THESE

présentée à

L'UNIVERSITE DE BORDEAUX 1

Ecole Doctorale Science et Environnements

par Lélia LAGACHE

pour l'obtention du grade de

DOCTEUR

SPECIALITE : Ecologie évolutive, fonctionnelle et des communautés

---

Hybridation et dynamique de la spéciation chez les chênes sessile (*Quercus petraea*) et pédonculé (*Quercus robur*)

Soutenue le 14 décembre 2012

Devant la commission d'examen formée de :

Mme Myriam HEUERTZ	Chargé de recherche, CIFOR-INIA, Madrid	Rapporteur
Mme Sophie ARNAUD-HAOND	Chargé de recherche, IFREMER, Sète	Rapporteur
Mme Brigitte CROUAU-ROY	Professeur à l'université de Toulouse	Présidente
M. Rémy PETIT	Directeur de Recherche, INRA, Bordeaux	Directeur de thèse

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

N° ordre : 4704

Hybridation et dynamique de la spéciation chez les chênes sessile et pédonculé

Les chênes sessiles et pédonculés sont deux espèces sympatriques interfertiles occupant des places distinctes dans la succession écologique. Elles constituent pour cela un bon modèle pour l'étude de la spéciation écologique. Malgré leur écologie contrastée, les deux espèces de chênes coexistent naturellement dans de nombreuses forêts, rendant possible l'étude de leur système de reproduction intra- et interspécifique. Des travaux précédents ont suggéré que l'hybridation entre ces deux espèces serait fréquence-dépendante. Elle dépendrait de la proportion de pollen allospécifique (c'est-à-dire de pollen de l'autre espèce) que reçoit l'arbre mère étudié. Ce phénomène d'hybridation fréquence-dépendante est appelé effet Hubbs, du nom d'un ichtyologue qui découvrit ce mécanisme en 1955. Mon travail a consisté à étudier, dans une parcelle mixte de chênes, les barrières à l'hybridation qui permettent la coexistence de ces deux espèces. Pour cela, j'ai effectué une étude de paternité de grande ampleur (près de 3500 individus typés sur 12 marqueurs microsatellites). Tout d'abord, je me suis intéressée à la délimitation des deux espèces en appliquant pour la première fois *in situ* le critère d'interfertilité. Je me suis ensuite concentrée sur les facteurs qui influencent l'hybridation au travers d'une modélisation des croisements à l'échelle de l'individu. Les résultats montrent que le maintien de ces deux espèces est régi par deux composantes environnementales importantes : la fréquence de chaque espèce et leur distribution, qui influencent la quantité de pollen reçue. Grâce à cette étude empirique et à l'approche de modélisation de ces croisements, nous avons désormais une bien meilleure vision de l'effet de l'environnement sur l'hybridation. J'ai par la suite comparé les caractéristiques du système de reproduction de chaque espèce (dispersion du pollen et fécondité mâle) en cherchant si un lien existait avec leur stratégie écologique. Les résultats suggèrent que les différences de dynamique écologique pourraient être à l'origine de la spéciation du fait de l'existence de compromis différents en termes d'allocation de ressources mais qu'à plus court terme la stabilité de l'environnement est essentielle au maintien des espèces.

Mots clés : spéciation écologique, *Quercus robur*, *Q. petraea*, hybridation, délimitation d'espèces, concept d'espèces, stratégies écologique, perturbations.

Hybridization and speciation dynamic of *Quercus petraea* and *Quercus robur*

Quercus petraea and *Q. robur* are two interfertile sympatric species. They occupy distinct stages during forest succession and constitute therefore good models for ecological speciation studies. Despite their differences, they often grow together in mixed stands, allowing the study of their intra- and interspecific reproductive system. Hybridization between these two oak species has been suggested to be frequency-dependent. The effect of the relative species abundance on hybridization is known as the "Hubbs' effect", from the name of an ichthyologist who described this mechanism in 1955. My work was to study the processes that limit hybridization between these two species, thereby allowing their coexistence. I conducted an extensive paternity analysis (almost 3500 individuals genotyped at 12 microsatellite markers). First, I tried to delimitate the two oak species by applying for the first time *in situ* the interfertility criterion. Then, I focused on interspecific crosses by studying those factors influencing hybridization. Results showed that the maintenance of these two species depends on the frequency of each species and their distribution, as both factors influence the quantity of pollen received by female flowers. Thanks to this empirical study and to this modeling approach, we have now a much better view of environmental effects on hybridization. I then compared the characteristics of the reproductive system of each species (pollen dispersal and male fertility) in relation with their ecological strategies. The results suggest that differences in species ecological dynamics are at the origin of the speciation process but that at a finer scale the stability of the environment is crucial for species' maintenance.

Keywords: ecological speciation, *Quercus robur*, *Q. petraea*, hybridization, species delimitation, species concept, ecological strategies, disturbance.

REMERCIEMENTS

Je tiens à remercier tout particulièrement mon directeur de thèse, Rémy Petit, pour avoir encadré cette thèse et surtout avoir supporté mon caractère. Merci pour ta patience, ta présence et ton aide tout au long de mon travail. Merci pour ton écoute lors des phases de doute concernant la suite de ma carrière professionnelle. Merci pour toutes ces discussions qui m'ont permis de prendre du recul sur mon travail de thèse. Je ne sais pas encore quel est le devenir de ma carrière professionnelle, mais j'espère que notre collaboration continuera.

Je remercie également l'ensemble des membres du jury et tout particulièrement Myriam Heuertz et Sophie Arnaud-Haond pour avoir accepté d'être les rapporteurs de mon travail de thèse.

Celui-ci est un travail collectif qui m'a permis d'être en interaction avec un certain nombre de personnes que je souhaite tout particulièrement remercier :

Merci à toi, Erwan, de m'avoir encadrée durant mon stage de Master 2 et de m'avoir appris les techniques de laboratoire : c'est un peu grâce à toi si j'en suis arrivée là aujourd'hui. Merci pour ta présence au laboratoire ainsi qu'au bureau pendant les quatre ans qui viennent de se dérouler. Merci d'être venu récolter tous les descendants de la P37 et surtout d'avoir bien voulu me suivre une troisième fois, pour les mesures de $\delta^{13}C$. Grâce à toutes les discussions techniques échangées, surtout pendant la dernière année de la thèse, j'ai pu échapper pendant quelques minutes à la rédaction et garder le contact avec le laboratoire qui me manquait.

Je souhaite remercier Etienne, sans qui deux chapitres de cette thèse n'auraient pu être écrits. Merci d'avoir pris le temps de m'expliquer ce qu'est un modèle de voisinage. Et même si je ne comprends pas encore tout, j'ai le sentiment d'avoir quand même beaucoup appris à tes côtés. Merci d'avoir pris le temps de m'accueillir en Avignon par trois fois et d'avoir toujours répondu à mes questions.

Je souhaite également te remercier, Corinne, pour l'étude sur les réseaux de reproduction et d'apparentement. Merci pour ton aide, tes critiques et surtout ta présence lors de l'écriture du second article de ma thèse. Merci de m'avoir présenté à Jean-Jacques et Jean-Benoist. Merci à eux pour m'avoir permis d'explorer le monde des réseaux et de leur modélisation.

Je souhaite remercier également Alexis et Jean-Marc pour leur mémoire et leur aide concernant les parcelles 26 et 37 de la Petite Charnie. Merci à Alexis d'être venu, malgré son calendrier chargé, aider à la première récolte de matériel végétal et d'avoir retrouvé dans ses nombreuses archives des données nécessaires à l'étude fine des croisements reproducteurs.

Merci à Patrick et Steffi pour toutes les discussions sur la mise au point et la lecture du kit 12plex. Merci pour le temps que vous m'avez accordé pour réaliser la double lecture du génotypage, me permettant ainsi de travailler sur un meilleur jeu de données. Merci à Steffi d'avoir risqué la vie de ses doigts de pied pour m'aider à la récolte des bourgeons en hiver, et d'avoir failli mourir d'insolation lors de la seconde récolte en été.

Merci à Pauline pour ta présence et ton aide lors de la conception de la puce ILLUMINA 384plex. Merci également de ton aide lors de la récupération des données et du matériel végétal des transects de Nancy.

Merci à François pour m'avoir éclairé sur les méthodes et le vocabulaire de la phylogénie, qui m'étaient complètement étrangers.

Merci à Benjamin d'avoir supporté tous mes problèmes (personnels ou non) pendant notre heure de covoiturage quotidienne. Merci de ne pas m'avoir abandonnée sur le bord de la route, quand bien même l'idée a pu te traverser l'esprit.

Merci à Audrey et Maïmiti pour nos discussions à l'heure du déjeuner, qui me permettaient de m'évader un peu.

Plus généralement, je souhaite remercier tous les pierrotonnais qui ont aidé de près ou de loin à la réalisation de ce travail. Merci à tous ceux avec qui j'ai pu discuter, de tout et de rien. Cela m'a permis de travailler dans des conditions de travail accueillantes et agréables.

Merci à mes parents, qui malgré mon échec au tout début de mes études post-bac ont continué à avoir confiance en moi. Je ne sais pas comment se poursuivra ma carrière professionnelle, mais j'espère que quels que soient mes futurs choix, vous les comprendrez et ne serez pas déçus.

Merci à ma petite sœur, Zélie, d'avoir été là pour moi lorsque le moral chutait. Merci aussi d'avoir écouté sagement les différentes présentations d'entrainement dont j'ai pu avoir besoin tout au long de ma thèse. D'ailleurs prépares toi, le plus long est à venir ^^.

Enfin merci à toi Christophe, pour tout ce que tu as fait pour moi. Merci de m'avoir suivi sur Bordeaux, de m'avoir soutenu les jours difficiles, et d'une manière générale, d'avoir toujours étais là pour moi. J'ai hâte d'être au 31 Décembre, même si cela passe par la soutenance de ma thèse (que je redoute, comme tu le sais). Saches que ce jour là ma réponse ne sera empreinte d'aucun doute.

http://lapinobservateur.over-blog.com

SOMMAIRE

INTRODUCTION

Parcelle mixte de chêne sessile et pédonculé dans la forêt de la Petite Charnie (P26, 1991)

Cette thèse a été financée par deux projets Européens, le réseau d'excellence européen EvolTree pour la première année et le projet LinkTree (ANR BIODIVERSA) pour les deux années suivantes. Elle a débuté en Janvier 2009 pour une durée légèrement inférieure à trois ans. Tous les travaux présentés dans cette thèse ont été réalisés au sein de l'UMR BIOGECO, dans l'équipe génétique des populations, sous la direction de Rémy Petit. Ce travail de thèse a fait l'objet de plusieurs collaborations, notamment avec Etienne Klein, Corinne Vacher, Jean-Jacques Daudin, Jean-Benoist Léger et Laurent Bouffier, qui m'ont permis d'exploiter au mieux les résultats. Ces derniers sont présentés sous forme d'une thèse sur articles. Après une introduction générale concise qui présente les objectifs de la thèse, trois chapitres sous forme d'articles rédigés en anglais, chacun suivis de leurs perspectives et compléments, en français, seront présentés, suivis d'une conclusion générale.

Au cours du temps, le monde biologique s'est diversifié pour donner naissance à différentes populations d'organismes plus ou moins ressemblants. Cette diversification a opéré, et continue de le faire de nos jours, sous la contrainte des forces évolutives que sont la mutation, la sélection, la dérive génétique et la migration. L'homme a pris conscience de ces patrons de diversité et de ressemblance et a tenté de classer ces organismes dans des catégories : les espèces. Historiquement, le principal cadre utilisé pour cela est celui fixiste de la Création (classification de Linné dont la première parution date de 1735). Selon ce cadre, les espèces n'évoluent pas : elles sont restées sous la même forme depuis leur création par Dieu et ont toujours été présentes sur terre. L'abandon d'une vision fixiste a incité les biologistes à définir l'espèce en fonction des processus évolutifs à l'origine de la divergence, aboutissant à une (trop) grande diversité de concepts (listés par Hausdorf 2011). La définition de l'espèce biologique de Ernst Mayr (1942), proche de celle initialement proposée par Theodosius Dobzhansky (1935; 1937), définit les espèces comme des ***"groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups"*.** C'est la plus connue et donc aussi la plus critiquée. La différenciation graduelle des espèces et la persistance de flux de gènes après la « spéciation » (mais voir Orr 2001; Noor 2002; Coyne & Orr 2004), ainsi que son caractère peu opérationnel (e.g. Donoghue 1985; de Meeûs *et al.* 2003), ont abouti aux mises en cause les plus sévères. Ainsi, la question « Qu'est ce qu'une espèce ? » est devenue la question la plus célèbre à l'interface entre biologie et philosophie (Hey 2001), au point d'engendrer chez les biologistes une certaine lassitude et un doute sur l'intérêt des discussions sur le sujet (ex. Coyne & Orr 2004). Par contre, les méthodes de délimitation utilisées pour classer les individus dans les espèces constituent le point de départ de la plupart des études empiriques (en génétique, écologie, …) et sont moins controversées (Dayrat 2005). Pourtant il existe autant, sinon plus, de méthodes de délimitation des espèces que de définitions de l'espèce (voir par exemple les revues de Sites & Jonathon 2004; de Queiroz 2007). Malgré cette grande diversité, aucune étude n'a à ce jour exploré en conditions naturelles la possibilité de délimiter les espèces sur la base du critère d'interfertilité, critère pourtant directement inspiré de la définition « biologique » de l'espèce de Theodosius Dobzhansky et Ernst Mayr, et considéré comme le plus pertinent pour l'étude de la spéciation (Coyne & Orr 2004).

En effet, le maintien des espèces n'est possible que si des barrières à l'hybridation existent. Ces barrières peuvent être de deux types : prézygotiques et postzygotiques. Chez les plantes, les barrières pré-zygotiques sont typiquement plus fortes que celles post-zygotiques (Lowry *et al.* 2008). Un exemple de barrière pré-zygotique est l'avantage au pollen conspécifique lors de la compétition pollinique. Toutefois, ces barrières peuvent être

suffisantes pour que les espèces restent distinctes, sans être pour autant absolues, aboutissant au phénomène d'hybridation. Dans ce cas, la délimitation des espèces peut devenir difficile. Une meilleure connaissance du fonctionnement de ces barrières ainsi que des mécanismes sous-jacents est nécessaire pour comprendre le maintien d'espèces entre lesquelles l'hybridation est possible.

Ces barrières ont longtemps été perçues comme uniquement contrôlées par des facteurs endogènes. Or des études préliminaires, essentiellement basées sur des observations *in situ*, ont mis en évidence un effet de la rareté des espèces sur leur taux d'hybridation au niveau populationnel, également appelé « effet Hubbs » (Focke 1881; Hubbs 1955; Lepais *et al.* 2009). Cet effet a également été décrit en conditions de croisements contrôlés au niveau individuel (ex. Rieseberg & Carney 1998). Dans ce genre d'études, différents mélanges de pollen allo- et conspécifiques sont injectés sur les fleurs d'une même plante (proportion de pollen allospécifique variant de 0 à 100%). Le taux d'hybridation augmente alors typiquement avec la proportion de pollen allospécifique (figure 1). De plus, cette augmentation n'est pas linéaire et c'est uniquement à partir d'une certaine proportion de pollen allospécifique que le taux d'hybridation devient notable (~50% dans le cas des croisements sur mère *Iris fulva* présentés dans la figure 1). Cet effet de la fréquence des espèces sur leur taux d'hybridation a été formalisé mathématiquement par Chan et Levin (2005), en prenant en compte la force des barrières à l'hybridation des espèces et leur abondance relative (Eq. 1). Cette modélisation de l'effet Hubbs peut s'appliquer à l'échelle populationnelle ou individuelle (figure 2). Selon ce modèle et pour le cas de barrières à l'hybridation symétriques, le taux d'hybridation total est maximal lorsque les deux espèces sont en mélange équilibré 50%-50% (figure 2). Par contre, pour une espèce donnée, le taux d'hybridation augmente quand son abondance relative diminue. L'hybridation est donc la résultante d'une interaction entre une barrière génétique endogène à l'individu et son environnement biotique. Une récente étude chez le chêne en conditions de croisements contrôlés a montré, en utilisant des individus clonés, que le taux d'hybridation d'un individu dépend également de l'environnement abiotique (Abadie *et al.* 2011).

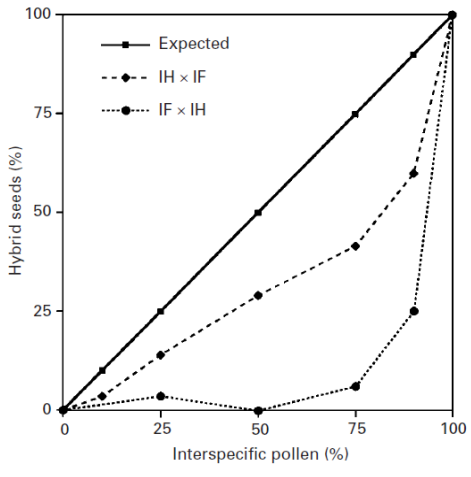


Figure 1 : Pourcentage de graines hybrides issues de *Iris hexagona* (IH) *et Iris fulva* (IF) en fonction du pourcentage de pollen allospécifique dans le mélange (Rieseberg & Carney 1998). *Pour chaque croisement (i.e. IH×IF ou IF×IH), la première espèce est la mère et la seconde est celle du donneur de pollen allospécifique. La ligne « expected » correspond au pourcentage d'hybrides attendu si la relation entre le taux d'hybridation et le pourcentage de pollen allospécifique était linéaire.*

$$Hyb_{tot} = \frac{h_{1<-2}q_2}{q_1 + h_{1<-2}q_2} + \frac{h_{2<-1}q_1}{q_2 + h_{2<-1}q_1} \qquad \text{(Eq. 1)}$$

Equation 1 : Modélisation de l'effet de la proportion de pollen allospécifique sur le taux d'hybridation (Chan & Levin 2005): *Hyb_{tot} est le pourcentage total d'hybrides dans une population constituée de deux espèces. A l'échelle populationnelle q_1 et q_2 sont les proportions relatives de l'espèce 1 et de l'espèce 2. $h_{1<-2}$ and $h_{2<-1}$ sont les deux barrières sexuelles de l'espèce 1 et de l'espèce 2 vis-à-vis du pollen allospécifique de l'espèce 2 et de l'espèce 1, respectivement. Le pourcentage total d'hybrides formés dans une population se décompose donc en deux termes, le premier est le nombre total d'hybrides dont les mères appartiennent à l'espèce 1 et le deuxième terme est le nombre total d'hybrides dont les mères appartiennent à l'espèce 2. Si on souhaite calculer le taux d'hybridation à l'échelle de l'individu, il suffit d'utiliser un des deux termes de cette équation, selon l'espèce de cet individu, avec q_1 et q_2 correspondant à la proportion de pollen de l'espèce 1 ou de l'espèce 2 reçue.*

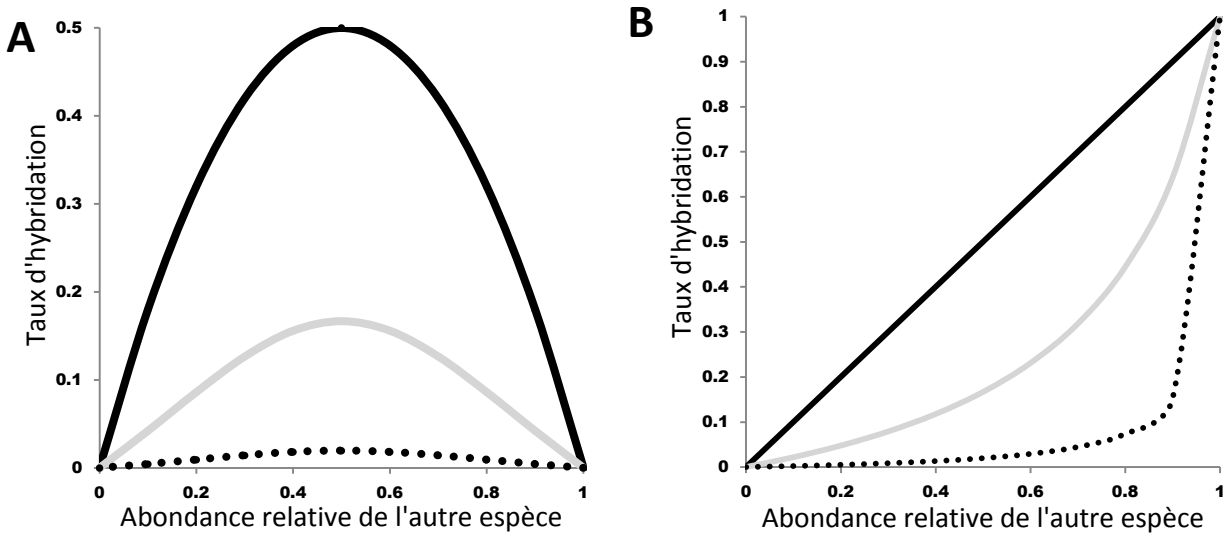


Figure 2: Taux d'hybridation en fonction de la proportion relative de chaque espèce selon le modèle de Chan et Levin (2005) au niveau populationnel (A) et individuel (B). *Les différentes courbes correspondent à différentes intensités pour les barrières endogènes à l'hybridation entre deux espèces (cas de barrières **symétriques**). La courbe noire représente le taux d'hybridation entre deux espèces en l'absence de barrière (panmixie), la courbe grise correspond à une barrière génétique faible (h=0.2) et la courbe noire en pointillés à une barrière reproductive forte (h=0.02).*

Plusieurs études ont mis en évidence une augmentation du taux d'hybridation dans des environnements dits « perturbés ». Les auteurs de ces études émettent l'hypothèse que c'est l'apparition d'un milieu aux caractéristiques intermédiaire entre ceux caractéristiques de chaque espèces qui serait l'origine de l'observation d'hybrides dans ces milieux perturbés (Zirkle 1935; Muller 1952; Vilà *et al.* 2000). Si l'effet Hubbs est considéré, une autre hypothèse est possible. En effet, l'augmentation d'hybridation en conditions naturelles dans des environnements dit « perturbés » pourrait s'expliquer par la modification de l'environnement biotique des espèces. Les perturbations des milieux pourraient alors remettre en contact des espèces habituellement séparées par leurs préférences écologiques. Ainsi, la modification de la répartition spatiale des espèces pourrait augmenter la proportion de pollen allospécifique que reçoit chaque arbre de chaque espèce car il se retrouverait avec plus de voisins allospécifiques, augmentant le taux d'hybridation. Des études fines,

spatialisées au niveau individuel, qui modélisent l'hybridation au travers de l'effet Hubbs, semblent donc nécessaires pour comprendre précisément les mécanismes modulant l'hybridation des individus. Ceci semble particulièrement important et urgent. Avec l'évolution du climat, les aires de répartition des populations changent (Vörösmarty *et al.* 2000), et certaines espèces qui jusqu'alors étaient séparées géographiquement peuvent se retrouver en contact et s'hybrider voire disparaitre par assimilation. Ce genre d'étude permettrait de mieux prédire leur hybridation et ses conséquences.

Pour mieux comprendre l'hybridation des espèces et surtout la mise en place des barrières reproductives interspécifiques, il faut considérer le contexte dans lequel les nouvelles espèces apparaissent. Traditionnellement, les premières études sur la spéciation ont distingué trois contextes géographiques pour la formation de nouvelles espèces. La spéciation allopatrique correspond au cas où les populations sont complètement séparées (voir figure 3). Deux populations peuvent également se différencier tout en étant partiellement ou totalement en contact, on parlera alors de spéciation parapatrique ou sympatrique (figure 3).

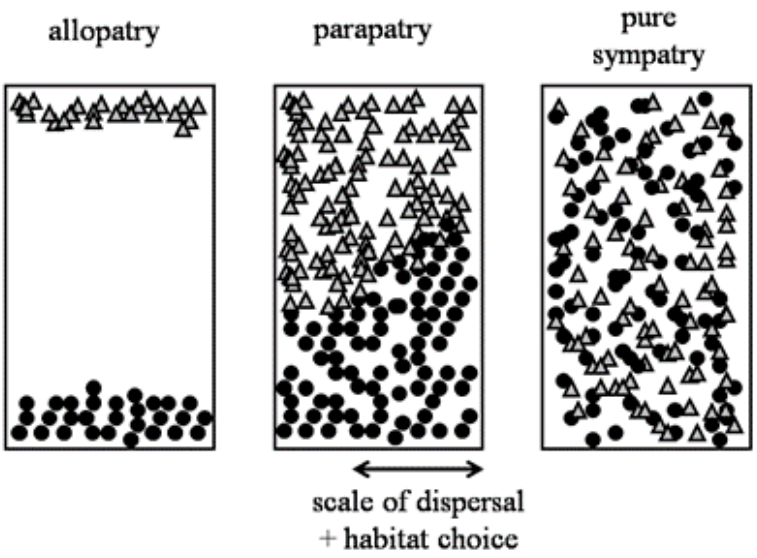


Figure 3 : Les différents types de spéciation en fonction des configurations spatiales des futures espèces (Mallet 2008; Mallet *et al.* 2009). *Les individus des deux populations correspondant aux futures espèces sont représentés par des ronds noirs ou des triangles blancs.*

La pertinence d'une vision purement géographique et à très large échelle de la spéciation a récemment été remise en cause (ex. Fitzpatrick *et al.* 2009). En effet, la spéciation est le résultat de processus micro-évolutifs et tout particulièrement de l'adaptation à des conditions environnementales différentes (Rundle & Nosil 2005; Mallet *et al.* 2009). Il convient donc d'étudier la spéciation à une échelle spatiale plus fine (voir par exemple l'étude de Schlüter 1994). De plus, restreindre l'étude de la spéciation à un cadre statique semble problématique car des études récentes montrent que la spéciation est plus rapide lorsque qu'on prend en compte la dynamique des populations (Aguilée *et al.* 2011). D'ailleurs, d'après l'étude de Ronce et Olivieri (1997), c'est dans ce cadre que différentes pressions de sélection vont apparaitre au sein des populations et favoriser la formation de populations aux stratégies écologiques contrastées. Burton *et al.* (2010) ont modélisé cet effet et montré que les populations en front de migration seront sélectionnées pour allouer plus de ressources pour la dispersion et moins pour la croissance, comparé aux populations en aval de ce front. L'adaptation des populations à ces différentes phases de la dynamique

de colonisation pourrait entrainer dès lors une sélection divergente sur de nombreuses composantes de la croissance et de la reproduction. Ce type de sélection multiple, également appelé « *multifarious divergent selection* » (Nosil *et al.* 2009), est un moteur puissant d'une spéciation dite « écologique ». En effet, la spéciation écologique est définie comme : ***"the process by which barriers to gene flow evolve between populations as a result of ecologically based divergent selection between environments"*** (Nosil 2012).

Cette divergence entre espèces sur de nombreux caractère est liée au « principe d'allocation des ressources » (Cody 1966). En effet un individu dispose de ressources limitées et ne pourra donc augmenter son investissement dans une fonction donnée (reproduction, dispersion, croissance…) sans diminuer de manière concomitante son investissement dans une autre. Par exemple, s'il accentue sa croissance, devenant ainsi meilleur compétiteur dans un environnement donné, il risque de le faire au détriment de sa reproduction (Obeso 2002). Les stratégies écologiques qui vont définir les espèces sont donc des compromis entre ces différents traits d'histoire de vie. Ces différentes stratégies ont été résumées selon un seul axe correspondant au modèle *r/K* (Macarthur & Wilson 1967), où *r* est la capacité à coloniser un nouvel environnement, et *K* est la capacité à tirer profit des conditions de croissance favorables. Ainsi, des individus *r*, adaptés à des milieux perturbés (c'est-à-dire à des milieux ouverts), seront typiquement dotés d'une forte aptitude à la dispersion les rendant capables de coloniser de nouveaux milieux. Les espèces pionnières, par rapport à des espèces post-pionnières, produisent souvent beaucoup de petites graines (Gaines *et al.* 1974), facilitant à la fois leur dispersion mais plus généralement la probabilité de coloniser des milieux favorables à leur croissance (milieux ouverts). Les individus *K* quant à eux sont des compétiteurs capables de s'installer dans un milieu déjà occupé par d'autres individus. Deux espèces proches capables de coexister en sympatrie correspondent donc à des compromis évolutifs différents et sont adaptées à des milieux différents entrainant un certain isolement spatial et/ou temporel, par exemple le long d'une succession écologique.

Suite à cette différentiation, une réduction partielle ou totale des croisements entre espèces apparait, soit parce que les stratégies mises en place sont si différentes qu'un individu hybride ne peut être qu'un compromis imparfait entre reproduction, croissance ou défenses, et est donc défavorisé, soit parce que les habitats diffèrent, soit parce que les gènes soumis à sélection divergente limitent directement les croisements interspécifiques (cas de la phénologie de la reproduction) ou sont liés à d'autres gènes limitant les croisements interspécifiques (Nosil 2012).

Le chêne (genre *Quercus*) est bien adapté à l'étude des mécanismes d'hybridation en conditions naturelles et constitue ainsi un bon modèle pour l'étude de la spéciation écologique. En effet, il existe dans le genre *Quercus* de nombreuses espèces (plus de 500 espèces, Nixon 1993) dont beaucoup peuvent s'hybrider (Cottam *et al.* 1982; Rushton 1993). Les chênes sessile et pédonculé (*Quercus robur* L. et *Q. petraea* (Matt.) Lielb.), appartenant au complexe d'espèces des chênes blancs, sont les plus répandus en Europe. Ces deux espèces ont des caractères morphologiques (figure 4) et une écologie très contrastés (Rameau *et al.* 1994; Kremer *et al.* 2002).

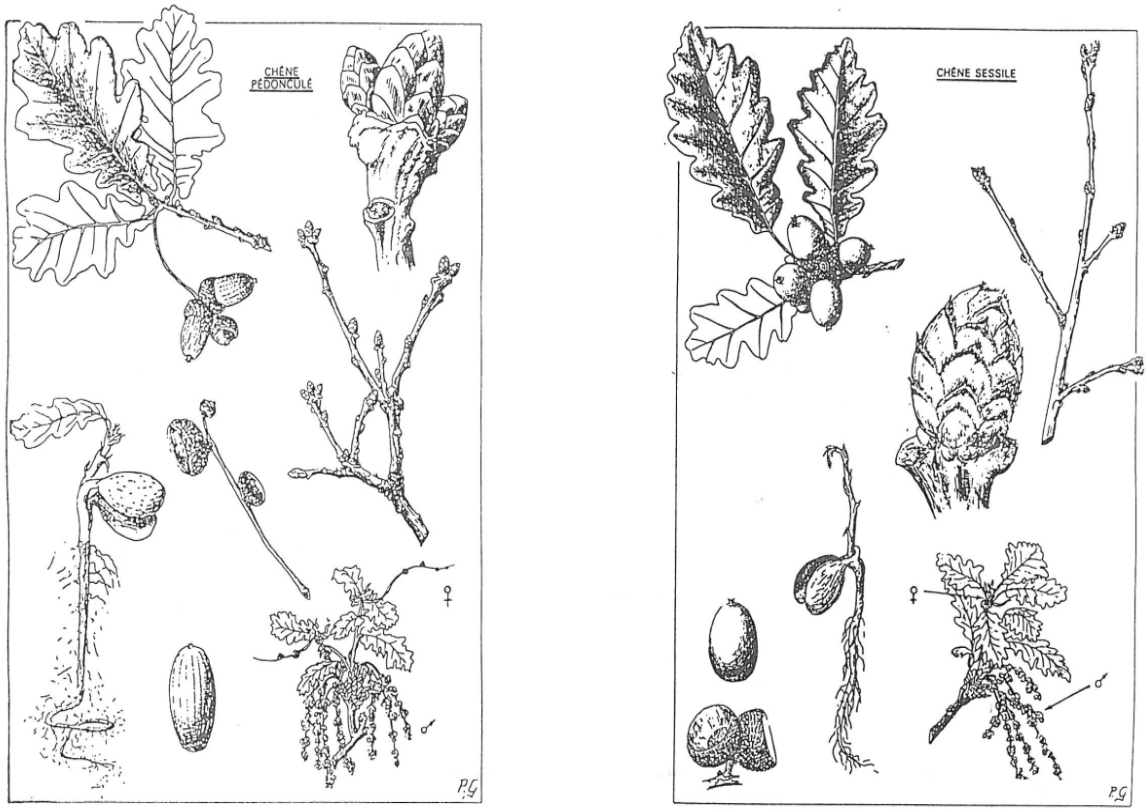


Figure 4: Morphologie de *Q. robur* (à droite) et *Q. petraea* (à gauche) (Jacamon 1979)

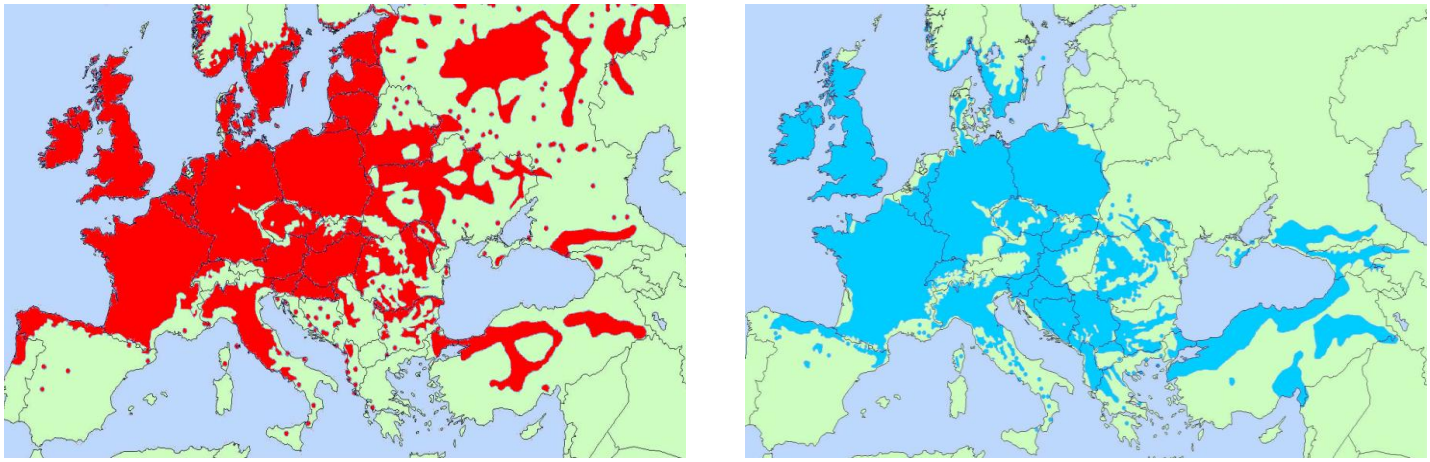


Figure 5: Aire de répartition de *Q. petraea* (en bleu) et de *Q. robur* (en rouge) en Europe (Ducousso & Bordacs 2004):

Elles coexistent dans une grande partie de leur aire de répartition (figure 5). Cependant, elles ont des préférences écologiques différentes qui limitent leur contact. En effet, *Quercus robur* est une espèce pionnière qui pousse dans des milieux plutôt ouverts et possède une grande capacité de dispersion des graines et du pollen. *Q. petraea* est au contraire une espèce post-pionnière qui possède une capacité de dispersion des graines et du pollen plus faibles (Rushton 1976; Pons & Pausas 2007; Jensen *et al.* 2009). De plus, *Q. petraea*, tolérant la sécheresse, se retrouve sur des sols plutôt secs alors que *Q. robur*, supportant l'ennoyage

racinaire mais craignant la sécheresse, se retrouve dans des zones plus humides (Parelle *et al.* 2006).

Chez la plupart des angiospermes, l'ADN chloroplastique a une hérédité maternelle. Son étude permet donc de reconstruire la dispersion des graines et donc les voies de recolonisation postglaciaires. L'étude de l'ADN chloroplastique de *Quercus robur* et *Q. petraea* ainsi que l'analyse de leur pollen fossile (même si celui-ci ne peut malheureusement permettre de distinguer à coup sûr les deux espèces) ont permis de retracer les grandes lignes de leur recolonisation de l'Europe après la dernière glaciation. En effet, ces deux espèces semblent être issues de trois refuges situés au sud de l'Europe (péninsules ibérique, italienne et balkanique ; Petit *et al.* 2002). Ces études ont également révélé que ces deux espèces possèdent localement le même haplotype chloroplastique, suggérant qu'elles sont issues d'une même lignée maternelle. De plus, l'hybridation de ces deux espèces est possible en croisement contrôlé (Steinhoff 1993; Lepais 2008; Abadie *et al.* 2011) et en conditions naturelles (Bacilieri *et al.* 1996; Jensen *et al.* 2009). Les croisements contrôlés interspécifiques ont mis en évidence une asymétrie de l'hybridation entre ces deux espèces : le pollen de *Q. petraea* féconde plus facilement une fleur de *Q. robur*, alors qu'un pollen de *Q. robur* fécondera plus difficilement une fleur de *Q. petraea* (Steinhoff 1993; Lepais 2008). Enfin, plusieurs types de marqueurs génétiques (isozymes (Gömöry *et al.* 2001), RAPD et SCAR (Bodénès *et al.* 1997), AFLP (Coart *et al.* 2002; Mariette 2002), microsatellites (Guichoux *et al.* 2011)) ont révélé une faible différentiation génétique entre ces deux espèces. Ces résultats ainsi que les différences écologiques connues de ces deux espèces ont conduit à l'élaboration d'un modèle d'introgression entre ces deux espèces (Petit *et al.* 2003). Selon ce modèle, *Q. robur* s'établit en premier du fait de son écologie et de ses plus grandes capacités à la dispersion (Figure 5 a). *Q. petraea* s'établit ensuite grâce notamment à son pollen, par hybridation avec *Q. robur* (sens d'hybridation préférentiel, Figure 5 b). Puis le cycle recommence avec la colonisation d'un autre milieu par *Q. robur* (Figure 5 c et d). Une récente étude basée sur de nouveaux marqueurs (SNP) a mis en évidence une forte différenciation entre ces espèces à certains locus avec la présence de marqueurs quasi-diagnostiques entre les espèces (Guichoux *et al.* 2012). Alors que *Q. petraea* possède des marqueurs privés, non partagés avec *Q. robur*, la plupart des variants trouvés chez *Q. robur* se retrouvent chez *Q. petraea*. Cette observation est cohérente avec le modèle de colonisation de ces deux espèces où le pollen de *Q. petraea* féconde une fleur femelle de *Q. robur*, puis par rétrocroisements successifs avec du pollen de *Q. petraea*, des individus principalement *Q. petraea* mais introgressés par *Q. robur* apparaissent. Le sens de la succession écologique explique que *Q. petraea* soit beaucoup plus introgressé et que les individus de cette espèce soient donc moins faciles à identifier. Ce modèle d'étude s'inscrit donc dans un cadre d'étude de la dynamique de la spéciation écologique le long d'une succession écologique.

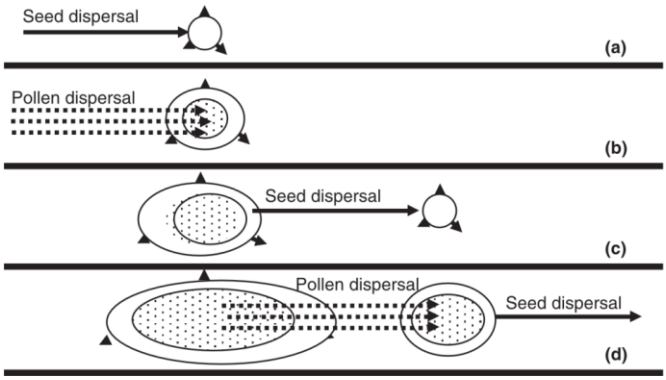


Figure 6: Modèle d'invasion de *Quercus robur* et *Q. petraea* (Petit *et al.* 2003). Le peuplement de *Q. robur* est symbolisé par des cercles vides, le peuplement de *Q. petraea* (ou mixte *Q. robur*/*Q. petraea*) par les cercles contenant des points.

Ce modèle de colonisation des milieux par hybridation récurrente de *Q. petraea* vers *Q. robur* (et donc d'introgression dans le sens opposé) explique bien le partage local quasi-systématique de l'ADN chloroplastique ainsi que la plus grande difficulté à identifier *Q. petraea* que *Q. robur*. Ces chênes constituent donc un modèle idéal pour traiter des problèmes de délimitation d'espèce et pour étudier précisément les mécanismes empêchant ou favorisant l'hybridation en conditions naturelles. Enfin, leur grande proximité génétique rend particulièrement informative la comparaison de leurs stratégies écologiques.

OBJECTIFS DE LA THESE

Les objectifs de cette thèse sont multiples. Dans un premier temps, j'ai cherché à comprendre l'origine des difficultés pour délimiter des espèces qui s'hybrident au travers de l'exemple *Quercus robur* et *Q. petraea.* Dans un deuxième temps, j'ai étudié, à l'échelle individuelle, les mécanismes permettant à ces deux espèces de chênes vivant en sympatrie et pouvant s'hybrider de se maintenir. En effet, une possibilité serait que ces deux espèces finissent par disparaitre complètement au profit d'hybrides lorsqu'elles sont en contact. Or aucune étude ne décrit des essaims d'hybrides (= « *hybrid swarms »*) entre ces deux espèces. Pour ce second objectif, j'ai étudié spécifiquement, à l'aide d'un modèle de voisinage, les mécanismes contrôlant les flux de gènes entre ces espèces. Dans un deuxième temps, à l'aide de ce même modèle de voisinage mais plus complet, j'ai étudié certaines des composantes des stratégies reproductives de ces deux espèces (essentiellement la fécondité mâle et la dispersion du pollen) permettant leur maintien en conditions naturelles.

