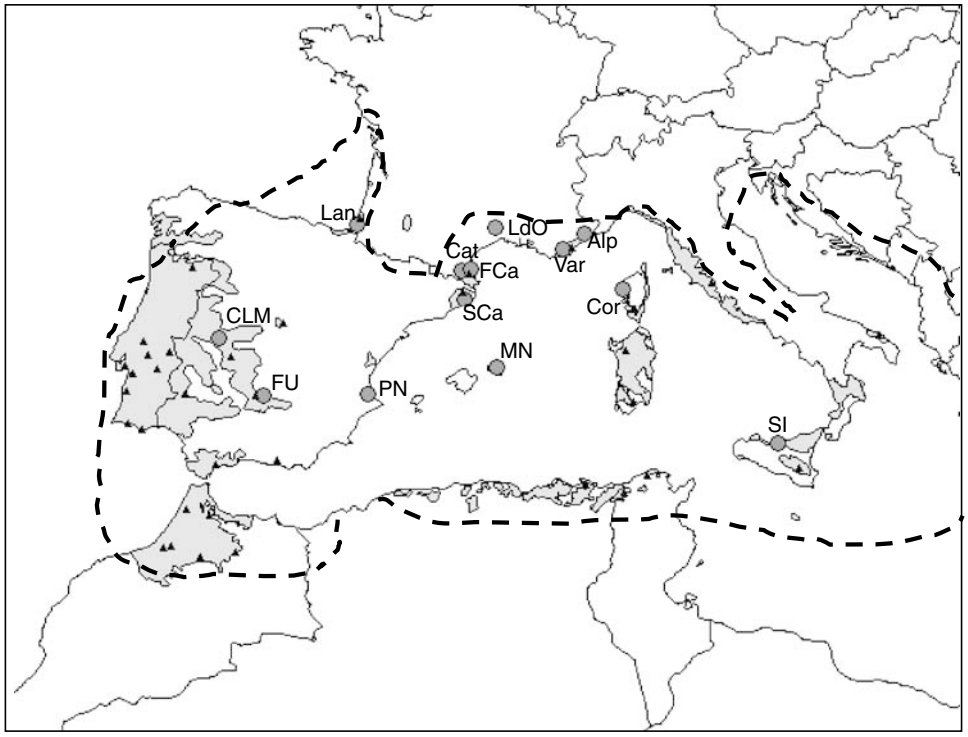


Figure 1 Sampling sites. Light gray, distribution range of *Quercus suber* (modified from http://www.bioversityinternational.org/networks/euforgen/); dashed line, distribution range of *Q. ilex*; triangles, populations included in the field trial and dark gray circles, *Q. ilex*, *Q. suber* and mixed stands (see Table 1 for population code).

Information content of microsatellites and genetic differentiation

Deviation from Hardy–Weinberg equilibrium and linkage disequilibrium (LD) was tested using FSTAT 2.9.3.2 (Goudet, 2001). To assess the diagnostic power of each marker, we estimated the allele frequency differential between the two species, δ (Shriver *et al*., 1997). For a given locus, δ is calculated as half the sum of the absolute value of allele frequency differences between species. F-statistics were also estimated for both species in each mixed population and in the whole set of individuals following the weighted analysis of variance method of Weir and Cockerham (1984). All analyses were carried out only with putative purebred individuals, selected after a preliminary screening for potential hybrids, as explained below.

Nuclear admixture analysis for hybrid identification

To identify hybrid individuals and estimate population-level hybridization, we carried out admixture analyses using two different Bayesian clustering approaches, as implemented in the programs STRUCTURE version 2 (Pritchard *et al*., 2000) and NEWHYBRIDS version 1.1 beta (Anderson and Thompson, 2002). Both methods were used to assign probabilistically individual multilocus genotypes to categories (clusters) by jointly inferring the parameters corresponding to each cluster and the cluster membership of each individual (that is, without *a priori* knowledge of the allele frequencies in the separate clusters). A Markov chain Monte Carlo simulation procedure provides the estimates from the posterior distribution reflecting the membership of each individual. In the STRUCTURE model, the posterior probability (q) describes the proportion of an individual genotype originating from each of K categories. In our case, setting $K=2$ corresponds to the assumption of two species contributing to the gene pool of the sample. Instead, NewHybrids model assumes that the sample is drawn from a mixture of pure individuals and hybrids (Anderson and Thompson, 2002). Under this model, q describes the probability that an individual belongs to each of different genotype frequency classes (in our case: parental purebreds, F1 hybrid and the two first backcrosses categories). Analyses were carried out for all individuals jointly and for each of the mixed populations separately. In all cases, no prior species information was used. With STRUCTURE, calculations were carried out under the admixture model assuming independent allele frequencies, given the high interspecific differentiation (see results). A burn-in of 50 000 steps followed by 100 000 iterations was used with each program, after verifying that results do not vary significantly across multiples runs and with longer cycles of burn-in/iterations.

When using these assignment approaches, an important decision is the choice of the optimal threshold value (Tq) for the q associated with the classification of each individual into purebred or hybrid (Vähä and Primmer, 2006). We used threshold values of 0.90 (Pritchard *et al*., 2000; Vähä and Primmer, 2006) and 0.75. With STRUCTURE, a value of q higher or equal to the threshold indicates a purebred genotype and a value of q lower than the threshold indicates an introgressed genotype. With NEWHYBRIDS, the threshold values can be used in three ways. In the most restrictive way (criterion 1) the threshold value is applied to each category (pure species, F1 hybrids, backcrosses) separately, by assigning only the

Table 1 Sample location, identifying code, type of population and sample size

Species	*Location*	*Code*	*Type*	N_{suber}/N_{ilex}
Quercus suber/Q. ilex	Catalonia (France)	FCa	Field	98/100
	Catalonia (Spain)	SCa	Field	73/74
	Castilla-La Mancha (Spain)	CLM	Field	95/95
	Minorca (Balearic Islands, Spain)	MN	Field	67/44
	Sicily (Italy)	SI	Field	63/66
Q. suber	Sierra Morena Oriental (Spain)	FU	Field	50/—
	Valencia (Spain)	PN	Field	69/—
	Var (France)	Var	Field	50/—
	Landes (France)	Lan	Field	32/—
	Vale do Tejo e Sado (Portugal)	PT1	Trial	11/—
	Vale do Tejo e Sado (Portugal)	PT2	Trial	11/—
	Vale do Tejo e Sado (Portugal)	PT3	Trial	11/—
	Vale do Tejo e Sado (Portugal)	PT4	Trial	11/—
	Alentejo e Beira Baixa (Portugal)	PT5	Trial	11/—
	Alentejo e Beira Baixa (Portugal)	PT6	Trial	10/—
	Sudoeste (Portugal)	PT7	Trial	11/—
	Sudoeste (Portugal)	PT8	Trial	11/—
	Sudoeste (Portugal)	PT9	Trial	10/—
	Tras-os-Montes e Beira Interior (Portugal)	PT10	Trial	11/—
	Sierra Morena Oriental (Spain)	ES1	Trial	10/—
	Madrid (Spain)	ES2	Trial	11/—
	Montes de Toledo (Spain)	ES3	Trial	9/—
	Sierra Morena Occidental (Spain)	ES4	Trial	11/—
	Sierra Nevada (Spain)	ES5	Trial	10/—
	Cádiz (Spain)	ES6	Trial	10/—
	Catalonia (Spain)	ES7	Trial	11/—
	Var (France)	FR1	Trial	11/—
	Landes (France)	FR2	Trial	11/—
	Pyrénées Orientales (France)	FR3	Trial	11/—
	Corsica (France)	FR4	Trial	11/—
	Lazio (Italy)	IT2	Trial	10/—
	Puglia (Italy)	IT3	Trial	10/—
	Sicily (Italy)	IT4	Trial	11/—
	Sardinia (Italy)	IT5	Trial	11/—
	Sardinia (Italy)	IT6	Trial	10/—
	Mekna (Tunisia)	TU1	Trial	11/—
	Fernana (Tunisia)	TU2	Trial	11/—
	Guerbès (Algeria)	AL	Trial	11/—
	Rif Atlantic (Morocco)	M1	Trial	11/—
	Rif Occidental (Morocco)	M2	Trial	11/—
	Maâmora (Morocco)	M3	Trial	11/—
	Maâmora (Morocco)	M4	Trial	11/—
	Plateau Central (Morocco)	M5	Trial	11/—
	Rif Oriental (Morocco)	M6	Trial	11/—
Q. ilex	Catalonia (France)	Cat	Field	—/55
	Languedoc (France)	LdO	Field	—/16
	Alpes-Maritimes (France)	Alp	Field	—/21
	Corsica (France)	Cor	Field	—/44

individuals with $q \geqslant Tq$ and leaving the others unassigned (Oliveira *et al*., 2007). Alternatively, q values for all hybrid categories (F1 hybrids, backcrosses) can be combined (Vähä and Primmer, 2006) to distinguish hybrids regardless of their category (criterion 2). A third option (criterion 3), the most relaxed, is to apply the threshold only to the purebred category, assuming that individuals with $q \geqslant Tq$ are purebreds and that all others are hybrids (this is the only case where no individual remains unassigned).