CHAPITRE 1: **"Putting the biological species concept to the test: using mating networks to delimit species"**. Cette étude a permis d'appliquer *in situ,* pour la première fois, le critère d'interfertilité pour délimiter des espèces. L'analyse du réseau des croisements des arbres de deux espèces d'une même parcelle (événements de reproduction constatés par recherche de paternité) a permis de délimiter les espèces. Ensuite, cette méthode de délimitation a été comparée à des méthodes déjà publiées inspirées d'autres concepts. Pour les individus purs étudiés, les résultats sont largement congruents. Néanmoins, il apparait que la délimitation d'espèces basée sur le critère d'interfertilité dépend de l'environnement pollinique. La question qui se pose finalement est la suivante : Comment résumer une diversité biologique continue mais non uniforme ? L'approche « réseau » fournit ici des pistes de réflexion originales.

CHAPITRE 2: **"Fine-scale environmental control of hybridization in oaks"**. Dans ce chapitre, je me suis concentrée spécifiquement sur l'hybridation de ces deux espèces de chênes. J'ai étudié trois facteurs influençant cette hybridation au travers d'une modélisation des croisements à l'échelle de l'individu. Le premier aspect concerne l'effet de l'abondance relative des espèces sur leur hybridation. Le deuxième aspect concerne l'effet de la distribution spatiale plus ou moins regroupée des espèces sur les opportunités d'hybridation. Enfin, le troisième aspect de ce travail concerne l'effet de la disponibilité du pollen sur l'hybridation de ces deux espèces

CHAPITRE 3: **"Mating system differences between two closely-related oak species with contrasted ecological strategies"**. Dans cette étude, j'ai comparé les caractéristiques du système de reproduction des deux espèces : dispersion du pollen et fertilité mâle au travers de ses déterminants (circonférence du tronc, hauteur de l'arbre, environnement). Je me suis interrogée sur le lien entre les stratégies écologiques de ces deux espèces et leur comportement reproducteur. Par exemple, l'espèce plutôt forestière disperse-t-elle moins bien son pollen ? Sa fécondité mâle est elle dépendante de l'environnement ? Dans cet article, nous verrons que la stratégie de reproduction intraspécifique semble bien constituer une réponse adaptative aux différences écologiques auxquelles les espèces sont confrontées.

Pour finir, je propose une conclusion sous forme d'une synthèse des résultats importants de cette thèse suivie de perspectives.

En fin de document sont annexés les papiers issus de travaux auxquels j'ai collaboré pendant mon master ou ma thèse. **L'annexe 1 : « Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (Quercus spp.) »** est un article publié dans *"Molecular Ecology Resources"* qui décrit deux kits multiplex de 12 et 8 marqueurs microsatellites élaborés dans notre laboratoire. J'ai fini le développement du kit 12plex, kit que j'ai ensuite utilisé pour les recherches de paternités qui sont à la base de tous les chapitres de cette thèse. **L'annexe 2 : « Current trends in microsatellite genotyping »** est une revue publiée dans *"Molecular Ecology Resources"* qui se base sur notre expérience collective dans l'utilisation de marqueurs microsatellites et le développement de kits multiplex. Ce papier déjà très cité est considéré comme un front de recherche dans le *Web of Science*. **L'annexe 3 : « Outlier loci highlight the direction of introgression in oaks »** a été publié dans *"Molecular Ecology".* Cet article traite de l'introgression du génome de *Q. petraea* par des allèles issus *Q. robur,* en relation avec leur dynamique écologique. J'ai contribué pour moitié à la conception de la puce 384 SNP et j'ai génotypé les individus utilisés dans cette étude. La réalisation de cette puce a permis d'affiner les affectations génétiques des individus parentaux étudiés dans cette thèse, ce qui était un pré-requis indispensable à l'étude des croisements intra- et interspécifiques. **L'annexe 4 : « Genetic diversity increases insect herbivory on oak saplings »** est un article publié dans *"Plos One"* qui traite de l'effet de la diversité génétique intraspécifique sur les dégâts d'herbivorie sur de jeunes plants de *Q. robur*. J'ai réalisé de la partie génétique de cette étude (lecture génotypage, calculs d'apparentements).

REFERENCES

Abadie P., Roussel G., Dencausse B., Bonnet C., Bertocchi E., Louvet J.M., Kremer A. & Garnier-Géré P. (2011). Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *J. Evol. Biol.*, 25, 157-173.

Aguilée R., Lambert A. & Claessen D. (2011). Ecological speciation in dynamic landscapes. *J. Evol. Biol.*, 24, 2663-2677.

Bacilieri R., Ducousso A., Petit R.J. & Kremer A. (1996). Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution*, 50, 900-908.

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.

Burton O.J., Phillips B.L. & Travis J.M.J. (2010). Trade-offs and the evolution of life-histories during range expansion. *Ecol. Lett.*, 13, 1210-1220.

Chan K.M.A. & Levin S.A. (2005). Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution*, 59, 720-729.

Coart E.C., Lamote V.L., De Loose M.D.L., Van Bockstaele E.V.B., Lootens P.L. & Roldán-Ruiz I.R.-R. (2002). AFLP markers demonstrate local genetic differentiation between two indigenous oak species [<SMALL>Quercus robur L</SMALL>. and <SMALL>Quercus petraea (Matt.) Liebl</SMALL>.] in Flemish populations. *TAG Theoretical and Applied Genetics*, 105, 431-439.

Cody M.L. (1966). A General Theory of Clutch Size. *Evolution*, 20, 174-184.

Cottam W.P., Tucker J.M. & Santamour F.S. (1982). *Oak hybridization at the University of Utah*. State Arboretum of Utah, Salt Lake City, USA.

Coyne J.A. & Orr H.A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass., USA.

Dayrat B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, 85, 407-415.

de Meeûs T., Durand P. & Renaud F. (2003). Species concepts: what for? *Trends Parasitol.*, 19, 425-427.

de Queiroz K. (2007). Species concepts and species delimitation. *Syst. Biol.*, 56, 879-886.

Dobzhansky T. (1935). A critique of the species concept in biology. *Philosophy of Science*, 2, 344-355.

Dobzhansky T.G. (1937). Genetics and the origin of species. In. New York : Columbia University Press.

Donoghue M.J. (1985). A critique of the Biological Species Concept and recommendations for a phylogenetic alternative. *The Bryologist*, 88, 172-181.

Ducousso A. & Bordacs S. (2004). *EUFORGEN Technical Guidelines for genetic conservation and use for pedunculate and sessile oaks (Quercus robur and Q. petraea).* International Plant Genetic Ressources Institute, Rome, Italy.

Fitzpatrick B.M., Fordyce J.A. & Gavrilets S. (2009). Pattern, process and geographic modes of speciation. *J. Evol. Biol.*, 22, 2342-2347.

Focke W.O. (1881). *Die Pflanzenmischlinge.* Bornträger, Berlin.

Gaines M.S., Vogt K.J., Hamrick J.L. & Caldwell J. (1974). Reproductive Strategies and Growth Patterns in Sunflowers (Helianthus). *Am. Nat.*, 108, 889-894.

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.

Guichoux E., Garnier-Géré P., Lagache L., Lang T., Bourry C. & Petit R.J. (2012). Outlier loci highlight the direction of introgression in oaks. *Mol. Ecol.*, *in press (MEC-12-0795.R1)*.

Guichoux E., Lagache L., Wagner S., Léger P. & Petit R.J. (2011). Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Mol. Ecol. Resour.*, 11, 578-585.

Hausdorf B. (2011). Progress toward a general species concept. *Evolution*, 65, 923-931.

Hey J. (2001). The mind of the species problem. *Trends Ecol. Evol.*, 16, 326-329.

Hubbs C.L. (1955). Hybridization between fish species in nature. *Syst. Zool.*, 4, 1-20.

Jacamon M. (1979). *Guide de dendrologie*. E.N.G.R.E.F., Nancy.

Jensen J., Larsen A., Nielsen L.R. & Cottrell J. (2009). Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark. *Ann. Forest Sci.*, 66.

Kremer A., Dupouey J.L., Deans J.D., Cottrell J., Csaikl U., Finkeldey R., Espinel S., Jensen J., Kleinschmit J., Dam B.V., Ducousso A., Forrest I., Heredia U.L.d., Lowe A.J., Tutkova M., Munro R.C., Steinhoff S. & Badeau V. (2002). Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Ann. For. Sci.*, 59, 777-787.

Lepais O. (2008). Dynamique d'hybridation dans le complexe d'espèces des chênes blancs européens (Ph. D.). In: *Ecole doctorale Sciences et Environnement*. Université Bordeaux 1 Talence, France. .

Lepais O., Petit R.J., Guichoux E., Lavabre J.E., Alberto F., Kremer A. & Gerber S. (2009). Species relative abundance and direction of introgression in oaks. *Mol. Ecol.*, 18, 2228-2242.

Linnæi C.L.C. (1735). *Systema naturæ per regna tria naturæ: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* Leiden (Netherlands).

Lowry D.B., Modliszewski J.L., Wright K.M., Wu C.A. & Willis J.H. (2008). The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philos. T. Roy. Soc. B.*, 363, 3009-3021.

Macarthur R.H. & Wilson E.O. (1967). *The Theory of Island Biodiversity*. Princeton University Press.

Mallet J. (2008). Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philos. T. Roy. Soc. B.*, 363, 2971-2986.

Mallet J., Meyer A., Nosil P. & Feder J.L. (2009). Space, sympatry and speciation. *J. Evol. Biol.*, 22, 2332-2341.

Mariette S.C., J.; Csaikl, U.M.; Goikoechea, P.; König, A.; Löwe, A.J.; Van Dam, B.C.; Barreneche, T.; Bodenes, C.; Streiff, R.; Burg, K.; Groppe, K.; Munro, R.C.; Tabbener, H.; Kremer, A (2002). Comparison of levels of genetic diversity detected with AFLP and microsatellite markers within and among mixed *Q-petraea (MATT.) LIEBL.* and *Q-robur L.* stands. *Silvae Genet.*, 51, 72-79.

Mayr E. (1942). *Systematics and the origin of species*. Columbia Univ. Press, New York.

Muller C.H. (1952). Ecological control of hybridization in *Quercus*: A factor in the mechanism of evolution. *Evolution*, 6, 147-161.

Nixon K. (1993). Infrageneric classification of Quercus (Fagaceae) and typification of sectional names. *Ann. For. Sci.*, 50, 25s-34s.

Noor M.A.F. (2002). Is the Biological Species Concept showing its age? *Trends Ecol. Evol.*, 17, 153-154.

Nosil P. (2012). *Ecological speciation*. Oxford University Press, Oxford.

Nosil P., Harmon L.J. & Seehausen O. (2009). Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, 24, 145-156.

Obeso J.R. (2002). The costs of reproduction in plants. *New Phytol.*, 155, 321-348.

Orr H.A. (2001). Some doubts about (yet another) view of species. *J. Evol. Biol.*, 14, 870-871.

Parelle J., Brendel O., Bodénès C., Berveiller D., Dizengremel P., Jolivet Y. & Dreyer E. (2006). Differences in morphological and physiological responses to water-logging between two sympatric oak species (*Quercus petraea* [Matt.] Liebl., *Quercus robur* L.). *Ann. For. Sci.*, 63, 849-859.

Petit R.J., Bodénès C., Ducousso A., Roussel G. & Kremer A. (2003). Hybridization as a mechanism of invasion in oaks. *New Phytol.*, 161, 151-164.

Petit R.J., Csaikl U.M., Bordács S., Burg K., Coart E., Cottrell J., van Dam B., Deans J.D., Dumolin-Lapègue S., Fineschi S., Finkeldey R., Gillies A., Glaz I., Goicoechea P.G., Jensen J.S., König A.O., Lowe A.J., Madsen S.F., Mátyás G., Munro R.C., Olalde M., Pemonge M.-H., Popescu F., Slade D., Tabbener H., Taurchini D., de Vries S.G.M., Ziegenhagen B. & Kremer A. (2002). Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecol. Manag.*, 156, 5-26.

Pons J. & Pausas J. (2007). Acorn dispersal estimated by radio-tracking. *Oecologia*, 153, 903-911.

Rameau J., Mansion D. & Dumé G. (1994). Flore forestière française.

Rieseberg L.H. & Carney S.E. (1998). Plant hybridization. *New Phytol.*, 140, 599-624.

Ronce O. & Olivieri I. (1997). Evolution of Reproductive Effort in a Metapopulation with Local Extinctions and Ecological Succession. *Am. Nat.*, 150, 220-249.

Rundle H.D. & Nosil P. (2005). Ecological speciation. *Ecol. Lett.*, 8, 336-352.

Rushton B. (1993). Natural hybridization within the genus *Quercus* L. *Ann. Forest Sci.*, 50, 73s-90s.

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

Schlüter D. (1994). Experimental evidence that competition promotes divergence in adaptive radiation. *Science*, 266, 798-804.

Sites J.J.W. & Jonathon C.M. (2004). Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. S.*, 35, 199-227.

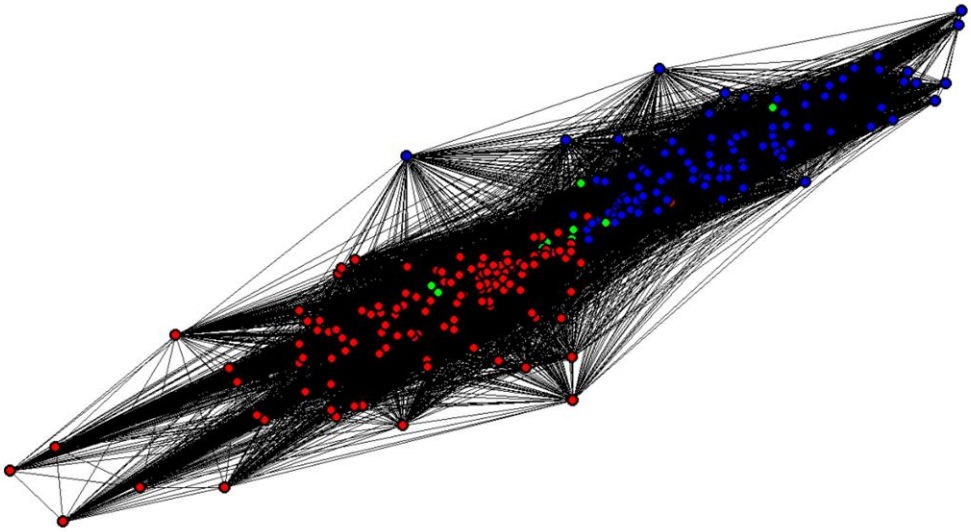
Steinhoff S. (1993). Results of species hybridization with *Quercus robur* L and *Quercus petraea* (Matt.) Liebl. *Ann. Forest Sci.*, 50, 137s-143s.

Vilà M., Weber E. & Antonio C.M.D. (2000). Conservation implications of invasion by plant hybridization. *Biological Invasions*, 2, 207-217.

Vörösmarty C.J., Green P., Salisbury J. & Lammers R.B. (2000). Global Water Resources: Vulnerability from Climate Change and Population Growth. *Science*, 289, 284-288.

Zirkle C. (1935). *The beginnings of plant hybridization*. University of Pennsylvania Press ; H. Milford : Oxford University Press, Philadelphia; London.

.



Putting the Biological Species Concept to the Test:

Using Mating Networks to Delimit Species

L. Lagache[1, 2], JB. Leger[3, 4], JJ. Daudin[3, 4], RJ. Petit[1, 2], C. Vacher[1, 2]

[1] INRA, UMR 1202 BioGeCo, F- 33610 Cestas, France

[2] Univ. Bordeaux, UMR1202 BioGeCo, F-33400 Talence, France

[3]INRA, UMR 518 MIA, F-75005 Paris, France

[4] AgroParistech, F-75005 Paris, France

(Article soumis)

INTRODUCTION

According to the biological species concept, the ability to interbreed (i.e. interfertility) is a defining property of species (1). Yet, to our knowledge, the interfertility criterion has never been used to delimit species on the basis of mating events observed under natural conditions. Only artificial crosses have been used for this purpose, including in fungi (e.g. 2), plants (3), or insects (4). However, this approach has been criticized (e.g. 5, 6) because artificial crosses bypass some pre-mating barriers to hybridization: mating events observed under artificial conditions might not reflect what would naturally occur. Hence, to date, there is no satisfactory example of the use of the interfertility criterion to delimit species. In fact, the methods used most frequently for species delimitation are not derived from the well-known biological species concept, but from other concepts such as the phylogenetic species concept, the genotypic species concept and the morphological species concept. Species definitions according to these concepts and possible associated criteria for species delimitation are listed in Table 1.

Species concept	**Species definition according to this concept**	**Possible criterion of species delimitation derived from this definition**	**Possible method of species delimitation using this criterion**	**First application of this method in the study site**
Biological species concept	Species are "*groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups*" (Mayr 1942). According to Hausdorf (2011), "natural populations" can be replaced by "individuals" in this formulation without change of meaning.	Higher natural interfertility between individuals within a species	Clustering of the network of natural mating events between individuals with C-SBM (Daudin *et al.* 2010).	this study
Phylogenetic species concept	A species is "*a diagnosable cluster of individuals within which there is a parental pattern of ancestry and descent, beyond which there is not, and which exhibits a pattern of phylogenetic ancestry and descent among units of like kind*" (Eldredge & Cracraft 1980).	Higher relatedness between individuals within a species	Clustering of the network of relatedness relationships between individuals with C-SBM (Daudin *et al.* 2010).	this study
Genotypic species concept	A species is a "*genotypic cluster [of individuals] that can overlap without fusing with its siblings*" (Mallet 1995; Hausdorf 2011)	Higher genotypic similarity between individuals within a species	Clustering of the individuals based on their multilocus genotype with STRUCTURE (Pritchard *et al.* 2000)	Guichoux et al. 2012
Morphological species concept	Species are "*the smallest detected samples of self-perpetuating organisms that have unique sets of characters*" (Nelson & Plantick 1981; Mishler 1985).	Higher morphological similarity between individuals within a species	Clustering of the individuals based on several morphological traits with a factorial discriminant analysis (Legendre & Legendre 1984).	Bacillieri et al. 1996

Table 1: Major species concepts with associated possible criterion for species delimitation

One potential method of species delimitation based on the interfertility criterion is the analysis of mating networks. Mating networks represent mating events between individuals (7). Nodes of the network represent the individuals and links connect the individuals between whom mating events have occurred. Applying methods of network clustering (8-10) to mating networks may allow the identification of subsets of strongly interconnected nodes that correspond to species. If the biological species concept is strictly interpreted, then a species should correspond to a connected component of the mating network (Fig. 1A). A connected component is a subset of nodes within the network that are directly or indirectly connected but are not connected to nodes not contained in the subset. According to a relaxed biological species concept, which allows for some level of hybridization between species (11-13), a species should correspond to a community in the mating network (Fig. 1B). Communities are subsets of nodes with a high density of links within the group and a lower density of links between different groups (8). It is in this latter case, when species hybridize, that species delimitation based on the interfertility criterion is particularly challenging and network analysis may be particularly useful.

The idea of analyzing mating networks to delimit species according to the biological species concept was proposed more than 40 years ago by Sokal and Crovello (14) but it does not appear to have been put into practice. Building a mating network is indeed a difficult task as it requires a very large data set of mating events collected under natural conditions. The species should be sympatric and have semi-permeable reproductive barriers so that the issue of species delimitation is relevant. The species should also be polygamous and have multiple offspring per generation so that actual mating events are representative of potential mating events between individuals at a given time (15-17). If such data were available, would the analysis of mating networks be an effective method to delimit species based on the interfertility criterion? Would the boundaries between species be the same as those obtained using other species delimitation criteria?

To answer these questions, we investigate the congruence between four methods of species delimitation, derived from the biological, morphological, genotypic and phylogenetic species concepts (Table 1), by applying them to two hybridizing tree species living in sympatry. The study site is a 5ha mixed stand of *Quercus robur* and *Q. petraea* comprising 298 adult trees originating from natural regeneration (18). As many other closely related plant species (19), these two oak species hybridize under natural conditions (20), including in the studied stand (21-23). To delimit species according to the interfertility criterion, we analyze the network of observed natural mating events between adult trees by using a method of network clustering. Each node of the mating network corresponds to an adult tree and each link corresponds to at least one mating event between two trees. To cluster individuals, we selected among available methods of network clustering (8-10) the Continuous Stochastic Block Model (C-SBM) recently introduced by Daudin *et al*. (24). As previous studies (23, 25) had shown that the forest stand is composed of two oak species, we expected to find two groups of interbreeding individuals. However, because the two species are also known to hybridize (22), we expected to find some individuals with a mixed reproductive behavior, that is, breeding with other individuals belonging to either species. Unlike many methods of network clustering, which assume that each node belongs to only one group, C-SBM allows model nodes to exhibit mixed connectivity behavior. This method is thus particularly suited to our study. The same method was used to delimit species based on genetic relatedness between individuals. In that case, each node of the network corresponds to an adult tree and links connect the individuals that are considered to be related based on their genotype. Finally, we compare individual assignments obtained by

analyzing the mating network and the relatedness network with those previously obtained in the same study site using criteria of morphological and genotypic similarities (23, 25). We then discuss how to summarize continuous but non-uniform variations in biological diversity.

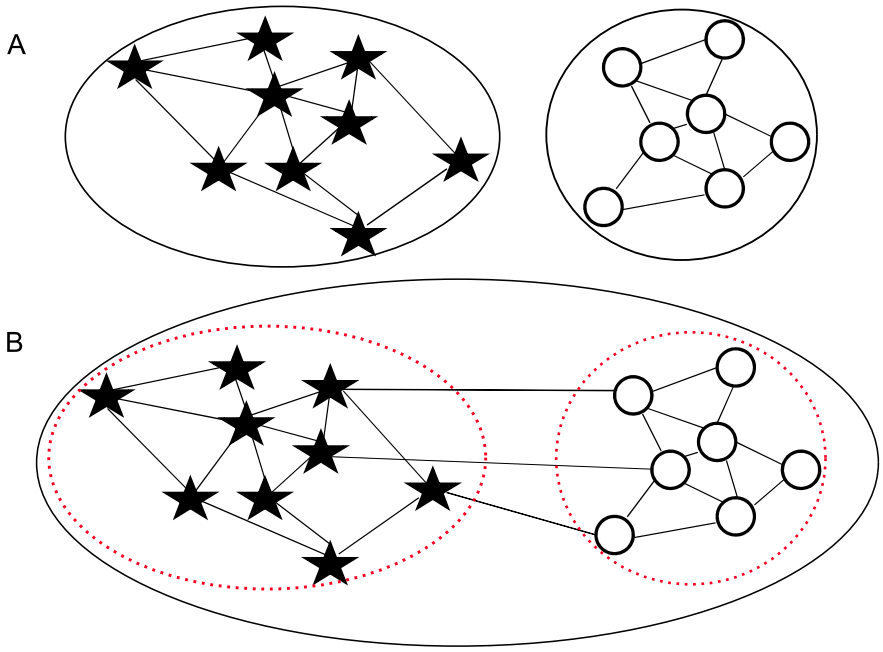


Figure 1: ***Example of mating networks with species boundaries.*** *Each node of the network, represented by a black star or a white circle, is an individual. Each link of the network, represented by a thin black line, corresponds to a mating event between two individuals. In A, there is no mating event between the two groups of individuals whereas in B, a few mating events occur between groups. Species boundaries according to a strict application of the biological species concept are indicated by a continuous thick black line. Species boundaries according to a relaxed interpretation of the biological species concept, allowing interspecific hybridization, are indicated by a broken red line. In network theory, the continuous black line delimits the connected components of the network whereas the broken red line delimits communities.*

RESULTS

Species Delimitation based on Interfertility

C-SBM (24) synthesizes the heterogeneity of a real network by producing a simplified version of the network composed of a few virtual nodes, called extremal hypothetical nodes (EHNs). According to the AIC criterion, the best model for the mating network was the one with four EHNs, followed by the models with five and three EHNs (Fig. S1 in SI Appendix). We selected the model with three EHNs because the two other models highlighted the structure of the sampling design (Text S2 in SI Appendix). According to the connectivity matrix for the EHNs (Fig. 2A), EHN0 corresponds to a virtual node not connected to the whole network. This EHN, which is systematically present in the network models produced by C-SBM (24), makes it possible to take into account the variation in the number of links attached to the nodes of the real network. The two other EHNs, called EHNB1 and EHNB2, were strongly connected within themselves and were not connected to the other EHNs.

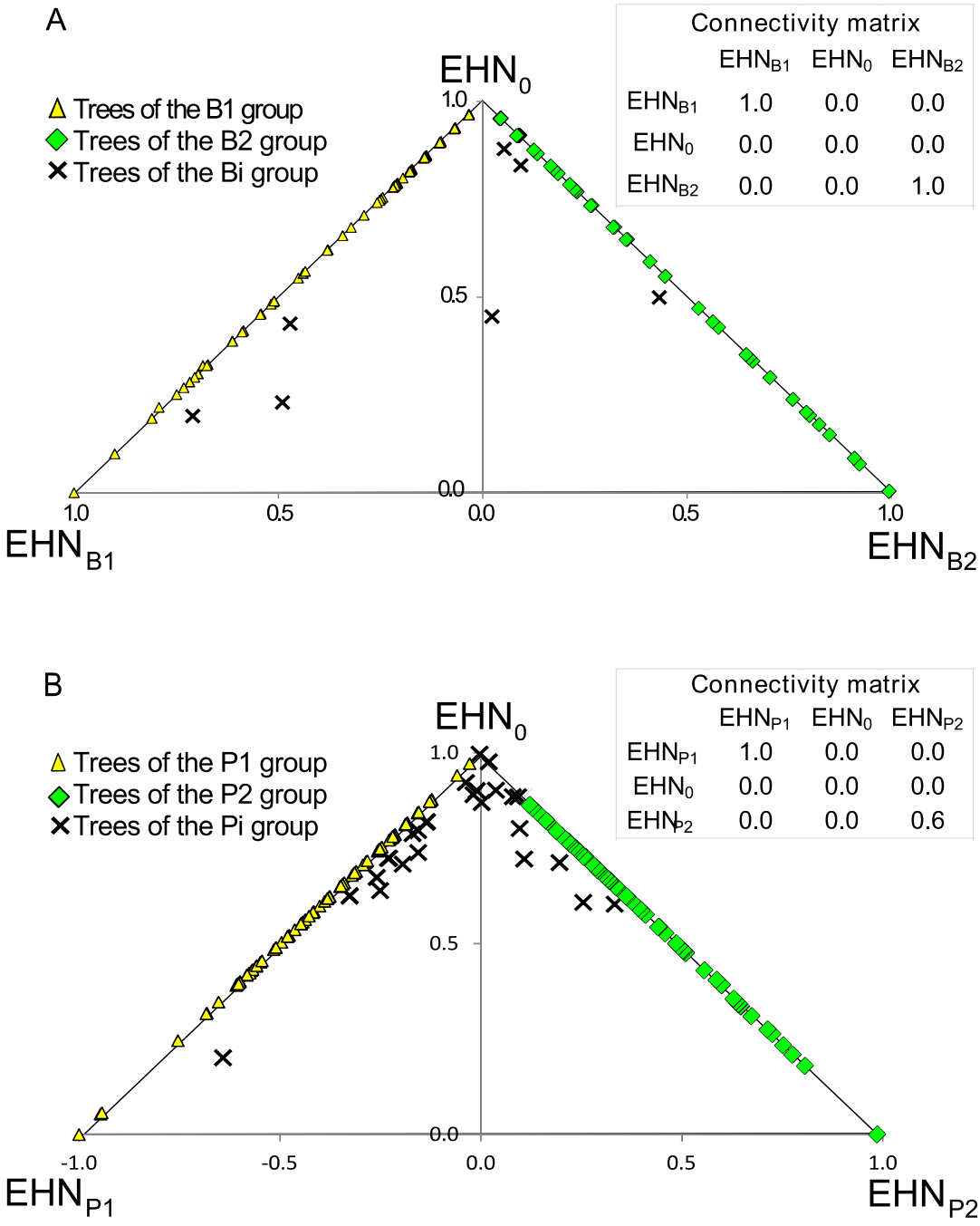


Figure 2: ***Triangular representation of the nodes of (A) the mating network and (B) the relatedness network,*** *indicating the mixture of EHNs for each node according to C-SBM. In A, nodes that are on the edge between EHN0 and EHNB1 are classified in group B1 whilst nodes on the edge between EHN0 and EHNB2 are classified in group B2. Other individuals are classified as intermediates (group Bi). In B, nodes that are on the edge between EHN0 and EHNP1 are classified in group P1 whilst nodes on the edge between EHN0 and EHNP2 are classified in group P2. Other individuals are classified as intermediates (group Pi). Connectivity matrixes for the EHNs are presented next to each triangular representation. Non-zero values are given in bold.*

C-SBM (24) assumes that each node of the real network is a mixture of the EHNs. The nodes of the mating network (each corresponding to an individual) were thus represented in a triangle, with one EHN at each point (Fig. 2A). The higher the proportion of a given EHN in the mixture of a node, the closer the node was to this EHN in the triangle. According to the connectivity matrix for the EHNs (Fig. 2A), the nodes that had a high proportion of EHN0 in their mixture were weakly connected to the mating network. The nodes that had a high proportion of EHNB1 in their mixture belonged to a group of nodes strongly connected to each other and weakly connected to nodes with a high proportion of EHNB2. Conversely, the nodes that had a high proportion of EHNB2 in their mixture belonged to a group of nodes strongly connected to each other and weakly connected to nodes with a high proportion of EHNB1. There were, therefore, two groups of adult trees in the mating network within which mating events were frequent and between which mating events were rare. The graphical representation of the network confirmed this result (Fig. 3A). According to the relaxed interpretation of the biological species concept, these two groups of individuals should correspond to two biological species (Fig. 1B).

In order to assign the individuals to the two species, we classified the nodes of the mating network according to their relative proportions of EHNB1 and EHNB2. We assumed that an individual belonged to species B1 if the corresponding node was a mixture of EHN0 and EHNB1, and only these two nodes. Conversely, we assumed that an individual belonged to species B2 if the corresponding node was a mixture of EHN0 and EHNB2. Other individuals were classified as being reproductively intermediate (group Bi). In the triangular representation (Fig. 2A), individuals assigned to species B1 were on the edge between EHN0 and EHNB1 (n=78 individuals) whilst individuals assigned to species B2 were on the edge between EHN0 and EHNB2 (n=121 individuals). Intermediate individuals were within the triangle (n=7 individuals). The three groups are shown in different colors in the network representation (Fig. 3A).

Species Delimitation based on Relatedness

According to the AIC criterion, the optimal number of EHNs in the relatedness network was six. Models with three, four, five and seven EHNs were also good models (Fig. S2 in SI Appendix). As we did not find any satisfactory way to identify the best model (Text S2 in SI Appendix), we selected the model with three EHNs to facilitate a comparison between the relatedness network structure and the mating network structure. According to the connectivity matrix for the EHNs (Fig. 2B), EHN0 corresponded to a virtual node not connected to the whole network. The two other EHNs, called EHNP1 and EHNP2, were strongly connected within themselves and were not connected to the other EHNs. Like the mating network, the individuals were, therefore, classified into three groups called P1, P2 and Pi. Group P1 (n=70 individuals located on the edge between EHN0 and EHNP1 in the triangular representation; Fig. 2B) and group P2 (n=108 individuals located on the edge between EHN0 and EHNP2; Fig. 2B) comprised individuals with high within-group and low between-group degrees of relatedness. The third group Pi (n=28 individuals located within the triangle; Fig. 2B) included trees related to both P1 and P2 individuals, and trees with few relatives. The three groups are shown in different colors in the network representation (Fig. 3B).

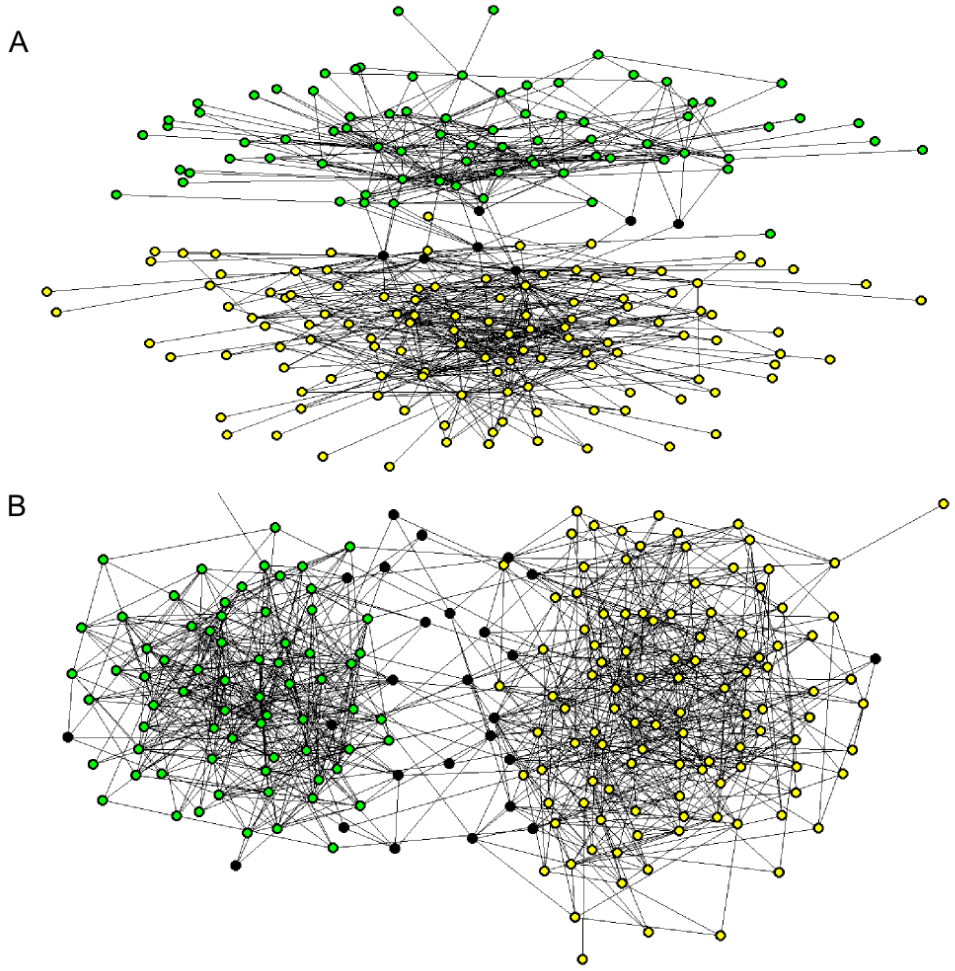


Figure 3: ***Graphical representation of (A) the mating network and (B) the relatedness network****, using the software PAJEK with the following parameters: Draw/Layout/Energy/Kamada-Kawai/Separate Components. Individuals classified into the B1 group (in A) or the P1 group (in B) are shown in green, individuals belonging to the B2 group (in A) or the P2 group (in B) are shown in yellow, and intermediate individuals are shown in black.*

Species Delimitation based on Morphology and Multilocus Genotypes

The morphological similarity criterion has previously been used by Bacilieri *et al.* (26) to identify all trees from the study site. Based on their results, we assigned the individuals to two pure morphological groups (called M1 and M2 in this study and corresponding to *Q. robur* and *Q. petraea*, respectively) and to a morphologically intermediate class (called Mi). Guichoux *et al.* (23) used genotypic similarity as a criterion to assign the trees of the study site to species. Based on their results, we classified the adult trees in two purebred groups (hereafter called G1 and G2) and one genetically intermediate class (Gi).

Congruence between the Four Methods of Species Delimitation

In order to assess the congruence between the four methods of species delimitation, we compared the spatial distribution of the three groups of individuals identified with each method. The results showed that the species boundaries were very similar (Fig. 4). Among the 206 adult trees included in the mating network and in the relatedness network, there were 97 trees classified consistently in the B1, P1, G1 and M1 groups and 63 trees classified consistently in the B2, P2, G2 and M2 groups. We therefore re-named groups B1, P1, G1 and M1 *Q. robur* and groups B2, P2, G2 and M2 *Q. petraea*. Based on this classification, there were only four species inversions associated with the delimitation methods (Table S1 in SI

Appendix). Among the 206 adult trees, 42 were classified as intermediates according to at least one method. Surprisingly, no individual was classified as intermediate according to all four methods. Therefore, 91% of the discrepancies between the four methods were caused by assignments to the intermediate class (Fig. S3 and Table S1 in SI Appendix).

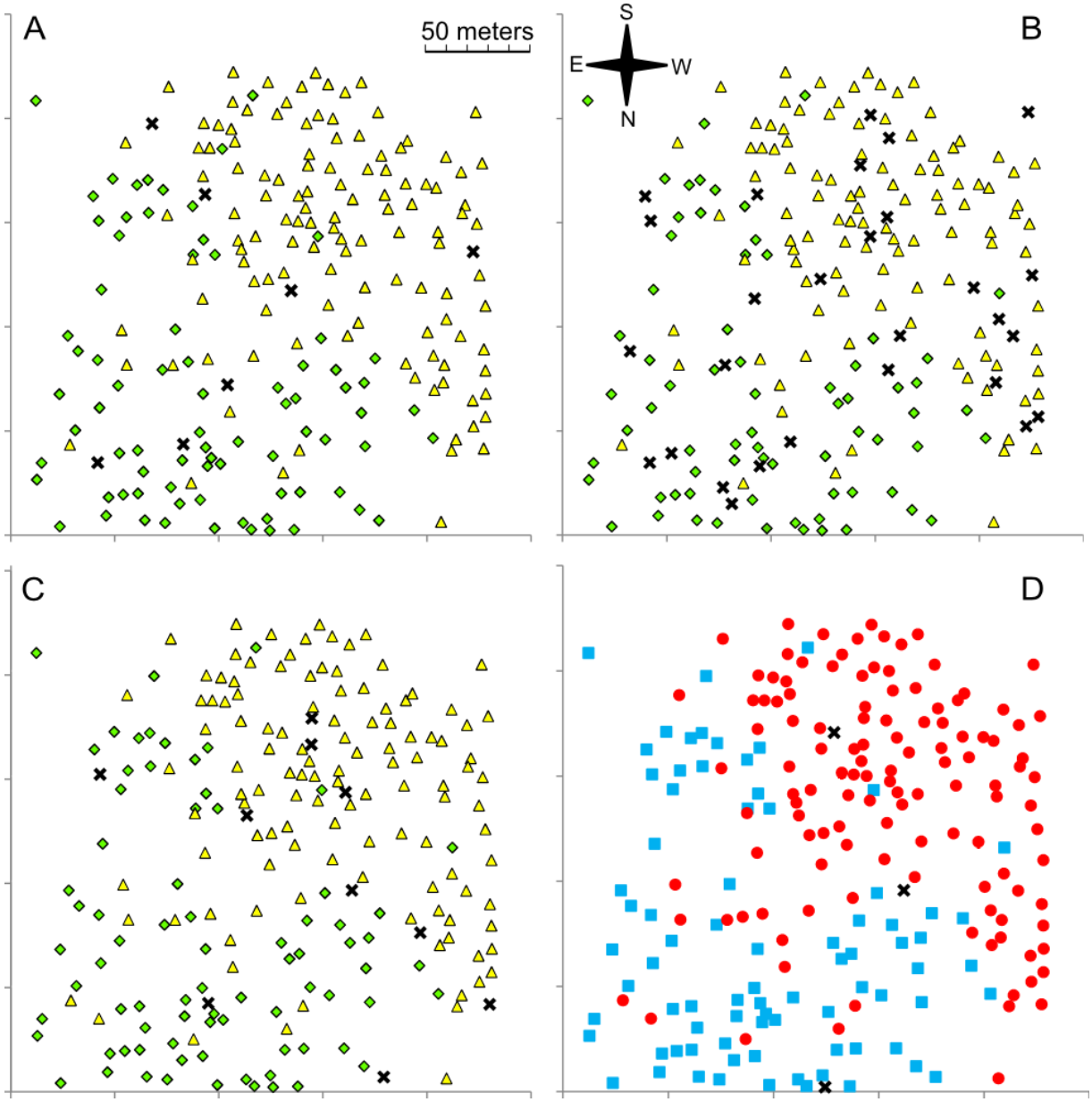


Figure 4: ***Species boundaries based on interfertility (A), relatedness (B), genotypic similarity (C) and morphological similarity (D) criteria****, represented on the map of the stand. In A, B and C, individuals classified into the B1, P1 or G1 species, respectively, are represented by yellow triangles. Individuals classified into the B2, P2 or G2 species are represented by green diamonds. Intermediate individuals are represented by black crosses. In D, individuals classified into M1 are shown in red, individuals classified into M2 in blue and morphologically intermediate individuals are indicated by black crosses. Individuals of the M1 group are assigned to Q. robur and individuals of the M2 group to Q. petraea on the basis of current taxonomical practices.*

There were nine discrepancies between the individual assignments according to the genotypic and morphological similarity criteria on the one hand and the interfertility criterion on the other hand. We investigated whether the biotic environment of the individuals might account for them. Our hypothesis is that the neighborhood of each tree influences its mating system and might thus influence its assignment to species based on the interfertility criterion, whereas it would hardly affect its assignment to species based on the genotypic and morphological criteria. We therefore examined the neighborhood of each tree for which the assignment to species based on genotypic and morphological similarity criteria were congruent (N= 192). For each tree, we calculated the proportion of allospecific neighbors within a radius of 69m (corresponding to the average distance of pollen dispersal within stand for *Q. petraea*, the species with the smallest dispersal ability; 22). We found, by performing a logistic regression, that the proportion of allospecific neighbors had a significant effect on the congruence between the individual assignments according to the genotypic and morphological similarity criteria on the one hand and the interfertility criterion on the other hand (χ^2=6.5, df=1, p-value=0.01). The individuals with congruent assignments had fewer allospecific neighbors on average (29%, versus 51% for individuals with incongruent assignments). Hence, individual species assignments based on the interfertility criterion were environment-dependent.

DISCUSSION

To our knowledge, this is the first time that the interfertility criterion is used successfully to delimit species under natural conditions. The analysis of a network of mating events between adult trees, constructed on the basis of a powerful paternity analysis of a large number of seedlings produced under natural conditions, allowed us to identify two groups of interfertile individuals, with only a few mating events between groups. The two groups that were delimited, corresponding to two species according to a relaxed interpretation of the biological species concept (Fig. 1B), were closely congruent with those obtained previously using morphological and genotypic similarity as criteria for species delimitation (23, 26). Indeed, 88% of the individuals were classified consistently according to the interfertility, morphological similarity and genotypic similarity criteria. Our results do not support earlier claims that the interfertility criterion cannot be applied in the field (e.g. 14, 15), particularly for the genus *Quercus* (27). They show instead that the analysis of mating networks can be used for delimiting species according to the biological species concept, as first suggested by Sokal and Crovello (14).

However this method of species delimitation has two main drawbacks. First, adequate network data are difficult to assemble. In our study we performed a paternity analysis on as many as 3046 offspring produced by 51 mothers in order to construct the mating network for adult trees. Despite the very large number of offspring, our network data did not allow us to assign all the individuals in the forest stand to species. Not all individuals sired offspring and some sired too few offspring to be reliably connected to the network. For example, three of the individuals whose assignment based on the interfertility criterion differed from that based on the three other criteria were represented by a single offspring in the progeny test. They were thus connected to the mating network through just a single link. Second, the sampling design may generate some heterogeneity in the network structure that blurs the biological heterogeneity caused by the existence of different species. This happened in our network data because we harvested the offspring of only 20% of the trees in the stand. The harvested trees (i.e. mother-trees), therefore, had more links in the mating network than the other trees. To solve both problems, one would have to harvest seeds

from all the individuals in the stand, assuming that all of them produced seeds. In principle, this goal could be achieved with our biological system by extending sampling over multiple years, because oak species are perennial and monoecious. However this would be impossible for annual or dioecious species. Another possibility to reduce the noise caused by sampling would be to introduce the sampling structure as a covariate in the statistical model (e.g 28). Unfortunately, the Continuous Stochastic Block Model (24), which was selected for this study because it allows modeling continuous variations in the connectivity properties of the nodes, does not currently allow the incorporation of covariates.

Our results further show that the analysis of the network of contemporary relatedness relationships is a relevant method for delimiting species. The two groups found in our study might be interpreted as corresponding to two different 'phylogenetic species' (29), if phylogenetic relationships are considered in a broad sense so as to include contemporary pedigree relationships. Methods of species delimitation derived from the phylogenetic species concept have almost exclusively focused on deep ancestry using tree-based phylogenetic methods (reviewed in 30, but see 31). These methods are not well-suited for delimiting hybridizing species because horizontal gene transfers between species, caused by hybridization and subsequent backcrossing events, produce conflicts between gene trees and species trees (32-33). Compared to data on mating events, data on relatedness were easier to acquire and there was no sampling issue. The analysis of the relatedness network revealed two groups of individuals with high within-group and low between-group degrees of relatedness. These two groups were highly congruent with those obtained using interfertility, morphological similarity and genotypic similarity as criteria, indicating that the analysis of relatedness networks may have potential for species delimitation. However, this method also has some drawbacks: the best model had five groups of related individuals and we did not find any hypothesis accounting for their origin; the number of species should thus be known in advance in order to apply this method.

By comparing the results obtained with the four criteria used for species delimitation (i.e. interfertility, relatedness relationships, morphological or genotypic similarities), we showed that the species boundaries were largely congruent across methods of species delimitation. Our analyses confirmed the existence of two groups of individuals that were both morphologically and genetically differentiated. We also showed that the individuals of each group preferentially mated and were more related with each other than with individuals from the other group. Therefore, there were two 'evolutionary lineages' in the studied stand. The Lineage Species Concept introduced by Simpson (34-35), then taken up by Wiley (36) and de Queiroz (16, 37-38), focuses on the question of congruence among methods of species delimitation. For these authors, modern species concepts (e.g. morphological, phylogenetic, genotypic and biological) assimilate, explicitly or implicitly, species *'to separately evolving (segments of) metapopulation lineages'* and are thus all by-products of the lineage species concept (16-17). This should account for the high degree of congruence among species delimitation methods.

Another important result of this comparison is that, irrespective of the criterion used for delimiting species, we found intermediate individuals that had features of both species. Interestingly, the individuals classified as intermediates often differed across methods. In particular, no individual was consistently classified as intermediate according to all four methods. These discrepancies might be explained by the thresholds that were chosen empirically to delimit purebred species and by data quality problems. As mentioned above, examining more offspring per parent tree may improve species delimitation based on the interfertility criterion. Similarly, a greater number of molecular markers (39-41) may improve

methods of species delimitation based on the genotypic and relatedness criteria. Likewise, a larger number of morphological markers (26) may improve morphological species delimitation. However, we believe that these discrepancies may also reflect a biological reality. Indeed, as shown in other studies (42-44), including in oaks (22, 45), species relative abundance affects hybridization rate. An individual tends to reproduce with its neighbors. If it is surrounded by numerous allospecifics and few conspecifics (e.g. 22, 44), this can result in much hybridization. Such an individual will tend to be assigned to another species or to a reproductively intermediate class, according to methods based on interfertility. Therefore, we expect some discrepancies in species assignments between methods based on environment-dependent criteria (such as that based on the interfertility criterion) and methods based on environment-independent criteria (such as that based on the genotypic similarity criterion).

CONCLUSION

Our results confirmed the existence of two differentiated groups of individuals at the study site, corresponding to two species: *Quercus robur* and *Q. petraea*. However, depending on the criterion used for assigning individuals to species (i.e. interfertility, relatedness, morphological or genotypic similarities), the boundary between species was not exactly the same. Most of the differences stem from assignment of individuals to an intermediate category. This finding illustrates the continuous nature of variation between species. The model we used, which belongs to a category called 'grade of membership models' (reviewed in 10) is appropriate for synthesizing continuous (but not uniform) variations in biological diversity. However, to get closer to the species concepts, which generally define species as groups of individuals, we finally classified the individuals into non-overlapping groups. Our approach, therefore, illustrates the influence of concepts on our (mis)representation of species and on our understanding of biological diversity. Frost and Hillis (46), as well as Mayr (47), proposed defining species as 'a whole' instead of as a group of individuals. According to our study, species could also be defined as an 'extreme point' to which individuals are more or less close, thus allowing the possibility of an individual being a mixture of two different species.

MATERIAL & METHODS

SPECIES DELIMITATION BASED ON INTERFERTILLITY

To construct the mating network for the adult trees, we made use of a progeny test involving 3046 offspring resulting from open pollination, harvested from 51 mother-trees distributed across the entire stand (Fig. S4 in SI Appendix). A paternity analysis was conducted (22) by genotyping all the offspring from the test and all the adults trees for which DNA was available, using 12 multiplexed microsatellite (SSR) markers developed by Guichoux *et al.* (48). According to the paternity analysis, 1575 offspring had only one possible father in the stand, 54 offspring had several potential fathers in the stand and 1417 offspring had no father in the stand (22). Based on the offspring for which only a single father was found, we identified 198 father-trees in the stand. These trees included 43 trees that were also mothers, because oak trees are monoecious. Based on these results, we reconstructed 1629 mating events between 206 adult trees within the stand. These mating events allowed us to identify 751 couples of trees that mated at least once, indicating that they were interfertile under natural conditions. These data were represented by an

undirected and unweighted network in which each of the 206 nodes corresponded to an adult tree and each of the 751 links corresponded to at least one mating event between two trees.

Then, the network was modeled with C-SBM (24). The parameters of the model are the connectivity coefficients between the EHNs and the coefficients of the mixture of EHNs for each node of the real network. For each possible number of EHNs, the parameters of the model were inferred by the maximum likelihood method, derived using the MATLAB program C-Mixnet (available at http://www.agroparistech.fr/mia/doku.php?id=productions:logiciels/). Then, the optimal number of EHNs in the network was determined by using the AIC criterion (24). The results were visualized with the software PAJEK (49).

Species Delimitation based on Relatedness

In order to build the relatedness network, we estimated the relatedness of the 206 adult trees included in the mating network. The estimation was performed with the software COANCESTRY (50), which offers seven different estimators of relatedness. As recommended by Wang (50), we used the 1629 offspring for which both parents were known to determine the most suitable estimator. The triadic likelihood estimator (51, denoted TrioML in COANCESTRY) was selected because it produced relatedness values closest to zero for unrelated offspring, closest to 0.25 for half-sibs and closest to 0.5 for full-sibs. With this estimator, the highest relatedness value between two unrelated offspring was 0.22. We therefore treated 0.22 as a threshold: trees whose relatedness value was higher than this were considered to be related individuals and the other trees were considered to be unrelated. The relatedness relationships were then represented by an unweighted and undirected network with 206 nodes, each corresponding to an adult tree, and 1078 links connecting the individuals considered to be related. As in the case of the mating network, we modeled the network structure using C-SBM (24) and we visualized the results with the software PAJEK (49).

Species Delimitation based on Morphology

The morphological similarity criterion has previously been used by Bacilieri *et al.* (26) to identify all trees from the study site. These authors performed a factorial discriminant analysis (FDA) based on 31 leaf morphological traits to delimit the species. Their study revealed the presence of two groups of individuals differing in their morphology. The first axis of the FDA accounted for 33% of the total variance and was highly correlated to the morphological markers traditionally used by taxonomists to distinguish *Q. robur* from *Q. petraea*. The distribution of the individuals along this axis was used to assign, graphically, the individuals to two pure morphological groups (called M1 and M2 in this study and corresponding to *Q. robur* and *Q. petraea* respectively) and to a morphologically intermediate class (called Mi). Among the 206 adult trees included in the mating and relatedness networks, 123 trees were assigned to M1, 80 to M2 and 3 to Mi (Fig. S5 in SI Appendix).

SPECIES DELIMITATION BASED ON MULTILOCUS GENOTYPES

Guichoux *et al.* (23) used genotypic similarity as a criterion to assign the trees of the study site to species. These authors genotyped the adult trees with the multiplex of 12 SSRs developed by Guichoux *et al.* (48) and with a chip of 262 single-nucleotide polymorphisms (SNP) enriched with markers highly differentiated between species (23). They used the software STRUCTURE (52) to group the individuals into genotypic clusters but did not formally determine the optimal number of genotypic clusters in the stand before performing the clustering. Here we used the ΔK statistic (53) to identify the number of genetically different groups. The optimal number of clusters was two (Fig. S6 in SI Appendix), as previously assumed by Guichoux *et al.* (23). The adult trees were therefore classified in two purebred groups and one genetically intermediate class. Among the 206 adult trees included in the mating and relatedness networks, 78 trees were assigned to the first purebred group (hereafter called G1), 118 to the second purebred group (G2) and 10 to the genetically intermediate class (Gi) (Fig. S7 in SI Appendix).

ACKNOWLEDGEMENTS

We are grateful to Alexis Ducousso who established the Petite Charnie progeny test and shared information about the stand, and for his help, together with that of Stefanie Wagner, during sampling. Patrick Léger greatly helped with microsatellite genotyping. The genotyping was performed at the Genome-Transcriptome facility of Bordeaux (grants from the Conseil Régional d'Aquitaine n°20030304002FA, n°20040305003FA and from the European Union, FEDER n°2003227). We are particularly grateful to François Hubert for helpful discussions about phylogeny and to Bastien Castagneyrol, Virgil Fievet, Cyril Dutech and Antoine Kremer for their constructive comments on the first version of the manuscript. Funding was provided by the LinkTree project (ANR BIODIVERSA) and by the EU Network of Excellence EvolTree.

AUTHOR'S CONTRIBUTIONS:

RJP initially conceived the study, which evolved significantly with the help of all the authors. LL performed the experiments, produced and analyzed the data with the help of CV. JBL and JJD performed the network modelling. LL wrote the paper with the help of CV and all four authors reviewed the complete manuscript.

REFERENCES

1. Mayr E (1942) *Systematics and the origin of species* (Columbia Univ. Press, New York).
2. Dettman JR, Jacobson DJ, Turner E, Pringle A, & Taylor JW (2003) Reproductive isolation and phylogenetic divergence in *Neurospora*: comparing methods of species recognition in a model eukaryote. *Evolution* 57(12):2721-2741.
3. Marcussen T & Borgen L (2011) Species delimitation in the Ponto-Caucasian *Viola sieheana* complex, based on evidence from allozymes, morphology, ploidy levels, and crossing experiments. *Plant Syst. Evol.* 291(3):183-196.
4. Marin J, Crouau-Roy B, Hemptinne J-L, Lecompte E, & Magro A (2010) *Coccinella septempunctata* (Coleoptera, Coccinellidae): a species complex? *Zool. Scr.* 39(6):591-602.
5. Hibbett DS, Fukumasa-Nakai Y, Tsuneda A, & Donoghue MJ (1995) Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA sequences. *Mycologia* 87(5):618-638.
6. Taylor JW*, et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* 31(1):21-32.
7. Fortuna MA, García C, Guimarães Jr PR, & Bascompte J (2008) Spatial mating networks in insect-pollinated plants. *Ecol. Lett.* 11(5):490-498.
8. Newman M (2003) The structure and function of complex networks. *SIAM Review* 45(2):167-256.
9. Fortunato S (2010) Community detection in graphs. *Phys. Rep.* 486(3–5):75-174.
10. Léger J-B, Vacher C, & Daudin J-J Detection of structurally homogeneous subsets in graphs. *Stat. Comput., in revision.*
11. Orr HA (2001) Some doubts about (yet another) view of species. *J. Evol. Biol.* 14(6):870-871.
12. Noor MAF (2002) Is the Biological Species Concept showing its age? *Trends Ecol. Evol.* 17(4):153-154.
13. Coyne JA & Orr HA (2004) *Speciation* (Sinauer Associates, Sunderland, Mass., USA).
14. Sokal RR & Crovello TJ (1970) The Biological Species Concept: a critical evaluation. *Am. Nat.* 104(936):127-153.
15. de Meeûs T, Durand P, & Renaud F (2003) Species concepts: what for? *Trends Parasitol.* 19(10):425-427.
16. de Queiroz (2005) Ernst Mayr and the modern concept of species. *P. Natl. Acad. Sci. USA* 102(Suppl 1):6600-6607.
17. Hausdorf B (2011) Progress toward a general species concept. *Evolution* 65(4):923-931.
18. Streiff R, Ducousso A, & Kremer A (1998) Spatial genetic structure and pollen gene flow in a mixed oak stand. (Translated from French) *Genet. Sel. Evol.* 30:S137-S152 (in French).
19. Rieseberg LH & Carney SE (1998) Plant hybridization. *New Phytol.* 140(4):599-624.
20. Petit RJ, Bodénès C, Ducousso A, Roussel G, & Kremer A (2003) Hybridization as a mechanism of invasion in oaks. *New Phytol.* 161(1):151-164.
21. Lepais O & Gerber S (2011) Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution* 65(1):156-170.
22. Lagache L, Klein EK, Guichoux E, & Petit RJ (2012) Fine-scale environmental control of hybridization in oaks. *Mol. Ecol. in press* (doi: 10.1111/mec.12121)
23. Guichoux E*, et al.* (2012) Outlier loci highlight the direction of introgression in oaks. *Mol. Ecol. in press (*doi: 10.1111/mec.12125).
24. Daudin J-J, Pierre L, & Vacher C (2010) Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network. *Biometrics* 66(4):1043-1051.

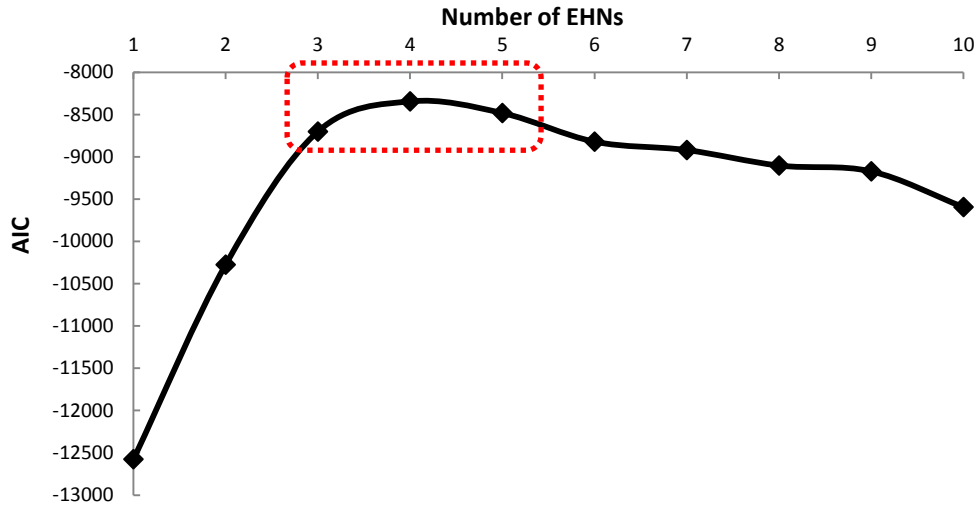
25. Bacilieri R, Ducousso A, Petit RJ, & Kremer A (1996) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution* 50(2):900-908.
26. Bacilieri R, Ducousso A, & Kremer A (1996) Comparison of morphological characters and molecular markers for the analysis of hybridization in sessile and pedunculate oak. *Ann. Sci. Forest* 53(1):79-91.
27. Donoghue MJ (1985) A critique of the Biological Species Concept and recommendations for a phylogenetic alternative. *The Bryologist* 88(3):172-181.
28. Mariadassou M, Robin S, & Vacher C (2010) Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.* 4(2):715-742.
29. Eldredge N & Cracraft J (1980) *Phylogenetic patterns and the evolutionary process* (Columbia Univ. Press, New York).
30. Sites JJW & Jonathon CM (2004) Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol.* 199-227.
31. Moalic Y, Arnaud-Haond S, Perrin C, Pearson G, & Serrao E (2011) Travelling in time with networks: revealing present day hybridization versus ancestral polymorphism between two species of brown algae, *Fucus vesiculosus* and *F. spiralis*. *BMC Evol. Biol* 11(1):33.
32. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution* 63(1):1-19.
33. Maddison WP (1997) Gene trees in species trees. *Syst. Biol.* 46(3):523-536.
34. Simpson GG (1951) The species concept. *Evolution* 5(4):285-298.
35. Simpson GG (1961) *Principles of animal taxonomy* (Columbia Univ. Press, New York).
36. Wiley EO (1978) The Evolutionary Species Concept reconsidered. *Syst. Biol.* 27(1):17-26.
37. de Queiroz K (2007) Species concepts and species delimitation. *Syst. Biol.* 56(6):879-886.
38. de Queiroz K (1998) *The General Lineage Concept of species, species criteria, and the process of speciation* (Oxford University Press).
39. 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(1):63-72.
40. Wang JL (2002) An estimator for pairwise relatedness using molecular markers. *Genetics* 160(3):1203-1215.
41. Norris AT, Bradley DG, & Cunningham EP (2000) Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. *Aquaculture* 182(1–2):73-83.
42. 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*. *J. Ecol.* 96(6):1198-1210.
43. Focke WO (1881) *Die Pflanzenmischlinge.* (Bornträger, Berlin).
44. Hubbs CL (1955) Hybridization between fish species in nature. *Syst. Zool.* 4(4):1-20.
45. Lepais O*, et al.* (2009) Species relative abundance and direction of introgression in oaks. *Mol. Ecol.* 18(10):2228-2242.
46. Frost DR & Hillis DM (1990) Species in concept and practice: herpetological applications. *Herpetologica* 46(1):86-104.
47. Mayr E (1992) A local flora and the Biological Species Concept. *Am. J. Bot.* 79(2):222-238.
48. Guichoux E, Lagache L, Wagner S, Léger P, & Petit RJ (2011) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Mol. Ecol. Resour.* 11(3):578-585.
49. Batagelj V & Mrvar A (1998) PAJEK - Program for large network analysis.
50. Wang JL (2010) Coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* 11(1):141-145.

51. Wang (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetical research* 89(3):135-153.
52. Pritchard JK, Stephens M, & Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945-959.
53. Evanno G, Regnaut S, & Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14(8):2611-2620.
54. Mallet J (1995) A species definition for the modern synthesis. *Trends Ecol. Evol.* 10(7):294-299.
55. Nelson G & Plantick N (1981) *Systematics and biogeography: cladistics and vicariance* (Columbia University Press, New York).
56. Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. *The Bryologist* 88(3):207-214.
57. Legendre G & Legendre P (1984) *Ecologie numérique* (Masson, Paris, France).

SUPPLEMENTARY INFORMATION

FIGURE S1: *OPTIMAL NUMBER OF EHNS IN THE MATING NETWORK ACCORDING TO THE AIC CRITERION*

The highest AIC values, circled with a red dotted line, correspond to the best models. The optimal number of EHNs is 4 (AIC=-8345.5), followed by 5 (AIC=-8443) and then 3 (AIC=-8702.3).



TEXT S1: *EFFECT OF THE SAMPLING DESIGN ON THE HETEROGENEITY OF THE MATING NETWORK*

There are two key elements for interpreting the heterogeneity of a real network modeled with C-SBM:

- The connectivity matrix between the EHNs
- The mixture of EHNs for each node of the network

Each element a_{ql} of the connectivity matrix A between the EHNs corresponds to the probability that there is a link between EHN_q and EHN_l. There is always one EHN, called EHN_0, which is not connected to itself and not connected to the other EHNs. In the mating network, the other EHNs, called $EHN_{Bj(1\leq j\leq k-1)}$, are strongly connected with themselves and not connected with the other EHNs.

k=3

	EHN_{B1}	EHN_0	EHN_{B2}
EHN_{B1}	**1.0**	0.0	0.0
EHN_0	0.0	0.0	0.0
EHN_{B2}	0.0	0.0	**1.0**

k=4

	EHN_{B1}	EHN_0	EHN_{B2}	EHN_{B3}
EHN_{B1}	**1.0**	0.0	0.0	0.0
EHN_0	0.0	0.0	0.0	0.0
EHN_{B2}	0.0	0.0	**1.0**	0.0
EHN_{B3}	0.0	0.0	0.0	**1.0**

k =5

	EHN_{B1}	EHN_0	EHN_{B2}	EHN_{B3}	EHN_{B4}
EHN_{B1}	**1.0**	0.0	0.0	0.0	0.0
EHN_0	0.0	0.0	0.0	0.0	0.0
EHN_{B2}	0.0	0.0	**1.0**	0.0	0.0
EHN_{B3}	0.0	0.0	0.0	**1.0**	0.0
EHN_{B4}	0.0	0.0	0.0	0.0	**1.0**

Connectivity matrices between the EHNs in the mating network, as a function of the number k of EHNs. The connectivity properties of EHN_0 are highlighted in grey. Non-zero values are in bold.

The more a node has a high proportion of a given EHN in the mixture, the more its connectivity properties resemble to those of this EHN. Therefore, the nodes with a high proportion of EHN_0 in the mixture are lowly connected to the network (i.e. they have a low degree). The nodes with a high proportion of one of the $EHN_{Bj(1\leq j\leq k-1)}$ in the mixture belong to a group of nodes strongly connected between them, and lowly connected with nodes of other groups (i.e. groups of interbreeding individuals, reproductively isolated from other groups). In order to delimit species based on the interfertility criterion, we therefore grouped together the nodes according to the proportion of each $EHN_{Bj(1\leq j\leq k-1)}$ in the mixture. We assumed that a node belongs to group B_j if it is a mixture between EHN_0 and EHN_{Bj}, and only between these two EHNs. Other individuals were classified as intermediate. We then analyzed the composition of the groups as a function of the sampling design.

k=3		
	Unsampled	Sampled
B_1	56	22
B_2	96	25

k=4		
	Unsampled	Sampled
B_1	97	**0**
B_2	**0**	8
B_3	54	22

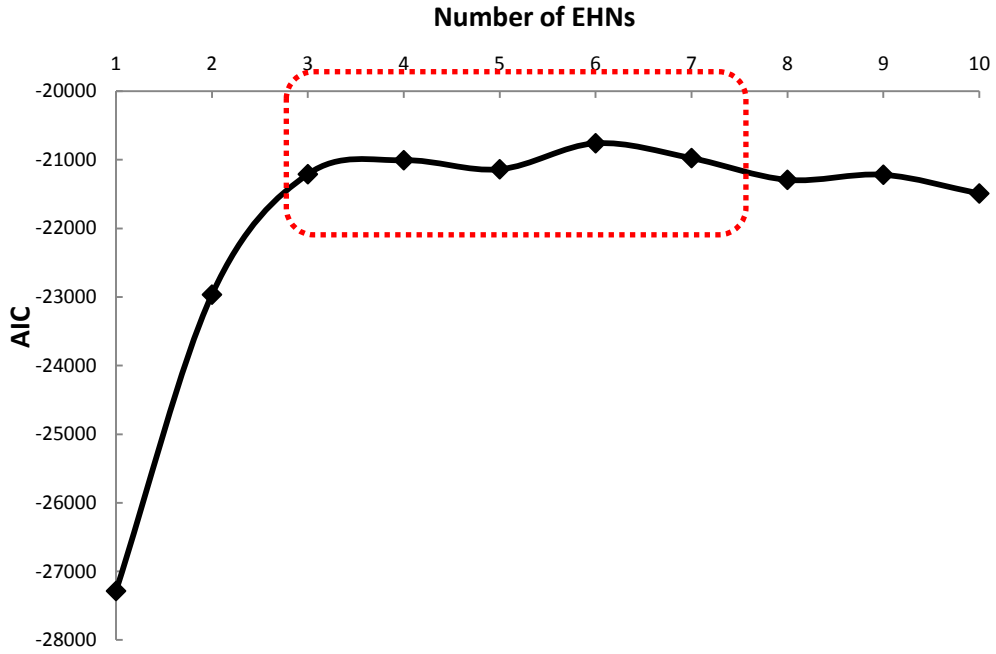
k=5		
	Unsampled	Sampled
B_1	18	4
B_2	**0**	3
B_3	52	**0**
B_4	48	21

Number of sampled and unsampled trees in each group as a function of k. The groups having only one type of tree are highlighted in grey.

In the models with 4 and 5 EHNs, one group was composed of samples trees only and another group was composed of unsampled trees only, indicating that the heterogeneous structure of the mating network is partly accounted for by the sampling design. Describing the heterogeneity due to sampling was not the purpose of our study so we selected the model with 3 EHNs, in which groups were independent from the sampling design.

FIGURE S2: *OPTIMAL NUMBER OF EHNS IN THE RELATEDNESS NETWORK ACCORDING TO THE AIC CRITERION*

The highest AIC values, circled with a red dotted line, correspond to the best models. The optimal number of EHNs is six (AIC = -20759), then seven (AIC = -20975), four (AIC = -21006), five (AIC = -21138) and finally three (AIC = -21210).



TEXT S2: *EFFECT OF THE SPATIAL STRUCTURE OF THE TREES ON THE HETEROGENEITY OF THE RELATNESS NETWORK*

According to the AIC criterion, the best model had six EHNs (Figure S6). According to the connectivity matrix between the EHNs, five of these EHNs (called EHN_{Pj} with $1 \leq j \leq 5$) were highly connected with themselves and not connected with the other EHNs.

k=6

	EHN_0	EHN_{p1}	EHN_{p2}	EHN_{p3}	EHN_{p4}	EHN_{p5}
EHN_0	0.0	0.0	0.0	0.0	0.0	0.0
EHN_{p1}	0.0	0.8	0.0	0.0	0.0	0.0
EHN_{p2}	0.0	0.0	1.0	0.0	0.0	0.0
EHN_{p3}	0.0	0.0	0.0	1.0	0.0	0.0
EHN_{p4}	0.0	0.0	0.0	0.0	1.0	0.0
EHN_{p5}	0.0	0.0	0.0	0.0	0.0	1.0

Connectivity matrix between the EHNs in the relatedness network, for k=6

Following the same method than for the mating network (Appendix S1), we thus classified the individuals into five groups of related individuals (called P1 to P5) and one group of intermediate individuals (called Pi). The limited dispersal of pollen grains and seeds in the studied oak species (1-3) might have generated a spatial structure in the relatedness relationships, with local subgroups of individuals strongly related to each other. Therefore, we investigated whether the five groups of related individuals corresponded to geographical groups. However, this was not the case (see below).

Map of the oak stand with individual assignments to groups based on the best model for the relatedness network

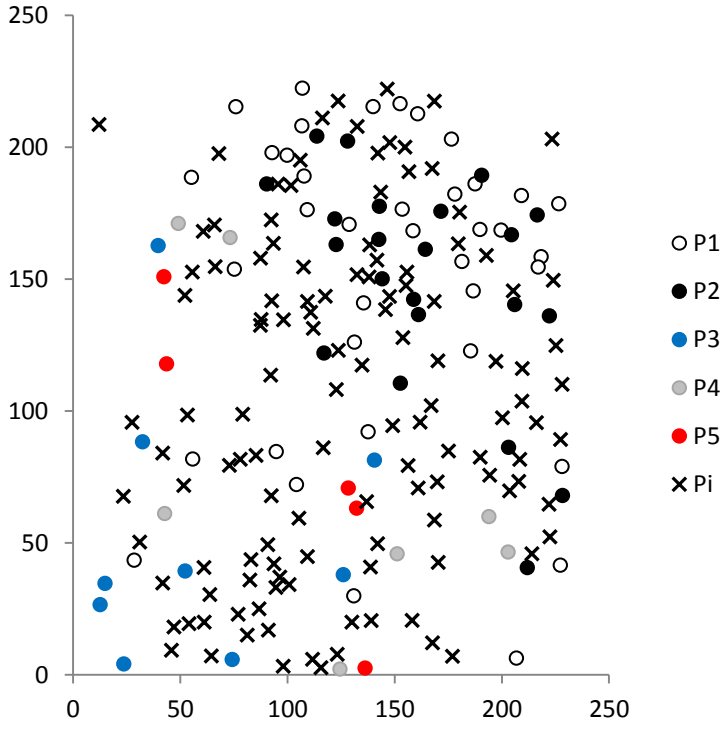


TABLE S1: *COMPARISON OF THE INDIVIDUAL ASSIGNMENTS TO SPECIES BASED ON INTERFERTILITY, RELATEDNESS, MORPHOLOGICAL AND GENOTYPIC SIMILARITIES CRITERIA*

Assignment based on the four criteria (i.e. interfertility, relatedness, genotypic and morphological similarities)	Number of individuals
The four criteria are in agreement	**160**
Qr Qr Qr Qr	97
Qp Qp Qp Qp	63
I I I I	0
Three criteria are in agreement	**42**
Qr Qr Qr I	26
Qp Qp Qp I	11
Qp Qp Qp Qr	2
Qr Qr Qr Qp	2
I I I Qr	1
Two criteria are in agreement	**4**
Qp Qp I I	2
Qr Qr I I	2

Qr: *Quercus robur*, Qp: *Quercus petraea*, I: intermediate.

FIGURE S3: *THE PERCENTAGE OF INDIVIDUALS ASSIGNED TO Q. PETRAEA (QP), THE INTERMEDIATE CLASS (I), AND Q. ROBUR (QR)* according to one criterion only (blue), or two (green), three (white) and four criteria (orange) consistently, out of all individuals assigned to this category by at least one criterion. The total number of individuals assigned to each category by at least one criterion is indicated above the bars.

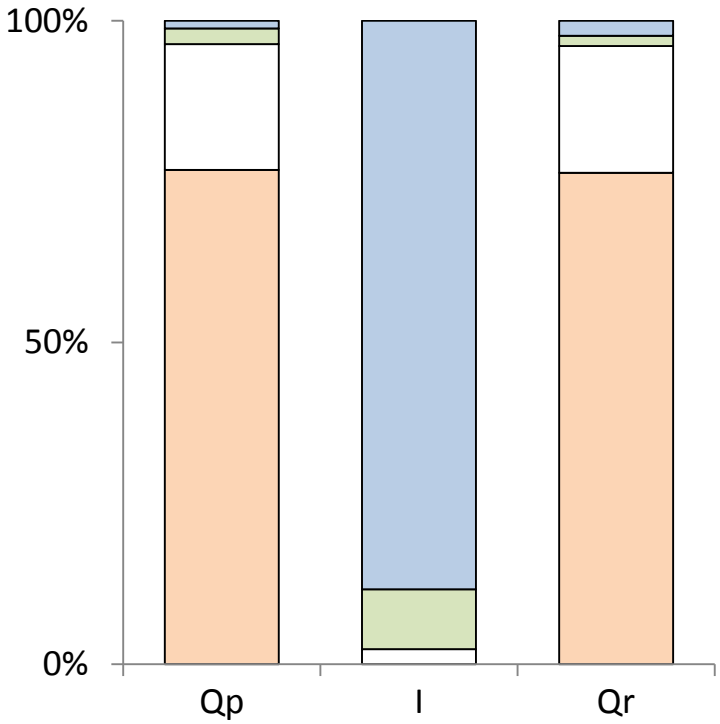


FIGURE S4: *MAP OF THE OAK STAND*

The stand was composed of 298 adult trees, among which 206 were genotyped and assigned to species by Guichoux *et al.* (4). Trees assigned to the *Quercus robur* species by Guichoux *et al.* (4) are represented by grey diamonds and trees that were assigned to the *Q. petraea* species are represented by black squares. Trees considered as hybrid trees by Guichoux *et al.* (4) are represented by white triangles. Trees on which acorns were sampled to set up the progeny test are circled.

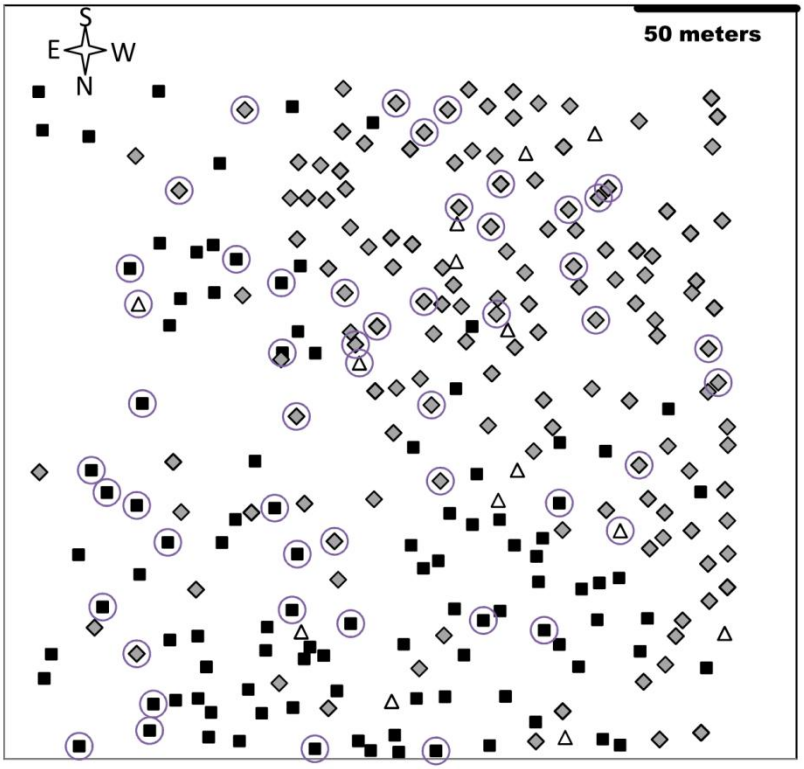


FIGURE S5: *MORPHOLOGICAL SPECIES DELIMITATION*

The 206 oak trees were ordered on the x-axis as a function of their value on the first axis of the Factorial Discriminant Analysis (FDA) performed by Bacilieri *et al.* (5), which was based on 31 morphological traits of leaves. Individuals were then graphically classified into two groups (M1 and M2) and an intermediate class. Individuals assigned to M1 are represented in red, those assigned to M2 are represented in blue, and trees with an intermediate morphology are represented in black. Leaf morphology reveals that individuals in M1 group are *Q. robur* and in M2 group are *Q. petraea* individuals. Morphological data can be found at http://bioinfo.orleans.inra.fr/TreePop/tmp/export_20121002141319506ada5f6da21.txt.

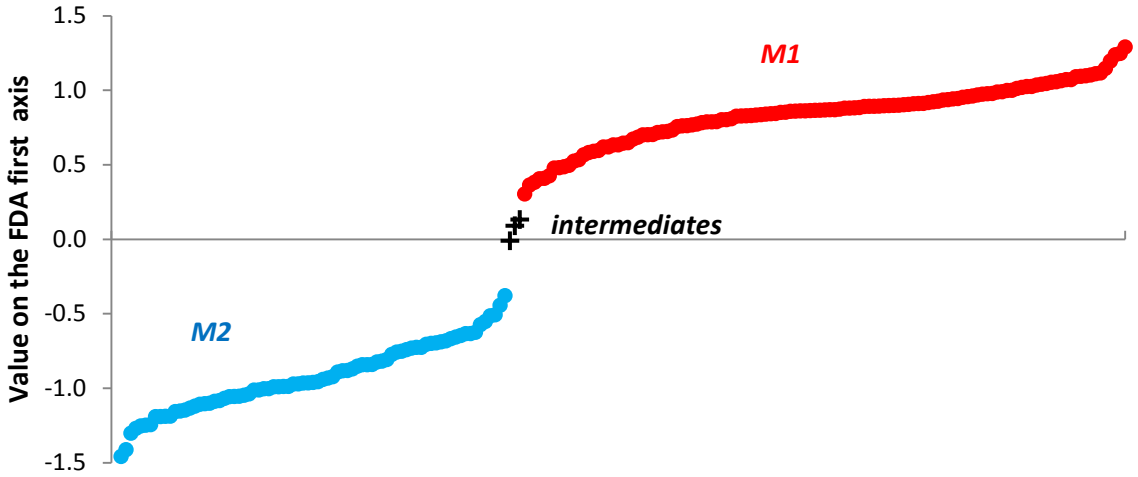


FIGURE S6: *OPTIMAL NUMBER OF GENOTYPIC CLUSTERS ACCORDING TO THE EVANNO'S CRITERION (6),* calculated by running STRUCTURE with the following parameters: 50000 burning, 50000 Markov chain with admixture, number of genotypic clusters (k) varying from 1 to 6 with five repetitions for each k values. The optimal number of clusters, indicated in red, is given by the highest Δk value. Microsatellite data can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.n50b4.

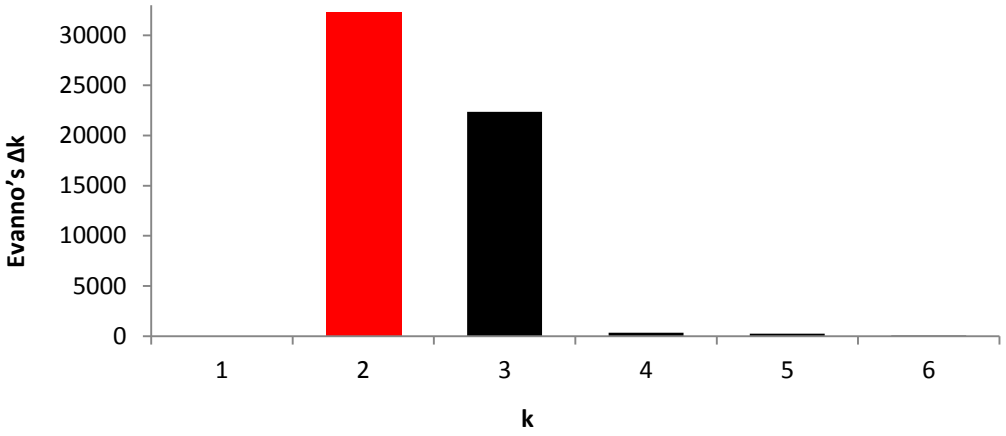
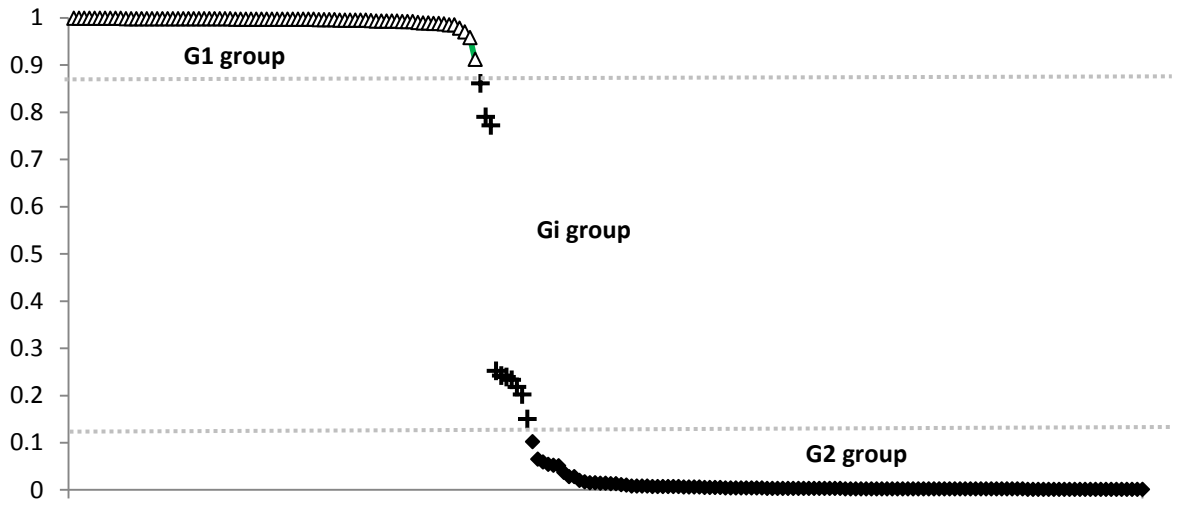


FIGURE S7: *GENOTYPIC SPECIES DELIMITATION*

The 206 individuals were ordered on the x-axis as a function of the admixture degree to the G1 group, obtained by Guichoux *et al.* (4). Individuals were classified into two groups and an intermediate class, by using the same thresholds than in Guichoux *et al.* (4). The lower and the higher thresholds represented by dotted grey lines, equal respectively 0.125 and 0.875. Individuals assigned to the G1 group are symbolized by white triangles, those classified into the G2 group by black diamonds and intermediate individuals (Gi group) are represented by black crosses. These three groups were respectively called *Q. petraea*, *Q. robur* and hybrids by Guichoux *et al.* (4). In the present study we preferred naming them G1, G2 and Gi, in order to differentiate them from the groups obtained through morphological data analysis.