Performance of the two admixture analyses

We used simulated data to assess which method provides the most reliable results with our experimental system (as suggested by Vähä and Primmer, 2006). Specifically, we tried to identify the Tq for the q to distinguish hybrids from purebreds. We also tested which of the criteria suggested for hybrid identification with NEWHYBRIDS performs best, and we evaluated the effect of different sample sizes.

Allele frequencies for parental species were estimated from the whole sample after taking out potentially introgressed individuals identified in preliminary runs of both STRUCTURE and NEWHYBRIDS (these are the individuals with $q < 0.90$ for pure species categories, which corresponds to the criterion 3 for NEWHYBRIDS). Ten thousand purebred genotypes were then generated with HYBRIDLAB 1.0 (Nielsen *et al*., 2006) for each species using these allele frequencies. In addition, three hybrid sets of 10 000 genotypes each were generated by randomly drawing alleles (random mating assumed)

from each of the simulated purebred genotypes for the F1 set and from simulated purebred genotypes and simulated F1 genotypes for each backcross set. Genotypes were sampled without replacement from the five simulated sets with POPTOOLS 2.6 (Hood, 2005) to create samples of 150 and 1500 individuals with two different proportions of hybrids (HP): 0 and 2%. The first figure corresponds to the complete lack of hybrids in the sample, whereas HP = 2% corresponds to 3 hybrids (one F1 and two F1 backcrosses to each parent species) and 30 hybrids (10 F1 and 10 of each of the two backcrosses), respectively, for $N = 150$ and 1500. Sample sizes and HPs have been chosen to represent the actual population samples. For each HP, 100 replicate data sets were generated for $N = 150$ and 10 replicates for $N = 1500$. Each simulated data set was analyzed with STRUCTURE and NEWHYBRIDS with the same setting conditions, threshold values and criteria described before.

The following measures were used to evaluate the performance of the methods:

(1) the hybrid proportion: number of individuals classified as hybrids over the total number of individuals in the sample;
(2) the power to detect the true hybrid/purebred status of individuals ('efficiency' *sensu* Vähä and Primmer, 2006): number of correctly identified individuals for a category over the actual number of individuals of that category in the sample;
(3) the accuracy (*sensu* Yang *et al*., 2005 and Vähä and Primmer, 2006): number of correctly identified individuals for a category over the total number of individuals assigned to that category; and
(4) the type I error: number of individuals wrongly identified as hybrids over the total number of actual purebreds in the sample.

Finally, we compared the power and accuracy of the clustering algorithms as a function of the number of molecular markers examined. We considered two sets of three combinations of molecular markers (2, 4 and 6 loci), with $N = 1500$ simulated genotypes. The first set was composed of three combinations of loci with decreasing value of δ, starting with the two most discriminating, MSQ13 and QpZAG9 (Table 2). The second set was composed of three combinations of loci with increasing value of δ, starting with the two with the least discriminatory power (that is, QpZAG36 and QrZAG20, Table 2). This provided approximate upper and lower bounds of the power and accuracy for different combinations of loci.

Table 2 Allele frequency differential (δ) between *Q. suber* and *Q. ilex* in mixed populations and in the whole sample for each of the eight microsatellite loci screened

Loci[a]	*FCa*	*SCa*	*CLM*	*MN*	*SI*	*Whole sample*
MSQ13	1.00	1.00	1.00	1.00	1.00	1.00
QpZAG9	0.99	0.99	0.97	0.95	0.97	0.96
MSQ4	1.00	1.00	0.78	0.90	1.00	0.96
QpZAG15	0.92	0.94	0.88	0.93	0.96	0.92
QpZAG46	0.83	0.76	0.97	0.64	0.87	0.84
QrZAG11	0.86	0.89	0.69	0.86	0.86	0.83
QpZAG36	0.91	0.88	0.75	0.68	0.80	0.78
QrZAG20	0.50	0.69	0.68	0.63	0.75	0.62

[a]Loci ranged according to decreasing values of δ for the whole sample.

Results

Information content of microsatellites and species differentiation

Although some loci showed significant homozygous excess (18 tests out of 144 with *P*-value <0.05) and LD (10 tests out of 504 with *P*-value <0.05), no consistent pattern was found across all populations and species (data not shown). All marker loci have high discriminatory power over the whole sample, with allele frequency differential ranging from $\delta = 0.62$ to $\delta = 1$ (Table 2). After removing putative hybrids to calculate δ, MSQ13 appears to be fully diagnostic. High and significant genetic differentiation between the two species was found over the whole sample as well as in each region (range wide $\theta = 0.41$, *P*-value = 0.001; minimum $\theta = 0.40$, Minorca; maximum $\theta = 0.44$, Spanish Catalonia). For comparison, intraspecific differentiation is 10 times lower (*Q. suber* $\theta = 0.05$; *Q. ilex* $\theta = 0.06$).

Hybrid detection and performance of the admixture analysis

Results of simulations performed with all eight loci for each sample size scenario (that is, 150 and 1500) were quite similar across methods (that is, STRUCTURE versus NEWHYBRIDS) and thresholds (that is, 0.90 versus 0.75). Nevertheless, higher power and accuracy and lower error rates were reached with the larger sample size (data not shown). Thus, results presented here refer exclusively to analyses of real data performed with all 1487 individuals jointly and of simulated data with the 1500 samples. With NEWHYBRIDS, criterion 2 (hybrid probability: sum of probabilities for F1 and backcrosses) was selected because it showed the best performance using simulated data (results not shown).

In the absence of hybrids, both Bayesian approaches used to infer the individual admixture proportions perform well, although STRUCTURE provides a small proportion of false hybrids with the 0.90 threshold (Table 3). On the contrary, when the simulated sample contains hybrid individuals, the best HP estimate is found with STRUCTURE and the 0.90 threshold; a slight underestimate is obtained with NEWHYBRIDS for both threshold values, and a strong underestimate with STRUCTURE and the 0.75 threshold (Table 3). Likewise, the power to correctly classify purebreds is higher than 99% in all cases, but the highest proportion of correctly identified hybrids is achieved when STRUCTURE is used with the 0.90 threshold (92%), followed by NEWHYBRIDS with thresholds of 0.75 and of 0.90. Compared to STRUCTURE, detection ability is lower with NEWHYBRIDS, because some individuals remain unassigned (for the empirical data set, nine genotypes are unassigned with $Tq = 0.90$ and four with $Tq = 0.75$), but accuracy in identifying hybrids is improved (>99% for a power >86% using both thresholds; Table 3). Thus, STRUCTURE provides power whereas NEWHYBRIDS provides accuracy.