REFERENCES:

1. Lagache L, Klein EK, Guichoux E, & Petit RJ (2012) Fine-scale environmental control of hybridization in oaks. *Mol. Ecol. (in press)*.
2. Jensen J, Larsen A, Nielsen LR, & Cottrell J (2009) Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark. *Ann. Forest Sci.* 66(7).
3. Pons J & Pausas J (2007) Acorn dispersal estimated by radio-tracking. *Oecologia* 153(4):903-911.
4. Guichoux E*, et al.* (2012) Outlier loci highlight the direction of introgression in oaks. *Mol. Ecol.* (*in press)*.
5. Bacilieri R, Ducousso A, & Kremer A (1996) Comparison of morphological characters and molecular markers for the analysis of hybridization in sessile and pedunculate oak. *Ann. Sci. Forest* 53(1):79-91.
6. Evanno G, Regnaut S, & Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14(8):2611-2620.

ORIGINE ET DEROULEMENT DE CE TRAVAIL

Il existe un grand nombre de définitions de l'espèce ainsi que de méthodes de délimitation des espèces (revus dans Sites & Jonathon 2004; de Queiroz 2007; Hausdorf 2011). Le concept le plus connu est le concept « biologique » de l'espèce d'Ernst Mayr (1942). Ce concept met en avant le critère d'interfertilité pour délimiter les espèces. De nombreux scientifiques travaillant sur la spéciation accordent une grande importance au critère d'interfertilité et aux mécanismes sous-jacents pour définir les espèces (Coyne & Orr 2004; Nosil 2012). Du point de vue de ces auteurs, le concept biologique de l'espèce doit avoir la primauté sur tous les autres concepts (ex. Coyne & Orr 2004). Ainsi les critères de délimitation communément utilisés pour délimiter les espèces (ex. similarités génétiques et morphologiques) servent typiquement de **substituts** à un critère plus difficile à utiliser : l'absence d'interfertilité des individus. D'autres auteurs ne sont pas d'accord avec ce principe de primauté du concept biologique de l'espèce (ex. Donoghue 1985) et soutiennent que la description des espèces devrait être basée sur les patrons de variation observés et non sur les processus sous-jacents. Bien que beaucoup de scientifiques critiquent le concept dit biologique de l'espèce, très peu d'études empiriques ont confronté expérimentalement des données d'interfertilité avec des données plus classiques de morphologie ou de marquage moléculaire (mais voir par exemple les étude de Taylor *et al.* 2000; et de Dettman *et al.* 2003 sur les champignons). Dans ce travail, j'ai précisément cherché à savoir si le critère d'interfertilité permettait de bien délimiter in situ les espèces en le comparant à d'autres critères plus communément utilisés.

Cet article est le premier présenté dans ce manuscrit mais est en fait le second que j'ai écrit durant ma thèse. Il s'appuie sur les recherches de paternité présentées au Chapitre 2, travaux qui faisaient suite à mon travail de Master II. Comme ce chapitre a une problématique différente des deux autres, j'ai préféré le présenter en premier.

La formation que j'ai suivie à la faculté de Bordeaux 1 a surtout été axée sur la biologie cellulaire, la physiologie et la génétique des plantes. C'est seulement en deuxième année de Master II que j'ai eu l'occasion d'aborder les notions de génétique des populations, d'évolution, d'écologie… Mon stage de Master II dans l'UMR BioGeCo m'a permis d'approfondir ces notions mais a aussi remis en question certaines de mes connaissances. Je me souviens d'un jour où, en pleine rédaction de mon rapport de stage, traitant de l'effet de l'environnement pollinique sur l'hybridation de deux espèces de chêne, j'ai demandé à Rémy « mais au final qu'est ce qu'une espèce ? ». A cette époque, je connaissais uniquement la définition de l'espèce d'Ernst Mayr selon laquelle une espèce est un groupe d'individus qui se reproduisent entre eux et qui ne se reproduisent pas avec les individus d'un autre groupe (ou espèce). En étudiant l'hybridation entre deux espèces de chênes : *Quercus robur* et *Q. petraea*, cette définition ne me convenait plus pour définir une espèce… Je me souviens qu'à l'époque Rémy a tenté de m'expliquer qu'il n'y a pas vraiment de définition universellement acceptée de l'espèce et que plutôt que de s'attarder à définir l'indéfinissable (cf. Darwin dans ses correspondances), il valait mieux se concentrer sur la question plus opérationnelle de la délimitation de nos espèces de chênes (à l'époque je n'avais utilisé que la méthode basée sur les similarités génotypiques). Je me souviens que cette réponse ne m'avait pas satisfaite et m'avait même plutôt frustrée…

Quand nous avons décidé de travailler sur les réseaux de reproduction de ces deux espèces afin de les délimiter selon le critère d'interfertilité, j'ai tout de suite pensé que cette étude m'apporterait les réponses que je cherchais concernant la définition de ces espèces. En faisant le tour de la littérature, je me suis aperçue que le concept biologique de l'espèce

n'était pas le seul concept d'espèce existant (listés dans de Queiroz 2007) et qu'il avait beaucoup été critiqué à cause son caractère peu opérationnel voire inopérant chez de nombreuses espèces, notamment dans le genre *Quercus* (ex. Donoghue 1985; de Meeûs *et al.* 2003). Avec l'étude de paternité que j'avais réalisée sur un grand nombre de descendants, j'avais la possibilité d'évaluer directement le critère d'interfertilité en conditions naturelles pour délimiter ces espèces de chênes. Je pensais ainsi apporter les réponses que je cherchais concernant l'utilisation de ce concept pour définir les espèces… Il ne me manquait alors qu'à trouver la méthode permettant de délimiter mathématiquement ces deux groupes d'individus partiellement interfertiles. C'est à ce moment là que Corinne m'a aidée en m'ouvrant les portes des méthodes de reconstruction des réseaux et en me permettant de rencontrer deux modélisateurs des réseaux : Jean-Jacques et Jean-Benoist. Ils ont alors tenté de comprendre la problématique de mon étude et m'ont apporté le soutien mathématique et de modélisation des réseaux dont j'avais besoin pour mener à bien cette étude. Cette étude réalisée, je me rends compte pourquoi il était si difficile à Rémy de répondre simplement à ma question.

PERSPECTIVES DE L'ETUDE

Dans cette étude j'ai comparé quatre méthodes de délimitation d'espèce. Je me suis aperçue que mise à part la catégorie des intermédiaires, ces quatre méthodes de délimitation d'espèce sont très congruentes (98% de congruence pour l'affectation des individus aux espèces pures entre les méthodes basées sur les critères morphologique, génétique et d'interfertilité); seule la délimitation du groupe d'individus intermédiaires est problématique. De ce point de vue, je trouve que le concept d'espèce de Simpson (1951; 1961) repris par Wiley (1978) et de Queiroz (1998; 2005; 2007), définissant l'espèce comme des lignées évolutives qui acquièrent au cours du temps des traits différenciés jusqu'à l'apparition d'un isolement reproducteur complet, est peut-être la définition qui se rapproche le plus de ce que j'ai observé dans cette parcelle où deux espèces de chêne (*Q. petraea* et *Q. robur*) s'hybrident. Toutefois il ne semble pas que l'isolement reproducteur total entre espèces soit nécessaire. En effet, des espèces peuvent durablement s'hybrider sans mettre en cause leur existence. Par exemple, une étude sur le peuplier de Eckenwalder (1984) a montré la présence d'hybrides fossiles très anciens entre des espèces qui s'hybrident encore de nos jours. Cette définition est donc celle qui me satisfait le plus à présent. Elle ne met pas en avant de critères particuliers pour délimiter les espèces (voir la critique de Hausdorf 2011 par exemple), ce qui ne me parait pas nécessaire vu que plusieurs critères aboutissent à des résultats sensiblement identiques et très répétables, du moins pour les individus purs.

Dans ce travail, j'ai observé que la catégorie des intermédiaires est la catégorie qui entraine le plus d'incohérences entre les méthodes. L'étude de Guichoux *et al.* (2012; et voir annexe 3) a mis en évidence une diminution du nombre d'intermédiaires déterminés à l'aide des analyses statistiques du logiciel STRUCTURE (Pritchard *et al.* 2000) quand le nombre de marqueurs génétiques différenciant les espèces était augmenté. De plus, à l'aide de simulations, nous avions observé que, du fait de la présence du nombre important d'individus purs et du petit nombre d'individus intermédiaires, le risque de prendre un individu pur pour un intermédiaire est plus grand que la réciproque. Je ne suis pas certaine qu'avec le nombre de marqueurs utilisés dans cette étude nous ayons suffisamment de puissance pour déterminer la catégorie des intermédiaires à l'aide de STRUCTURE sur la base des génotypes multilocus. Ceci pourrait être à l'origine de certaines incohérences entre cette méthode de délimitation et les trois autres. Je pense qu'une première perspective à ce

travail serait de poursuivre l'effort pour établir des jeux de marqueurs très discriminants et préciser encore un peu plus les affectations des individus issus de croisements backcross. L'utilisation de marqueurs très polymorphes pourrait également aider à préciser l'apparentement entre individus (Wang 2002). En effet, l'estimation de l'apparentement est largement perfectible: sur la base des apparentements connus par recherche de paternité des 3046 descendants, la valeur maximale d'apparentement pour des individus non apparentés était de 0.22, soit une valeur proche de celle attendue pour des demi-frères.

Puisqu'il apparaissait que c'était la classe des individus intermédiaires qui était responsable des incohérences entre méthodes de délimitation, je me suis demandé ce qui se passerait si je supprimais cette classe. J'ai alors établi un seuil à 0,5 et considéré que tous les individus ayant des affectations variant entre les seuils 1 et 0,5 appartenaient à *Q. petraea* et tous ceux ayant des affectations variant entre 0 et 0,5 appartenaient à *Q. robur*. Je me suis alors aperçue qu'il n'existait quasiment plus d'incohérences entre les méthodes (seules les inversions d'espèces dues à des problèmes d'identification morphologique des individus ou à un nombre trop faible de descendants échantillonnés pour déterminer l'interfertilité restaient). J'ai été surprise par la disparition de ces incohérences. Ne serait ce pas lié à la construction artificielle d'une classe d'« intermédiaires », qui serait ainsi responsable des incohérences entre méthodes de délimitation? En fait, il n'existe pas dans cette parcelle particulière d'individu avec une affectation à mi chemin entre *Q. robur* et *Q. petraea* (= individus hybrides ou F1). Je pense donc que tous les individus de cette parcelle affectés à la classe intermédiaire (quelque soit le critère) ont déjà acquis par rétrocroisement certaines propriétés des espèces pures. Ils sont alors considérés comme purs ou intermédiaires selon les critères étudiés en fonction du degré d'admixture pour chaque caractère. Ceci explique la quasi-disparition des incohérences une fois supprimée la classe des intermédiaires. Une seconde perspective à ce travail serait alors de réaliser des études similaires sur ces deux espèces et sur d'autres afin de vérifier que les méthodes de délimitation d'espèce donnent majoritairement des résultats congruents en termes de nombre d'espèces et d'affectation individuelle. L'idéal serait que ces études intègrent de nombreux individus hybrides de première génération afin d'étudier dans un second temps leur affectation par les différentes méthodes de délimitation d'espèce.

L'étude que j'ai réalisée ne permet pas de déterminer si une méthode de délimitation est plus pertinente qu'une autre pour ces deux espèces. On serait tenté dans un premier temps de comparer les méthodes deux à deux et de choisir celle pour laquelle on trouve le moins d'incongruences avec les trois autres. Cependant, cela ne remplace pas une vraie classification de référence et revient à considérer que les méthodes les plus proches de la moyenne sont les plus pertinentes, ce qui est discutable. Une troisième perspective à cette étude serait de réaliser une classification de référence par simulation et d'effectuer les mêmes comparaisons que celles de cette étude. Pour réaliser cette référence il faudrait prendre les individus les plus purs (c'est-à-dire avec une affectation proche des extrêmes), quelque soit le critère, et de simuler un même nombre d'individus dans chaque catégorie : espèce pures 1 et 2, hybrides et backcross. Techniquement il est aisé de produire des génotypes multilocus correspondant à ces différentes classes (voir l'étude de Guichoux *et al.* 2011 par exemple). Par contre, il est beaucoup plus compliqué de simuler la morphologie des individus purs, hybrides de première génération ou issus de backcross (cela supposerait qu'on connaisse très bien le déterminisme génétique et la plasticité de ces caractères). Enfin, il est à ce jour impossible de savoir quel comportement reproducteur aurait un hybride de première génération ou un individu issu de backcross. Pour les deux espèces étudiées dans cette thèse, les études précédentes en conditions naturelles et au niveau populationnel ont

montré que les intermédiaires se reproduisaient préférentiellement avec les espèces parentales (Lepais & Gerber 2011). Dans mon étude, je n'ai pas pu étudier finement la reproduction des individus intermédiaires car leur nombre était trop faible et ils semblaient pour la plupart être issus de backcross (apparemment pas d'individu hybride de première génération). Pour réaliser cette référence, une connaissance approfondie de la biologie de la reproduction de chaque classe d'arbres : purs, F1 et issus de backcross, serait donc nécessaire. Des individus F1 issus du programme de croisements contrôlés menés au sein de l'UMR sont désormais disponibles, les premiers backcross ont été produits mais ils ne sont pas encore matures.

Un résultat important de cette étude est l'effet de l'environnement sur les affectations individuelles. En effet, j'ai mis en évidence un effet de la proportion de voisins allospécifiques sur l'affectation des individus basée sur le critère d'interfertilité. D'autres études ont montré un effet de l'environnement sur la morphologie des feuilles de ces deux espèces (Sork *et al.* 1993; Bacilieri *et al.* 1995; Kremer *et al.* 2002). Cet effet peut entrainer des différences avec les affectations basées sur des critères indépendants de l'environnement (affectations basées sur les génotypes multilocus et sur l'apparentement). Une autre perspective de ce travail serait, je pense, d'étudier spécifiquement ces individus à la morphologie ou à la reproduction atypique mais affectés à l'une ou l'autre espèce pure sur la base des génotypes multilocus, pour comprendre les désaccords entre méthodes. On se rendrait peut-être compte que les méthodes qui intègrent le contexte dans lequel un individu se trouve sont des méthodes « plus fines » de délimitation car elles intègrent, en conditions naturelles, des informations sur l'environnement des individus, une information très pertinente sur le comportement des individus.

RÉFÉRENCES

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

Coyne J.A. & Orr H.A. (2004). *Speciation*. Sinauer Associates, Sunderland, Mass., USA.

de Meeûs T., Durand P. & Renaud F. (2003). Species concepts: what for? *Trends Parasitol.*, 19, 425-427.

de Queiroz (2005). Ernst Mayr and the modern concept of species. *P. Natl. Acad. Sci. USA*, 102, 6600-6607.

de Queiroz K. (1998). *The General Lineage Concept of species, species criteria, and the process of speciation*. Oxford University Press.

de Queiroz K. (2007). Species concepts and species delimitation. *Syst. Biol.*, 56, 879-886.

Dettman J.R., Jacobson D.J., Turner E., Pringle A. & Taylor J.W. (2003). Reproductive isolation and phylogenetic divergence in *Neurospora*: comparing methods of species recognition in a model eukaryote. *Evolution*, 57, 2721-2741.

Donoghue M.J. (1985). A critique of the Biological Species Concept and recommendations for a phylogenetic alternative. *The Bryologist*, 88, 172-181.

Guichoux E., Garnier-Géré P., Lagache L., Lang T., Bourry C. & Petit R.J. (2012). Outlier loci highlight the direction of introgression in oaks. *Mol. Ecol.*, *in press (MEC-12-0795.R1)*.

Guichoux E., Lagache L., Wagner S., Léger P. & Petit R.J. (2011). Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Mol. Ecol. Resour.*, 11, 578-585.

Hausdorf B. (2011). Progress toward a general species concept. *Evolution*, 65, 923-931.

Kremer A., Dupouey J.L., Deans J.D., Cottrell J., Csaikl U., Finkeldey R., Espinel S., Jensen J., Kleinschmit J., Dam B.V., Ducousso A., Forrest I., Heredia U.L.d., Lowe A.J., Tutkova M., Munro R.C., Steinhoff S. & Badeau V. (2002). Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Ann. For. Sci.*, 59, 777-787.

Lepais O. & Gerber S. (2011). Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution*, 65, 156-170.

Mayr E. (1942). *Systematics and the origin of species*. Columbia Univ. Press, New York.

Nosil P. (2012). *Ecological speciation*. Oxford University Press, Oxford.

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

Simpson G.G. (1951). The species concept. *Evolution*, 5, 285-298.

Simpson G.G. (1961). *Principles of animal taxonomy*. Columbia Univ. Press, New York.

Sites J.J.W. & Jonathon C.M. (2004). Operational criteria for delimiting species. *Annu. Rev. Ecol. Evol. S.*, 35, 199-227.

Sork V.L., Stowe K.A. & Hochwender C. (1993). Evidence for local adaptation in closely adjacent subpopulations of northern red oak (*Quercus rubra L.*) expressed as resistance to leaf herbivores. *Am. Nat.*, 142, 928-936.

Taylor J.W., Jacobson D.J., Kroken S., Kasuga T., Geiser D.M., Hibbett D.S. & Fisher M.C. (2000). Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.*, 31, 21-32.

Wang J.L. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, 160, 1203-1215.

Wiley E.O. (1978). The Evolutionary Species Concept reconsidered. *Syst. Biol.*, 27, 17-26.

Afin de comprendre quels paramètres affectent les croisements intra- et interspécifiques dans cette parcelle mixte, j'ai étudié plus en détail dans les **chapitres 2 et 3** le système de reproduction de ces deux espèces. Pour ces études il était nécessaire d'affecter les individus aux espèces. J'ai choisi d'affecter les individus aux espèces sur la base de leur génotype et de n'étudier que la reproduction des individus purs. Nous venons de voir dans ce chapitre que l'environnement biotique joue un rôle sur l'hybridation de ces espèces. Dans le chapitre suivant, j'ai décidé d'aller un peu plus loin dans l'étude de l'effet de l'environnement biotique sur l'hybridation de ces deux espèces en modélisant à l'aide d'un modèle de voisinage (modèle spatialisé) les croisements intra- et interspécifiques de ces deux espèces.

Fine-scale environmental control of hybridization in oaks

Lélia Lagache[1, 2], Etienne K. Klein[3], Erwan Guichoux[1, 2], Rémy J. Petit[1, 2]

[1] INRA, UMR 1202 Biogeco, F- 33610 Cestas, France

[2] Univ. Bordeaux, UMR1202 Biogeco, F-33400 Talence, France

[3]INRA, UR546, Biostatistique et Processus Spatiaux (BioSP), F-84914 Avignon, France

(Article paru dans *Molecular Ecology* : doi: 10.1111/mec.12121)

.

INTRODUCTION

Hybridization has become a major research topic in evolutionary biology and conservation biology because of its prevalence and potentially important consequences for biodiversity. First, the phenomenon is central to our understanding of the speciation process (Arnold 1997). Second, it raises important conservation issues. Rates of hybridization are increasing worldwide, as a consequence of biological invasions, range fragmentation, homogenization of natural environments and climate-triggered phenological changes. This has raised concerns for biodiversity in many plant and animal groups (Allendorf *et al.* 2001; Rhymer & Simberloff 1996). Therefore, predicting when and where hybridization is likely to occur represents an important research objective.

Hybridization events are often quite rare on ecological timescales, making their direct investigation challenging. However, some progress has been made towards a more quantitative description of hybridization potential (Field *et al.* 2011; Heinze 2011). Hybridization has been shown to depend not only on the intrinsic characteristics of the species involved but also on the environmental context (Hersch & Roy 2007; Lamont *et al.* 2003; Seehausen *et al.* 2008). The studies of Focke (1881) on plants and of Hubbs (1955) on fishes first showed that hybridization is frequency dependent; the scarcity of conspecifics would increase hybridization rates as a consequence of mate recognition errors by females of the rare species. Recently, this process, sometimes called Hubbs' effect, has received renewed interest, with a growing number of empirical studies reporting that rates of hybridization vary with interspecific mating opportunities (Field *et al.* 2008; Lepais *et al.* 2009; Wirtz 1999). However, these studies only provide a macroscopic view of the effect of relative species abundance on hybridization; what really matters is the proportion of allospecific versus conspecific mates available to each individual. In plants, pollen dispersal is limited by distance (Adams 1992) and hybridization rates are therefore expected to differ according to the degree of species intermixing. Moreover, a few individuals could contribute disproportionately to the overall hybridization rate (e.g. Bacilieri *et al.* 1996a; Streiff *et al.* 1999). Hence, explaining differences in hybridization rate under stable or disturbed environments requires both spatially-explicit and individual-based analyses of mating events.

Chan & Levin (2005) have proposed a simple mass action model of hybridization that accounts for species relative abundance and for the intensity of sexual barriers. However, their model is not spatially explicit and only considers the overall proportion of each species. In contrast, Burczyk *et al.* (2002) and Oddou-Muratorio *et al.* (2005), expanding on earlier efforts (Adams 1992), have developed spatially-explicit mating models to predict mating events based on parentage analysis. These models have the advantage of considering the immediate environment of each adult but their use has been largely restricted to the analysis of intraspecific crosses. Moreover, in these models, immigration from outside the neighbourhood is assumed to occur at a constant rate, a clear limitation. Current neighbourhood models are therefore poorly adapted to study Hubbs' effect. The combination of a mass action hybridization model and of an improved version of the neighbourhood model, in which mating partners originating from outside the neighbourhood are allowed to compete with those from within, seems more appealing. Such a model could contribute to a better understanding of the effects of environmental context on natural hybridization while simultaneously estimating intrinsic sexual barriers to hybridization.

Due to their sessile nature, high reproductive output and high propensity for hybridization, seed plants are good models for studying environmental effects on hybridization (Rieseberg & Carney 1998). Forest trees in general, and oaks (*Quercus* spp.) in

particular, have been the focus of numerous hybridization studies (reviewed in Rushton 1993). The genus *Quercus* is species-rich and many closely related oak species can be found in sympatry. The two most widespread European oak species, *Q. robur* L. and *Q. petraea* (Matt.) Liebl., have been intensively investigated (Petit *et al.* 2004). Controlled crossing experiments have shown that hybridization is possible between these species and that *Q. petraea* has stronger post-pollination hybridization barriers than *Q. robur* (Steinhoff 1993). Genetic analyses of open-pollinated progenies indicate that hybridization also occurs under natural conditions (Jensen *et al.* 2009; Streiff *et al.* 1999). In these oaks, the most important sexual barriers are prezygotic (pollen competition and pollen–pistil interactions), with a lower but significant contribution of postzygotic barriers (seed germination and progeny fitness-related traits, Abadie *et al.* 2011). This species pair seems therefore well suited to study the extrinsic factors controlling hybrid production.

To understand the influence of local environment on hybridization, we studied mating events in a mixed stand of *Q. robur* and *Q. petraea* using a combination of Chan & Levin's mass action model and a revised neighbourhood model. In particular, we compared observed hybridization rates with predicted ones and tested if species distribution in the stand influences hybridization rates. Altogether, our results suggest that disturbances, which typically decrease species clustering and density, should increase hybridization rates. We conclude by outlining the need to model hybridization at the scale at which the relevant biological processes take place.

MATERIAL & METHODS

MATERIAL

The investigated 5-ha even-aged oak stand, which has been intensively studied for more than 20 years, is located in the Petit Charnie State forest in western France (latitude: 48.08° N, longitude: 0.17° W). It contains both *Q. robur* and *Q. petraea* trees (Fig. 1) and probably originates from natural regeneration as both oak species show a clear spatial genetic structure (Bacilieri *et al.* 1994; Streiff *et al.* 1998). The species status of all 298 trees growing in the stand had been determined using leaf morphological characters, indicating that *Q. robur* is more abundant than *Q. petraea* (Bacilieri *et al.* 1996a). Mature neighbouring stands are also dominated by *Q. robur* (Supporting Information 1). A 3-years survey of the stand showed that the flowering periods of the two oak species are largely synchronous (Bacilieri *et al.* 1995). In 1995, seeds were harvested on 51 open-pollinated mother trees distributed throughout the stand (Fig. 1). After germination in a nursery, the resulting offspring (3780) were transplanted close to the adult oak stand at wide spacing (1.5 x 3 m) to delay the onset of competition. Buds or leaves from all 3213 surviving offspring were harvested in 2009 for genotyping. The adult trees could not be directly sampled for genotyping as they had been felled in 1998. However, grafts of 256 out of the 298 parental trees (86%) were available for this study.

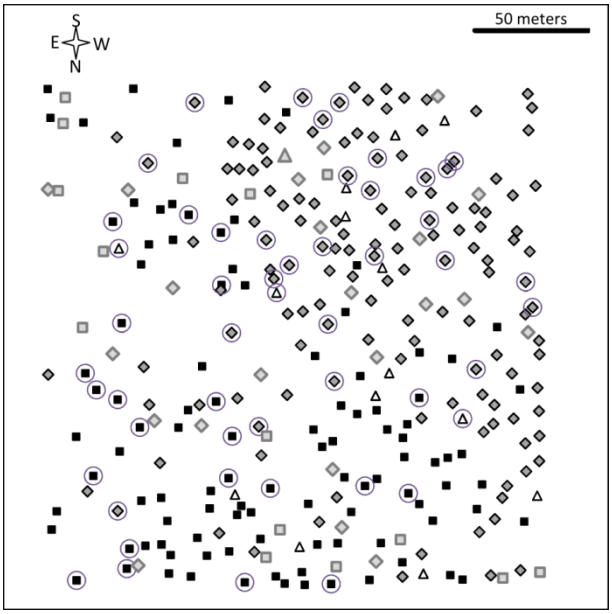


Figure 1: ***Map of the studied mixed oak stand****. Quercus robur genotyped trees are represented by grey diamonds, Q. petraea trees by black squares and intermediate trees by white triangles (species assignment based on multilocus genotypes). Sampled mother trees are circled. Q. robur, Q. petraea and intermediate ghost trees are represented by light grey diamonds, light grey squares and light grey triangles, respectively (morphological species assignment).*

GENOTYPING

DNA was isolated from leaves or buds of the 256 parental grafts and 3213 offspring using the Invisorb DNA plant HTS 96 kit (Invitek, Berlin, Germany) and a 12plex microsatellite kit was used to genotype all individuals (256 parents and 3213 offspring, Guichoux *et al.* 2011a). All the recommendations given by Guichoux *et al.* (2011b) were followed: genotype double-blind reading, allele binning, positive and negative controls and blind duplicate samples. In addition, a single nucleotide polymorphism assay (384plex) developed by Guichoux *et al.* (2012), enriched with markers showing high interspecific differentiation, was used to characterize all adults and a subset of 306 offspring (six from each of the 51 families). This data helped improve parental species assignment (Guichoux *et al.* 2012). Comparisons between parental and offspring genotypes made it possible to correct a number of genotyping errors (Supporting Information 2).

GENETIC DATA ANALYSES

We first relied on genotypes at all microsatellite and single-nucleotide polymorphism loci to assign adult trees to species using version 2.3.3 of STRUCTURE (Pritchard *et al.* 2000). The number of groups was set at two (corresponding to the two species). The admixture model with correlated allele frequencies was used. A burn-in of 50,000 steps was followed by a Markov chain Monte Carlo repetition of 50,000 steps. Following Guichoux *et al.* (2012), individuals were grouped into three classes: *Q. petraea* purebreds (admixture values between 0 and 0.125), admixed trees (0.125-0.875) and *Q. robur* purebreds (0.875-1). These threshold values were chosen because they are optimal for distinguishing between purebreds and first generation backcrosses, which was deemed sufficient for this study (Guichoux *et al.* 2012).

The species identity of the 38 adult trees in the stand with no available genotype (so-called 'ghost trees') was assigned by relying on morphological criteria (Bacilieri *et al.* 1996a). To determine observed hybridization rates, we used two different methods. For offspring whose fathers were detected within the plot (the mother's identity was known for all seedlings from the original seed collection), the species of both parents were determined using STRUCTURE. For offspring with no identified father, we computed the two likelihoods of observing the diploid genotype of the offspring conditional on the mother's genotype, assuming that the father belongs either to *Q. robur* or to *Q. petraea* (i.e. excluding the admixed category), using overall allelic frequencies of each species as reference data. The father species was assumed to be that with the highest likelihood of producing the observed offspring genotype. For comparison purposes, this procedure was also used for those offspring whose father had been identified using paternity analysis.

MODELLING AND PARAMETER ESTIMATION

Spatially-explicit mating model: We used a spatially-explicit mating model to investigate pollen flow inside and outside the studied stand (Burczyk *et al.* 2002; DiFazio *et al.* 2012; Oddou-Muratorio *et al.* 2005). Spatially-explicit mating models integrate genotypic, spatial and phenotypic information into a likelihood function that is maximized to directly estimate the parameters of interest. Three important processes were included in this model. First, the effect of the distance between trees on their probability of mating was modelled using a pollen dispersal kernel that featured two parameters for each species (mean distance and shape parameter). Second, sexual barriers between species were introduced by considering that allospecific and conspecific pollen grains that land on a stigma have different probabilities of fertilizing the ovules. Third, pollen immigration from outside the stand was modelled using a mass action law.

We thus considered a spatially-explicit mating model in which it was assumed that the amounts of immigrant *Q. robur* or *Q. petraea* pollen received are constant across the different mother-trees. In turn, immigration rates can vary across mothers following the mass action law (Holsinger 1991). In this model, the probability that a seed *o* from mother j_o has genotype g_o is

$$P\left(g_o \middle| g_{j_o}\right) = sT\left(g_o \middle| g_{j_o}, g_{j_o}\right) + (1-s)\left[mig_{jP}\, T\left(g_o \middle| g_{j_o}, AFP\right) + mig_{jR}\, T\left(g_o \middle| g_{j_o}, AFR\right) + \sum_{k:candidates} \pi_{jk}\, T\left(g_o \middle| g_{j_o}, g_k\right)\right]$$

(Eq. 1)

where *s* is the selfing rate, $T(g_o|.,.)$ are the Mendelian probabilities of generating the offspring's genotype g_o from the known genotypes of the two parents, *AFR* and *AFP* are the microsatellite allelic frequencies of *Q. robur* and *Q. petraea*, π_{jk} is the relative contribution of candidate father *k* in the pollen pool of mother *j, and* mig_{jR} and mig_{jP} correspond to the two migration rates (*Q. robur* and *Q. petraea*) for mother *j* (these parameters are detailed below). In contrast to most previously published spatially-explicit mating models, migration rates can vary across mothers due to the amount of local pollen they receive and their proximity to the 38 ghost trees (Fig. 1).

Parameters: The relative contribution π_{jk} of the candidate father *k* in the pollen pool of mother *j* results from the competition with pollen from all other candidate fathers but also with pollen from all ghost fathers and with immigrant pollen:

$$\pi_{jk} = \frac{K_{sp_k}(d_{jk})\, Hyb_{jk}}{\sum_{l:candidates} K_{sp_l}(d_{jl})\, Hyb_{jl} + \sum_{l:ghosts} K_{sp_l}(d_{jl})\, Hyb_{jl} + q_P Hyb_{jP} + q_R Hyb_{jR}} \quad \text{(Eq. 2)}$$

where *K* is a dispersal kernel that accounts for distance d_{jk} (or d_{jl}) between mother-tree *j* and father-tree *k* (or *l*). Different fathers *k* disperse their pollen following one of three different exponential power kernels K_{sp_k} depending on their species sp_k (= R (*Q. robur*), P (*Q. petraea*), or H (admixed category)). Exponential power kernels are two-parameters functions recognized as sufficiently flexible to characterize pollen dispersal. They are described in detail in e.g. in Austerlitz *et al.* (2004) and are specified with two parameters: δ, the mean dispersal distance, and *b*, the shape parameter (smaller *b*, fatter tail).

Reproductive barriers Hyb_{jk}, Hyb_{jP} and Hyb_{jR} represent the post-dispersal relative fertilization successes of one pollen grain from father *k*, from *Q. robur* immigration and from *Q. petraea* immigration. They are obtained as:

$$Hyb_{jk} = h_{sp_j, sp_k},\ Hyb_{jP} = h_{sp_j, P},\ Hyb_{jR} = h_{sp_j, R}$$

from the following matrix (to be estimated) where each row corresponds to the mother species and each column corresponds to the father species:

$$\begin{pmatrix} 1 & h_{PR} & h_{PH} \\ h_{RP} & 1 & h_{RH} \\ h_{HP} & h_{HR} & 1 \end{pmatrix} \quad \text{(Eq. 3)}$$

This model can be viewed as an extension of the model used by Chan & Levin (2005) to account for three groups instead of two. Note however that our main interest was to determine sexual barriers between pure individuals from each species, not the mating behaviour of trees from the admixed category.

Finally, the two migration rates mig_{jR} and mig_{jP} in Eq. 2 also result from the mass-action law as

$$mig_{jR} = \frac{\sum_{l:robur\ \ ghosts} K_{sp_l}(d_{jl})\, Hyb_{jl} + q_R Hyb_{jR}}{\sum_{l:candidates} K_{sp_l}(d_{jl})\, Hyb_{jl} + \sum_{l:ghosts} K_{sp_l}(d_{jl})\, Hyb_{jl} + q_R Hyb_{jR} + q_P Hyb_{jP}} \quad \text{(Eq. 4)}$$

$$mig_{jP} = \frac{\sum_{l:petraea\,ghosts} K_{sp_l}\left(d_{jl}\right) Hyb_{jl} + q_P\, Hyb_{jP}}{\sum_{l:candidates} K_{sp_l}\left(d_{jl}\right) Hyb_{jl} + \sum_{l:ghosts} K_{sp_l}\left(d_{jl}\right) Hyb_{jl} + q_R Hyb_{jR} + q_P\, Hyb_{jP}} \quad \text{(Eq. 5)}$$

where q_P and q_R are the amounts of *Q. petraea* and *Q. robur* pollen coming from outside the stand, assumed to be constant across all mother trees. To keep the model flexible, we did consider that q_P and q_R are independent of the local pollen dispersal kernels, as often done in neighbourhood models.

These equations imply that we consider as 'migrant pollen' both pollen originating from outside the study site and pollen originating from the ghost trees; in the latter case we account for the position of the tree relative to the sampled mother tree. The offspring sired by local pollen coming from ghost trees cannot be distinguished from offspring sired by immigrant pollen on an individual basis. However, the explicit inclusion of ghost trees in Eq. 4 and Eq. 5 enabled to accurately estimate the amounts of true immigrant pollen q_P and q_R after statistically removing the pollen from ghost trees from the set of unassigned seeds.

Likelihood tests using sub-models: Sub-models investigating different biological hypotheses by omitting or fixing different parameters of interest were also fitted to the data. Likelihood-ratio tests were then used to test the hypotheses (i.e. investigate whether the fixed parameters are significant in the full model), following Oddou-Muratorio *et al*. (2005). First, the effect of dispersal on mating events was studied by contrasting the full model with an unlimited dispersal model (also named mean field model). Second, the full model was compared with a model where *Q. robur* and *Q. petraea* had the same pollen dispersal kernel ("Homogeneous dispersal across species"). Third, we contrasted the full model with one with no hybridization barrier (matrix Eq. 3 with all parameters *h* set to 1). Fourth, we compared the full model with one where the barriers were symmetric between *Q. robur* and *Q. petraea* ($h_{PR} = h_{RP}$ in Eq. 2-5). Fifth, we tested if different amounts of *Q. robur* and *Q. petraea* pollen come from outside the stand by fitting a model with the same amounts of immigrant pollen for the two species ($q_P=q_R$). Sixth, we compared the full model with the traditional spatially-explicit mating model where migration rates are constant across mother trees (i.e. each seed has a probability *s* to result from selfing, probabilities mig_{jR} and mig_{jp} to result from an immigrant *Q. robur* or *Q. petraea* pollen, respectively, and a probability $(1 - s - mig_{jp} - mig_{jR})\pi_{jk}$ to have been sired by a pollen donor tree within the studied stand). Finally we tested wether including the ghost trees as sources of unknown pollen improved the fit.

Parameter estimation: The log-likelihood of the full genotypic dataset was computed by summing the logarithm of Eq. 1 for all 3213 genotyped offspring. All computations necessary to calculate the likelihood were conducted with MATHEMATICA 8.1 (Wolfram Research Inc. 2010). We maximized the log-likelihood using a quasi-Newton algorithm to obtain maximum likelihood estimates for all parameters considered. Confidence intervals for the parameters were derived using 500 bootstrap datasets obtained by re-sampling mother trees at random while keeping constant the number of seeds from each species (*Q. robur*, *Q. petraea* and admixed trees).

Predictions of hybridization rates: We first computed the expected hybridization rate in the stand on the basis of actual species proportions using Chan & Levin's mass action model. In this "mean-field model", all trees receive a proportion of conspecific and allospecific pollen corresponding to the species' relative abundance:

$$Hb_{tot} = \frac{h_{PR} q_R}{q_P + h_{PR} q_R} + \frac{h_{RP} q_P}{q_R + h_{RP} q_P} \quad \text{(Eq. 6)}$$

where Hb_{tot} is the total hybridization rate of a stand composed of a proportion q_R of *Q. robur* and q_P of *Q. petraea*, and h_{PR} and h_{RP} are the sexual barriers against allospecific pollen for *Q. petraea* and for *Q. robur* mother trees, respectively. The computation is based on all trees growing in the stand (including ghost trees) and assumes that intermediate trees are compatible with mother trees from both species.

To evaluate the effect of species distribution on hybridization, we generated 100 simulated stands in which the relative abundance of each species and the trees' geographical coordinates were identical to those in the real stand but the species identity of all trees were randomly permuted. We also used the original tree distribution and created a configuration where the species are fully spatially segregated (Supporting Information 3). We then used our spatially-explicit mating model with all estimated parameters to predict hybridization rates for each oak species under each spatial configuration of the two species (i.e. intermixed at random, fully separate, and corresponding to the original stand). For these simulations, we assumed that all trees produce equal numbers of seeds and pollen and that pollen is not limiting.

Theoretical expectations under pollen limitation: To illustrate hybridization rates expected under different levels of pollen limitation, we assumed that (i) an average of $n_{p/o}$ pollen grains compete to fertilize each ovule, (ii) if the relative abundance of the hybridizing species is p, the numbers of auto- and allospecific pollen grains N_S and N_H competing for each ovule are random and follow Poisson distributions with means (1-p) x $n_{p/o}$ and p x $n_{p/o}$, respectively, (iii) given N_S and N_H, and given that $N_S + N_H > 0$, the probability to produce an hybrid offspring is given by Chan & Levin's formula:

$$\frac{hN_H}{N_S + hN_H} \quad \text{(Eq. 7)}$$

where h measures the success of allospecific pollen relative to conspecific pollen, and (iv) when $N_S + N_H = 0$, no seed is produced. Averaging over all values for N_S and N_H provides the expected hybridization rate.

RESULTS

MICROSATELLITE GENOTYPING

All individuals analysed (256 parents and 3213 offspring) were successfully genotyped, with a mean proportion of typed loci per individual of 99.5%. The average number of alleles per locus was 13.7 (range 7-23) and the observed heterozygosity was 0.71 (range 0.46-0.88). A total of 167 offspring (5.1%) were excluded from the analysis because their genotype did not match with that of their putative mother. Parent-offspring genotype comparisons resulted in the correction of 8 mother trees, 12 father trees and 120 offspring; the genotype of four unavailable mother trees could be unambiguously reconstructed from the genotypes of their offspring (Supporting Information 2). In the final dataset, no null alleles were found to segregate and the error rates based on the 163 duplicate samples (which do not benefit

from corrections based on parentages) were very low (no error detected at five loci, one error at six loci, and three errors at one locus).

SPECIES ASSIGNMENT

Among the 260 available genotypes for the adult stand, we identified 142 *Q. robur* (54.6%), 104 *Q. petraea* (40.0%) and 14 admixed individuals (5.4%) using STRUCTURE. Among them, there were 26 *Q. robur*, 22 *Q. petraea* and 3 admixed mother trees. The correspondence between morphological assignment and genetic assignment based on STRUCTURE was high (up to 99% for purebreds). The 38 ghost trees were composed of 22 *Q. robur*, 15 *Q. petraea* and 1 intermediate tree, according to morphological data.

PATERNITY ANALYSES

Simple exclusion tests for the 3046 offspring identified a single compatible father for 51.7% of the offspring (855 *Q. robur* and 615 *Q. petraea*) and two or more compatible fathers for 1.8% of the offspring (31 *Q. robur* and 22 *Q. petraea*). The remaining individuals (46.5% of all offspring: 885 *Q. robur* and 427 *Q. petraea*) had no compatible father among the 260 adult trees studied (Table 1).

	Q. robur ♀	*Q. petraea* ♀	Total[1]
Offspring sired by immigrant pollen[2]	885 (50.0%)	427 (40.1%)	1417
Interspecific crosses[3] with immigrant pollen[2]	35 (4.0%)	13 (3.0%)	-
Offspring with a single compatible father in the stand	855 (48.3%)	615 (57.8%)	1575
Offspring with two or three compatible fathers in the stand	31 (1.7%)	22 (2.1%)	54
Interspecific crosses[3] in the stand	13 (1.5%)	0 (0.0%)	-
Crosses[4] with admixed trees located in the stand	35 (4.0%)	12 (1.9%)	-

[1]The results for the 211 offspring from the three admixed mother trees are not detailed but are included in the totals.
[2]Immigrant pollen includes pollen from ghost trees and from trees located outside the stand
[3]Indirect paternal species assignments; see text for explanation.
[4] Direct paternal species assignments using STRUCTURE.