As expected, both the power and accuracy increase with the number of loci (Figure 4). This increase is higher for the identification of hybrids than for the identification

Table 3 Results of STRUCTURE and NEWHYBRIDS analyses with simulated samples of $N = 1500$

Simulated HP (%)	*No. of repetitions*	*No. of hybrids in the sample*	*Method*	*Tq*	*Mean No. of hybrids (s.d.)*	*Estimated HP (%)*	*Mean squared error*	*Power*		*Accuracy*		*Type I error*	*Not assigned*
								Hybrids	*Purebreds*	*Hybrids*	*Purebreds*		
0	10	0	STRUCTURE	0.9	2 (1.07)	0.13	0.000	—	0.998	—	1.000	0.000	
				0.75	0	0.00	0.000	—	1.000	—	1.000	0.000	
				0.9	0	0.00	0.000	—	1.000	—	1.000	0.000	0
			NEWHYBRIDS 2nd	0.75	0	0.00	0.000	—	1.000	—	1.000	0.000	0
2	10	30	STRUCTURE	0.9	29.5 (2.92)	1.97	0.035	0.920	0.999	0.936	0.998	0.001	
				0.75	19.1 (3.04)	1.27	0.564	0.640	1.000	1.000	0.993	0.000	
				0.9	26 (2.83)	1.73	0.103	0.867	0.997	1.000	0.999	0.000	71
			NEWHYBRIDS 2nd	0.75	27.2 (2.70)	1.81	0.064	0.867	0.999	0.989	0.999	0.000	23

Abbreviations: HP, hybrid proportions; *Tq*, threshold *q*-value.
With NEWHYBRIDS, for each individual the *Tq* was applied to the sum of posterior probability for all hybrid classes used as one estimate (criterion 2, see text).

of purebreds (results not shown). The simulations show that the four most discriminant loci suffice to reach high power in identifying hybrids with STRUCTURE and high accuracy with NEWHYBRIDS, values comparables with those obtained using eight loci (Figure 4). However, a higher number of individuals remain unassigned with NEWHYBRIDS when only four loci are used (112, including 39 hybrids, compared to 71, including 22 hybrids, with all eight markers).

When applied to our experimental data set, both methods separated the 1487 individuals examined into two well-defined groups congruent with the observed *suber* and *ilex* phenotypes. Both methods also identified a very low total number of putative hybrids, most of them in mixed populations. Some differences were found between both methods, in agreement with the results of the simulations. With STRUCTURE, 17 potential hybrids were detected with a threshold $Tq = 0.90$ (that is, an HP = 1.1%), but this estimate drops to 4 with $Tq = 0.75$ (HP = 0.03%; Figure 2a). All remaining individuals have a very high probability to belong to the purebred species (*Q. suber*: range 0.903–0.998; *Q. ilex*: range 0.925–0.998). With NEWHYBRIDS, five individuals were identified as hybrids with $Tq = 0.90$ (HP = 0.20%) and seven with $Tq = 0.75$ (HP = 0.34%; Figure 2b). Again, putative purebreds present high *q*-values (*Q. suber*: range 0.901–1.000; *Q. ilex*: range 0.960–1.000). Surprisingly, three individuals morphologically identified as *Q. suber*, from Minorca (one) and from Sicily (two), have been classified by molecular analysis as pure *Q. ilex*.

Genetic composition of hybrid/introgressed individuals

STRUCTURE detected a total of 17 individuals with *q* between 0.10 and 0.90 (Figure 2a); 8 of them had been classified in the field as *Q. suber* and 9 as *Q. ilex*. However, NEWHYBRIDS assigns six of them to purebred categories with $q > 0.95$ (two *Q. suber* and four *Q. ilex*, matching field identification) (Figure 2b). In view of the high accuracy provided by NEWHYBRIDS and the false positive rate associated with STRUCTURE (when HP = 2%, type I error = 0.001; Table 3), the hybrid nature of those six individuals is uncertain. In contrast, the hybrid nature of the remaining 11 trees appears more consistent and for 7 of them very well supported. Only two individuals, one from the SCa population (*suber* SCa70) and one from the MN population (*ilex* MN36), showed intermediate proportions compatible with an F1 genotype with both methods (Figure 2), although a backcross status cannot be excluded. In fact, simulations showed that all F1 hybrids are always correctly classified as hybrids (that is, none was assigned to any pure species) whichever method and threshold is used (data not shown), but some of them present a pattern of admixture indistinguishable from that of backcrosses (Figure 3). The remaining nine individuals (SCa95, MN32, MN39, MN45, TU2 *suber* morphotype, CLM48, SCa36, SCa84 and SI2 *ilex* morphotype) probably result from one or more generations of backcross. Among them, SCa95, SCa36 and SI2 have the phenotype of one species despite having a large assignment probability to the other species (Figure 2).

Discussion

Evidence and rate of hybridization between cork and holm oaks

The microsatellite loci chosen for this work were highly differentiated between species ($\theta = 0.41$) and had good diagnostic power ($\delta = 0.62$–1.0). In fact, both Bayesian clustering approaches used (implemented in STRUCTURE and NEWHYBRIDS) assigned nearly all individuals with high probability to each of two genetically defined groups, resulting in an almost perfect match with the observed morphotypes. Very few hybrid genotypes have been detected (0.027–1.14% of the total sample, using the most and least restrictive conditions, respectively; Figure 2). Using simulated data, we have quantified the resolution level achieved and the uncertainty attached to the experimental system and threshold values for two posterior probabilities (0.90 and 0.75). These results indicate that, although the correct identity of hybrid individuals cannot be guaranteed in all cases, it is possible to get a good estimate of the actual proportion of hybrids in our sample (see estimated and simulated HP in Table 3). Simulations also showed that we could achieve similar results with half of the loci (Figure 4) by

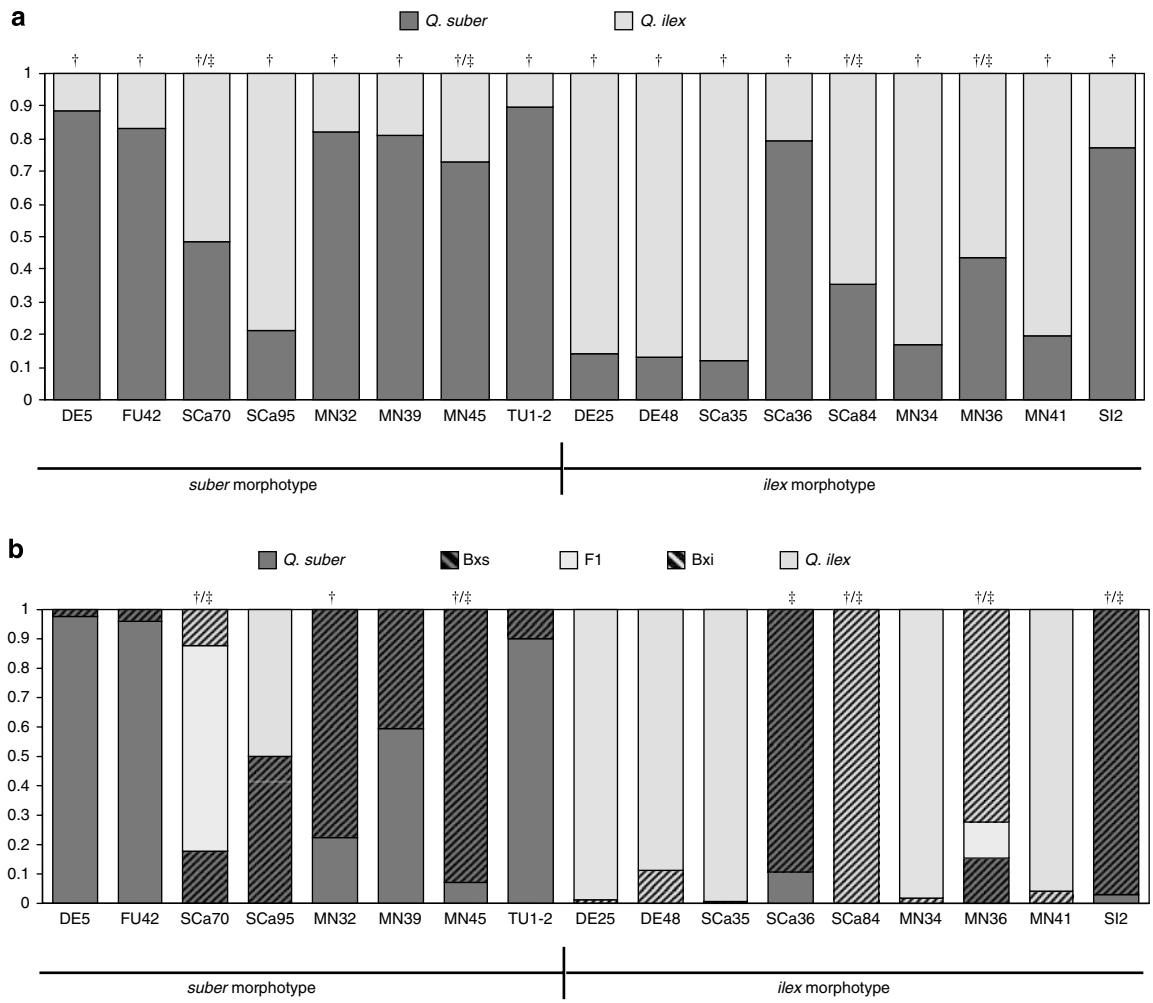


Figure 2 Posterior probability (q) for all individuals identified as putative hybrids by at least one of the method–threshold (Tq) combination. Each individual is represented by a vertical bar partitioned into segments, the length of which describes (**a**) the estimated membership proportions for each parental species (*Q. suber* and *Q. ilex*) by STRUCTURE and (**b**) the estimated probability of belonging to the parental species and the three hybrid classes (F1, first backcross with each of the parental species) by NEWHYBRIDS. Individuals are identified by a population code (see Table 1) and ID number. †Classified as hybrid with $Tq = 0.90$, ‡Classified as hybrid with $Tq = 0.75$.

selecting those with the highest discriminatory power, in agreement with Boecklen and Howard (1997). This may suggest a rapid method to distinguish hybrids from pure holm oaks and cork oaks. However, this conclusion has to be taken with caution, because simulations relies on simplifying assumptions (for example, symmetrical introgression, limited type of backcross categories) likely not fulfilled by natural populations. Hence, we consider a worth effort increasing the number of molecular markers to improve the level of resolution, even if highly diagnostic markers are available.

The low frequency (<2%) of contemporary gene exchange detected between *Q. suber* and *Q. ilex* is consistent with the available knowledge on nuclear variability for the species. A low number of hybrids has been reported in previous surveys of isozyme diversity (Elena-Rosselló *et al*., 1992; Toumi and Lumaret, 1998; Lumaret *et al*., 2002; Staudt *et al*., 2004). Nevertheless, the extensive sharing of chloroplast DNA haplotypes between *Q. suber* and *Q. ilex* in some regions has led some authors to hypothesize widespread introgressive hybridization events in the past (Belahbib *et al*., 2001; Lumaret *et al*., 2002; Jiménez *et al*., 2004; López de Heredia *et al*., 2005). Such findings are not incompatible, given that even a low fraction of hybrids can have considerable evolutionary impact because of the cumulative effect of introgression through time (Ellstrand *et al*., 1996; Mallet, 2005) and the possibility for introgressed genes to become amplified by demographic growth (Currat *et al*., 2008). In this respect, López de Heredia *et al*. (2007a) suggested that the acidophilous *Q. suber* was able to colonize the calcareous area of eastern Iberia (where chloroplast introgression has been reported), thanks to the hybridization with *Q. ilex*, which is largely indifferent to soil nature. It is noteworthy that we found a higher proportion of early generation hybrids in Catalonia and Minorca, located within the area of chloroplast introgression and where soils are mostly formed on more or less decarbonated calcarenites and dolomites, unfavorable to cork oak. This would be consistent with the 'environmental emasculation' hypothesis proposed by Williams *et al*. (2001), according to which environmental stress, at the margins of the suitable habitat of a species, can lead to a decrease in the competitive ability of its pollen, thus favoring hybridization. Alternatively, the process could be driven (exclu-

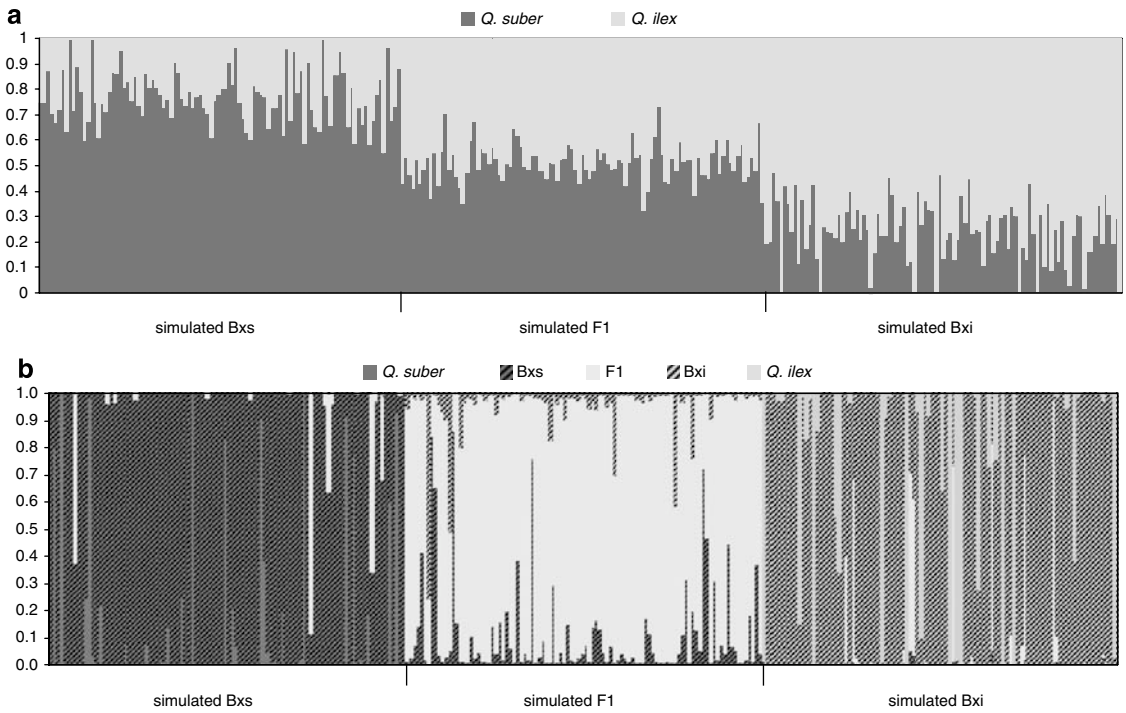


Figure 3 Posterior probability (q) for simulated hybrid individuals analyzed in 10 repetitions of simulated samples with $N = 1500$ and 2% hybrid proportions each with (**a**) STRUCTURE and (**b**) NEWHYBRIDS. Number of hybrids: 100 backcrosses to *Q. suber* (Bxs), 100 first-generation hybrids (F1) and 100 backcrosses to *Q. ilex* (Bxi).

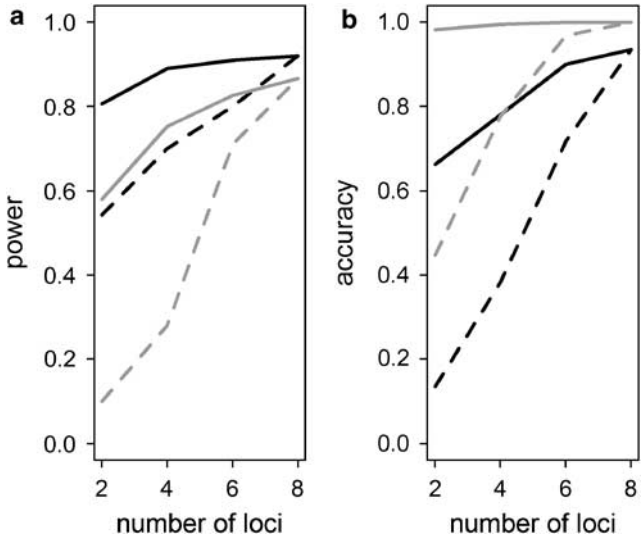


Figure 4 Power (**a**) and accuracy (**b**) in detecting hybrid individuals for 10 simulated samples of $N = 1500$, analyzed with STRUCTURE (black line) and NEWHYBRIDS (gray line), as a function of the number of microsatellite markers ($Tq = 0.90$). Loci have been combined according to their decreasing (solid line) and increasing (dashed line) value of frequency differential δ (see text).

sively or complementarily) by demographic factors, due to demographic imbalance during colonization, as suggested by Currat *et al*. (2008). In Minorca, for instance, *Q. suber* population size is limited to the 67 individuals we sampled.

As shown by results from controlled crosses (Boavida *et al*., 2001), *Q. suber* likely acts as the pollen donor in interspecific mating events with *Q. ilex*. This finding is supported by the discovery of widespread introgression of *ilex*-type cpDNA in *Q. suber* populations (Belahbib *et al*., 2001; Jiménez *et al*., 2004; Lumaret *et al*., 2005), whereas the opposite situation (that is, *Q. ilex* trees showing *suber* chlorotypes) is considerably less frequent. However, no evidence of unidirectional gene flow has been found in this study, because we detected a similar number of backcrosses to each species (Figure 2). Artificial crosses involving F1s and the parental species would help determine the direction of introgression and the nature of barriers to random mating. For instance, Olrik and Kjaer (2007) showed that *Q. robur*–*Q. petraea* unidirectional hybridization does not imply necessarily asymmetric backcrossing to the parental species, after performing controlled crosses with an F1 tree of known pedigree.