***Table 1:* Results of the simple exclusion paternity analysis and average proportions of hybrids in the progenies resulting from immigrant and local pollen.**

Parameter[1]	Estimate	Confidence intervals
Selfing		
Selfing rate (*s*)	0.0016	0.0000-0.0040
Dispersal		
Mean dispersal distance *Q. petraea* (δ_P)	69	56-107
Mean dispersal distance *Q. robur* (δ_R)	151	105-167
Shape parameter *Q. petraea* (b_P)	0.68	0.37-1.20
Shape parameter *Q. robur* (b_R)	0.23	0.16-0.32
Immigration		
Q. petraea immigrant pollen amount (q_P)	0.0010	0.0007-0.0013
Q. robur immigrant pollen amount (q_R)	0.0017	0.0013-0.0020
Hybridization barriers		
Hybridization barrier on *Q. petraea* mothers (h_{PR})	0.0019	0.0000-0.0074
Hybridization barrier on *Q. robur* mothers (h_{RP})	0.023	0.005-0.043

[1]A total of 15 parameters were estimated, but the 6 parameters for intermediate trees (h_{IP}, h_{IR}, h_{PI}, h_{RI}, δ_I and b_I) are not shown in this table.

***Table 2:** Parameters estimated from the spatially-explicit mating model.*

PARAMETER ESTIMATES BASED ON THE SPATIALLY-EXPLICIT MATING MODEL

The estimated selfing rate was very low (0.2%; only five selfed individuals detected; Table 2). *Quercus robur* was found to disperse its pollen over greater distances within the study site (mean pollination distance: 152m and fatter-tailed dispersal kernel, b=0.23) than *Q. petraea* (69m, b=0.68). Sexual barriers were estimated to be on average ten times lower for *Q. robur* than for *Q. petraea* (h_{RP} = 0.023 versus h_{PR} = 0.0019), although confidence intervals for these estimates were large.

The processes with the greatest influence on observed mating patterns were pollen dispersal and interspecific incompatibility (models that did not include these processes had much lower likelihoods; Table 3). The inclusion of ghost trees in the analysis also improved model fit (ΔAIC=184). We found significant support for differences in dispersal kernels between species (p=0.0006), for heterogeneous immigration rates across mother trees ($p<10^{-4}$) and for asymmetric hybridization rates (p=0.013). As expected given the greater

abundance of *Q. robur* in the surroundings of the stand, the abundance of immigrant *Q. robur* pollen was estimated to be higher than that of immigrant *Q. petraea* pollen (m_R=0.0017 vs m_P=0.0010, p=0.03, Table 2 and 3).

Model[1]	-LogL	Δ -LogL	ddl[2]	*p*-value[3]	AIC[4]
Full model	61927	-	15	-	123883
Hybridization parameters					
No hybridization barrier	63195	1268	6	$<10^{-4}$	2524
Symmetric hybridization barriers	61933	6	1	0.013	10
Dispersal parameters					
Unlimited dispersal ("mean field")	63237	1310	6	$<10^{-4}$	2608
Homogenous dispersal across species	61942	15	2	0.0006	26
Immigration parameters					
Same amounts of immigrant pollen from both species	61931	4	1	0.034	7
Constant immigration rate across mother-trees (classical mating model)	61993	67	4	$<10^{-4}$	125
Not considering ghosts in the immigrant pollen pools	62018	92	0	-	184

[1] *All models listed are based on the full model modified in one respect to yield the corresponding submodel.*
[2] *ddl provides the number of estimated parameters for the full model and the number of parameters that are fixed and not estimated in the corresponding submodel.*
[3] *p-values lower than 0.05 indicate that the full model is significantly more informative than the tested* ***submodel.***
[4] *AIC is the Akaike Information Criterion, computed as 2xddl – 2xlogL. The AIC value is provided for the full model and the ΔAIC= AIC(submodel) - AIC (Full Model) for each submodel.*

***Table 3:** Likelihood-ratio test of the significance of each sub-model component*

POLLEN IMMIGRATION RATES

Predicted and observed pollen immigration rates were high and varied considerably among progenies; the range for the predicted rates was 32-79% (including ghost tree contributions) while that for the observed rates was 6-92% (Supporting Information 4A). These high immigration rates support the existence of long distance dispersal in oaks. Average dispersal distances are thus certainly higher than the mean pollination distance of 152m and 69m estimated on the basis of mating events taking place between trees located within the study site. For each mother tree, there was a strong correlation between predicted and observed numbers of offspring sired by immigrant pollen, suggesting that the model accurately predicted the observed heterogeneity of immigration rates across progenies (R^2=0.86; Supporting Information 4B). Ghost trees were estimated to have sired 6.4% of the offspring (1.9 to 13.6%, depending on the mother tree and ghost tree positions in the stand; see Fig. 1).

HYBRIDIZATION RATES

Predicted hybridization rates were 1.2% for *Q. robur* (ranging from 0.6 to 3.0% across the 26 mother trees) and as little as 0.18% for *Q. petraea* (0.13 to 0.29% across the 22 mother trees). Similar hybridization rates are predicted when assuming that all trees in the stand reproduce (Table 4), suggesting that the sample of mother trees is representative. Simulations based on different stand structure showed that the predictions clearly depend on the spatial distribution of the two species in the stand. The highest hybridization rates are predicted for the case of random species distribution. Lower hybridization rates are predicted when using actual species distribution within the stand and even lower hybridization rates are predicted when using a stand with fully separate species distribution. Mean-field estimates were close to those of the spatially-explicit mating model (Table 4).

The range of observed hybridization rates, estimated directly from the paternity analysis, varied considerably across maternal progenies, ranging from 0 to 15%. Some of the variation in observed individual hybridization rates was explained by the model in *Q. robur* (R^2=0.23), but not in *Q. petraea* (R^2=0.004) (Fig. 2, top panel). Interestingly, observed hybridization rates were larger than predicted for both species: 2.7% for *Q. robur* (more than twice the predicted value) and 1.2% for *Q. petraea* (about seven times the predicted value). This excess of hybrids in both species is due primarily to the larger hybridization rates with immigrant pollen compared with expectations (see Fig. 2, middle and bottom panels). In *Q. robur* progenies, hybrids represent 1.5% of the offspring sired by known fathers (i.e. close to predictions) but as much as 4.0% of those sired by immigrant pollen (Table 1). For *Q. petraea* progenies, the same trend exists; while there were no hybrid offspring sired by *Q. robur* trees in the stand, in line with predictions of low hybridization rate, there were 3.0% of hybrids sired by *Q. robur* trees not found in the stand (Table 1). These differences still hold when using the same method to determine hybridization events for both types of crosses, indicating statistical support for an excess of hybridization with distant fathers in each species (Supporting Information 5). As some offspring sired by unknown fathers could have been sired by neighbouring ghost trees and not only by distant trees, the comparison between both types of crosses is in fact conservative.

Model		Hybridization rate on *Q. robur* ♀ (%)	Hybridization rate on *Q. petraea* ♀ (%)
mean field		1.51	0.23
spatially explicit mating model	randomized distribution	1.70 (1.65-1.74)[1]	0.24 (0.23-0.25)[1]
	original distribution	1.33	0.19
	separate distribution	0.87	0.13

[1] *95th percentile from the 100 input files obtained by randomization of species status*

***Table 4:** Prediction of hybridization rates in the stand under different scenarios*

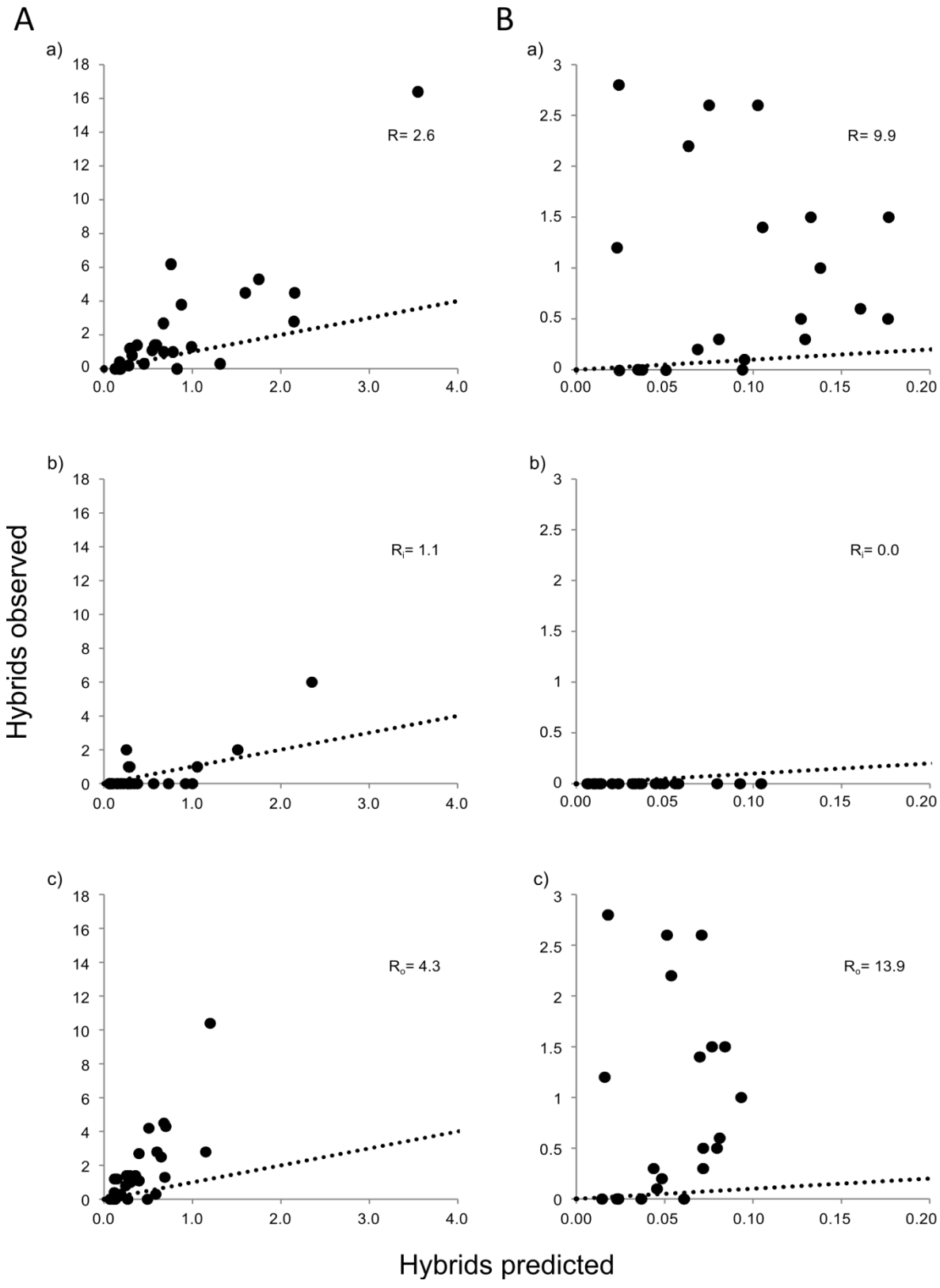


Figure 2: ***Comparison of observed and predicted numbers of hybrids in the progeny of each mother tree.*** *A) Q. robur. B) Q. petraea. a: total number of hybrids, b: hybrids originating from pollen produced by trees found inside the stand. c: hybrids originating from pollen produced by trees not found inside the stand. R is the ratio of the observed number of hybrids compared with the expected number (R: global deviation, Ri: trees found inside the stand only and Ro: trees not found inside the stand). Dotted line corresponds to identical observed and predicted hybridization rates.*

EFFECT OF POLLEN LIMITATION

Pollen competition is a frequency-dependent process. As a consequence, pollen limitation should have a strong effect on hybridization rate. Consider for instance a mother tree receiving a pollen pool with 20% of conspecific pollen and 80% of allospecific pollen, such that the relative fertilization success of an allospecific pollen grain is h = 0.05 compared to a conspecific pollen grain. Our computations suggest that in such a case the realized hybridization rate should be 17% when pollen is non-limiting, 21% for an average density of 20 competing pollen grains per ovule, 44% with five pollen grains per ovule and up to 72% (i.e. close to the maximum value of 80% expected in the absence of barrier) when there is on average only one pollen grain per ovule (Fig. 3).

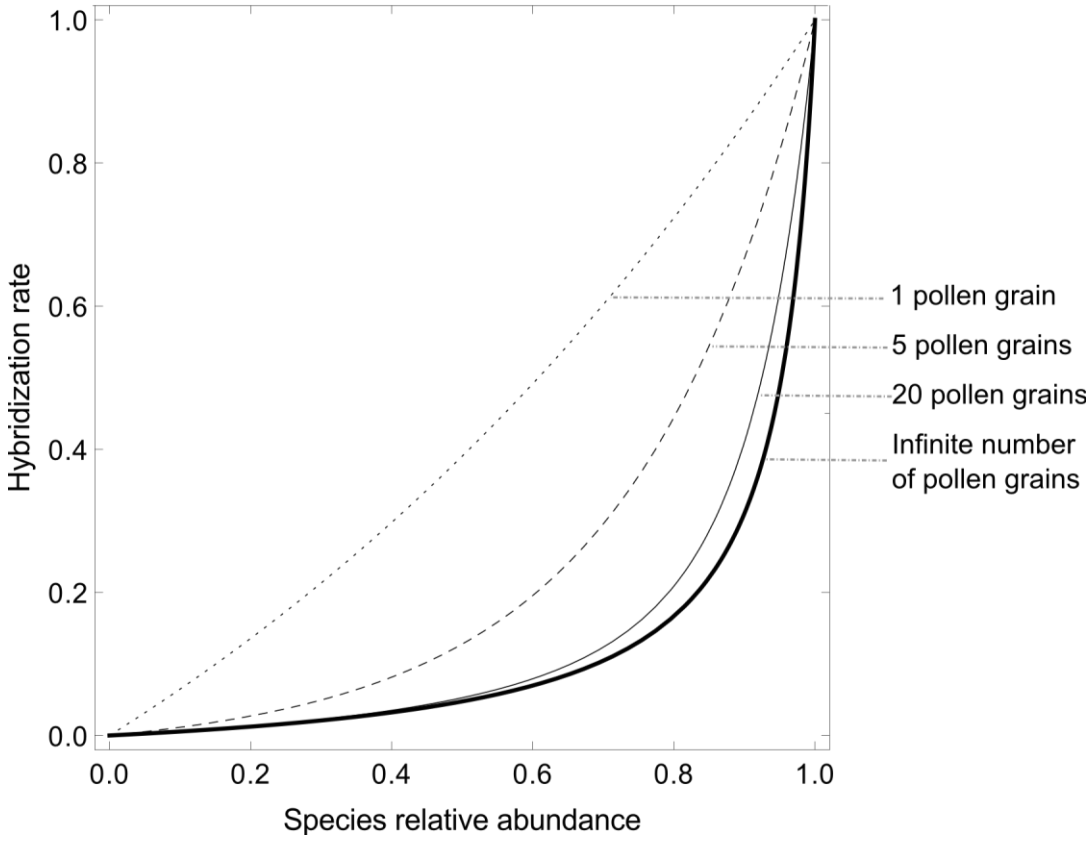


***Figure 3:* Effect of pollen quantity on hybridization rate**. *The thick black curve represents the hybridization rate of a mother tree when pollen is non-limiting and hybridization barrier is h=0.05. The solid, dashed and dotted curves represent hybridization rates for mother trees receiving limited numbers of pollen grains per ovule (20, 5 and 1 pollen grain(s) per ovule, respectively).*

DISCUSSION

Our study combines features of spatially-explicit mating models and of mass action models to predict intra- and interspecific crosses at the individual mother tree level. This requires estimating pollen dispersal curves for each species and taking into account the locations of all individuals to compute the composition of the pollen pools sampled by the mother trees. While some studies have compared individual pollen pool composition (e.g. Lepais & Gerber 2011; Smouse *et al.* 2001), empirical studies directly focusing on individual differences in pollen immigration and hybridization rates are rare. One exception is the study of Field et al. (2011) who investigated the effect of local demographic factors and flowering phenology on hybridization between two species of *Eucalyptus*. Here, we have moved one step further by modelling explicitly individual differences in mating system. In particular, our model allows immigrant pollen to compete with pollen produced by local trees. This resulted in a better fit of observations with predictions compared to a model assuming homogenous pollen immigration rate across mother trees. Focusing on individual mating behaviour illustrates the considerable tree-to-tree heterogeneity in immigration and hybridization rates. As immigration and hybridization rates should depend in a complex way on the spatial distribution of trees, as well as on stand size and configuration, this indicates that comparisons of the observed proportions of mating events across studies are unlikely to be informative. Instead, efforts should focus on comparisons among estimated model parameters, as their estimation can take into account stand characteristics. In what follows, we illustrate the power of the modelling approach proposed for understanding gene flow and hybridization in plants.

Pollen dispersal: Accurate estimation of pollen dispersal is needed to determine the spatial scale at which neighbours influence hybridization rate. We found that the pollen of *Q. robur* travels greater distances than that of *Q. petraea*. On average, a pollen grain of *Q. robur* was estimated to travel twice as far inside the stand as a pollen grain of *Q. petraea*. Jensen *et al*. (2009) reported a similar difference between species in pollen dispersal. Measurements of pollen grain sizes are consistent with this finding as *Q. robur* pollen is smaller than *Q. petraea* pollen (Rushton 1976), and is therefore more likely to travel greater distances (Niklas 1985). In addition, spatial genetic structure is lower at short distances in *Q. robur* than in *Q. petraea*, further pointing to higher pollen dispersal ability in *Q. robur* than in *Q. petraea* (Jensen *et al.* 2009; Streiff *et al.* 1998).

Sexual barriers: We found asymmetric barriers between these oak species, with *Q. robur* mother trees characterized by a lower barrier to hybridization than *Q. petraea* mother trees, confirming previous findings of asymmetric hybridization in the field (e.g. Bacilieri *et al.* 1996b). However, inferring sexual barriers from field-based data on hybridization without explicitly accounting for species relative abundance is risky, as relative abundance has a major effect on hybridization rate (Lepais *et al.* 2009). Similarly, studies based on controlled crosses (e.g. Steinhoff 1993) cannot provide accurate estimates of the strength of sexual barriers under natural conditions if they do not consider pollen competition, a major component of sexual isolation (e.g. Abadie *et al.* 2011; Howard 1999). In contrast, our field study naturally incorporates the pollen environment of each mother tree in the estimates of the sexual barriers. As the offspring analysed were saplings rather than seeds, the estimated parameters correspond in principle to the action of both prezygotic and postzygotic factors. Previous investigations have shown that in these oak species prezygotic barriers are more important than postzygotic barriers (seed germination and progeny fitness-related traits) (Abadie *et al.* 2011). This pattern is quite general among flowering plants. Lowry *et al.* (2008)

reviewed 19 cases and showed that on average prezygotic isolation is approximately twice as strong as postzygotic isolation. As a consequence, we expect that observed hybridization patterns should largely reflect processes acting at the time of mating, even though differences in seed and seedling viability might also play a role. To our knowledge, our study is the first to use parentage studies to directly estimate sexual barriers between hybridizing species in the field.

Effect of species distribution in the stand: The mean-field model yields estimates of hybridization rates close to predictions based on the spatially-explicit mating model. Compared with model predictions for random intermixing of the two species, the expected hybridization rate for the studied stand (i.e. using observed species distribution) is reduced. It would be further reduced if the two species had been completely spatially segregated. The stand encompasses a slight slope with *Q. robur* trees mainly at the bottom of the slope and *Q. petraea* trees at the top (Bacilieri *et al.* 1995). The two species have well-known differences in drought tolerance, *Q. robur* trees being less resistant to drought and better adapted to anoxia (Le Provost *et al.* 2011). This difference in species ecological niches probably caused the patchy spatial distribution, which in conjunction with limited pollen dispersal constitutes a first extrinsic barrier to hybridization. These results point to the need to move beyond the traditional sympatric versus allopatric speciation debate and to estimate the actual interspecific mating opportunities in hybridization studies (Mallet *et al.* 2009).

Hybridization rate and pollen limitation: We found a low rate of hybridization compared with previous studies of these species (e.g. Jensen *et al.* 2009; Lepais & Gerber 2011) or with other studies on related tree species (e.g. Gallo 1997; Kennington & James 1997; Marchelli & Gallo 2001). Several non exclusive factors could account for this result. First, from a technical standpoint, the precision of species assignment could affect the apparent rate of hybrid production, as shown in Vähä and Primmer (2006). Second, postzygotic selection might have further reduced the proportion of hybrids in our study, although there are indications that oaks hybrids are not strongly counter-selected, as discussed above. Third, the studied stand has an even species composition and is part of a large and relatively continuous mixed oak forest, which should ensure abundant production of conspecific pollen from both species and thus limit hybridization, compared to other situations where species differ more in relative abundance or where the stand is more isolated (see e.g. Field *et al.* 2008; Jensen *et al.* 2009; Lepais & Gerber 2011; Lepais *et al.* 2009). To counterbalance the lack of power caused by the low hybridization rate, we sampled a very large cohort of offspring and carried out a high-resolution paternity analysis.

Overall, our model, which accurately predicted the bulk of mating events, underestimated hybridization rates. This could be due to the omission of important individual parameters concerning trees inside the site such as flowering phenology (Slavov *et al.* 2005) or fecundity (Oddou-Muratorio *et al.* 2005). However, only predictions of hybridization rate with pollen donors located outside the stand were problematic. When pollen donors were located inside the stand, the observations regarding hybridization were in line with predictions. The ratio of observed versus expected hybridization rate was 1.1 on *Q. robur* mothers, i.e. very close to the expected value of 1. Furthermore, no hybrids sired by local fathers were produced on *Q. petraea* mother trees, as expected given the very strong sexual barriers in this species. So we have no reason to believe that the failure of the model is caused by the omission of some important characteristics of the adult trees located in the stand. Interestingly, our study was not the first to report higher hybridization rates with immigrant pollen than with local pollen (see e.g. Field *et al.* 2011 in Eucalyptus; Lepais & Gerber 2011 in oaks). In principle, higher

hybridization rates with immigrant pollen could be due to different species relative abundance inside and outside the stand. In our study, however, the specific composition of the immigrant pollen pool was accounted for when predicting hybridization rates and yet observed hybridization rates with immigrant pollen were larger than predictions in each of the two species. High pollen immigration rate is indicative of lower pollen load size as immigration rate is inversely proportional to the amount of pollen produced locally. We therefore suggest that the larger hybridization rate found with immigrant pollen than with local pollen in both species results at least partly from local pollen limitation.

Oaks, like many other plants, are known to experience pollen limitation (Knapp *et al.* 2001; Koenig & Ashley 2003). Pollen limitation can be caused by a variety of factors, including low density of conspecifics, poor climatic conditions or early/late flowering of seed trees (Knight *et al.* 2005). Effects of pollen load size on intraspecific pollen competition have been investigated experimentally (e.g. Mitchell 1997) and theoretically (Gregorius 1989). Both approaches indicate that the success rate of poor quality pollen increases at low pollen density. Basically, when low numbers of pollen grains compete for fertilization of a single ovule, pollen competition becomes less effective as chance pollination events start to predominate over deterministic competition (El-Kassaby & Ritland 1992). Our results suggest that this mechanism could also apply to the competition between allospecific and conspecific pollen, thus explaining the underestimation of hybridization rate with immigrant pollen.

In our study, all model parameters were estimated using the mass action law, i.e. assuming that pollen is not limiting. As this assumption was probably not met, the estimated sexual barriers probably depend not only on intrinsic factors but also on extrinsic factors such as pollen density. Recent experimental results show that oak sexual barriers depend both on genetic effects and on plastic responses to micro-environmental heterogeneity (Abadie *et al.* 2011). A challenging solution to disentangle the effect of these factors on hybridization would be to model absolute pollen amounts received by female flowers using an approach similar to that illustrated in Fig. 3. Note however that our conclusion regarding the effect of stand configuration (i.e. clustered or random species distribution) on hybridization rate, which is based on the hypothesis that pollen is not limiting, should remain qualitatively the same.

Importance of individual-based approaches in hybridization studies: A fundamental issue in the analysis and modelling of any system is the choice of the level of details (1994). For practical reasons, ecological data are often analysed at a scale different from that at which the process of interest operates. While in some conditions details of the system can safely be ignored, in other situations it is crucial to focus on discrete individuals in a spatially explicit way. Specifically, when interactions among individuals are nonlinear, ignoring spatial heterogeneity or using deterministic approaches that implicitly assume large population size can provide results different from those of corresponding individual-based models (1994). For instance, focusing on species niches, Clark *et al.* (2011) showed that the analysis of aggregated data at a higher scale than that at which the critical ecological process operates can be very misleading. They concluded that *the ideal solution is often to 'analyse, then aggregate', rather than 'analyse the aggregate'*. In our study, we found that, due to the non-random distribution of the trees in the stand, the average hybridization rate of the population differs from the hybridization rate of a tree exposed to the average pollen environment. Such findings support conclusions that spatial (as well as environmental or genetic) variations among individuals cannot always be ignored in demographic models (Boyce *et al.* 2006; Hughes *et al.* 2008). To account for the effect of pollen limitation, we

further showed that the finer scale of pollen grains competing for individual ovules should be considered. As pollen density decreases, the assumption of large population size no longer holds: selective forces underlying sexual isolation become less important relative to stochastic forces. Specifically, the concave relationship predicted between hybridization and the proportion of allospecific pollen becomes increasingly linear (Fig. 3), resulting in considerably higher hybridization rates than when pollen is not limiting. Therefore, our study illustrates the importance of using both individual-based and spatially explicit approaches to develop a mechanistic understanding of hybridization.

CONCLUSIONS

In the wake of human-induced global change, species distribution, proportions and abundances will be modified. Such conditions should generally increase hybridization rates (Allendorf *et al.* 2001). An approach to understanding hybridization that explicitly integrates environmental factors is therefore warranted. However, while the importance of an ecological approach to hybridization is increasingly acknowledged (e.g. The Marie Curie Speciation Network 2012), few empirical studies go beyond characterizing patterns. Our approach, based on the estimation of process parameters for discrete individuals in a spatially explicit context, helps explain why hybridization rates often increase under altered conditions. First, most types of disturbance should disrupt pre-existing spatial structure resulting from colonization history or local adaptation. This should typically decrease the relative abundance of conspecifics around mother plants, thus increasing hybridization rate. Specifically, our simulations show that randomizing species distribution (which we assimilate to a strong disturbance) should increase hybridization rates, compared to the absence of disturbance (i.e. leaving the original stand configuration untouched). Second, most types of disturbances should decrease the absolute density of conspecifics. Our study suggests that this should also increase hybridization rate through a decreased intensity of pollen competition, one of the major sexual barrier in plants (Howard 1999). In general, in most organisms, an improved appreciation of the importance of the environmental context of hybridization should benefit to speciation studies, given the prevalence of dispersal limitation and its consequences for mating opportunities.

ACKNOWLEDGEMENTS

We are grateful to Alexis Ducousso who established the Petite Charnie progeny test and shared information on the stand, and for his help, together with Stefanie Wagner, during sampling. Patrick Léger greatly helped with microsatellite genotyping. Pauline Garnier-Géré provided support during the design of the 384plex SNP chip. The genotyping was performed at the Genome-Transcriptome facility of Bordeaux (grants from the Conseil Régional d'Aquitaine n°20030304002FA, n°20040305003FA and from the European Union, FEDER n°2003227). We thank Christophe Boury for his contribution with SNP genotyping. Funding was provided by the LinkTree project (ANR BIODIVERSA) and by the EU Network of Excellence EvolTree. We thank Pauline Garnier-Géré, Sylvie Oddou-Muratorio, Alain Franc and Sophie Gerber for helpful discussions and Arndt Hampe, Olivier Lepais, Stéphanie Mariette and Réjane Streiff for critical reading of the manuscript.

AUTHOR'S CONTRIBUTIONS:

RJP initially conceived the study, which evolved significantly with the help of all the authors. EG analyzed genotyping data on parental ramets and checked their identity. LL performed the experiments, produced and analyzed the data. EK performed the modelling. LL wrote the paper with the help of RJP, EK wrote part of the Methods and Results section and all four authors reviewed the complete manuscript.

REFERENCES

Abadie P, Roussel G, Dencausse B*, et al.* (2011) Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *Journal of Evolutionary Biology* **25**, 157-173.

Adams WT (1992) Gene dispersal within forest tree populations. *New Forests* **6**, 217-240.

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology & Evolution* **16**, 613-622.

Arnold ML (1997) *Natural hybridization and evolution* Oxford University Press.

Austerlitz F, Dick CW, Dutech C*, et al.* (2004) Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology* **13**, 937-954.

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

Bacilieri R, Ducousso A, Kremer A (1996a) Comparison of morphological characters and molecular markers for the analysis of hybridization in sessile and pedunculate oak. *Annales Des Sciences Forestieres* **53**, 79-91.

Bacilieri R, Ducousso A, Petit RJ, Kremer A (1996b) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution* **50**, 900-908.

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

Boyce MS, Haridas CV, Lee CT, the NSDWG (2006) Demography in an increasingly variable world. *Trends in Ecology & Evolution* **21**, 141-148.

Burczyk J, Adams WT, Moran GF, Griffin AR (2002) Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. *Molecular Ecology* **11**, 2379-2391.

Chan KMA, Levin SA (2005) Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution* **59**, 720-729.

Clark JS, Bell DM, Hersh MH*, et al.* (2011) Individual-scale variation, species-scale differences: inference needed to understand diversity. *Ecology Letters* **14**, 1273-1287.

DiFazio SP, Leonardi S, Slavov GT*, et al.* (2012) Gene flow and simulation of transgene dispersal from hybrid poplar plantations. *New Phytologist* **193**, 903-915.

Durrett R, Levin S (1994) The importance of being discrete (and spatial). *Theoretical Population Biology* **46**, 363-394.

El-Kassaby YA, Ritland K (1992) Frequency-dependent male reproductive success in a polycross of Douglas fir. *Theoretical and Applied Genetics* **83**, 752-758.

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.

Field DL, Ayre DJ, Whelan RJ, Young AG (2011) The importance of pre-mating barriers and the local demographic context for contemporary mating patterns in hybrid zones of *Eucalyptus aggregata* and *Eucalyptus rubida*. *Molecular Ecology* **20**, 2367-2379.

Focke WO (1881) *Die Pflanzenmischlinge.* Bornträger, Berlin.

Gallo LAM, P.; Breitembucher, A. (1997) Morphological and allozymic evidence of natural hybridization between two southern beeches (*Nothofagus spp*.) and its relation to heterozygosity and height growth. *International Journal of Forest Genetics* **4**, 15-23.

Gregorius HR (1989) *Characterization and analysis of mating systems* Ekopan Verlag, Witzenhausen.

Guichoux E, Garnier-Géré P, Lagache L*, et al.* (2012) Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, *in press (MEC-12-0795.R0791)*.

Guichoux E, Lagache L, Wagner S*, et al.* (2011a) Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**, 591-611.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011b) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources* **11**, 578-585.

Heinze B (2011) Towards a quantitative description of landscape, demography and flowering phenology effects on realized hybridization potential. *Molecular Ecology* **20**, 2233-2235.

Hersch EI, Roy BA (2007) Context-dependent pollinator behavior: an explanation for patterns of hybridization among three species of Indian paintbrush. *Evolution* **61**, 111-124.

Holsinger KE (1991) Mass-action models of plant mating systems: the evolutionary stability of mixed mating systems. *The American Naturalist* **138**, 606-622.

Howard DJ (1999) Conspecific sperm and pollen precedence and speciation. *Annual Review of Ecology and Systematics* **30**, 109-132.

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

Hughes AR, Inouye BD, Johnson MTJ, Underwood N, Vellend M (2008) Ecological consequences of genetic diversity. *Ecology Letters* **11**, 609-623.

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**.

Kennington WJ, James SH (1997) The effect of small population size on the mating system of a rare clonal mallee, *Eucalyptus argutifolia (Myrtaceae)*. *Heredity* **78**, 252-260.

Knapp EE, Goedde MA, Rice KJ (2001) Pollen-limited reproduction in blue oak: implications for wind pollination in fragmented populations. *Oecologia* **128**, 48-55.

Knight TM, Steets JA, Vamosi JC*, et al.* (2005) Pollen limitation of plant reproduction: pattern and process. *Annual Review of Ecology, Evolution, and Systematics* **36**, 467-497.

Koenig WD, Ashley MV (2003) Is pollen limited? The answer is blowin' in the wind. *Trends in Ecology & Evolution* **18**, 157-159.

Lamont BB, He T, Enright NJ, Krauss SL, Miller BP (2003) Anthropogenic disturbance promotes hybridization between *Banksia* species by altering their biology. *Journal of Evolutionary Biology* **16**, 551-557.

Le Provost G, Sulmon C, Frigerio JM*, et al.* (2011) Role of waterlogging-responsive genes in shaping interspecific differentiation between two sympatric oak species. *Tree Physiology* **32**, 119-134.

Lepais O, Gerber S (2011) 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.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 3009-3021.

Mallet J, Meyer A, Nosil P, Feder JL (2009) Space, sympatry and speciation. *Journal of Evolutionary Biology* **22**, 2332-2341.

Marchelli P, Gallo LA (2001) Genetic diversity and differentiation in a southern beech subjected to introgressive hybridization. *Heredity* **87**, 284-293.

Mitchell RJ (1997) Effects of pollen quantity on progeny vigor: evidence from the Desert Mustard *Lesquerella fendleri*. *Evolution* **51**, 1679-1684.

Niklas K (1985) The aerodynamics of wind pollination. *The Botanical Review* **51**, 328-386.

Oddou-Muratorio S, Klein EK, Austerlitz F (2005) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. II. Pollen dispersal and heterogeneity in mating success inferred from parent–offspring analysis. *Molecular Ecology* **14**, 4441-4452.

Petit RJ, Bialozyt R, Garnier-Géré P, Hampe A (2004) Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecology and Management* **197**, 117-137.

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

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27**, 83-109.

Rieseberg LH, Carney SE (1998) Plant hybridization. *New Phytologist* **140**, 599-624.

Rushton B (1993) Natural hybridization within the genus *Quercus* L. *Annals of Forest Science* **50**, 73s-90s.

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

Seehausen OLE, Takimoto G, Roy D, Jokela J (2008) Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Molecular Ecology* **17**, 30-44.

Slavov GT, Howe GT, Adams WT (2005) Pollen contamination and mating patterns in a Douglas-fir seed orchard as measured by simple sequence repeat markers. *Canadian Journal of Forest Research* **35**, 1592-1603.

Smouse PE, Dyer RJ, Westfall RD, Sork VL (2001) Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution* **55**, 260-271.

Steinhoff S (1993) Results of species hybridization with *Quercus robur* L and *Quercus petraea* (Matt.) Liebl. *Annals of Forest Science* **50**, 137s-143s.

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.

The Marie Curie Speciation Network (2012) What do we need to know about speciation? *Trends in Ecology & Evolution* **27**, 27-39.

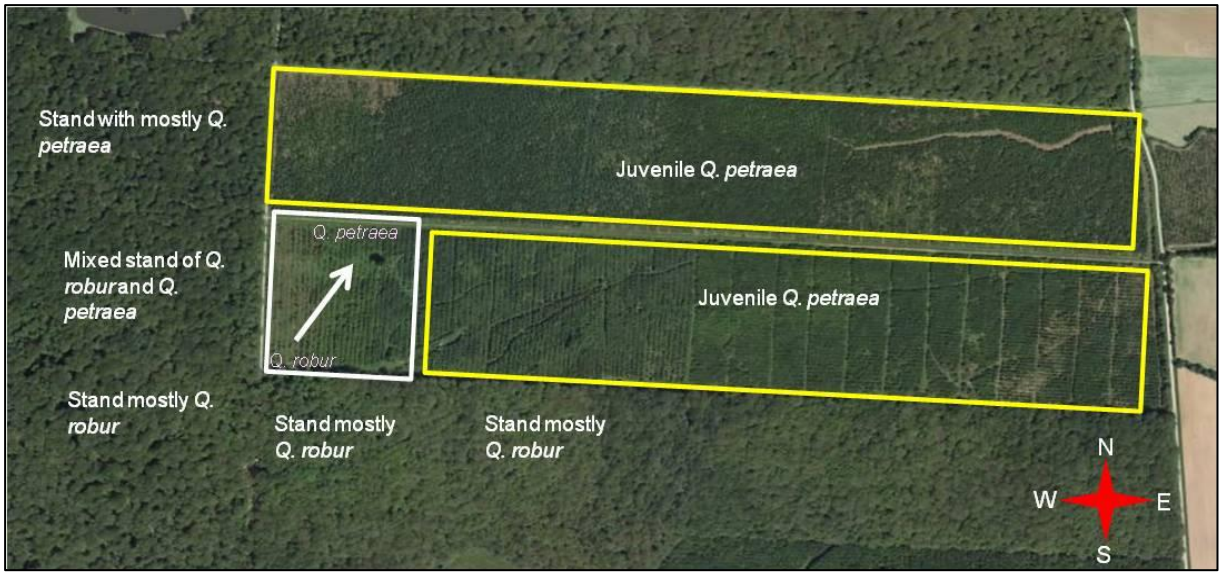
Vähä J-P, Primmer CR (2006) Efficiency of model-based bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology* **15**, 63-72.

Wirtz P (1999) Mother species–father species: unidirectional hybridization in animals with female choice. *Animal Behaviour* **58**, 1-12.

Wolfram Research Inc. (2010) *Mathematica Edition: Version 8.0* Wolfram Research, Inc., Champaign, Illinois.

SUPPORTING INFORMATION

SUPPORTING INFORMATION 1: AERIAL PHOTOGRAPH OF THE PARENTAL STAND AND OF NEIGHBORING STANDS WITH INDICATIONS ON THE DOMINANT SPECIES COMPOSITION



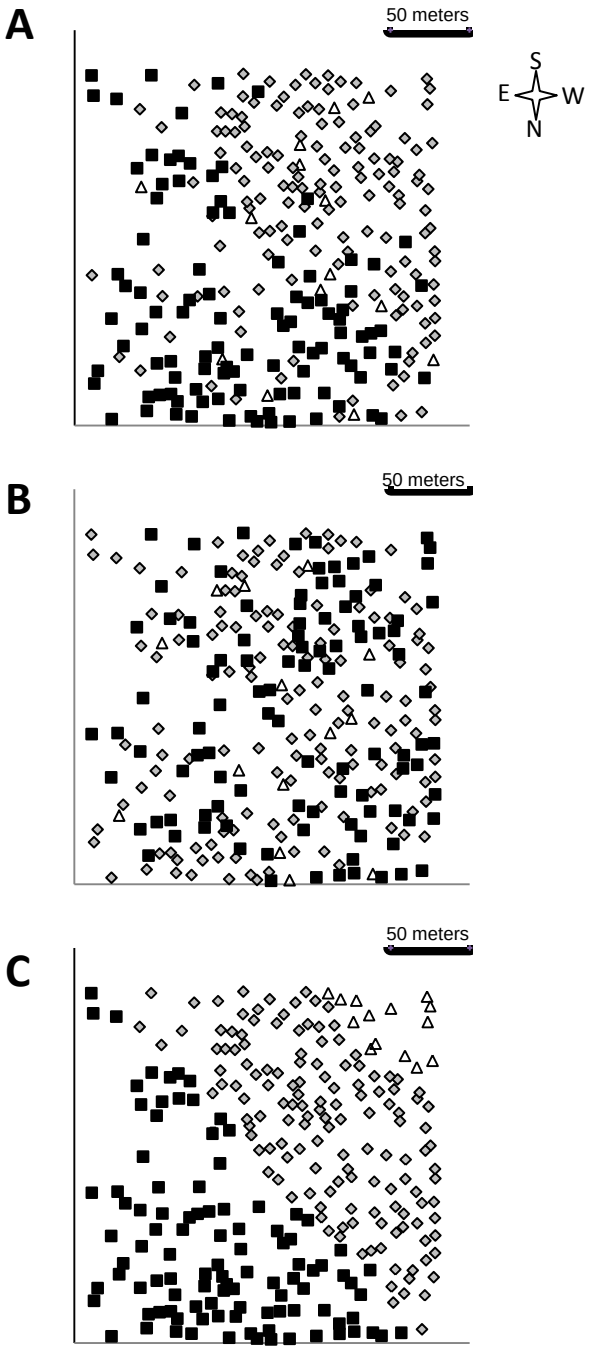
The white square corresponds to the parental stand and the yellow rectangles to recently naturally regenerated stands.

SUPPORTING INFORMATION 2: VALIDATION OF SSR GENOTYPES THROUGH MATERNITY AND PATERNITY ANALYSES

Four of the 51 mother trees were not represented by grafts in the nursery but their genotype could be unambiguously reconstructed from the genotypes of their offspring. For the remaining progenies, we checked if the genotype of each offspring matched with the corresponding mother genotype. If a mismatch was found at more than one locus (out of 12), the offspring was excluded from the analysis. If only one mismatch was found, we re-genotyped and read again the SSR profiles of the mother/offspring couple. If these new analyses validated the mismatch, we excluded the offspring from the final dataset.
We then used CERVUS (Marshall *et al.* 1998) to carry out a first paternity analysis. When the most likely father identified had a high SSR LOD score (>8) and exactly one mismatch with its offspring, we genotyped again the father/offspring pair to correct for any possible genotyping errors. We then used the subset of offspring genotyped at SNPs in a second paternity analysis with CERVUS. We found 144 trios (father/mother/offspring) with a high SNP LOD score (>40). If an inferred paternity relationship had a poor SSR LOD score (<8) and one or more mismatches with a given offspring but had a high SNP LOD score, we re-genotyped the corresponding father/offspring pair using SSRs. In this way, the SSR genotype of 90 fathers out of 260 (i.e. most of those with high reproductive success) could be controlled and corrected if needed.
In principle, discarding potential parental trees on the basis of a single mismatch could prove unjustified, as mutations can occur at SSRs. Actually, by comparing with the paternity analysis based on SNPs, we identified one father-tree with a high SNP LOD score that had two offspring presenting the same mismatch at one SSR locus. We interpreted this as a result of a mutation in the crown of the adult paternal tree; for subsequent analyses, we assigned to this tree the genotype matching with the two offspring. In any case, discarding completely a few offspring on the basis of a single mismatch with its mother is a conservative procedure that might eliminate a few legitimate offspring as well as many illegitimate ones but that should not create biases.
One possible concern with our approach is that, while all offspring benefitted from a quality control on the basis of allelic transmission from their mother, only part of them could benefit from a similar quality control from the father's side, as the father of only a fraction of the offspring (50-60%) could be identified.