Hybrid identity

Among the 17 putative hybrids, we could distinguish at least 2 putative F1s (*suber* SCa70 and *ilex* MN36) and 5 backcrosses (*suber* MN32, MN45, *ilex* SCa36, SCa84 and SI2) with very high probability (Figure 2). The reliability of their hybrid identity is supported by the coincident assignation with two different methods and by the high accuracy and low error observed in data-based simulations with two different thresholds values (Table 3). Although we can be reasonably sure that these seven individuals are not purebreds, and that any actual F1 present in the sample would not have been classified as purebred, we cannot exclude that the two putative F1s are backcrosses. Similarly, some uncertainty is involved in the identity of the 10 remaining individuals (Figure 2). Our simulations indicated that the identification of backcrosses is more problematic than that of F1 hybrids, because they can be confused with pure individuals and

vice versa, as already shown with other hybrid systems (Barilani *et al*., 2007). The extent of incorrect classification can be minimized by choosing an appropriate threshold to improve accuracy (and reduce error), whereas accepting the associated trade off of loss of power (Vähä and Primmer, 2006). We found through simulations that $Tq = 0.90$ is an appropriate threshold for this purpose (Table 3). With STRUCTURE, we obtain a good estimate of the proportion of hybrids in the sample (>90%) with a very low associated error. With NEWHYBRIDS, reliable results on true hybrid identity (accuracy = 1.000) are obtained with virtually no error. Hence, the joint use of these Bayesian approaches is suggested to improve the resolution in hybrid identification, especially for studies relying on the prior identification of hybrid plants (for example, controlled crosses or detailed phenotypic observations of hybrids compared to parental species). We note that the present study was based on a very limited number of loci. In admixture zones that are already many generations old, both power and accuracy of hybrid detection will increase greatly if a much larger, genome-wide panel of diagnostic marker loci is used, especially if linkage between loci is accounted for during the estimation of hybrid ancestry (Falush *et al*., 2003).

No general rule about morphological features of hybrid individuals between *Q. suber* and *Q. ilex* can be deduced from previous studies. Putative hybrids with parental morphology (Toumi and Lumaret, 1998; Belahbib *et al*., 2001; Staudt *et al*., 2004), intermediate morphology (Toumi and Lumaret, 1998; Lumaret *et al*., 2002; Bellarosa *et al*., 2005) or leaf morphology skewed toward *Q. ilex* (Staudt *et al*., 2004) have been reported. Bark cannot be used as a discriminating feature because F1s are considered to lack cork and, thus, they could be confused with pure *Q. ilex* (Lumaret *et al*., 2002; Bellarosa *et al*., 2005). The existence of morphologically cryptic hybrids seems to be the only certainty. In any case, results from the studies cited above are hardly comparable among them and with the present one, due to the different sample designs and type of genome variability observed. Moreover, in all of these studies the identification of genetic diagnostic elements is dependent on the morphological determination of pure species. In contrast, the Bayesian approach used here allows us to define the genetic boundaries of pure species independently from any feature other than genetic data, thus allowing more accurate estimates of species status (Duminil *et al*., 2006) and gene exchange. We found a similar proportion of each parental morphotype among the putative hybrid individuals (Figure 2) and very good correspondence between morphotype and molecular-based assignation for the putative purebreds. However, discrepancy was detected for a few individuals, because three trees identified in the field as *Q. suber* were assigned to pure *ilex* using microsatellites and three putative hybrid individuals (*suber* SCa95 and *ilex* SCa36, SI2) were morphologically similar to one species but assigned with greater probability to the other species (Figure 2). In contrast, there was no ambiguous assignment with simulated data; that is, backcrosses to *Q. suber* (Bxs) were never assigned to *Q. ilex* with $q > 0.50$ by STRUCTURE (Figure 3a), and they were never assigned to *Q. ilex* nor to backcrosses with *Q. ilex* (Bxi) with $q > 0.10$ by NewHybrids (Figure 3b). The same was found with backcrosses to *Q. ilex* (Figures 3a and b). Thus, we consider that the discrepancy mentioned above is not due to the lack of resolution of the methods but reveals instead either the lack of correspondence between the phenotype and nuclear genotype (expected after several backcrosses, that is, 'advanced' introgressed individuals) or mislabeling of samples during their collection and processing (although this is unlikely for backcrosses, given their extremely low frequency in the sample).

Conclusions

The strength of our approach relied on the combination of two complementary Bayesian methods and on their validation by systematic simulations precisely adjusted to the empirical data investigated. The whole procedure is recommended to gain precision and accuracy in the identification of F1 hybrids and backcrosses for every real-case study, regardless of the level of hybridization. We expect that future studies of hybrids in natural populations will achieve even greater accuracy and power by increasing genomic coverage and accounting for linkage between loci. In the case of *Q ilex* and *Q. suber*, the identification of hybrid types has been addressed for the first time in this study. Our results suggest a very low rate of bidirectional gene flow between *Q. ilex* and *Q. suber*. Further studies are required to understand the geographic distribution and possible adaptive function of hybridization between these two species through time and space. Powerful and accurate detection of adult hybrid and introgressed individuals will be particularly valuable to address the adaptive differences among hybrid classes and the reproductive behavior of hybrid individuals.

Acknowledgements

This work was partially funded by the EU Project QLRT-2001-01594 (CREOAK) and by the Spanish Ministry of Environment (DGB) through the covenant 'Evaluación y conservación de los recursos genéticos de los *Quercus* esclerófilos mediterráneos en España'. We thank Stella Mérola, Unai Lopez de Heredia and Carmen García Barriga for their help with field and laboratory work, and P Montoya for English revision. We also thank Miguel Navascués for constructive discussions during the development of the work and on earlier versions of the paper. Suggestions by anonymous reviewers and editor were much appreciated.

References

Amaral Franco J (1990). *Quercus* L. In: Castroviejo S (ed). *Flora Ibérica. Plantas vasculares de la Península Ibérica e Islas Baleares*. Real Jardín Botánico, CSIC: Madrid. vol. II.

Anderson EC, Thompson EA (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**: 1217–1229.

Barilani M, Sfougaris A, Giannakopoulos A, Mucci N, Tabarroni C, Randi E (2007). Detecting introgressive hybridisation in rock partridge populations (*Alectoris graeca*) in Greece through Bayesian admixture analyses of multilocus genotypes. *Conserv Genet* **8**: 343–354.

Belahbib N, Pemonge MH, Ouassou A, Sbay H, Kremer A, Petit RJ (2001). Frequent cytoplasmic exchanges between oak species that are not closely related: *Quercus suber* and *Q. ilex* in Morocco. *Mol Ecol* **10**: 2003–2012.