However, we decided to correct all possible genotypes as we were interested in improving the estimates of model parameters, which depend on the likelihood of the inferred parentages (following prescription of Oddou-Muratorio *et al.* 2003). Since the model used relies on fractional paternity analyses, true fathers having remaining mismatches with a given offspring (caused by genotyping errors or by a mutation) would still play a role in parameter estimates.

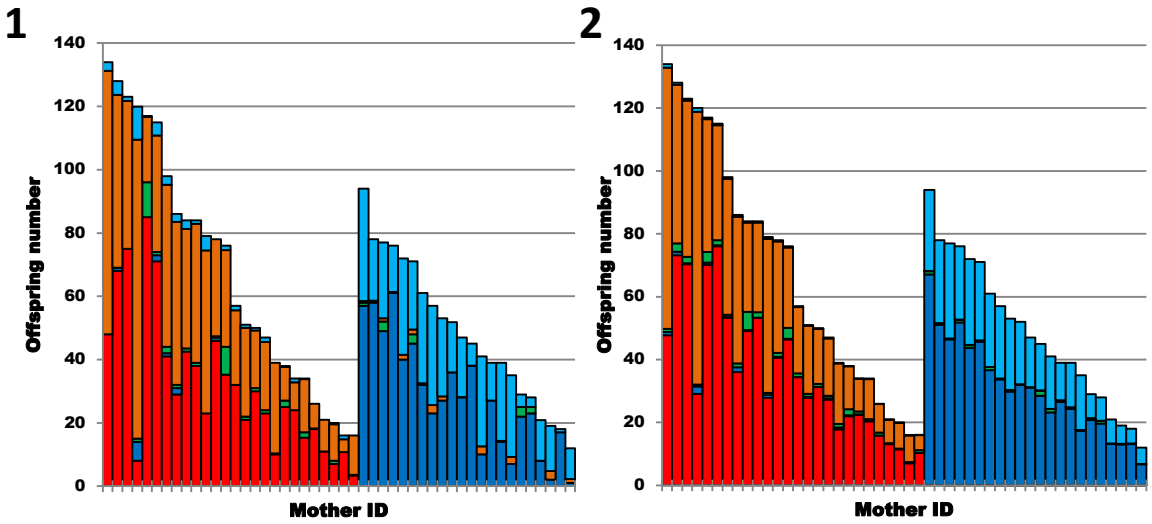
Marshall T.C., Slate J., Kruuk L.E.B. & Pemberton J.M. (1998). Statistical confidence for likelihood based paternity inference in natural populations. *Mol. Ecol.*, 7, 639-655.
Oddou-Muratorio S., Houot M.L., Demesure-Musch B. & Austerlitz F. (2003). Pollen flow in the wildservice tree, Sorbus torminalis (L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. *Mol. Ecol.*, 12, 3427-3439.



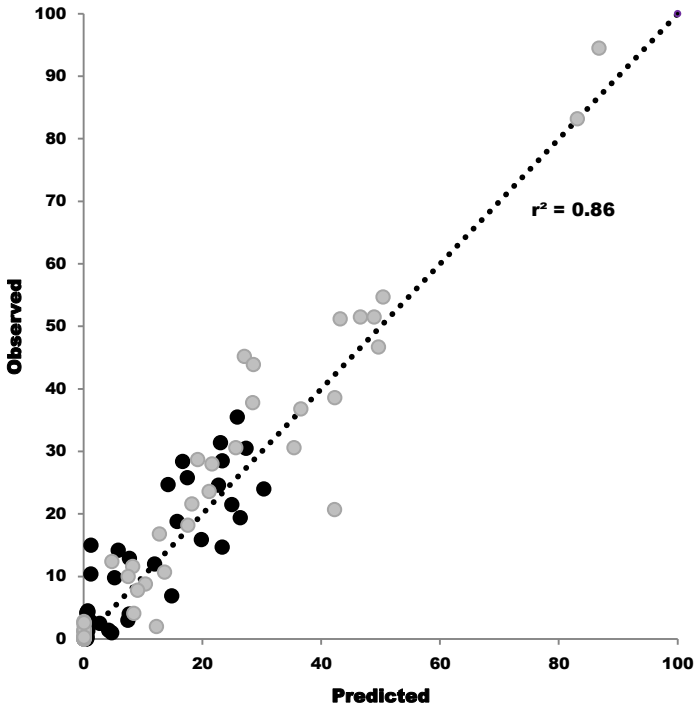
Quercus robur trees are represented by grey diamonds, *Q. petraea* trees by black squares and intermediate trees by white triangles (genetic species assignment). **A:** Actual spatial distribution of the trees. **B:** One of the 100 input files corresponding to random spatial distribution of species. **C:** Disjoint spatial distribution of species in the stand.

SUPPORTING INFORMATION 4: COMPARISON OF OBSERVED AND PREDICTED NUMBERS OF IMMIGRANT AND HYBRID OFFSPRING

A: Numbers of offspring of each mother tree resulting from *Q. petraea* pollen coming from inside or outside the stand (red and orange bars, respectively), of *Q. robur* pollen coming from inside or outside the stand (dark and light blue bars, respectively), and of pollen from admixed trees coming from inside the stand (green). 1) Observed results. 2) Predicted results.



B: Comparison between observed and predicted numbers of offspring originating from immigration for each mother tree.



Black circles represent the progeny of *Q. petraea* and grey circles those of *Q. robur*.

SUPPORTING INFORMATION 5: CHECKING THAT OBSERVED HYBRIDIZATION RATE IS STRONGER WITH IMMIGRANT POLLEN THAN WITH LOCAL POLLEN

While fathers located inside the stand had been genotyped with a larger number of markers (12 SSRs + 262 SNPs) and could be directly assigned to species, the paternal species assignment of offspring produced by unknown fathers is only based on the likelihood for the father to belong to one species knowing the genotype of the offspring, the genotype of the mother and the allelic frequencies of the two species. In particular, some of the hybrids identified with this method might have admixed fathers. It was thus necessary to check that the difference in hybridization rates between local crosses and crosses involving immigrant pollen is not an artifact of the difference in the resolution of the methods of species assignment. For this purpose, we checked species assignment of all fathers using the same method (i.e. we used the indirect approach to assign the species of the father, even when the father had been identified through paternity analysis). We then used a χ^2 test to compare the observed and expected number of hybrids as a function of the type of cross for the two species (Table 1). By comparing χ^2_{robur} and $\chi^2_{petraea}$ to $\chi^2_{theoretical}$ (ddl=1) we could conclude that there was significantly more hybrids from immigrant than from local pollen for both species.

Table 1: χ^2 test of the hypothesis that interspecific crosses are more frequent with immigrant pollen than with local pollen

Q. robur ♀	Interspecific crosses	Intraspecific crosses
Crosses with local pollen	19	867
Crosses with immigrant pollen	35	850

χ^2_{robur}= 4.9 (p<0.05).

Q. petraea ♀	Interspecific crosses	Intraspecific crosses
Crosses with local pollen	4	633
Crosses with immigrant pollen	13	414

$\chi^2_{petraea}$= 9.5 (p<0.005).

These results cannot be influenced by paternal assignment quality. They support the idea that hybrid offspring are more likely to have been sired by immigrant pollen than by locally produced pollen.

ORIGINE ET DEROULEMENT DE CE TRAVAIL

L'hybridation entre espèces végétales a parfois été perçue (notamment par les zoologistes) comme un phénomène artificiel. En effet, cette hybridation paraissait se produire surtout dans des environnements dits « perturbés » (ex. Wiegand 1935; Anderson 1948; Rieseberg *et al.* 1989). L'hybridation entre espèces semblait ainsi devoir être liée à l'action de l'homme (ex. Lamont *et al.* 2003; Parsons & Hermanutz 2006). Des perturbations édaphiques ont été particulièrement fréquemment évoquées. En modifiant l'environnement physique (sols remués etc.) on créerait ainsi des « environnements intermédiaires » favorables à des individus eux-mêmes intermédiaires entre les deux espèces parentales (Anderson 1948; Muller 1952). Parallèlement, d'autres observations ont mis en évidence un effet de l'abondance relative des espèces sur leur hybridation (ex. les observations de Focke 1881 sur les plantes; et de Hubbs 1955 sur les poissons). Cependant, pendant de nombreuses années, l'effet de l'environnement biotique sur l'hybridation a été largement négligé, peut-être parce que l'idée que l'hybridation est artificielle était si fortement ancrée qu'il ne semblait pas « fondamentalement » intéressant de comprendre dans quelles conditions elle apparaissait. Ce n'est que récemment que des études ont redécouvert l'effet fréquence-dépendant de la proportion des espèces sur leur hybridation (ex. McGowan & Davidson 1992; Travnichek *et al.* 1996; Young *et al.* 2001; Lepais *et al.* 2009). Cet effet a été modélisé de manière très simple par Chan & Levin (2005). Ces auteurs ont montré que l'hybridation dépend des proportions des deux espèces (ex. lorsque les barrières entre espèces sont symétriques, le nombre d'hybrides atteint un maximum pour une proportion de 50/50). Il est ainsi devenu progressivement de plus en plus clair que le contexte local devait être pris en compte pour comprendre les mécanismes d'hybridation entre espèces (Field *et al.* 2011; Heinze 2011; Jahner *et al.* 2011). Cependant les études sur l'effet de la proportion relative des espèces sur l'hybridation sont restées à une échelle populationnelle. Or, dans un peuplement d'arbres par exemple, chaque individu perçoit son environnement différemment et des études plus fines, au niveau individuel et non populationnel, sont nécessaires pour comprendre précisément les mécanismes d'hybridation en conditions naturelles (Clark *et al.* 2011). Mon étude s'intègre dans cette problématique. J'ai étudié à une échelle très fine, celle de l'individu, voire de la fleur, les mécanismes (exogènes et endogènes) contrôlant le croisement entre deux individus d'espèces différentes.

Lors de mon stage de Master II, j'ai terminé l'élaboration d'un kit 12plex de marqueurs microsatellites hautement polymorphes (Annexes 1 et 2) afin d'étudier à partir de 1200 descendants l'effet de l'environnement pollinique sur l'hybridation de deux espèces de chêne. Pour cette étude, j'avais dans un premier temps estimé à partir d'une recherche de paternité et des événements réalisés de reproduction les courbes de dispersion du pollen des deux espèces. Puis dans un second temps, j'avais utilisé ces courbes de dispersion pour estimer la proportion de pollen de chaque espèce arrivant sur les arbres mères étudiés. Enfin, j'avais mis en évidence une corrélation entre le nombre d'hybrides dans la descendance de chaque mère et la proportion de pollen allospécifique qu'elle recevait. Dans le cadre de ma thèse, nous souhaitions poursuivre ces travaux et les affiner avec un nombre plus important de descendants et l'ajout d'un paramètre de fécondité mâle (car jusqu'ici nous faisions une hypothèse forte que chaque arbre produit la même quantité de pollen) afin d'estimer le plus précisément possible la proportion de pollen de chaque espèce reçue par chaque arbre mère. En exposant ces résultats et ces perspectives lors du premier meeting du projet européen Linktree qui s'est déroulé deux mois avant le début de ma thèse, Sylvie Oddou-Muratorio (INRA Avignon) m'a conseillé de prendre contact avec

Etienne Klein pour l'analyse de ces données. Une fois l'échantillonnage des descendants complété, le génotypage et les lectures des génotypes effectués, j'ai pris contact avec Etienne pour lui exposer ce que je souhaitais faire. Il m'a alors proposé son aide pour modéliser les événements de reproduction dans notre parcelle, ce qui permettait d'améliorer de deux façons l'étude que j'avais faite précédemment. Dans un premier temps, les courbes de dispersion que j'avais construites à partir de croisements réalisés seraient modélisées en prenant en compte la distribution des arbres mères et **de tous les pères potentiels**. Dans un deuxième temps, cette modélisation permettrait de prendre en compte et d'estimer **simultanément** (et non séquentiellement comme je souhaitais le faire au départ) de nombreux paramètres affectant les croisements, tels que l'affectation aux espèces des individus, les courbes de dispersion du pollen, la fécondité mâle… Ainsi je pourrais connaitre assez précisément la proportion de pollen de chaque espèce que reçoit un arbre mère étudié.

J'ai donc passé par trois fois un peu moins d'une semaine à Avignon au cours de ma thèse pour discuter des objectifs de mon étude avec Etienne. Dans un premier temps, j'ai décidé d'élaborer un modèle avec le plus de paramètres possibles pour expliquer la reproduction intra- et interspécifique des deux espèces et prédire la proportion de pollen de chaque espèce que reçoit chaque arbre mère et ainsi le nombre d'hybrides. J'ai rassemblé toutes les données disponibles sur la reproduction des arbres de la parcelle (telle que la phénologie, le diamètre, la hauteur ; Voir **Chapitre 3**) auprès d'Alexis Ducousso et Jean-Marc Louvet (UMR Biogeco). Curieusement, la prise en compte de tous ces paramètres n'a pas permis de modéliser de façon satisfaisante l'hybridation dans le cas où le pollen provenait de l'extérieur de la parcelle. Les prédictions du nombre de descendants hybrides issus de pollen immigrant étaient en effet bien inférieures aux observations. De plus, la plus forte proportion qu'attendue d'hybrides issus de pollen immigrant (bien plus élevée que la proportion d'hybrides produits à partir de pollen local) m'a amené à considérer l'effet non seulement de la proportion de pollen de chaque espèce reçue par une fleur femelle, mais aussi sa quantité. Une faible quantité de pollen ne permet pas à la compétition pollinique de se manifester pleinement, les effets stochastiques devenant trop importants. Ce résultat, ainsi que des observations d'augmentation du taux d'hybridation dans des populations fragmentées suite à des perturbations environnementales (voir l'étude de Rieseberg *et al.* 1989 par exemple), m'ont permis d'affiner l'objectif de l'étude en me focalisant uniquement sur l'effet des modifications environnementales sur le taux d'hybridation de ces deux espèces. C'est ainsi que nous avons simplifié le modèle et ajouté des simulations sur l'effet de la distribution des espèces sur leur hybridation. J'ai alors décidé de garder les résultats du modèle complet pour une autre étude (celle du **chapitre 3**) dont l'objectif serait de comparer les comportements reproducteurs des deux espèces et d'étudier si ces différences peuvent être attribuées à leurs stratégies écologiques respectives.

PERSPECTIVES DE L'ETUDE

J'ai souhaité poursuivre cette étude en testant l'hypothèse soulevée par ce travail selon laquelle le pollen limitant diminuerait l'efficacité de la compétition pollinique et donc augmenterait l'hybridation. La diminution de la compétition pollinique devrait en principe aussi bien affecter les croisements intra- qu'interspécifiques (Charlesworth 1988). J'ai donc souhaité tester si la diminution supposée de compétition pollinique au niveau des croisements intraspécifiques avait un effet sur la vigueur des descendants. Suite à l'étude de Mitchell (1997) , j'ai émis l'hypothèse selon laquelle les descendants issus du pollen local seraient en moyenne plus compétitifs que ceux issus de pollen immigrant. Le fait qu'un

pollen immigrant ait fécondé la fleur pourrait en effet signifier que la quantité de pollen local était (en moyenne) plus faible et donc la compétition moins efficace. Pour tester cette hypothèse, j'ai souhaité analyser le test de descendance contenant les descendants, maintenant âgés de 17 ans des arbres de la parcelle étudiée. Ce test est constitué de douze blocs mono-spécifiques (six *Q. robur* et six *Q. petraea*) dans lesquels toutes les descendances d'une même espèce sont représentées au moins une fois par une parcelle unitaire de six individus ayant la même mère et étant donc au moins demi-frères. La hauteur de chaque descendant du dispositif a été mesurée en 2007. La hauteur observée pour chaque descendant peut être décomposée en une part génétique et une part environnementale. Il est donc possible, au travers d'un modèle approprié qui prend en compte les effets environnementaux sur la croissance des descendants, d'estimer la composante génétique de ce trait, et ainsi tester l'hypothèse selon laquelle le pollen local, supposé plus fortement soumis à la compétition pollinique, a conféré aux descendants une meilleure croissance et/ou une variance plus faible pour leur hauteur par rapport aux descendants issus de pollens immigrants.

Pour tester si l'origine du pollen (local ou immigrant) confère aux descendants une valeur génétique moyenne différente ou une variance génétique différente pour leur hauteur, il est nécessaire d'élaborer un modèle qui prenne en compte à la fois les effets environnementaux (Parcelles unitaires et Blocs), les effets génétiques (Apparentement entre individus) et la provenance des grains de pollen (local ou immigrant). Pour cela, j'ai suivi au cours de ma thèse une formation d'une semaine au logiciel ASReml, un logiciel statistique qui permet d'élaborer un modèle linéaire mixte (avec des effets fixes et des effets aléatoires) et d'estimer par maximum de vraisemblance les effets génétiques et environnementaux expliquant la variable étudiée. Cette formation m'a permis de me familiariser avec le logiciel. Cependant elle fût difficile à suivre, en partie parce que j'ai très peu abordé la génétique quantitative durant mes études et mes stages, et parce que l'utilisation du logiciel demande de s'approprier un langage assez complexe. Pour réaliser les modèles qui vont suivre, j'ai sollicité l'aide de Laurent Bouffier qui a soutenu une thèse intitulée « Evolution de la variabilité génétique dans les populations d'amélioration du Pin maritime et conséquences pour la sélection» au sein de l'unité. Il a notamment utilisé ce logiciel pour estimer les valeurs génétiques des arbres de la population d'amélioration de pin maritime.

J'ai souhaité que l'espèce de la mère soit incluse dans ce modèle pour pouvoir étudier l'effet de la compétition pollinique sur la croissance des descendants séparément pour chaque espèce. La limite de ce dispositif initialement mis en place pour comparer les deux espèces est que les blocs sont mono-spécifiques. Il y a donc une confusion des effets « blocs » et « espèce de la mère » (Figure 1). J'ai donc redéfini trois nouveaux blocs de plus grande taille (Figure 1) en essayant de prendre en compte l'hétérogénéité observée sur le terrain (pente nord / sud). Nous avons alors élaboré le modèle suivant :

Modèle A :

Hauteur(2007)= moyenne + NV bloc + ESP mère + Orig pollen + ID mère+ PU + résiduelle

Les paramètres en vert sont considérés comme des effets fixes, ceux en bleu comme aléatoires
Hauteur (2007)= *hauteur totale mesurée en 2007 (cm)*
NV bloc : *nouveaux blocs définis a posteriori*
ESP mère *: espèce de la mère (Pédonculé, Sessile ou Intermédiaire)*
Orig pollen : *Origine du pollen (immigrant ou local)*
ID mère : *identité de la mère, permettant de définir la famille*
PU : *Parcelle unitaire de 6 arbres consécutifs d'une même famille*

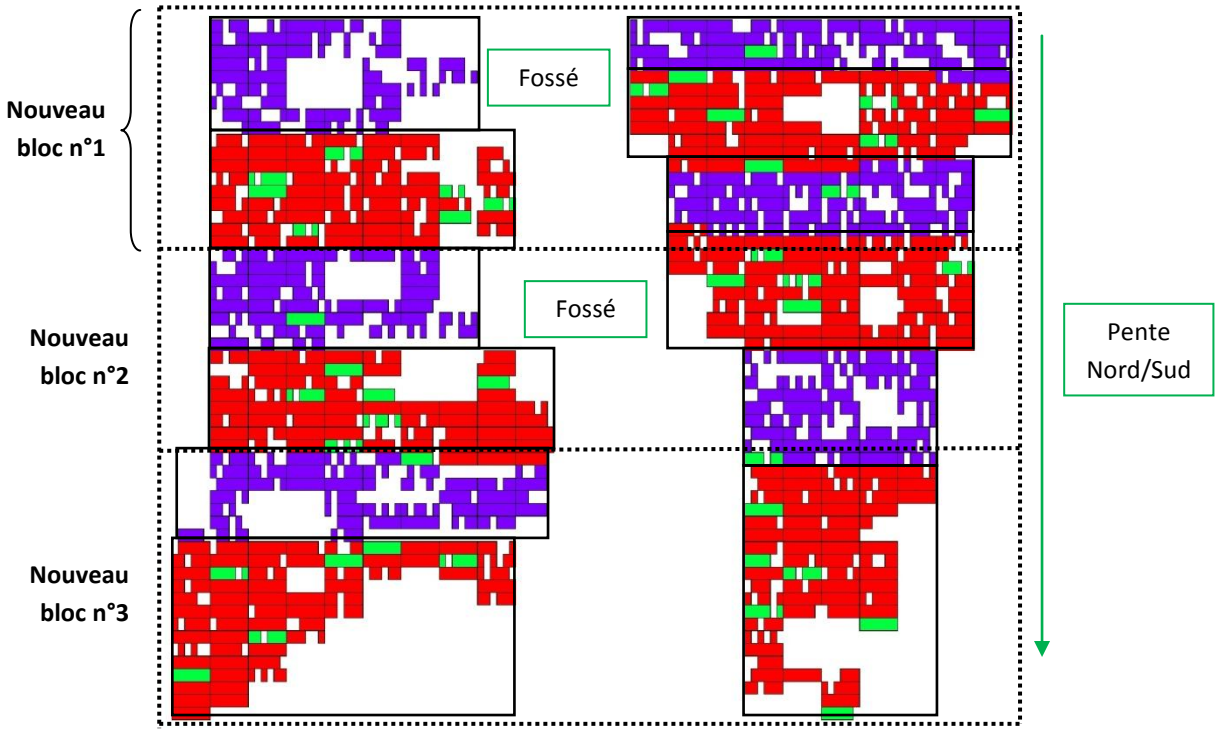


Figure 1 : Affectation des mères de tous les descendants du test de descendance*. Chaque petit rectangle constitue une parcelle unitaire d'au maximum 6 descendants issus de la même mère. Les parcelles unitaires en bleu représentent les descendants issus de mère sessile, celles en rouge, les descendants issus de mère pédonculé et celles en vert regroupent les descendants issus de mères intermédiaires. Les grands rectangles au contour noir représentent les blocs définis lors de la plantation. Horizontalement entre deux blocs se trouve un fossé et il existe une pente nord/sud dans le dispositif. Les rectangles aux contours en pointillé sont les trois nouveaux blocs définis a posteriori.*

Effets environnementaux :

Lors des récoltes sur le terrain effectuées durant ma thèse, j'ai remarqué un fort effet de l'environnement au sein d'un même bloc sur la hauteur des descendants. Mes premières observations suggéraient que les descendants proches des bords du dispositif avaient tendance à être plus grands que ceux à l'intérieur du dispositif au sein d'un même bloc. Ceci est confirmé par la représentation spatiale des hauteurs (Figure 2). Je me suis aperçue qu'effectivement la hauteur des individus au sein d'un même bloc (défini *a priori* ou *a posteriori*) n'est pas homogène. Il existe de nombreux effets : bordure, fossé, pente…, qui influencent la croissance des arbres. La définition des nouveaux blocs semble donc insuffisante pour prendre en compte l'effet de l'environnement sur la croissance des arbres. Une meilleure modélisation de l'effet de l'environnement sur la croissance des arbres apparaît donc nécessaire pour estimer correctement la composante génétique des croissances observées.

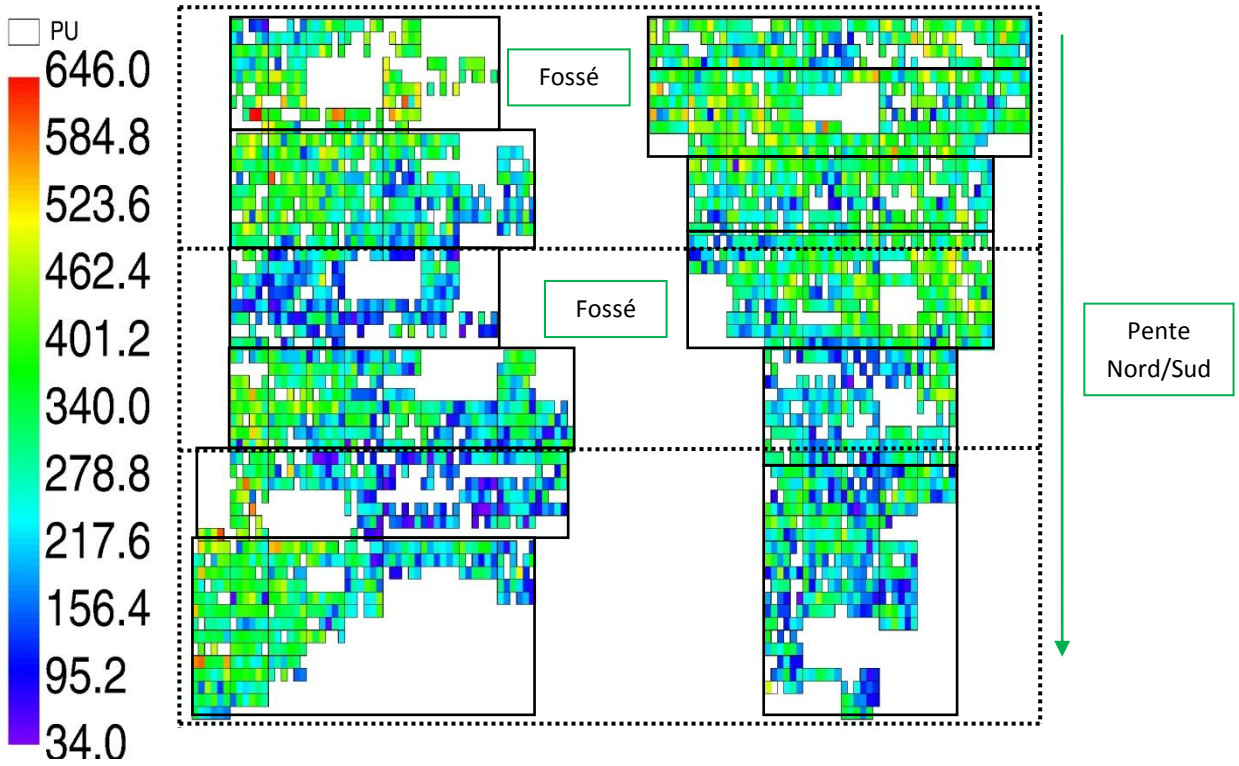


***Figure 2 :* Données brutes individuelles des hauteurs mesurées en 2007.** *Les couleurs déterminent la hauteur des arbres, plus elles sont chaudes et plus l'arbre est grand. Les grands rectangles au contour noir représentent les blocs définis lors de la plantation. Les nouveaux blocs définis a posteriori pour contenir les deux espèces sont symbolisés par des rectangles noirs en pointillé.*

J'ai donc demandé à Laurent s'il pouvait modéliser l'effet de l'environnement grâce à un ajustement spatial pour prendre en compte la non-indépendance des résidus du modèle A. Ce nouveau modèle intègre le fait que la croissance des arbres est partiellement corrélée à celle de leurs voisins. En effet, deux arbres proches ayant le même micro environnement

auront une croissance plus ou moins corrélée alors que deux arbres éloignés, possédant des microenvironnements différents, auront des croissances indépendantes. A l'inverse, lorsqu'il existe une forte compétition entre arbres, des arbres proches auront des croissances très contrastées alors que des arbres éloignés auront des croissances équivalentes. L'ajustement spatial pourra donc être négatif (entre -1 et 0) lorsque des arbres proches ont des croissances totalement différentes ou positif (entre 0 et 1) lorsque des arbres proches ont des croissances équivalentes. La résiduelle du modèle A est ici décomposée en deux composantes : une composante indépendante entre arbres et une composante spatiale qui intègre les auto-corrélations entre arbres. Cet ajustement est réalisé dans les deux dimensions du dispositif (X et Y). Le modèle par ajustement spatial (Figure 3) a une vraisemblance (LogL=-5125.9) nettement supérieure à celle du modèle A qui ne prend en compte que les effets « blocs » (LogL=-5321.5; test du χ^2 avec 3 ddl >>3.9). J'ai donc retenu le modèle intégrant un ajustement spatial.

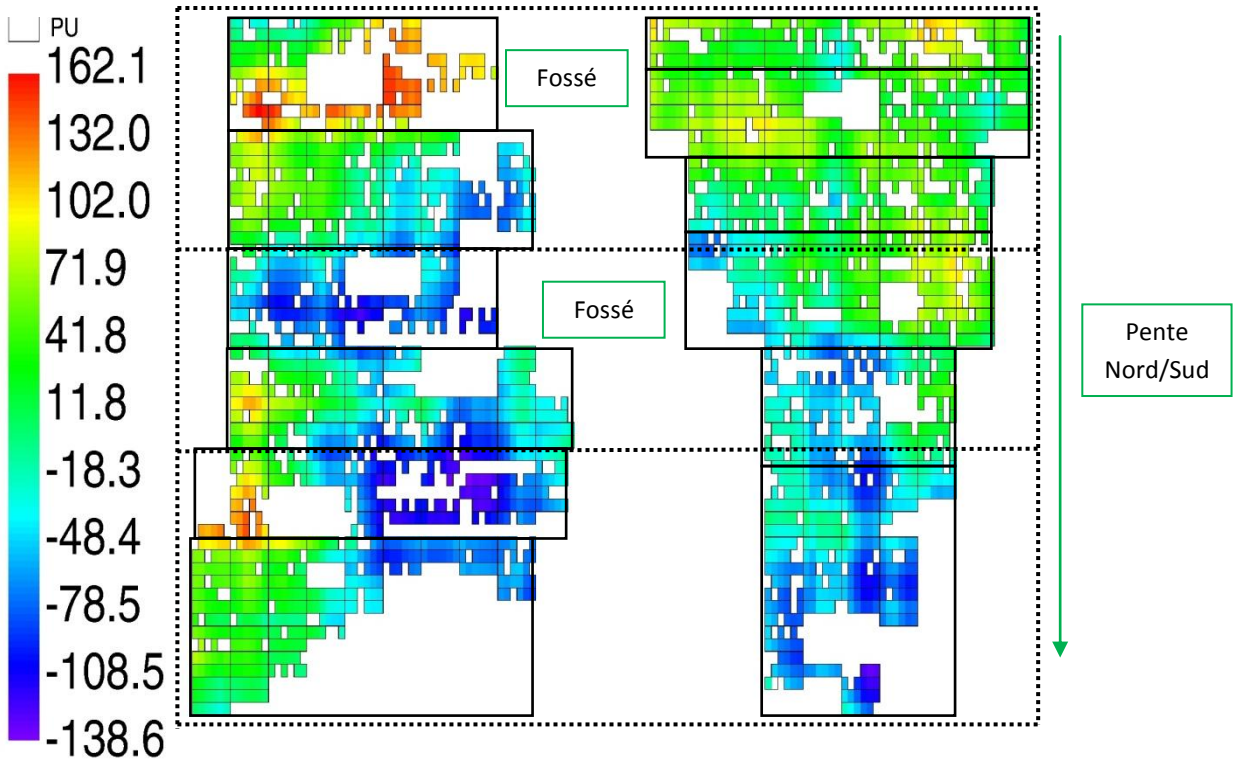


Figure 3 : ***Visualisation de la composante spatiale estimée par le modèle B pour la hauteur mesurée en 2007.*** *Les couleurs déterminent l'effet de l'environnement, plus elles sont chaudes et plus l'arbre est dans un environnement favorable par rapport à l'environnement moyen du dispositif. L'ajustement spatial permet d'estimer une auto-corrélation de 0, 89 selon l'axe des X et de 0,95 selon l'axe des Y.*

Modèle avec ajustement spatial : Modèle B

Hauteur(2007)= moyenne + NV bloc + ESP mère + Orig pollen + ID mère+ PU + AJUST spatial + résiduelle

Les paramètres en vert sont considérés comme des effets fixes, ceux en bleu comme aléatoires
Hauteur (2007)= *hauteur totale mesurée en 2007 (cm)*
NV bloc : *nouveaux blocs définis a posteriori*
ESP mère *: espèce de la mère (Pédonculé, Sessile ou Intermédiaire)*
Orig pollen : *Origine du pollen (immigrant ou local)*
ID mère : *identité de la mère, permettant de définir la famille*
PU : *Parcelle unitaire de 6 arbres consécutifs d'une même famille*
AJUST spatial : *composante spatiale*

Effet génétique :

Par défaut, pour estimer la composante génétique de la croissance des individus, on considère que tous les descendants d'une même mère sont issus de pères différents, c'est-à-dire qu'ils sont tous demi-frères (effet aléatoire ID mère du modèle). Or, il existe des cas où ces individus sont plus apparentés que des demi-frères. Connaissant les relations d'apparentement complètes pour la moitié des descendants (ceux pour lesquels les pères ont été retrouvés), j'ai proposé à Laurent d'intégrer dans le modèle l'apparentement réel des descendants pour l'améliorer. Il a alors construit un modèle mixte individuel (l'effet aléatoire n'est plus l'effet « famille » (ID mère) mais un effet « individu » (ID seeds)) où tous les individus sont reliés entre eux par une matrice d'apparentement. Le modèle intégrant tous les apparentements connus (modèle C décrit ci-dessous) présente une vraisemblance (LogL=-5108.6) supérieure au modèle précédent (LogL=-5125.9). J'ai donc choisi de conserver ce modèle car il permet de valoriser l'information complète des pédigrées des descendants.

Modèle avec apparentement réel : Modèle C

Hauteur(2007)= moyenne + NV bloc + Orig pollen + IDseeds + PU + AJUST spatial + résiduelle

Les paramètres en vert sont considérés comme des effets fixes, ceux en bleu comme aléatoires
Hauteur (2007)= *hauteur totale mesurée en 2007 (cm)*
NV bloc : *nouveaux blocs définis a posteriori*
Orig pollen : *Origine du pollen (immigrant ou local)*
ID seeds : *paramètre remplaçant* ***ID mère****, les individus sont reliés entre eux par une matrice d'apparentement dérivée du pédigrée de la population étudiée (ce pédigrée intègre l'espèce des mères donc* ***ESP mères*** *n'est plus déclaré comme effet fixe)*
PU : *Parcelle unitaire de 6 arbres consécutifs d'une même famille*
AJUST spatial : *composante spatiale*

Ayant ajouté tous les paramètres qui me semblaient importants pour modéliser la hauteur des descendants de ce test, le modèle C permet de tester si la composante génétique moyenne pour la hauteur est significativement différente entre les descendants issus de pollen local et les descendants issus de pollen immigrant. L'effet fixe « **Orig pollen** » n'étant pas significatif (P_{valeur}=0.17), nous pouvons conclure que l'origine du pollen (local ou immigrant) n'est pas responsable d'une différence de hauteur moyenne en 2007 entre les descendants, et ceci pour les deux espèces étudiées. L'origine du pollen ne semble donc pas modifier la composante génétique de la croissance des descendants.

Nous pouvons également nous demander si la variance génétique pour ce caractère dépend de l'origine du pollen. Pour cela, nous avons construit un nouveau modèle (Modèle D) qui permet de définir une variance génétique spécifique pour les descendants issus de pollen local et une autre pour les descendants issus de pollen immigrant. Le modèle D présente une vraisemblance similaire à celle du modèle C (LogL=-5111.2 pour le modèle D et LogL=-5108.6 pour le modèle C). Ces deux modèles ne sont pas significativement différents. Nous retenons donc le modèle C qui est le modèle le plus parcimonieux (celui faisant intervenir le moins de paramètres). Ainsi ces données ne permettent pas de conclure à une différence de variabilité génétique entre les descendants issus de pollen immigrant et ceux issus de pollen local pour le trait croissance.

Modèle avec apparentement réel : Modèle D

Hauteur(2007)= moyenne + NV bloc + Orig pollen*IDseeds + PU + AJUST spatial + résiduelle

Les paramètres en vert sont considérés comme des effets fixes, ceux en bleu comme aléatoires
Hauteur (2007)= *hauteur totale mesurée en 2007 (cm)*
NV bloc : *nouveaux blocs définis a posteriori*
Orig pollen×ID seeds : Orig pollen *n'apparaît plus dans les effets fixes car on considère que l'origine du pollen ne modifie pas la valeur génétique moyenne de la croissance des arbres. En revanche ce modèle permet d'estimer des variances génétiques pour le caractère étudié différentes en fonction de l'origine du pollen.*
PU : *Parcelle unitaire de 6 arbres consécutifs d'une même famille*
AJUST spatial : *composante spatiale*

Tableau récapitulatif des résultats :

Modèle	A	B	C	D
LogL	-5321,5	-5125,9	-5108,6	-5111,2
Var génétique	1835,6	1984,8	1872,8	1604,4 (pollen immigrant)
				1779,7 (pollen local)
Var PU	2993	235	269	302
Var Spatiale	-	3429	3855	3966
Var résiduelle	7122	6568	5156	5252
Effet Orig pollen (Pvalue)	0.21	0.17	0.17	-

Aucun effet de l'origine du pollen n'est donc visible sur les données de croissance des arbres. Ceci pourrait être dû à différents facteurs. Tout d'abord, les descendants ont été récoltés en 1995 et avaient 12 ans lors des mesures de hauteurs (2007). Il est possible qu'une partie des descendants morts au cours de ces 12 ans aient été des individus issus de pollen immigrant, à la croissance lente et qui auraient été rapidement contre-sélectionnés. Toutefois, le taux de mortalité n'est pas considérable dans cette expérimentation (15%). A l'inverse, peut-être que les différences de croissances entre les arbres ne se sont pas totalement exprimées en 12 ans et qu'au fur et à mesure des années les écarts vont se creuser entre les individus (une nouvelle campagne de mesure de croissance est prévue pour l'année prochaine). L'autre facteur est l'imprécision des conditions de pollinisation et donc de l'intensité de sélection présidant à la formation de chaque descendant. J'ai considéré que les descendants issus de pollen immigrant ont été produits dans des conditions où le pollen était plus limitant, et inversement pour ceux issus de pollen local. Or, certains descendants issus de pollen immigrant sont des descendants dont le père se trouve juste à côté de la parcelle mais non échantillonné (comme c'est le cas des descendants issus de pères fantômes dont l'ADN n'est pas disponible). Il serait donc intéressant d'affiner ces groupes, peut-être en considérant l'apparentement des descendants issus de pollen immigrant et local. Ceux qui possèdent des plein-frères ou des demi-frères par le père dans la parcelle sont a priori issus d'un père proche de la parcelle, ce qui pourrait être un signe que le pollen était moins limitant. En réalité, en fonction de la phénologie, chaque fleur de chaque arbre mère va recevoir des quantités de pollen différentes, même quand le pollen fécondant vient du même arbre père. Sur une même mère, certains glands pourront être formés en condition de pollen limitant et d'autre non. Le nombre de demi-frères par le père et de plein frères ne semble pas être une solution pour affiner la catégorie des descendants issus de conditions de faibles densité de pollen, d'autant plus qu'en moyenne le nombre de descendants par père dans la parcelle est faible : 7 pour *Q. robur* et 8 pour *Q. petraea* (avec une médiane à 4 pour les deux espèces). Une autre possibilité est qu'il n'existerait pas d'effet de la diminution de la quantité de pollen sur la compétition pollinique intraspécifique et qu'elle n'affecterait que la compétition entre pollen allospécifique et conspécifique. Il est donc nécessaire pour tester cette hypothèse de l'étudier au niveau des croisements interspécifiques. La première solution serait de tester s'il y a bien augmentation du taux d'hybridation lorsque la quantité de pollen diminue en condition de croisements contrôlés, en faisant varier les quantités de pollen que reçoit un arbre mère et en calculant son taux d'hybridation en fonction des proportions et des quantités reçues. La deuxième possibilité serait d'étudier le taux d'hybridation des arbres d'un peuplement naturel mixte sur plusieurs années parallèlement à un suivi de l'abondance de pollen produit chaque année.

RÉFÉRENCES

Anderson E. (1948). Hybridization of the habitat. *Evolution*, 2, 1-9.

Chan K.M.A. & Levin S.A. (2005). Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution*, 59, 720-729.

Charlesworth D. (1988). Evidence for pollen competition in plants and its relationship to progeny fitness: A comment. *Am. Nat.*, 132, 298-302.

Clark J.S., Bell D.M., Hersh M.H., Kwit M.C., Moran E., Salk C., Stine A., Valle D. & Zhu K. (2011). Individual-scale variation, species-scale differences: inference needed to understand diversity. *Ecol. Lett.*, 14, 1273-1287.

Field D.L., Ayre D.J., Whelan R.J. & Young A.G. (2011). The importance of pre-mating barriers and the local demographic context for contemporary mating patterns in hybrid zones of *Eucalyptus aggregata* and *Eucalyptus rubida*. *Mol. Ecol.*, 20, 2367-2379.

Focke W.O. (1881). *Die Pflanzenmischlinge.* Bornträger, Berlin.

Heinze B. (2011). Towards a quantitative description of landscape, demography and flowering phenology effects on realized hybridization potential. *Mol. Ecol.*, 20, 2233-2235.

Hubbs C.L. (1955). Hybridization between fish species in nature. *Syst. Zool.*, 4, 1-20.

Jahner J.P., Shapiro A.M. & Forister M.L. (2011). Drivers of hybridization in a 66-generation record of *Colias* butterflies. *Evolution*, no-no.

Lamont B.B., He T., Enright N.J., Krauss S.L. & Miller B.P. (2003). Anthropogenic disturbance promotes hybridization between *Banksia* species by altering their biology. *J. Evol. Biol.*, 16, 551-557.

Lepais O., Petit R.J., Guichoux E., Lavabre J.E., Alberto F., Kremer A. & Gerber S. (2009). Species relative abundance and direction of introgression in oaks. *Mol. Ecol.*, 18, 2228-2242.