Bellarosa R, Simeone MC, Papini A, Schirone B (2005). Utility of ITS sequence data for phylogenetic reconstruction of Italian *Quercus* spp. *Mol Phylogenet Evol* **34**: 355–370.
Boavida LC, Silva JP, Feijó JA (2001). Sexual reproduction in the cork oak (*Quercus suber* L.). II. Crossing intra- and inter-specific barriers. *Sex Plant Reprod* **14**: 143–152.
Boecklen WJ, Howard DJ (1997). Genetic analysis of hybrid zones: numbers of markers and power of resolution. *Ecology* **78**: 2611–2616.
Burgarella C, Navascués M, Soto A, Lora González A, Fici S (2007). Narrow genetic base in forest restoration with holm oak (*Quercus ilex* L.) in Sicily. *Ann Forest Sci* **64**: 757–763.
Burgess KS, Morgan M, DeVerno LL, Husband CB (2005). Asymmetrical introgression between two *Morus* species (*M. alba*, *M. rubra*) that differ in abundance. *Mol Ecol* **14**: 3471–3483.
Currat M, Ruedi M, Petit R, Excoffier L (2008). The hidden side of invasions: massive introgression by local genes. *Evolution* **62**: 1908–1920.
Curtu AL, Gailing O, Finkedley R (2007). Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. *BMC Evol Biol* **7**: 218.
Dodd RS, Afzal-Rafii Z (2004). Selection and dispersal in a multispecies oak hybrid zone. *Evolution* **58**: 261–269.
Dow B, Ashley M, Howe H (1995). Characterization of highly variable (GA/CT)n microsatellites in the bur oak, *Quercus macrocarpa*. *Theor Appl Genet* **91**: 137–141.
Duminil J, Caron H, Scotti I, Cazal S-O, Petit RJ (2006). Blind population genetics survey of tropical rainforest trees. *Mol Ecol* **15**: 3505–3513.
Dumolin-Lapègue S, Kremer A, Petit RJ (1999). Are chloroplast and mitochondrial DNA variation species independent in oaks? *Evolution* **53**: 1406–1413.
Dumolin S, Demesure B, Petit R (1995). Inheritance of chloroplast and mitochondrial genomes in pedunculate oak investigated with an efficient PCR method. *Theor Appl Genet* **91**: 1253–1256.
Elena-Rosselló JA, Lumaret R, Cabrera E, Michaud H (1992). Evidence for hybridization between sympatric holm-oak and cork oak in Spain based on diagnostic enzyme markers. *Vegetatio* **99–100**: 115–118.
Ellstrand NC, Whitkus R, Rieseberg LH (1996). Distribution of spontaneous plant hybrids. *Proc Natl Acad Sci* **93**: 5090–5093.
Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
Fernández-Manjarrés JF, Gerard PR, Dufour J, Raquin C, Frascaria-Lacoste N (2006). Differential patterns of morphological and molecular hybridization between *Fraxinus excelsior* L. and *Fraxinus angustifolia* Vahl (Oleaceae) in eastern and western France. *Mol Ecol* **15**: 3245–3257.
González-Rodríguez A, Arias DM, Valencia S, Oyama K (2004). Morphological and RAPD analysis of hybridization between *Quercus affinis* and *Q. laurina* (Fagaceae), two Mexican red oaks. *Am J Bot* **91**: 401–409.
Goudet J (2001). FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available at http://www2.unil.ch/popgen/softwares/fstat.htm.
Heuertz M, Carnevale S, Fineschi S, Sebastiani F, Hausman JF, Paule L *et al.* (2006). Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Mol Ecol* **15**: 2131–2140.
Hood GM (2005). PopTools version 2.6.6. Available at http://www.cse.csiro.au/poptools.
Howard DJ, Preszler RW, Williams JH, Fenchel S, Boecklen WJ (1997). How discrete are oak species? Insights form a hybrid zone between *Quercus grisea* and *Quercus gambelii*. *Evolution* **51**: 747–755.
Jensen RJ, Hokanson SC, Isebrands JG, Hancock JF (1993). Morphometric variation in oaks of the Apostle Islands in Wisconsin: evidence of hybridization between *Quercus rubra* and *Q. ellipsoidalis*. *Am J Bot* **80**: 1358–1366.
Jiménez MP, Lopez de Heredia U, Collada C, Lorenzo Z, Gil L (2004). High variability of chloroplast DNA in three Mediterranean evergreen oaks indicates complex evolutionary history. *Heredity* **93**: 510–515.
Kampfer S, Lexer C, Glössl J, Steinkellner H (1998). Characterization of (GA)n microsatellite loci from *Quercus robur*. *Hereditas* **129**: 183–186.
Kothera L, Ward SM, Carney SE (2007). Assessing the threat from hybridization to the rare endemic *Physaria bellii* Mulligan (Brassicaceae). *Biol Conserv* **140**: 110–118.
Lexer C, Fay MF, Joseph JA, Nica MS, Heinze B (2005). Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. *Mol Ecol* **14**: 1045–1057.
Lexer C, Kremer A, Petit RJ (2006). Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Mol Ecol* **15**: 2007–2012.
López de Heredia U, Carrión JS, Jiménez P, Collada C, Gil L (2007a). Molecular and palaeobotanical evidence for multiple glacial refugia for evergreen oaks on the Iberian Peninsula. *J Biogeogr* **34**: 1505–1517.
López de Heredia U, Jiménez P, Collada C, Simeone MC, Bellarosa R, Schirone B *et al.* (2007b). Multi-marker phylogeny of three evergreen oaks reveals vicariant patterns in the Western Mediterranean. *Taxon* **56**: 1199–1209.
López de Heredia U, Jiménez P, Díaz-Fernández PM, Gil L (2005). The Balearic Islands: a reservoir of cpDNA genetic variation for evergreen oaks. *J Biogeogr* **32**: 939–949.
Lumaret R, Mir H, Michaud H, Raynal V (2002). Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* L.). *Mol Ecol* **11**: 2327–2336.
Lumaret R, Tryphon-Dionnet M, Michaud H, Sanuy A, Ipotesi E, Born C *et al.* (2005). Phylogeographical variation of chloroplast DNA in cork oak (*Quercus suber*). *Ann Bot* **96**: 853–861.
Mallet J (2005). Hybridization as an invasion of the genome. *Trends Ecol Evol* **20**: 229–237.
Manos PS, Doyle JJ, Nixon KC (1999). Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). *Mol Phylogenet Evol* **12**: 333–349.
Manos PS, Zhou Z, Cannon CH (2001). Systematics of Fagaceae: phylogenetic tests of reproductive trait evolution. *Int J Plant Sci* **162**: 1361–1379.
Martín Vicente Á, Fernández Alés R (2006). Long term persistence of dehesas. Evidences from history. *Agroforest Syst* **67**: 19–28.
Muir G, Schlötterer C (2005). Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). *Mol Ecol* **14**: 549–561.
Nason JD (1992). Patterns of hybridization and introgression in populations of oaks, manzanitas, and irises. *Am J Bot* **79**: 101–111.
Nielsen EE, Bach LA, Kotlick P (2006). HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. *Mol Ecol Notes* **6**: 971–973.
Oliveira P, Custódio AC, Branco C, Reforço I, Rodrigues F, Varela MC *et al.* (2003). Hybrids between cork oak and holm oak: isoenzyme analysis. *Forest Genet* **10**: 283–297.
Oliveira R, Godinho R, Randi E, Ferrand N, Alves P (2007). Molecular analysis of hybridisation between wild and domestic cats (*Felis silvestris*) in Portugal: implications for conservation. *Conserv Genet* **9**: 1–11.
Olrik DC, Kjaer ED (2007). The reproductive success of a *Quercus petraea* x *Q. robur* F1-hybrid in back-crossing situations. *Ann Forest Sci* **64**: 37–45.

Palmé AE, Su Q, Palsson S, Lascoux M (2004). Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among *Betula pendula*, *B. pubescens* and *B. nana*. *Mol Ecol* **13**: 167–178.

Petit RJ, Bialozyt R, Garnier-Gere P, Hampe A (2004). Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecol Manag* **197**: 117–137.

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A (1997). Chloroplast DNA footprints of postglacial recolonization by oaks. *Proc Natl Acad Sci* **94**: 9996–10001.

Plieninger T, Pulido FJ, Konold W (2003). Effects of land-use history on size structure of holm oak stands in Spanish dehesas: implications for conservation and restoration. *Environ Conserv* **30**: 61–70.

Potts B, Reid JB (1988). Hybridization as a dispersal mechanism. *Evolution* **42**: 1245–1255.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Rieseberg LH, Carney SE (1998). Plant hybridization. *New Phytol* **140**: 599–624.

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T *et al.* (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**: 1211–1216.

Rushton BS (1993). Natural hybridization within the genus *Quercus* L. *Ann Sci Forest* **50** (Suppl 1): 73–90.

Seehausen O (2004). Hybridization and adaptive radiation. *Trends Ecol Evol* **19**: 198–207.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R *et al.* (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Gen* **60**: 957–964.

Soto A, Lorenzo Z, Gil L (2003). Nuclear microsatellites markers for the identification of *Quercus ilex* L. and *Quercus suber* L. hybrids. *Silvae Genet* **52**: 63–66.

Soto A, Lorenzo Z, Gil L (2007). Differences in fine-scale genetic structure and dispersal in *Quercus ilex* L. and *Q. suber* L.: consequences for regeneration of Mediterranean open woods. *Heredity* **99**: 601–607.

Staudt M, Mir C, Joffre R, Rambal S, Bonin A, Landais D *et al.* (2004). Isoprenoid emissions of *Quercus* spp. (*Q.suber* and *Q.ilex*) in mixed stands contrasting in interspecific genetic introgression. *New Phytol* **163**: 573–584.