McGowan C. & Davidson W.S. (1992). Unidirectional Natural Hybridization between Brown Trout (Salmo trutta) and Atlantic Salmon (S. salar) in Newfoundland. *Canadian Journal of Fisheries and Aquatic Sciences*, 49, 1953-1958.

Mitchell R.J. (1997). Effects of pollen quantity on progeny vigor: evidence from the Desert Mustard *Lesquerella fendleri*. *Evolution*, 51, 1679-1684.

Muller C.H. (1952). Ecological control of hybridization in *Quercus*: A factor in the mechanism of evolution. *Evolution*, 6, 147-161.

Parsons K. & Hermanutz L. (2006). Conservation of rare, endemic braya species (Brassicaceae): Breeding system variation, potential hybridization and human disturbance. *Biological Conservation*, 128, 201-214.

Rieseberg L.H., Zona S., Aberbom L. & Martin T.D. (1989). Hybridization in the island endemic, Catalina Mahogany. *Conservation Biology*, 3, 52-58.

Travnichek V.H., Maceina M.J., Smith S.M. & Dunham R.A. (1996). Natural hybridization between black and white crappies (*Pomoxis*) in 10 Alabama reservoirs. *American Midland Naturalist*, 135, 310-316.

Wiegand K.M. (1935). A taxonomist's experience with hybrids in the wild. *Science*, 81, 161-6.

Young W.P., Ostberg C.O., Keim P. & Thorgaard G.H. (2001). Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss irideus*) and coastal cutthroat trout (*O-clarki clarki*). *Mol. Ecol.*, 10, 921-930.

Dans ce chapitre j'ai donc choisi d'étudier les mécanismes affectant l'hybridation entre *Quercus robur* et *Q. petraea*. Les barrières à l'hybridation de chaque espèce, la répartition spatiale des espèces en lien avec leur préférence écologique ainsi que la proportion de pollen conspécifique que reçoit chaque fleur femelle permettent le maintien de ces deux espèces. Dans le chapitre suivant, j'ai décidé d'étudier l'effet de leur dynamique écologique (i.e. espèce plus pionnière pour *Q. robur* et espèce plus forestière pour *Q. petraea*) sur les stratégies reproductives et de croissance de ces deux espèces afin d'illustrer comment ces différences de dynamique peuvent entrainer une sélection divergente sur de nombreux caractères entre les deux espèces, aboutissant à la spéciation écologique.

CHAPITRE 3

Mating system differences between two closely-related oak species with contrasted ecological strategies

Lélia Lagache[1,2] , Etienne K. Klein[3], Alexis Ducousso[1,2], Rémy J. Petit[1,2]

[1] INRA, UMR 1202 Biogeco, F- 33610 Cestas, France

[2] Univ. Bordeaux, UMR1202 Biogeco, F-33400 Talence, France

[3]INRA, UR546, Biostatistique et Processus Spatiaux (BioSP), F-84914 Avignon, France

INTRODUCTION

An ecological strategy is a specific viable combination of life history traits evolved in a given environment in response to multiple selective pressures and trade-offs among them (Burton *et al.* 2010; Newell & Tramer 1978; Westoby *et al.* 2002). Following the "principle of allocation (Cody 1966), these trade-offs exist because species have a limited pool of energy to allocate to different biological functions. Important biological functions include growth, survival and reproduction (Obeso 2002). The reproductive strategy of a species is, therefore, an important component of its ecological strategy that should not be considered separately from the rest (Harper & Ogden 1970). In general, comparisons of ecological strategies should focus on closely related species, to avoid confounding factors. In the case of perennial plants with a long life cycle, studies often deal with seed number, seed provisioning and seed dispersal to investigate selective pressures favoring dispersal versus persistence (e.g. Gaines *et al.* 1974). The observed variation has been interpreted in the framework of the *r*/*K* model (Macarthur & Wilson 1967), where *r* is the capacity to colonize new environment (i.e. species adapted to disturbed environments) and *K* is the capacity to take advantage of favorable growth conditions (i.e. competitive species). In contrast, studies dealing with male fecundity and pollen dispersal are virtually absent (but see van Kleunen & Burczyk 2008). Yet tradeoffs should also exist between male components of the reproductive system, i.e. pollen production, pollen size and dispersal. Indeed energy allocated to reproductive strategy can be divided in three parts: energy allocated to male sexual function, to female sexual function and to subsequent fruit development. In anemophilous species, a larger share of the energy investment is allocated to the male function rather than to the female function (Friedman & Barrett 2009). For such species, it is therefore especially important to examine in detail the different components of the male reproductive strategy.

The two most common oak species in Europe (*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.) are closely related and have similar geographic distribution but contrasting colonization dynamics (Petit *et al.* 2003). A well-investigated difference is that *Q. robur* supports root waterlogging and grows on wet and rich soil, whereas *Q. petraea* requires well drained soils and can grow in poorer and drier soils (Gérard *et al.* 2009; Parelle *et al.* 2006; Ponton *et al.* 2002). More generally, *Q. robur* can be considered as a pioneer species, preferring open environments and well-adapted to environmental perturbations. In contrast, *Q. petraea* is a late-successional post-pioneer species forming dense populations with a close canopy (Bacilieri *et al.* 1996b). These two species should thus be good models for a comparative study of their ecological strategies. Indeed, *Q. robur* is expected to invest a larger share of its energy in female and male reproduction and in seed and pollen dispersal than *Q. petraea*. In contrast, *Q. petraea* is expected to invest a larger share of its energy than *Q. robur* in growth and survival. Previous studies have shown that *Q. robur* seeds and pollen are better dispersed than those of *Q. petraea* (Jensen *et al.* 2009; Lagache *et al.* 2012; Petit *et al.* 2003). However, detailed information on the differences in mating system between the two species is lacking. Here we address this issue by providing comparative information on intra- and interspecific reproductive strategies of these two species and by discussing if these differences in reproductive and growth strategies can be attributed to differences in their ecological dynamics.

To compare the mating system of these two species, a precise modeling of intra- and interspecific crosses is needed. The spatially-explicit individual-based mating model introduced by Lagache *et al.* (2012) was used as a starting point for predicting, at a very fine scale, intra- and interspecific crosses. However this previous model did not consider

important features of their reproductive strategy, such as variation in male fecundity caused by tree size (Oddou-Muratorio *et al.* 2005) and variation in inter-plant compatibility caused by phenology (Slavov *et al.* 2005; Wendt *et al.* 2011). Moreover, the abiotic environment could in principle also affect male fecundity (e.g. De Cauwer *et al.* 2012). Thus, a rigorous comparison should explicitly include it. Overall, our aims are therefore to i) define which traits affect the mating system of these two oak species, ii) study if these traits differ between the two species, iii) and if so, investigate whether the differences can be interpreted in terms of adaptive responses to the species colonization dynamics.

MATERIAL & METHODS

STUDY SITE

The study site is a mixed oak stand of 5ha located in the Petite Charnie State forest in western France (latitude: 48.08° N, longitude: 0.17° W). It contains both *Q. petraea* and *Q. robur* (Figure 1). We determined the taxonomic status of 260 out of 298 trees using assignment methods on the basis of multilocus genetic data obtained using a 12plex microsatellite and a 384plex single nucleotide polymorphism assay (Guichoux *et al.* 2012; Guichoux *et al.* 2011; Lagache *et al.* 2012). In the few cases when DNA was not available, we relied on morphological markers (for the 38 remaining trees; Bacilieri *et al.* 1996a). A smaller proportion of *Q. petraea* than of *Q. robur* was present in the stand (40% versus 55%). In 1995, 3780 seeds were harvested on 51 open-pollinated mother-trees distributed throughout the entire stand (Figure 1). Seeds were collected on all adult trees that had produced a significant acorn crop, i.e. on 22 *Q. petraea* mother-trees, 26 *Q. robur* and 3 admixed mother-trees. The resulting seedlings were grown in a nursery and subsequently planted in a progeny test located nearby the adult stand.

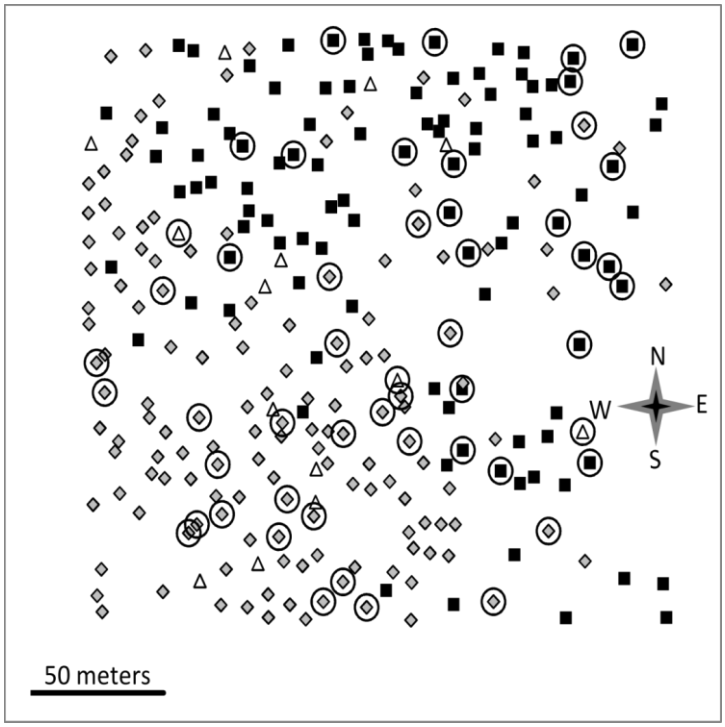


***Figure 1:* Map of the study stand.** *Map of the studied mixed oak stand. Quercus robur genotyped trees are represented by grey diamonds, Q. petraea trees by black squares and intermediate trees by white triangles (species assignment based on multilocus genotypes). Sampled mother trees are circled.*

PHENOTYPIC DATA

Bacilieri *et al*. (1995) studied the flowering phenology of these two species during three years (i.e. 1989, 1991 and 1992). Oak is protandrous, i.e. on a given tree, pollen shedding from catkins precedes female flowers receptivity. As phenology is highly heritable in oaks (Alberto *et al.* 2011), phenology data from one year should help predict compatible mating events in other years. We chose to use the phenological data recorded in 1992 because it was the most precise (i.e. it included most records). Each week during two months, two notations of the phenological stages (one for female flowers and one for male flowers) were made for each of the 298 individuals of the stand. For each pair of candidate trees, we computed the number of weeks where male flowering of the candidate father *k* overlapped female flowering of the mother *j*: OP_{jk} varied between 1 and 4. For all adult trees (i.e. 298 individuals), the height (H) and the circumference at 1.3m of the ground ($Cir_{1.3}$) were also measured in 1998 before cutting all the trees. We used the circumference as a proxy of male fecundity and the H/DBH index (where $DBH=Cir_{1.3}/(2\pi)$) as a life-long indicator of tree competition (Becker 1992). Trees with high H/DBH index are considered to have suffered more from competition than other trees. The number of rings at three different heights (from 6 to 25 meters) was counted after tree cutting. By extrapolation an estimate of the age of the tree can be deduced (Supporting information 1).

TERRAIN ELEVATION

The stand presents a regular slope that creates an ecological gradient, from humid clay soil in the lower part up to relatively dry silt and sandy soil in the upper part. Bacilieri *et al*. (1995) had produced a fine-scale topographical map of the stand by measuring terrain elevation every 25 meters across the stand (Supporting information 2). We can thus deduce the approximate terrain elevation at which each tree was growing, summarized by a class score between 1 and 10 (1 being the lowest class, Supporting information 2).

PATERNITY ANALYSIS

Simple paternity exclusion tests for the 3046 offspring for which genotypic data is accessible revealed a single father for 51.7% of the offspring (615 *Q. petraea* and 855 *Q. robur* and two or more compatible fathers for 1.8% of the offspring (22 *Q.* petraea and 31 *Q. robur*). The remaining individuals, 46.5% of all offspring (427 *Q. petraea* and 885 *Q. robur*) had no compatible father among the 260 adult trees studied (details of the paternity analysis are provided in Lagache *et al.* 2012). As in this previous study, we decomposed offspring whose fathers were not found (i.e. immigration) in two parts: offspring sired by fathers inside the stand for which circumference and terrain elevation data were available but for which DNA was not available (called “ghost trees”), and offspring sired by fathers outside the stand, for which no information at all was available.

SPATIALLY-EXPLICIT MATING MODEL

We used a spatially-explicit individual-based hybridization model (De Cauwer *et al.* 2012; DiFazio *et al.* 2012; Oddou-Muratorio *et al.* 2005) to investigate intra- and interspecific mating events with pollen from inside and outside the studied stand. Following Lagache *et al.* (2012), we modeled pollen immigration from outside the stand using a mass action law (Holsinger 1991). With this model, the probability that a seed *o* from mother j_o has genotype g_o was:

$$P\left(g_o \middle| g_{j_o}\right) = sT\left(g_o \middle| g_{j_o}, g_{j_o}\right) + (1-s)\left[mig_{j_oP}T\left(g_o \middle| g_{j_o}, AFP\right) + mig_{j_oR}T\left(g_o \middle| g_{j_o}, AF_R\right) + \sum_{k:candidates} \pi_{j_o k} T\left(g_o \middle| g_{j_o}, g_k\right)\right]$$

(Eq. 1)

Where *s* is the selfing rate, T(g_o|.,.) are the Mendelian probabilities of generating the offspring's genotype g_o from the known genotypes of the two parents, *AFR* and *AFP* are the microsatellite allelic frequencies of *Q. robur* and *Q. petraea*, mig_{jR} and mig_{jP} correspond to the two migration rates (*Q. robur* and *Q. petraea*) on mother *j* and π_{jk} is the relative contribution of the candidate father *k* in the pollen pool of mother *j* (detailled below). As in Lagache *et al.* (2012), migration rates can vary across mothers due to the amount of pollen from local trees (including from ghost trees) that they receive.

Modeling of the relative contributions of the candidate fathers to the pollen pools (π_{jk})

The relative contribution π_{jk} of the candidate father *k* in the pollen pool of mother *j* results from the competition with pollen from all other candidate fathers but also with pollen from all ghost fathers and with immigrant pollen. Following Smouse & Sork (2004), we considered two kind of effects determining the pollen pool available to each mother-tree *j*: effects affecting the male fecundity of each father tree *k* of the stand (F_k) and effects affecting the cross compatibility between each mother *j* and father *k* ($Compat_{jk}$):

$$\pi_{jk} = \frac{F_k \times Compat_{jk}}{\sum_{l:candidates} F_l \times Compat_{jl} + \sum_{l:ghosts} F_l \times Compat_{jl} + q_P(DE_J)Hyb_{jP} + q_R(DE_j)Hyb_{jR}}$$

(Eq. 2)

where $q_P(DE_j)$ and $q_R(DE_j)$ are the amounts of *Q. petraea* and *Q. robur* pollen coming from outside the stand and decreasing with the distance of the mother tree *j* to the edges of the plot (detailed below). Hyb_{jP} and Hyb_{jR} represent the post-dispersal relative fertilization successes of one pollen grain from *Q. robur* immigration (Hyb_{jR}) and from *Q. petraea* immigration (Hyb_{jP}). As in Lagache *et al.* (2012), they are obtained as:

$$Hyb_{jP} = h_{sp_j,P},\ Hyb_{jR} = h_{sp_j,R}$$

(sp_j is the species of the mother *j*: R (*Q. robur*), P (*Q. petraea*), or H (admixed category))

The fecundity component (F_k) in the equation (2) includes the effects of circumference and terrain elevation following:

$$F_k = \exp\left(\alpha_{Cir_{1.3},sp_k}\left(Cir_{1.3,k} - \overline{Cir_{1.3}}\right) + \frac{\exp\left(\alpha_{sp_k} + b_{sp_k}EL_k\right)}{1+\exp\left(\alpha_{sp_k} + b_{sp_k}EL_k\right)}\right)$$

(Eq. 3)

where $Cir_{1.3,k}$, EL_k are the circumference and terrain elevation for tree *k*, and $\overline{Cir_{1.3}}$ is the average circumference of trees in the study site. One set of parameters (*a, b,* $\alpha_{Cir_{1.3}}$) applies to each of the three categories (*Quercus petreae*, *Q. robur* and admixed trees). Finally, sp_k is the species of tree *k*.

The compatibility component in equation (2) included effects of phenology overlap, interspecific sexual barriers and spatial distance between parents:

$$compat_{jk} = \gamma_{PO_{jk}} h_{sp_j, sp_k} f_{EP}(d_{jk}, az_{jk}; \delta_{sp_k}, b_{sp_k}, \kappa_{sp_k}, \theta_{sp_k})$$

(Eq. 4)

where PO_{jk} d_{jk}, and az_{jk} are phenological overlap, distance and azimuth between trees *j* and *k* and f_{EP} is the anisotropic exponential power dispersal kernel (detailed below). Different sets of parameters (δ, *b,* κ, θ) apply to the three species *Quercus petreae*, *Q. robur*, admixed trees and must be estimated. γ and *h* are also parameters to estimate.

For pollen dispersal modeling, we took into account two parameters: the distance between the trees and the direction of mating pairs. We therefore improved the model of pollen dispersal commonly used (e.g. De Cauwer *et al.* 2012; Lagache *et al.* 2012; Oddou-Muratorio *et al.* 2005) by incorporating two parameters that model polarity of pollen dispersal (Torimaru *et al.* 2012). As in Lagache *et al.* (2012), the pollen dispersal curves of the two species differed. We therefore modeled different anisotropic exponential power pollen dispersal kernels for each species:

$$f_{EP}(d, az; \delta, b, \kappa, \theta) \propto \exp\left[-\left(\frac{d\Gamma(3/b)}{\delta\Gamma(2/b)}\right)^b\right] \exp\left[\kappa(az - \theta)\right],$$

(Eq. 5)

where δ is the mean dispersal distance, *b* the shape parameter, κ the intensity of anisotropy and θ (in radian with east as starting point) the main direction of anisotropy.

Our main interest was to study pure individuals from each species and compare their fecundity and mating behavior. The admixed category is indeed very heterogenous, with some individuals in this category that are very similar to either *Q. robur* or to *Q. petraea*. The results for this admixed category will therefore not be developed here.

Modeling pollen immigration

As in Lagache et al. (2012), our model allowed immigrant pollen to compete with pollen produced inside the stand, resulting in immigration rates that can vary across mother

trees following the mass action law (Holsinger 1991). To take into account possible edge effects (i.e. a mother tree near the edge of the stand might receive more immigrant pollen than a mother tree located in the middle), we considered that the amount of immigrant pollen q_R and q_P decrease with the distance to the edges of the plot. We therefore computed DE_j the minimum distance of mother tree *j* to the closest edge of the stand and defined:

$$q_P(DE_j) = q_P \frac{\exp(a_{m,P} + b_{m,P} DE_j)}{1+\exp(a_{m,P} + b_{m,P} DE_j)} \text{ and } q_R(DE_j) = q_R \frac{\exp(a_{m,R} + b_{m,R} DE_j)}{1+\exp(a_{m,R} + b_{m,R} DE_j)}$$

(Eq.6)

where $q_p(0)$ and $q_r(0)$ are the amount of pollen of *Q. petraea* and *Q. robur* at the edge of the plot (i.e. $DE_j=0$). The parameters q_p and q_r correspond to the amount of pollen received per unit of area relative to the total emission of a tree with a fecundity equal to 1. a_{mP}, b_{mP}, a_{mR}, b_{mR} are parameters of pollen dilution with distance of the mother *j* to the edge of the plot.

The *Q. robur* and *Q. petraea* immigration rates on a mother tree *j* (i.e. mig_{jR} and mig_{jP}) were then calculated as follows:

$$mig_{jR} = \frac{\sum_{l:robur\ ghosts} F_l \times Compat_{jl} + q_R(DE_j) Hyb_{jR}}{\sum_{l:candidates} F_l \times Compat_{jl} + \sum_{l:ghosts} F_l \times Compat_{jl} + q_P(DE_J) Hyb_{jP} + q_R(DE_j) Hyb_{jR}}$$

(Eq. 7)

$$mig_{jP} = \frac{\sum_{l:petraea\ ghosts} F_l \times Compat_{jl} + q_P(DE_j) Hyb_{jP}}{\sum_{l:candidates} F_l \times Compat_{jl} + \sum_{l:ghosts} F_l \times Compat_{jl} + q_P(DE_J) Hyb_{jP} + q_R(DE_j) Hyb_{jR}}$$

(Eq. 8)

Parameter estimation

The log-likelihood of the full genotypic dataset was computed by summing the logarithm of Eq. 1 for all 3213 genotyped offspring. All computations necessary to derive the likelihood were conducted with MATHEMATICA 8.1 (Wolfram Research Inc. 2010). We maximized the log-likelihood using a quasi-Newton algorithm to obtain maximum likelihood estimates for all parameters considered. The maximization was repeated several times using contrasted initial values to be more confident that we had reached a global maximum.

Likelihood ratio tests

Sub-models investigating different biological hypotheses by omitting or fixing different parameters of interest were also fitted to the data. Likelihood-ratio tests were then

used to test the hypotheses (i.e. investigate whether the fixed parameters are significant in the full model), following Oddou-Muratorio *et al*. (2005). First, the general effect and species-specific effect of circumference on male fecundity were studied by contrasting the full model with a model without this effect of circumference on male fecundity and with a model with the same effect of circumference on the male fecundity for both species. Second, the full model was compared with a model without the effect of terrain elevation on male fecundity and with a model where the effect of terrain elevation on male fecundity was the same for both species. Third, we contrasted the full model with one with no phenological effect on crossing probability. Fourth we compared the full model with one where the barriers were symmetric between *Q. robur* and *Q. petraea* ($h_{PR} = h_{RP}$). Fifth, the effect of dispersal on mating events was studied by contrasting the full model with an unlimited dispersal model and with a model with the same dispersal parameters for both species. Sixth, the full model was compared with a model where there was no anisotropic effect on pollen dispersal and with a model where this anisotropic effect was the same for both species. Seventh, we tested if different amounts of *Q. robur* and *Q. petraea* pollen come from outside the stand by fitting a model with the same amounts of immigrant pollen for the two species (q_{PO}=q_{RO}). Eighth, we compared the full model with a model without edge effect on the amount of immigrant pollen, assuming that whatever the position of a mother-tree, it received the same amount of *Q. robur* or *Q. petraea* immigrant pollen. Finally we compared the full model with a model where there was neither a difference in the amount of immigrant pollen for *Q. robur* and *Q. petraea* (q_{PO}=q_{RO}) nor an edge effect.

EXPECTED MALE FECUNDITY

To compare whether species differ in male fecundities evenness, we first estimated the fecundity of each tree with the complete model. This fecundity is influenced by tree circumference and terrain elevation. As the absolute amounts of pollen produced by individuals were not known, we present relative fecundities by reference to a tree growing in terrain elevation class 10 and of average size (1.73m of circumference). We can then compare the distribution (evenness) of relative male fecundities within each species. For a given census size, species having more heterogeneous male fecundity distributions should have a smaller effective population size. The difference between the two distributions was tested with a Kruskal-Wallis test.

EFFECTIVE PATERNITY NUMBER

The effective paternity number (K_e) corresponds to the number of offspring that need to be examined in the same progeny to find two offspring sired by the same father (Nielsen *et al.* 2003). We used the formula proposed by Starr (1984) to compute the effective paternity number in the progeny of a tree:

$$K_e = \frac{1}{\sum_{k=1}^{n} p_k^2} \qquad \text{(Eq. 9)}$$

where p_k is the proportion of offspring sired by a given father tree k in the progeny of a given mother-tree. In our study, we used the model predictions for the 51 mother trees to calculate average K_e for each species. This allows overcoming the imbalance of seed production of each mother-tree (and of each species) on K_e calculation. We used the Kruskal-

Wallis test to evaluate the significance of the difference between species. Note that in the computation of K_e we excluded fathers located outside of the stand but included ghost trees. For a full picture of the diversity of offspring in the maternal progeny of each individual, differences in immigration rates ought to be accounted for. Here these were investigated separately through a comparison of immigration parameters.

RESULTS

DIRECT COMPARISON OF GROWTH, SEED PRODUCTION AND PHENOLOGY

In this mixed oak stand, *Quercus petraea* trees are on average larger ($Cir_{1.3}$: 1.84m; height: 26.5m) than *Q. robur* trees ($Cir_{1.3}$: 1.67m; height: 25.1m; both *p*-values $<10^{-4}$; Table 1). The H/D index also differs between the two species (14.4 for *Q. petraea* versus 15.3 for *Q. robur*; *p*-value = 0.002; Table 1), suggesting that *Q. petraea* trees suffer less than *Q. robur* trees from competition. According to the number of rings of each tree (extrapolated at height 0), *Q. petraea* trees are slightly younger than *Q. robur* trees (118 years for *Q. petraea* versus 124 years for *Q. petraea, p-value* $<10^{-4}$; Table 1). *Q. petraea* is more frequently encountered in high terrain elevation classes than *Q. robur* (6.9 versus 4.7, *p*-value$<10^{-4}$; Table 1). There is also a slight phenological shift between the two species in this stand: *Q. petraea* trees flower slightly later than *Q. robur* trees (first record of mature male and female flower for *Q. petraea:* 3.5 and 3.9 versus 3.0 and 3.3 for *Q. robur*; *p-value* $<10^{-4}$; Table 1). In 1995, a larger proportion of *Q. petraea* trees produced seeds (22/119, i.e. 19%) compared to *Q. robur* trees (26/164, i.e. 16%), but *Q. petraea* trees tended to have fewer offspring than *Q. robur* trees (Table 1).

Phenotypic data	***Q. petraea*** [1]	***Q. robur*** [1]	**p-*value*** [2]
Circumference(m)	**1.84 (0.42)**	**1.67 (0.24)**	**$<10^{-4}$**
Height (m)	**26.5 (1.6)**	**25.1 (1.6)**	**$<10^{-4}$**
H/D	**14.4 (2.0)**	**15.3 (1.8)**	**0.002**
Age	**118 (12)**	**124 (16)**	**0,0004**
Terrain elevation	**6.9 (1.3)**	**4.7 (1.7)**	**$<10^{-4}$**
1st record of a mature ♂ flower	**3.5 (0.6)**	**3.0 (0.7)**	**$<10^{-4}$**
1st record of a mature ♀ flower	**3.9 (0.4)**	**3.3 (0.7)**	**$<10^{-4}$**
Mean # offspring /mother	**48 (23)**	**68 (39)**	**0.09**

[1] *Mean values with standard deviation in brackets*
[2] *Kruskal-Wallis test for independent samplings*

Table 1: ***Direct comparisons between Q. petraea and Q. robur***

COMPARISONS BASED ON THE SPATIALLY-EXPLICIT MATING MODEL

Male fecundity parameters

We found that individual male fecundity depends on tree circumference (*p*-value < 10^{-4}; Table 2). Moreover, this effect differed between the two species (*p*-value = 0.027; Table 2). The two coefficients modulating the effect of circumference on male fecundity for each species were estimated in the model: $\alpha_{Cir1.3\_Q\_petraea}$=0.0075 and $\alpha_{Cir1.3\_Qrobur}$=0.0161 (Table 3). Circumference had a stronger effect on *Q. robur* than on *Q. petraea* (Figure 2A). For example, if we consider trees with circumferences ranging from 1.55m to 2m (corresponding to the most frequent values for both species; Supporting information 3A), fecundity increases by 40% in *Q. petraea* (relative fecundity of 0.87 for trees with a circumference of 1.55m and 1.16 for trees with a circumference of 2m; Figure 2A) whereas it increases by as much as 101% in *Q. robur* (0.74 for trees with a circumference of 1.55m and 1.49 for trees with a circumference of 2m; Figure 2A).

The model was also improved by taking into account the effect of terrain elevation on tree male fecundity (*p*-value <10^{-3}; Table 2). Moreover, this effect differed for the two species (*p*-value <10^{-4}; Table 2). The two coefficients modulating the effect of terrain elevation on male fecundity for each species were estimated as follows: $a_{Q.\ petraea}$= -5.7, $a_{Q.\ robur}$= -2.5 and $b_{Q.\ petraea}$= 0.81, $b_{Q.\ robur}$= 2.17 (Table 3). Trees growing in the bottom of the stand had reduced male fecundity (Figure 2B). However, the fecundity of *Q. petraea* was more strongly affected than that of *Q. robur*. For example, if we consider terrain elevations ranging between 2 and 5 (corresponding to the most frequent classes; Supporting information 3B), the fecundity of *Q. petraea* was increased by 530% (relative fecundity of 0.03 for terrain elevation class of 2 and 0.19 for terrain elevation class 5; Figure 2B) whereas the fecundity of *Q. robur* was increased by 19% (0.84 for terrain elevation class 2 and 1.00 for terrain elevation class 5; Figure 2B).

Model[1]	-LogL	Δ -LogL	ddl[2]	*p*-value[3]
Full model	61423	-	37	- .
Male fecundity				
Diameter effect on male fecundity	61513	90	3	$<10^{-4}$
Species-specific diameter effect on male fecundity	61432	9	1	0.027
Terrain elevation effect on male fecundity	61448	25	6	$<10^{-3}$
Species-specific terrain elevation effect on male fecundity	61442	19	2	$<10^{-4}$
Phenology				
Phenology effect on cross compatibility	61448	25	3	$<10^{-4}$
Hybridization				
Asymmetric hybridization	61431	8	1	0.0047
Pollen dispersal				
Distance effect on cross compatibility	62565	1142	6	$<10^{-4}$
Species-specific pollen dispersal	61433	10	2	0.0067
Anisotropic pollen dispersal	61606	183	6	$<10^{-4}$
Species-specific anisotropy of pollen dispersal	61495	72	2	$<10^{-4}$
Immigration				
Species-specific pollen immigration rates	61432	9	1	0.0027
Immigration rate with edge effects	61449	26	4	$<10^{-4}$
Species-specific edge effects	61432	9	2	0.011

[1] *All models listed are based on the full model modified in one respect to yield the corresponding submodel.*
[2] *ddl provides the number of estimated parameters for the full model and the number of parameters that are fixed and not estimated in the corresponding submodel.*
[3] *p-values lower than 0.05 indicate that the full model is significantly more informative than the tested submodel.*

Table 2: ***Likelihood-ratio test of the significance of each sub-model component***

Parameters[1]	Values
Diameter parameters	
$\alpha_{DBH\_Q.\ petraea}$	0.0075
$\alpha_{DBH\_Q.\ robur}$	0.0161
Terrain elevation parameters	
$a_{Q.\ petraea}$	-5.7
$a_{Q.\ robur}$	-2.5
$b_{Q.\ petraea}$	0.81
$b_{Q.\ robur}$	2.17
Phenology	
$\gamma_{POJk=1}$	0.57
$\gamma_{POJk=2}$	0.69
$\gamma_{POJk=3}$	1
$\gamma_{POJk=4}$	0.68
Hybridization barriers	
$h_{Q.\ petraea\_Q.\ robur}$	0.001
$h_{Q.\ robur\_Q.\ petraea}$	0.035
Dispersal curves	
$\delta_{Q.\ petraea}$ (m)	97
$\delta_{Q.\ robur}$ (m)	137
$b_{Q.\ petraea}$	0.48
$b_{Q.\ robur}$	0.25
Anisotropic parameters	
$K_{Q.\ petraea}$	1.21
$K_{Q.\ robur}$	-0.66
$\theta_{Q.\ petraea}$	0.17
$\theta_{Q.\ robur}$	4.28
Immigration	
$q_p(0)$	0.0010
$q_r(0)$	0.0018
a_{mP}	-8.3
b_{mP}	-0.017
a_{mR}	-4.1
b_{mR}	-0.005

[1]A total of 37 parameters were estimated, but the 11 parameters for intermediate trees (h_{IP}, h_{IR}, h_{PI}, h_{RI}, δ_I, b_I, K_I, ϑ_I, a_I, b_I and α_{DBH_I}) are not shown in this table.

***Table 3**: Parameters estimated from the spatially-explicit mating model*

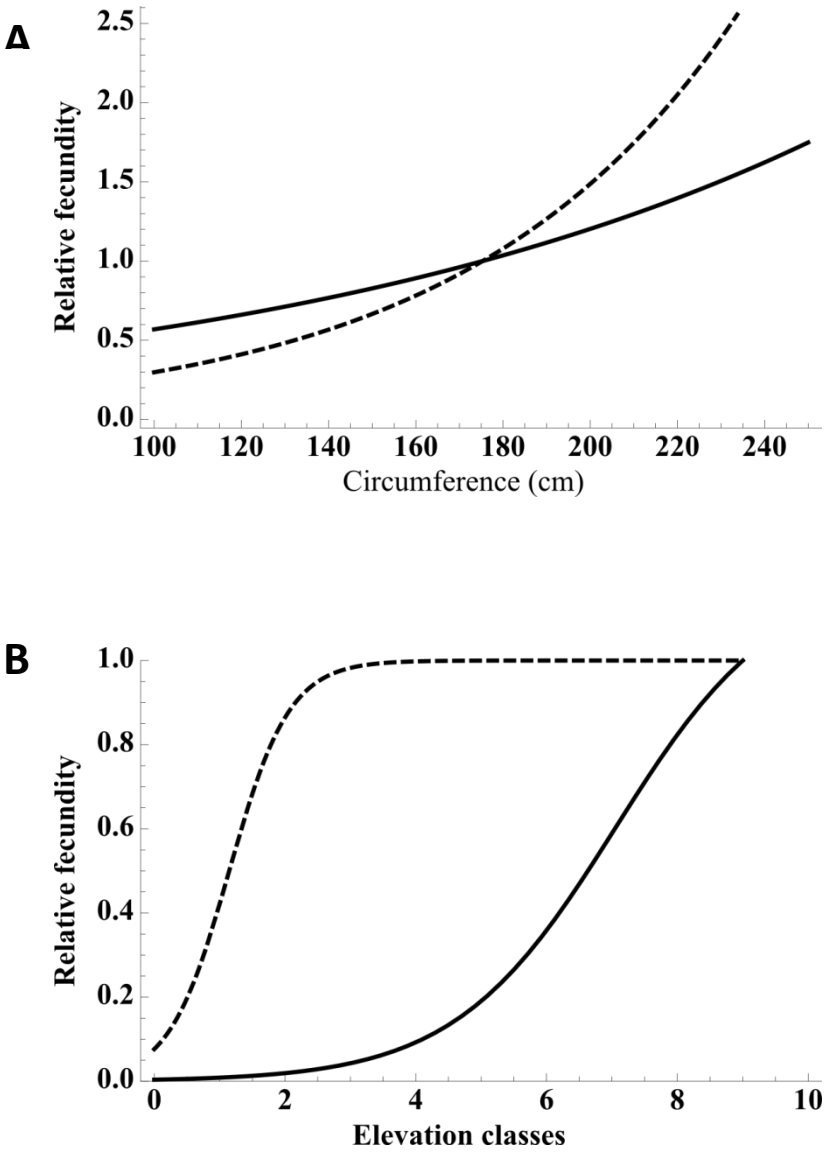


Figure 2: ***Effect of tree circumference and terrain elevation on*** **Q. robur** ***and*** **Q. petraea** ***father-trees fecundity***. Q. robur *relative fecundity is symbolized by dotted line and by full line for* Q. petraea*. In A, the reference is the fertility of a tree with a circumference equal to the mean circumference of the population (1.7cm). In B, the reference is the fertility of a tree growing in terrain elevation class 10.*

Mating compatibility between trees

Taking into account the phenological overlap between mother and father trees significantly improved the model (p-value $<10^{-4}$; Table 2). As expected, the longer the overlap between mature male and female flowers, the greater was the likelihood of mating between the trees (except for one overlap class that was underrepresented; Table 3). Furthermore, the sexual barriers were on average 35 times lower for *Q. robur* than for *Q. petraea* (h_{RP} = 0.035 versus h_{PR} = 0.001, p-value$<10^{-4}$; Table 2 and Table 3).

Distance between trees had the greatest influence on observed mating patterns (the model that did not include distance-dependent mating success had the lowest likelihood; Table 2). Moreover pollen dispersal curves were not the same for both species (p-value = 0.007; Table 2). Indeed, *Q. petraea* was found to disperse its pollen over shorter distances

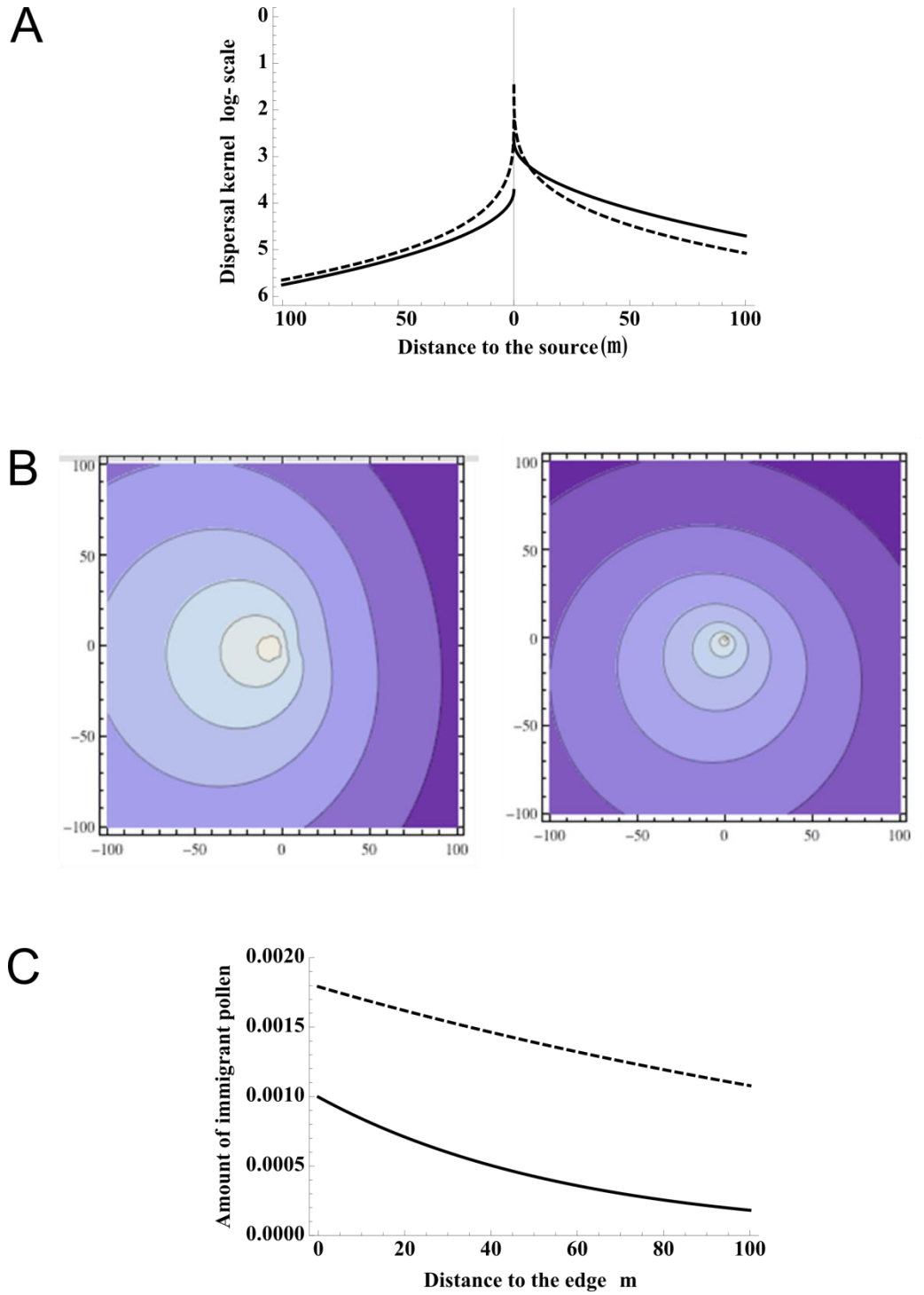
within the study site (mean pollination distance was 97m, with a weaker-tailed dispersal kernel: b=0.48) than *Q. robur* (137m, b=0.25) (Table 3 and Figure 3A). Pollen dispersal had a preferential direction in both species (models without anisotropy had a lower likelihood than models where anisotropy was included, Table 2) but the polarity of crosses was higher for *Q. petraea* than for *Q. robur* ($K_{Q.\ petraea}$ =1.21 > $K_{Q.\ robur}$ =- 0.66; *p*-value $<10^{-4}$; Table 2). The preferential direction was from east to west in *Q. petraea* ($\theta_{Q.\ petraea}$ =0.17rad; Table 2 and Figure 3B) and from north-east to south-west for *Q. robur* ($\theta_{Q.\ robur}$ =4.28rad; Table 2 and Figure 3B).

Immigration

Q. petraea immigrant pollen was nearly twice less abundant than *Q. robur* immigrant pollen (at the edge of the stand: $q_p(0)$=0.0010 vs $q_r(0)$=0.0018, *p*-value=0.003, Tables 2 and 3). In both species, immigration rates were larger at the edge than inside the stand (*p*-value $<10^{-4}$; Table 2 and Figure 3C). However, this difference was more marked for *Q. petraea* than for *Q. robur* (*p*-values=0.011 Table 2): between 0 and 100m from the edge, the amount of immigrant *Q. petraea* pollen decreased 5.5 times ($q_p(100)$=0.00019, Figure 3C) whereas the amount of immigrant pollen of *Q. robur* decreased 1.6 times ($q_r(100)$=0.0011, Figure 3C).