Steinkellner H, Fluch S, Turetschek E, Lexer C, Streiff R, Kremer A *et al.* (1997). Identification and characterization of (GA/CT)n—microsatellite loci from *Quercus petraea*. *Plant Mol Biol* **3**: 1093–1096.

Toumi L, Lumaret R (1998). Allozyme variation in cork oak (*Quercus suber* L.): the role of phylogeography and genetic introgression by other Mediterranean oak species and human activities. *Theor Appl Genet* **97**: 647–656.

Toumi L, Lumaret R (2001). Allozyme characterization of four Mediterranean evergreen oak species. *Biochem Syst Ecol* **29**: 799–817.

Tovar-Sanchez E, Oyama K (2004). Natural hybridization and hybrid zones between *Quercus crassifolia* and *Quercus crassipes* (Fagaceae) in Mexico: morphological and molecular evidence. *Am J Bot* **91**: 1352–1363.

Vähä J-P, Primmer CR (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Mol Ecol* **15**: 63–72.

Valbuena-Carabaña M, Gonzalez-Martinez SC, Hardy OJ, Gil L (2007). Fine-scale spatial genetic structure in mixed oak stands with different levels of hybridization. *Mol Ecol* **16**: 1207–1219.

Varela MC, Valdiviesso T (1996). Phenological phases of *Quercus suber* L. flowering. *Forest Genet* **3**: 93–102.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Whittemore AT, Schaal BA (1991). Interspecific gene flow in sympatric oaks. *Proc Natl Acad Sci* **88**: 2540–2544.

Williams JH, Boecklen WJ, Howard DJ (2001). Reproductive processes in two oak (*Quercus*) contact zones with different levels of hybridization. *Heredity* **87**: 680–690.

Yang BZ, Zhao H, Kranzler HR, Gelernter J (2005). Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. *J Epidemiol* **28**: 302–312.

ANNEXE 5

Diagnostic tests for species identification

METHODS

Diagnostic tests are widely used in the field of medicine and are performed to aid in the diagnosis or detection of a disease (Leeflang *et al.*, 2008; Asimaki *et al.*, 2009; Uyeki *et al.*, 2009). Each test can be positive or negative. When compared to the actual status of the patient, four possibilities occur: true positive (sick patient correctly identified as sick), false positive (healthy patient incorrectly identified as sick), true negative (healthy patient correctly identified as healthy) and false negative (sick patient incorrectly identified as healthy). Therefore, these tests can be summarized in 2x2 contingency tables (Attia, 2003) and associated diagnostic measures, such as sensitivity or specificity, can be obtained. Sensitivity measures how well the test detects the disease when it is really there. Hence, a sensitive test has few false negatives. On the other hand, specificity measures how well the test rules out disease when it is really absent. Hence, a specific test has few false positives (Attia, 2003). Diagnostic tests, unlike assignments methods, are based on *a priori* information. Two types of errors can occur in conformity tests: false positive and false negatives. False positives (or type I error) are samples which are falsely declared non-conforms. False negatives (or type II error) are samples which are falsely declared conforms (Deguilloux *et al.*, 2003). Depending on the aim of the test, one might want to minimize one type of error more than the other. For example, when controlling seed trade or illegal logging, type I error should be minimized, so as not to falsely accuse somebody. Conversely, when conformity is critical, whereas falsely excluding some *bona fide* samples has only limited consequences, type II error should be minimized. Regardless of the practical objective, a major advantage of conformity tests over assignment methods is their ability to quantify associated risks (type I and II errors).

Positive and negative diagnostic likelihood ratios

We applied diagnostic tests used in medicine to develop diagnostic tests, based on diagnostic likelihood ratios, for two species (in our case two close related oak species, *Quercus robur* and *Q. petraea*). We replaced counts of sick and healthy patients in contingency tables by allelic frequencies at each SNP, for both species (see **Chapter 4**). Diagnostic likelihood ratios are

directly deduced from sensitivity and specificity (Attia, 2003). Positive diagnostic likelihood ratio (PDLR) represents the ratio of the odds that the most frequent allele in one targeted species will be observed in this species compared to the odds that the same allele will be observed in the other species. It can be expressed as $\left[\frac{\text{sensitivity}}{1-\text{specificity}}\right]$. Negative diagnostic likelihood ratio (NDLR) represents the odds ratio that a rare allele will be observed in the targeted species compared to the odds that the same allele will be observed in the other species. It can be expressed as $\left[\frac{1-\text{sensitivity}}{\text{specificity}}\right]$.

Conformity test for *Q. robur*		True diagnostic	
		Q. robur	***Q. petraea***
Test result	**Conform**	a =0.99	b=0.4
	Nonconform	c=0.01	d=0.6

Table 1: example of contingency table based on allelic frequencies for a conformity test for *Q. robur*. "a" is the most frequent allele in *Q. robur* and "b" is the corresponding allele in *Q. petraea*.
Positive diagnostic likelihood ratio (PDLR) = [a/(a+c)]/[b/(b+d)]=0.99/0.4=2.475
Negative diagnostic likelihood ratio (NDLR) = [c/(a+c)]/[d/(b+d)]=0.01/0.6=0.017

The most powerful SNPs will have NDLR close to 0 and PDLR with highest possible values. The ideal SNP to test conformity to either species is fixed for one allele in species A and fixed for the other allele in species B. All SNPs getting close to this ideal marker should be powerful for diagnostic tests. In one preliminary approach, allelic frequencies at each SNP for both species were used to complete contingency tables, and to calculate associated diagnostic likelihood ratios. Afterwards, we used genotypic frequencies to improve diagnostic likelihood ratios as rare alleles in targeted species are generally found in heterozygous genotypes. To allow NDLR and PDLR measures on genotypic frequencies, we had to combine the two most frequent genotypes in the targeted species to complete contingency tables. For the other species, we kept the same group of two genotypes to calculate the two ratios. This way, test is more powerful for the targeted species, but is less powerful for the other one, confirming the necessity to develop different tests if targeting different species. Based on 855 samples genotyped at 262 SNPs (see **Chapter 4**), PDLR and NDLR were calculated from three datasets that were created based on the admixture level of each sample, with the objective to target one specific category at once. NDLR and PDLR for

Q. robur were calculated using *Q. robur* purebreds frequencies in comparison to all other samples. With the same method, we calculated NDLR and PDLR for *Q. petraea* and for F1 hybrids, by targeting on category against all the others. Hence, conformity was more stringent as intermediates (F1 hybrids and backcrosses) were considered as non-conform. In cases were intermediates can be tolerated for conformity testing, they should be included in the targeted category, but associated errors would be higher. PDLR and NDLR were calculated on genotypic frequencies for each category (*Q. robur*, *Q. petraea* and intermediates) and for each locus.

RESULTS

Diagnostic likelihood ratios for both species were estimated and ranked according to the PDLR/NDLR ratio ("accuracy" in medical diagnostic tests). "Accuracy" of conformity tests differed between species, with the best SNPs for *Q. robur* conformity outperforming those selected for *Q. petraea* (Figure 1).

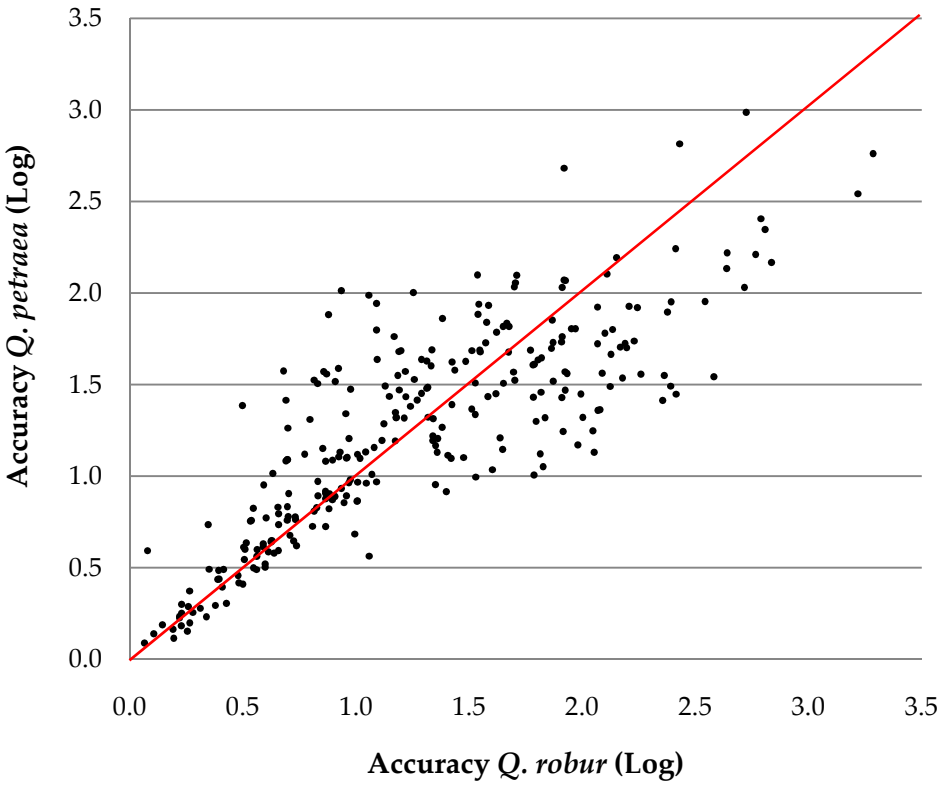


Figure 1: Accuracy (=PDLR/NDLR) of each SNP for species conformity (*Q. robur* and *Q. petraea*). Most of the loci with the highest accuracy values (Log>2) target *Q. robur*.

Among the best 50 SNPs, only 15 (30%) were for *Q. petraea*. Hence, associated errors will be generally lower if the targeted species is *Q. robur*. For SNPs with the best NDLR (nonconformity tests), the proportion of markers targeting *Q. petraea* decreased to 11 (22%) among the 50 best loci. On the contrary, for SNPs with the best PDLR (conformity tests), proportion were reversed. For these 50 loci, mean NDLR was twice higher for *Q. robur* than for *Q. petraea*, whereas mean PDLR was twice higher for *Q. petraea* than for *Q. robur* (Table 2).

Category	PDLR	NDLR	Accuracy
Quercus robur	27.3	0.014	287.9
Quercus petraea	44.8	0.027	152.5
Intermediates	7.9	0.262	10.2

Table 2: Mean PDLR, NDLR and accuracy (=PDLR/NDLR) for the best 50 SNPs for *Q. robur*, *Q. petraea* and intermediates (F1 hybrids and backcrosses).

Intermediates had low values for all ratios, which underlined the difficulty to develop accurate tests to identify these samples. In conclusion, it will be easier to declare a *Q. robur* nonconform or a *Q. petraea* conform than a *Q. robur* conform or a *Q. petraea* nonconform. Depending on the problematic, appropriate sets of loci must be chosen for conformity or nonconformity testing of each species separately.

APPLICATIONS

Diagnostic likelihood ratios are widely used for diagnostic tests in medicine but had, to our knowledge, never been used for species conformity tests. When adapted to molecular markers such as SNPs to identify species, they allow direct visualization of the diagnostic power of each locus. The use of genotype frequencies instead of allelic frequencies to complete contingency tables helped to increase the test power. At a larger scale, loci can be classified on the basis of their ability for conformity or nonconformity tests. But conformity tests target only one question at a time: is this sample conform or nonconform to this specific species? Different subsets of loci need to be used for each question. But one major advantage of this simple approach is the possibility to estimate associated errors (type I and type II), which can be relevant in some fields (control of illegal logging, certification of seed lots used for plantations) where accurate estimation of risks to declare conformity or nonconformity

are requested. Conformity tests relying on SNPs rather than SSRs have the advantage to be more easily applicable to degraded DNA present in dry wood samples (Asari *et al.*, 2009). With such techniques, only few appropriate SNP can allow accurate control of wood species.

REFERENCES

Asari M, Watanabe S, Matsubara K, Shiono H, Shimizu K (2009) Single nucleotide polymorphism genotyping by mini-primer allele-specific amplification with universal reporter primers for identification of degraded DNA. *Analytical Biochemistry* **386**, 85-90.

Asimaki A, Tandri H, Huang HD, *et al.* (2009) A new diagnostic test for arrhythmogenic right ventricular cardiomyopathy. *New England Journal of Medicine* **360**, 1075-1084.

Attia J (2003) Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests. *Australian Prescriber* **26**, 111-113.

Deguilloux MF, Pemonge MH, Bertel L, Kremer A, Petit RJ (2003) Checking the geographical origin of oak wood: molecular and statistical tools. *Molecular Ecology* **12**, 1629-1636.

Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM (2008) Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* **149**, 889.

Uyeki TM, Prasad R, Vukotich C, *et al.* (2009) Low sensitivity of rapid diagnostic test for influenza. *Clinical Infectious Diseases* **48**, E89-E92.

Erwan GUICHOUX - 2011
UMR BIOGECO 1202, INRA, 69 route d'Arcachon, 33612 CESTAS Cedex
CRPR, 120 avenue du Maréchal Foch, 94015 CRÉTEIL Cedex

Au cours du vieillissement, les caractéristiques organoleptiques du vin se modifient au contact du bois de chêne. Le composé aromatique le plus important, la whisky-lactone, aux notes noix de coco et boisé, est facilement détectable et apprécié par les consommateurs. *Quercus petraea* et *Q. robur*, les deux principales espèces européennes de chêne utilisées pour le vieillissement des vins, ont des profils aromatiques très contrastés, particulièrement pour la whisky-lactone. Parvenir à identifier l'espèce de chêne permettrait de fournir aux tonnelleries des lots de bois plus homogènes. L'objectif de cette étude est d'identifier l'espèce de chêne à partir de bois sec, à l'aide de marqueurs moléculaires utilisables dans un contexte industriel. Le bois sec est un tissu mort dans lequel l'ADN est très dégradé et donc difficilement accessible. Pour optimiser l'extraction d'ADN à partir de ce tissu, nous avons développé une méthode de PCR en temps-réel ciblant l'ADN chloroplastique, permettant ainsi d'évaluer l'efficacité des différents protocoles d'extraction. Nous avons également développé des marqueurs moléculaires (SSRs et SNPs) fortement différenciés entre espèces et particulièrement bien adaptés au bois. Grâce à des protocoles d'extraction d'ADN optimisés et ces marqueurs performants, nous avons pu identifier l'espèce sur des lots de bois séchés pendant deux ans. De plus, par l'étude de 262 SNPs dont la moitié est fortement différenciée entre espèces, nous avons démontré que les gènes sélectionnés (loci « outlier ») sont très performants pour délimiter ces deux espèces proches. Ils permettent également de détecter des processus démographiques fins (flux de gènes intra- et interspécifiques), alors que les gènes a priori non-sélectionnés (loci neutres) se révèlent peu informatifs.

Mots-clés: *Quercus spp.*, ADN dégradé, bois, loci outliers, méthodes d'affectation, multiplex SSRs, SNPs

Most of aromatic compounds in wine are directly induced during maturation by the contact with oak wood. For example, whisky-lactone, the most important aromatic compound, which gives a coconut and woody taste, is easily detected and appreciated by consumers. *Quercus petraea* and *Q. robur*, the two major European oak species used for wine maturation, have very contrasted aromatic patterns, especially for whisky-lactone. Identifying the species used for cooperage will facilitate the maturation process, for instance by providing wineries with more homogenous batches of barrels. The objective of our study is to characterize the oak species directly from dry wood, using molecular markers that will be applicable in an industrial context. Unfortunately, dry wood is a dead tissue in which DNA is highly degraded and difficult to access. To optimize DNA recovery from dry wood, we developed a quantitative PCR protocol based on chloroplast DNA to evaluate the efficiency of DNA isolation protocols. We identified and developed molecular markers (SSRs and SNPs) adapted to dry wood that are particularly diagnostic. Using an optimized DNA isolation protocol and these powerful markers, the species identity from wood samples dried during two years could be successfully characterized. Using 262 SNPs highly differentiated between the two species, we also demonstrate that genes under selection (outlier loci) have outstanding power to delimitate the two oak species and provide unique insights on intra- and interspecific gene flow, whereas genes lacking such a signature (putatively neutral loci) provide little or no resolution.

Keywords: *Quercus spp.*, degraded DNA, wood, outlier loci, assignment methods, SSR multiplex, SNPs