Distribution of male fecundities and effective number of males

Evenness of male fecundities differed between species. Indeed, the amplitude of *Q. petraea* male fecundities (A_{qp}=2.2) is lower than for *Q. robur* (A_{qr}=2.5), suggesting more homogenous distribution of male fecundities in *Q. petraea* (Figure 4). In view of the more even male fecundities for *Q. petraea* but greater pollen dispersal (and immigration) rates for *Q. robur*, it is not clear if *Q. robur* and *Q. petraea* ovules are sired by a more diverse cohort of males. We therefore computed the effective number of different fathers that contribute to the progeny of a given mother tree in each species. This number was high in both species but lower for *Q. petraea* (56) than for *Q. robur* (93) (*p*-value=0.04, Kruskal-Wallis test), indicating that the effect of greater pollen dispersal in *Q. robur* was predominant over that of more even male fecundity in *Q. petraea*. If pollen was not too limiting (cf. discussions in Lagache *et al.* 2012), the conditions for intensive intra- and interspecific pollen tube competition would therefore be met in both species, but more so in *Q. robur*.



<u>Figure 3</u>: Anisotropic pollen dispersal curves, anisotropic modeling for both species and edge effect on immigrant pollen. *In A and C, Q. robur relative fecundity is symbolized by a dotted line and Q. petraea relative fecundity by a full line. In B, anisotropic pollen dispersal modeling for* Q. robur *(right) and* Q. petraea *(left). In A, the position 0 symbolizes a father tree and anisotropy of pollen dispersal is from left to right.*

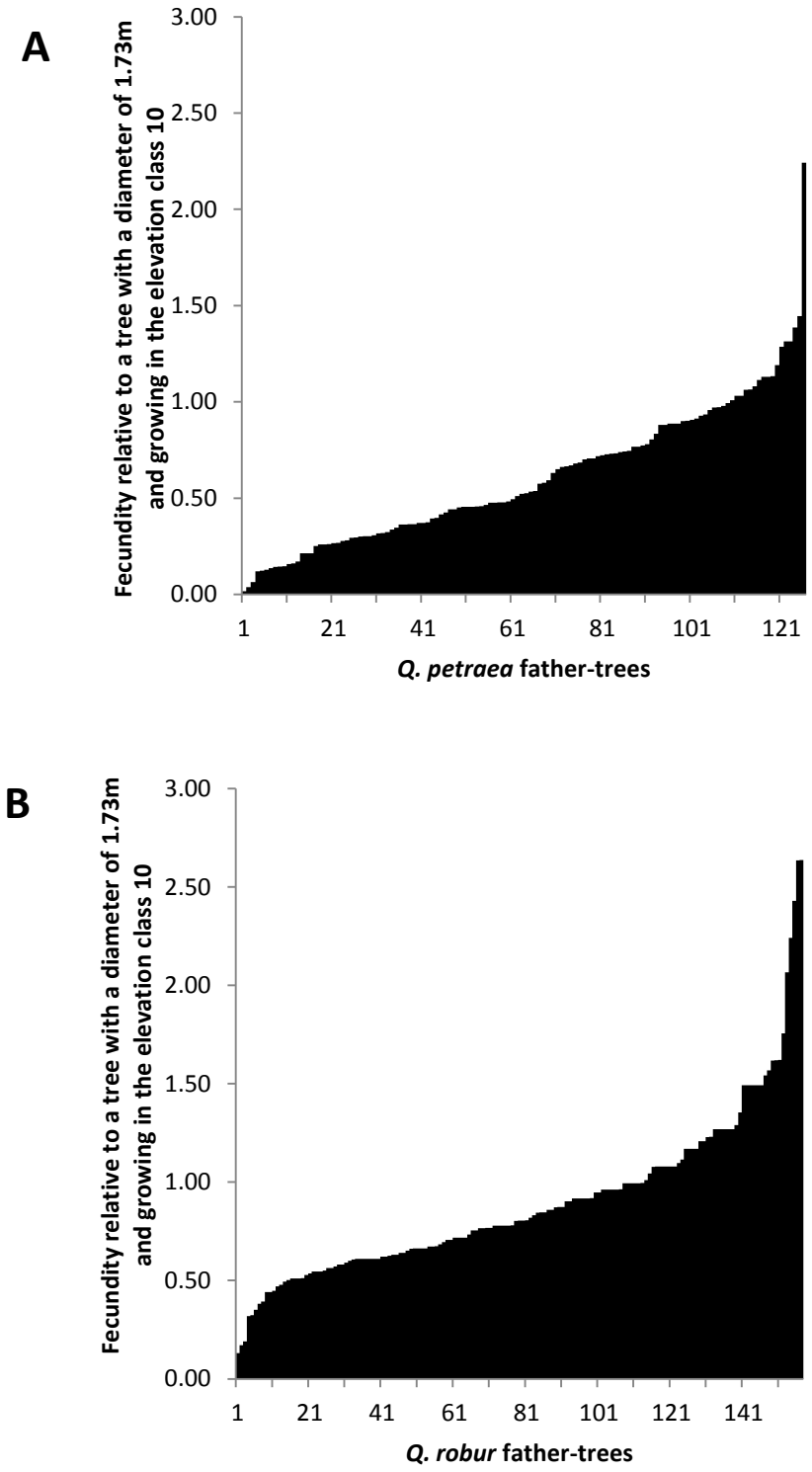


***Figure 4**: Individual male fecundities predicted by the model for* **Q. petraea** *(A) and* **Q. robur** *(B).* *The fertility reference is a tree with a circumference of 1.7cm and growing in terrain elevation class 10.*

DISCUSSION

We modelled mating events in a mixed oak stand of *Q. petraea* and *Q. robur*. The model relies on individual characteristics of trees or of pairs of trees (circumference, inter-tree distance, phenological overlap etc.) and an accurate paternity analysis. This allowed the estimation of important yet rarely available reproductive parameters that could then be compared between species. Our goal here is to discuss the meaning of these characters and attempt to interpret them in the light of theoretical predictions for life history strategies.

QUERCUS ROBUR *AND* Q. PETRAEA*: TWO SPECIES WITH CONTRASTING ECOLOGICAL DYNAMICS*

Our study confirms that these two species have different ecological niches and highlights an effect of environment on their fecundity. First, *Q. petraea* and *Q. robur* grow at different places within the studied stand. *Q. petraea* is found preferentially on the top of the slope and is completely absent at the bottom whereas *Q. robur* is found everywhere but is most abundant at the bottom. This finding is consistent with the ecological preferences of these two species: *Q. robur,* unlike *Q. petraea*, supports root waterlogging and can thus be found in wet soils at the bottom of a slope or basin (Parelle *et al.* 2006; Ponton *et al.* 2002). A new result of this study is that this fine scale variation in the abiotic environment can affect male fecundity, even when circumference is controlled for. We identified a decrease of male fecundity for all trees growing in low terrain elevation classes. However *Q. petraea* male fecundity was more sensitive to this effect than *Q. robur*. Several factors of the environment (e.g. soil water content, nutrients, risks of frosts or waterlogging) could affect male fecundity (e.g. Byron *et al.* 1994). The reduced sensitivity of *Q. robur* to the environment is in agreement with the fact that *Q. robur* is described as a pioneer species that is adapted to a broader spectrum of abiotic environments (Lepais 2008), as confirmed in our study by the larger range of environments where *Q. robur* grows (cf. Supporting information 3B).

QUERCUS PETRAEA ECOLOGICAL STRATEGY: ENERGY PREFERENTIALLY INVESTED IN GROWTH

Our results support the hypothesis that *Q. petraea* is more competitive than *Q. robur*, as suggested by previous studies (reviewed in Petit *et al.* 2004). First, greater investment of energy by *Q. petraea* in its growth is supported by our finding of a significantly higher mean tree diameter and tree height in this species, compared to *Q. robur*. Differences in tree ages between species cannot explain this difference as *Q. petraea* trees were estimated to be slightly younger than *Q. robur* trees. Jensen's (2000) study based on an interspecific provenance test also found a higher growth for *Q. petraea* than for *Q. robur*. Note that for seedlings, the opposite is observed: *Q. robur* seedlings grow about twice as fast as *Q. petraea* seedlings, which should facilitate establishment in the more pioneer species *Q. robur* (Landergott *et al.* 2012). Second, our results confirm that *Q. petraea* is more competitive than *Q. robur*, as suggested by Bacilieri *et al.* (1996b). Indeed, *Q. petraea*, which has a lower H/DBH index than *Q. robur,* appears to have less suffered from tree-competition throughout its life than *Q. robur* (Becker 1992). These results are consistent with the study of Landergott *et al.* (2012). These authors have shown that, in paired seedlings grown side by side, competition boosts the growth of *Q. petraea* superior seedlings but not that of *Q. robur* superior seedlings. Simultaneously, competition has little effect on *Q. petraea* inferior seedlings but strongly affects *Q. robur* inferior seedlings. Third, the neighbourhood model indicates that tree circumference less strongly affects *Q. petraea* male fecundity than *Q. robur*, suggesting that male fecundity is more stable in the more competitive species. This

is confirmed by measurements of the evenness of male fecundity. All these results illustrate the greater investment of *Q. petraea* in growth and are consistent with its greater competitive ability (Petit *et al.* 2004).

QUERCUS ROBUR ECOLOGICAL STRATEGY FAVOURS DISPERSAL

The energy not invested in growth by *Q. robur* could instead be invested in reproduction and more especially in pollen (and seed) dispersal. To evaluate this prediction, it would be interesting to compare pollen production between these two species. Unfortunately, no information is available to test this prediction. Note that even if the two species invested equal amounts of resources in pollen production, the smaller average size of pollen grains in *Q. robur* (Niklas 1985; Rushton 1976) would imply that more numerous pollen are produced in this species. Regarding pollen dispersal *per se*, we found that mean pollen dispersal distance of *Q. robur* is 1.4 times greater than that of *Q. petraea*. Second, immigrant pollen is biased towards *Q. robur* and dilution with distance from the edge of the stand is three times lower for *Q. robur* than for *Q. petraea,* also suggesting that *Q. robur* pollen travel greater distances and is therefore less quickly diluted. These findings support those of Lagache et al. (2012) on the basis of the same data but using a more simple model and those of Jensen et al. (2009) based on realized pollen dispersal distances in another study plot during two different mating episodes. This asymmetry in pollen dispersal between the two species might be caused, at least in part, by differences in pollen size for the two species: the smaller pollen grains of *Q. robur* are predicted to travel over greater distances than those of *Q. petraea* (Niklas 1985). Another non exclusive hypothesis is that leaves exert a stronger blocking effect in *Q. petraea* than in *Q. robur* due to different timing of pollen release and leaf development for the two species. If leaves develop more rapidly in *Q. petraea*, the dispersal of its pollen could be more strongly hampered by foliage.

In this study stand and during this flowering episode (1995), pollen dispersal was anisotropic in both species. This polarity of crosses can be explained in different ways. A straightforward interpretation is that wind is responsible for this polarity (e.g. Pluess *et al.* 2009; Torimaru *et al.* 2012). Alternatively, the particular stand configuration could favour some crosses (e.g. from father in the top to mothers in the bottom). Unfortunately, no data for wind direction for this particular year are available. We found a preferential direction of crosses from east (north-east) to west (or south-west). Yet winds from east or north-east are very rare in this area in spring. Dominant winds blow instead from west to east (Meteo France database). While some authors have reported anistropic pollen dispersal correlated with predominant wind direction (e.g. Burczyk & Prat 1997; Shen *et al.* 1981), others have found the opposite pattern (e.g. Burczyk *et al.* 2004; Robledo-Arnuncio & Gil 2005), indicating that our findings is not isolated. These latter authors suggested that topography and vegetation surrounding the study stand could modify the expected wind direction. In our study, it is possible that terrain elevation was responsible for this greater cross polarity because it coincides with the direction of the slope. It is conceivable that pollen grains tend to more easily attain flowers located lower than their starting point. Indeed, larger pollen grains might fertilize flowers located upslope with more difficulty than smaller pollen grain. Another reason could be that westerly winds are more humid and tend to stick pollen grain while the rarer easterly wind, being drier, eases the take off of pollen grains (Whitehead 1969). This could explain polarity of crosses does not predominant wind direction in the stand.

The fact that the degree of polarity of crosses differed for each species was unexpected. It may be due to differences in pollen size. A pollen grain with a large size needs a stronger wrenching force to take off than a small pollen grain (Aylor & Parlange 1975; Friedman &

Barrett 2009). Larger pollen grain size also increases momentum, thereby increasing the chance of breaking away from deflected streamlines to collide with stigmas (Aylor & Parlange 1975; Friedman & Barrett 2009).

Whatever the reason for this asymmetry of pollen dispersal polarity, it has consequences for mating opportunities. *Q. petraea* tends to have directional crosses while *Q. robur* can disperse well in all directions. A more polarized dispersal should reduce mating opportunities, especially at low tree densities, giving *Q. robur* a clear edge under open conditions, at least if this greater polarity has a biological rather than environmental basis.

Previous studies have reported greater seed dispersal in *Q. robur* than in *Q. petraea* (mostly indirectly, by comparing genetic structure at maternally inherited markers; reviewed in Petit et al. 2003). Our study, limited to one year and one stand, found slightly higher acorn production for *Q. robur* than for *Q. petraea*, even though *Q. petraea* trees are on average larger than *Q. robur* trees. Greater seed dispersal ability and greater establishment success, due to the greater reserves in the cotyledons of *Q. robur* acorns (Dupouey & Le Bouler 1989; Gérard *et al.* 2009; Landergott *et al.* 2012) should combine with greater seed production to result in greater seed dispersal (Gaines *et al.* 1974). Hence, *Q. robur* has all the conditions for better pollen and seed dispersal than *Q. petraea*. These findings are consistent with the scheme of environment colonisation of both species with first installation of *Q. robur* in disturbed environments and subsequent arrival of *Q. petraea* in stands already colonized by *Q. robur* (Bacilieri *et al.* 1996b; Petit *et al.* 2003), provided that *Q. petraea* has sufficiently strong sexual barriers against *Q. robur* so as not to be swamped out when it is still at low density, surrounded by many allospecifics.

INTERSPECIFIC SEXUAL BARRIERS

Contrary to the conclusion of Bacilieri *et al.* (1995), these two species have asynchronous phenologies, with both *Q. petraea* female and male flowers maturing later on average than those of *Q. robur*. This phenological shift contributes therefore to limit interspecific crosses between these two species. As both species are protandrous, interspecific crosses should be more strongly reduced on *Q. petraea* mother-trees than on *Q. robur* mother-trees. Indeed, there will be few *Q. robur* trees still shedding pollen when the female flowers of the later flowering *Q. petraea* will finally become receptive, whereas when the first female flowers of *Q. robur* will start to be receptive, the first male flowers of *Q. petraea* should already have started to shed pollen. This phenological shift could be responsible for the asymmetry of interspecific sexual barriers observed under natural condition between these two species (Bacilieri *et al.* 1996b; Lagache *et al.* 2012). However, our study identified asymmetric sexual barriers for both species in addition to the effect contributed to by the phenological shift between the two species. This interspecific sexual barrier on *Q. petraea* mother trees was 35 times stronger than on *Q. robur* mother trees. While this finding might in principle be explained by imperfect phenological data, it is consistent with results from of interspecific controlled crosses conducted on these two species, which pointed to asymmetric mating compatibility in the same direction (e.g. Lepais 2008; Steinhoff 1993).

The smaller effective number of fathers in *Q. petraea* compared to *Q. robur* suggests that the greater homogeneity of individual male fecundities of *Q. petraea* is not sufficient to counterbalance its less effective pollen dispersal (in terms of distance and radial spread). This does not support the idea that the larger sexual barrier in *Q. petraea* is caused by a more competitive pollen environment in this species (i.e. greater sexual selection). Other

interpretations are called for. One possible interpretation is that superior siring ability of *Q. petraea* pollen on *Q. robur* stigmas, compared to the reciprocal cross, is caused by the larger pollen grain size of *Q. petraea*. Alternatively, this asymmetric prezygotic barrier could be a result of (asymmetric) reinforcement, due to differences in interspecific mating opportunities, as found in *Drosophila* (Yukilevich 2011).

INTERPRETATION OF INDIVIDUAL-BASED NEIGHBORHOOD MODELS

The neighbourhood model allows a rapid and simultaneous estimation of all the parameters influencing male fecundity and pollen dispersal (e.g. De Cauwer *et al.* 2012; DiFazio *et al.* 2012; Oddou-Muratorio *et al.* 2003). However it is based on a classical likelihood approach that is prone to false positives compared to Bayesian approaches (Klein *et al.* 2011). This information should be kept in mind when interpreting the results. More importantly, one should always recall that parameter estimation depends on the thoroughness of the model used. Omitting some important processes could greatly modify the value of other parameters of interest. For example, by comparing the mean pollen dispersal distances of the two species found in this study with those found by Lagache *et al.* (2012), we see that mean pollen dispersal distance of the two species differ although they are based on the same dataset. This difference is due in part to the addition of anisotropy of pollen dispersal in our model. This resulted in a significant model improvement over the previous simpler version. Indeed, in the model without anisotropy, mean pollen dispersal of both species is under-estimated, especially for *Q. petraea*, as found by Austerlitz *et al.* (2007) in a study of another oak species.

Similarly, interspecific sexual barriers depend on whether or not phenology is taken into account in the model. Parameter estimates obtained when phenology is modelled cannot be compared with estimates obtained when this process is not accounted for. In the latter case, the overall sexual barriers also include the effect of phenological differentiation between the two species.

CONCLUSION

Our study of the mating system and male fitness of these two closely related oak species suggested many links with their ecological dynamics. In turn, such differences in species dynamics could play an important role in triggering ecological speciation (Nosil 2012). A recent model by Burton *et al*. (2010) has shown how different strategies are selected for during range invasions. The differentiation between these two closely related oak species along the *r*/*K* axis of ecological strategies (Macarthur & Wilson 1967) clearly involve many traits related to growth, competition, pollen and seed dispersal. Such cases of multiple selection pressures (i.e. of “multifarious divergent selection”) corresponding to many interrelated selection trade-offs represent one of the most favorable situation conducive to ecological speciation and to the maintenance of newly formed species (Nosil *et al.* 2009).

ACKNOWLEDGEMENTS

We are grateful to Stefanie Wagner, for her help during sampling and Patrick Léger during microsatellite genotyping. Pauline Garnier-Géré provided support during the design of the 384plex SNP chip. We thank Erwan Guichoux to share data on the number of rings at different height for all the trees and Didier Bert for his help on the estimate of the age of each tree. The genotyping was performed at the Genome-Transcriptome facility of Bordeaux (grants from the Conseil Régional d'Aquitaine n°20030304002FA, n°20040305003FA and from the European Union, FEDER n°2003227). We thank Christophe Boury for his contribution with SNP genotyping. Funding was provided by the LinkTree project (ANR BIODIVERSA) and by the EU Network of Excellence EvolTree.

AUTHOR'S CONTRIBUTIONS:

RJP initially conceived the study, which evolved significantly with the help of all the authors. LL performed the experiments, produced and analyzed the data. EK performed the modelling. LL wrote the paper with the help of RJP, EK wrote part of the Methods. AD established the progeny test and shared information about tree characteristics and all four authors reviewed the complete manuscript.

REFERENCES

Alberto F, Bouffier L, Louvet JM*, et al.* (2011) Adaptive responses for seed and leaf phenology in natural populations of sessile oak along an altitudinal gradient. *Journal of Evolutionary Biology* **24**, 1442-1454.

Austerlitz F, Dutech C, Smouse PE, Davis F, Sork VL (2007) Estimating anisotropic pollen dispersal: a case study in Quercus lobata. *Heredity* **99**, 193-204.

Aylor DE, Parlange JY (1975) Ventilation required to entrain small particles from leaves. *Plant Physiology* **56**, 97-99.

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

Bacilieri R, Ducousso A, Kremer A (1996a) Comparison of morphological characters and molecular markers for the analysis of hybridization in sessile and pedunculate oak. *Annales Des Sciences Forestieres* **53**, 79-91.

Bacilieri R, Ducousso A, Petit RJ, Kremer A (1996b) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution* **50**, 900-908.

Becker M (1992) Deux indices de compétition pour la comparaison de la croissance en hauteur et en diamètre d'arbres aux passés sylvicoles variés et inconnus. *Ann. For. Sci.* **49**, 25-37.

Burczyk J, Lewandowski A, Chalupka W (2004) Local pollen dispersal and distant gene flow in Norway spruce (Picea abies [L.] Karst.). *Forest Ecology and Management* **197**, 39-48.

Burczyk J, Prat D (1997) Male reproductive success in Pseudotsuga menziesii (Mirb.) Franco: the effects of spatial structure and flowering characteristics. *Heredity* **79**, 638-647.

Burton OJ, Phillips BL, Travis JMJ (2010) Trade-offs and the evolution of life-histories during range expansion. *Ecology Letters* **13**, 1210-1220.

Byron BL, Rees RG, Witkowski ETF, Whitten VA (1994) Comparative size, fecundity and ecophysiology of roadside plants of *Banksia hookeriana*. *Journal of Applied Ecology* **31**, 137-144.

Cody ML (1966) A General Theory of Clutch Size. *Evolution* **20**, 174-184.

De Cauwer I, Arnaud JF, Klein EK, Dufay M (2012) Disentangling the causes of heterogeneity in male fecundity in gynodioecious Beta vulgaris ssp. maritima. *New Phytologist* **195**, 676-687.

DiFazio SP, Leonardi S, Slavov GT*, et al.* (2012) Gene flow and simulation of transgene dispersal from hybrid poplar plantations. *New Phytologist* **193**, 903-915.

Dupouey J, L., Le Bouler H (1989) Discrimination morphologique des glands de chênes sessile (*Quercus petraea* (Matt.) Liebl.) et pédonculé (*Quercus robur* L.). *Ann. For. Sci.* **46**, 187-194.

Friedman J, Barrett SCH (2009) Wind of change: new insights on the ecology and evolution of pollination and mating in wind-pollinated plants. *Annals of Botany* **103**, 1515-1527.

Gaines MS, Vogt KJ, Hamrick JL, Caldwell J (1974) Reproductive Strategies and Growth Patterns in Sunflowers (Helianthus). *The American Naturalist* **108**, 889-894.

Gérard B, Alaoui-Sossé B, Badot P-M (2009) Flooding effects on starch partitioning during early growth of two oak species. *Trees - Structure and Function* **23**, 373-380.

Guichoux E, Garnier-Géré P, Lagache L*, et al.* (2012) Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, *in press (MEC-12-0795.R0791)*.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources* **11**, 578-585.

Harper JL, Ogden J (1970) The Reproductive Strategy of Higher Plants: I. The Concept of Strategy with Special Reference to Senecio Vulgaris L. *Journal of Ecology* **58**, 681-698.

Holsinger KE (1991) Mass-action models of plant mating systems: the evolutionary stability of mixed mating systems. *The American Naturalist* **138**, 606-622.

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**.

Jensen JS (2000) Provenance variation in phenotypic traits in *Quercus robur* and *Quercus petraea* in danish provenance trials. *Scandinavian Journal of Forest Research* **15**, 297-308.

Klein EK, Carpentier FH, Oddou-Muratorio S (2011) Estimating the variance of male fecundity from genotypes of progeny arrays: evaluation of the Bayesian forward approach. *Methods in Ecology and Evolution* **2**, 349-361.

Lagache L, Klein EK, Guichoux E, Petit RJ (2012) Fine-scale environmental control of hybridization in oaks. *Molecular Ecology*, *in press*.

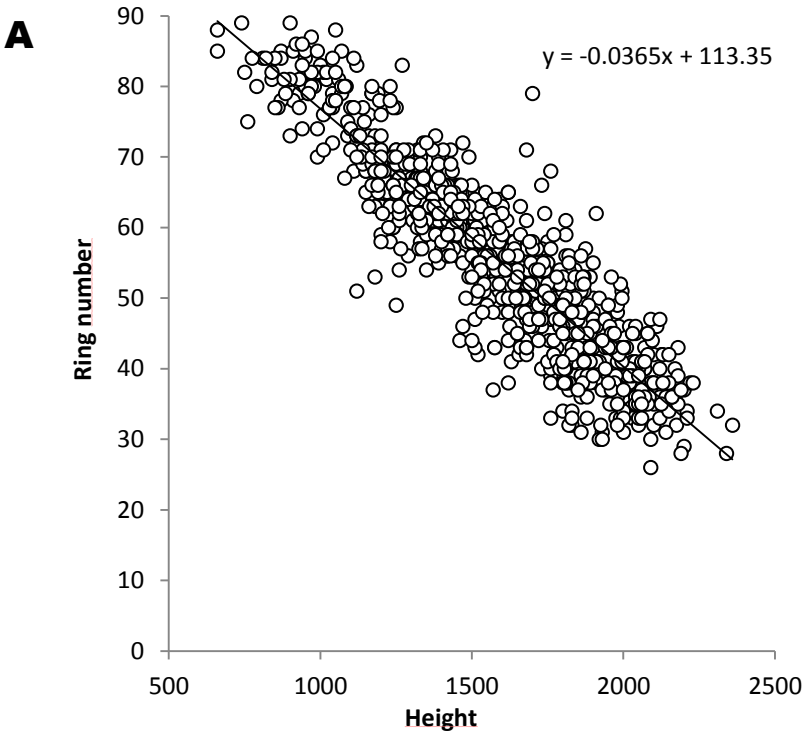
Landergott U, Gugerli F, Hoebee SE, Finkeldey R, Holderegger R (2012) Effects of seed mass on seedling height and competition in European white oaks. *Flora - Morphology, Distribution, Functional Ecology of Plants*.

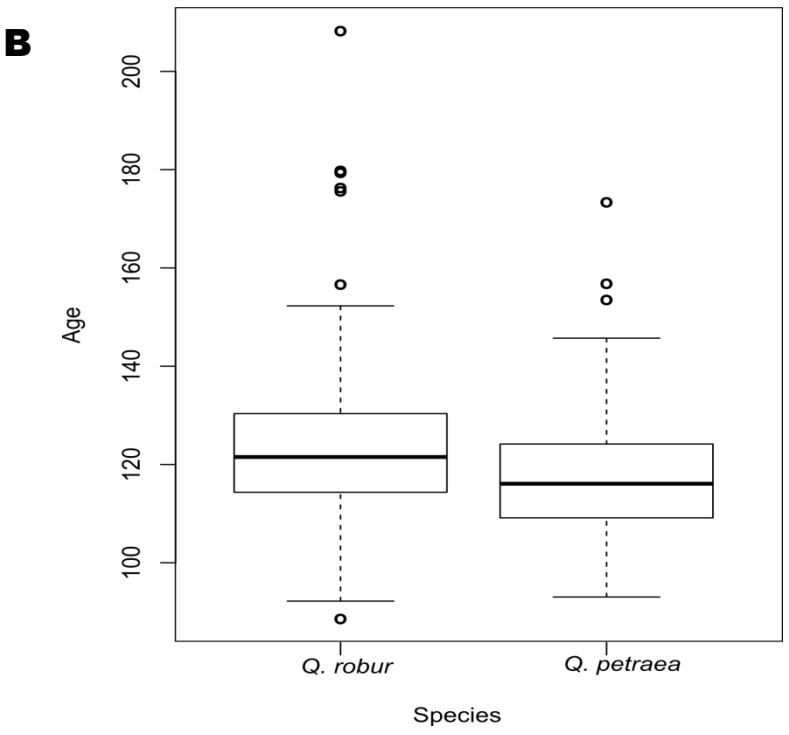
Lepais O (2008) *Dynamique d'hybridation dans le complexe d'espèces des chênes blancs européens (Ph. D.)*, Université Bordeaux 1.
Macarthur RH, Wilson EO (1967) *The Theory of Island Biodiversity* Princeton University Press.
Newell SJ, Tramer EJ (1978) Reproductive Strategies in Herbaceous Plant Communities During Succession. *Ecology* **59**, 228-234.
Nielsen R, Tarpy DR, Reeve HK (2003) Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology* **12**, 3157-3164.
Niklas K (1985) The aerodynamics of wind pollination. *The Botanical Review* **51**, 328-386.
Nosil P (2012) *Ecological speciation* Oxford University Press, Oxford.
Nosil P, Harmon LJ, Seehausen O (2009) Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution* **24**, 145-156.
Obeso JR (2002) The costs of reproduction in plants. *New Phytologist* **155**, 321-348.
Oddou-Muratorio S, Houot ML, Demesure-Musch B, Austerlitz F (2003) Pollen flow in the wildservice tree, Sorbus torminalis (L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. *Molecular Ecology* **12**, 3427-3439.
Oddou-Muratorio S, Klein EK, Austerlitz F (2005) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. II. Pollen dispersal and heterogeneity in mating success inferred from parent–offspring analysis. *Molecular Ecology* **14**, 4441-4452.
Parelle J, Brendel O, Bodénès C*, et al.* (2006) Differences in morphological and physiological responses to water-logging between two sympatric oak species (*Quercus petraea* [Matt.] Liebl., *Quercus robur* L.). *Ann. For. Sci.* **63**, 849-859.
Petit RJ, Bialozyt R, Garnier-Géré P, Hampe A (2004) Ecology and genetics of tree invasions: from recent introductions to Quaternary migrations. *Forest Ecology and Management* **197**, 117-137.
Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2003) Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164.
Pluess AR, Sork VL, Dolan B*, et al.* (2009) Short distance pollen movement in a wind-pollinated tree, Quercus lobata (Fagaceae). *Forest Ecology and Management* **258**, 735-744.
Ponton S, Dupouey J-L, Bréda N, Dreyer E (2002) Comparison of water-use efficiency of seedlings from two sympatric oak species: genotype × environment interactions. *Tree Physiology* **22**, 413-422.
Robledo-Arnuncio JJ, Gil L (2005) Patterns of pollen dispersal in a small population of Pinus sylvestris L. revealed by total-exclusion paternity analysis. *Heredity* **94**, 13-22.
Rushton BS (1976) Pollen grain size in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. *Watsonia* **11**, 137-140.
Shen H, Rudin D, Lindgren D (1981) Study of the pollination pattern in a Scots pine seed orchard by means of isozyme analysis. *Silvae Genetica*, 7-15.
Slavov GT, Howe GT, Adams WT (2005) Pollen contamination and mating patterns in a Douglas-fir seed orchard as measured by simple sequence repeat markers. *Canadian Journal of Forest Research* **35**, 1592-1603.
Smouse PE, Sork VL (2004) Measuring pollen flow in forest trees: an exposition of alternative approaches. *Forest Ecology and Management* **197**, 21-38.
Starr CK (1984) Sperm competition, kinship, and sociality in the aculeate Hymenoptera. *Sperm Competition and the Evolution of Animal Mating Systems*.
Steinhoff S (1993) Results of species hybridization with *Quercus robur* L and *Quercus petraea* (Matt.) Liebl. *Annals of Forest Science* **50**, 137s-143s.
Torimaru T, Wennstrom U, Lindgren D, Wang XR (2012) Effects of male fecundity, interindividual distance and anisotropic pollen dispersal on mating success in a Scots pine (Pinus sylvestris) seed orchard. *Heredity* **108**, 312-321.
van Kleunen M, Burczyk J (2008) Selection on floral traits through male fertility in a natural plant population. *Evolutionary Ecology* **22**, 39-54.
Wendt T, da Cruz DD, Demuner VG, Guilherme FAG, Boudet-Fernandes H (2011) An evaluation of the species boundaries of two putative taxonomic entities of Euterpe (Arecaceae) based on reproductive and morphological features. *Flora - Morphology, Distribution, Functional Ecology of Plants* **206**, 144-150.
Westoby M, Falster DS, Moles AT, Vesk PA, Wright IJ (2002) Plant Ecological Strategies: Some Leading Dimensions of Variation between Species. *Annual Review of Ecology and Systematics* **33**, 125-159.
Whitehead DR (1969) Wind pollination in the angiosperms: evolutionary and environmental considerations. *Evolution* **23**, 28-35.
Wolfram Research Inc. (2010) *Mathematica Edition: Version 8.0* Wolfram Research, Inc., Champaign, Illinois.
Yukilevich R (2011) Asymmetrical patterns of speciation uniquely support reinforcement in *Drosophila*. *Evolution*, no-no.

SUPPORTING INFORMATION

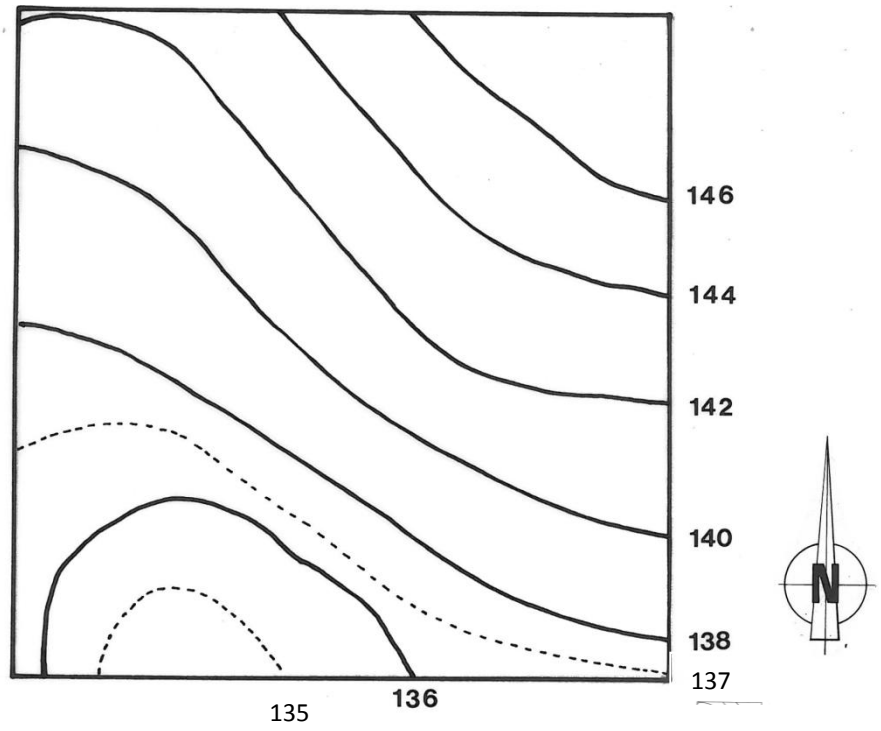
SUPPORTING INFORMATION 1: DETERMINING THE AGE OF TREES BASED ON THE NUMBER OF RINGS (TAKEN AT THREE DIFFERENT HEIGHTS) FOR EACH TREE OF THE STAND.

A: Number of rings as a function of tree height. B: Age distribution of each individual for each species



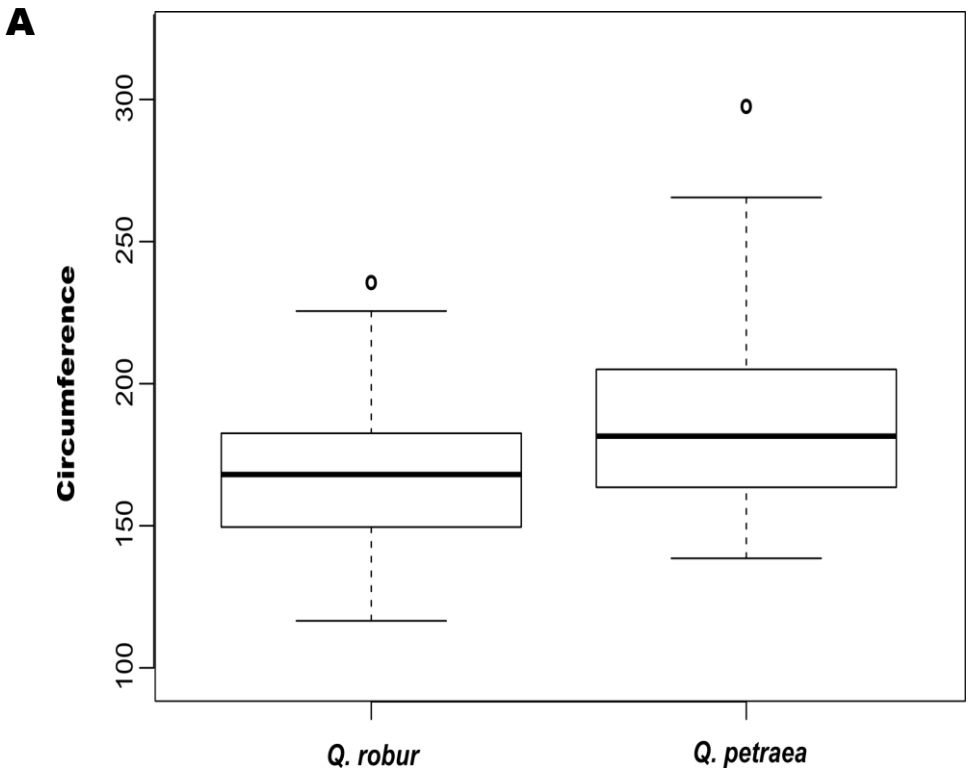


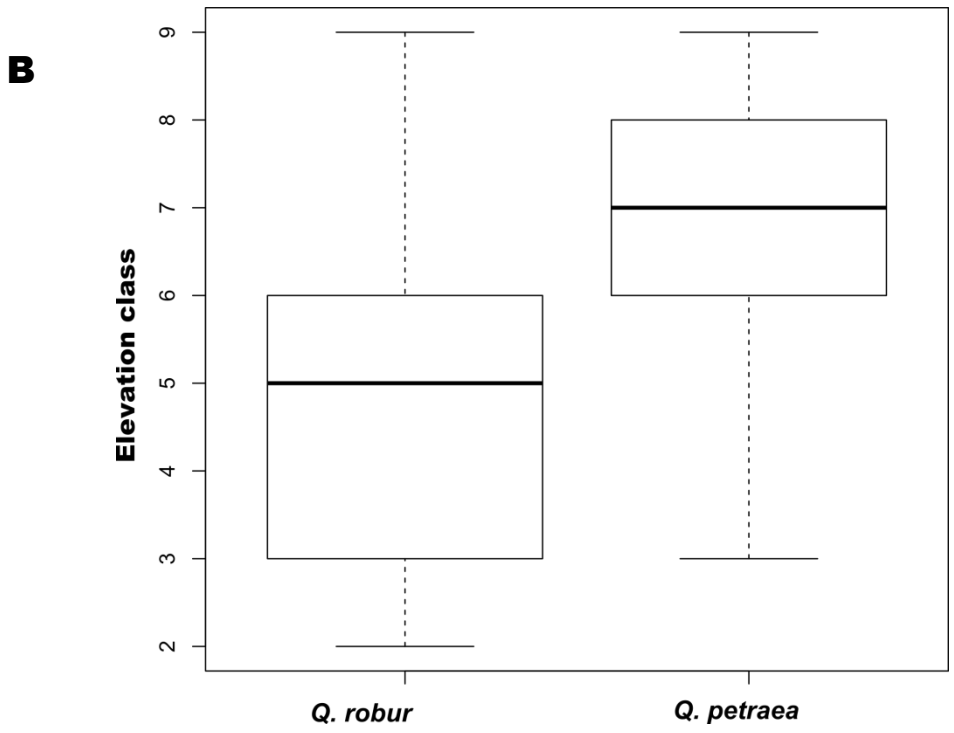
<u>SUPPORTING INFORMATION 2:</u> TOPOGRAPHIC MAP OF THE STAND AND DEFINITION OF TERRAIN ELEVATION CLASSES.



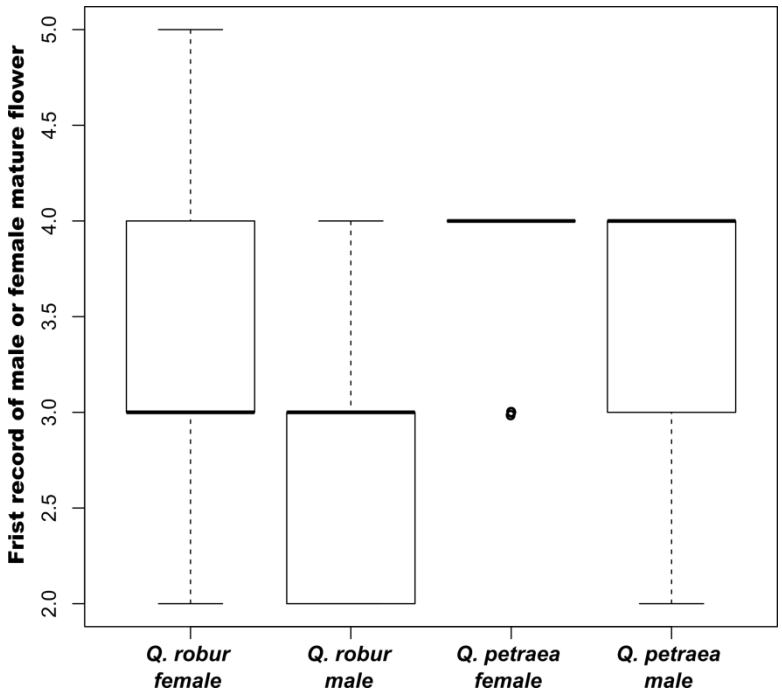
Terrain elevation classes	Terrain elevation (m)
1	less than 135
2	135-136
3	136-137
4	137-138
5	138-140
6	140-142
7	142-144
8	144-146
9	more than 146

<u>SUPPORTING INFORMATION 3:</u> DISTRIBUTION OF INDIVIDUAL CIRCUMFERENCE (A), TERRAIN ELEVATION CLASS (B) AND FIRST RECORD OF MALE AND FEMALE MATURE FLOWERS (C) FOR EACH SPECIES.





c


Frist record of male or female mature flower
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Q. robur female
Q. robur male
Q. petraea female
Q. petraea male

ORIGINE ET DEROULEMENT DE CE TRAVAIL

Les travaux des écologistes forestiers ont montré que *Q. robur* est une espèce pionnière colonisatrice de milieux ouverts alors que *Q. petraea* est une espèce post-pionnière plus compétitrice arrivant plus tard au cours de la succession (in Bacilieri *et al.* 1996; Petit *et al.* 2003). Ces deux espèces, bien que pouvant s'hybrider en conditions de croisements contrôlés (ex. Steinhoff 1993) et en conditions naturelles (e.g. Jensen *et al.* 2009), ont par ailleurs des stratégies reproductives différentes. Par exemple, *Q. robur* disperse son pollen (Jensen *et al.* 2009) et ses graines (Jones 1959; Petit *et al.* 2003) plus loin que *Q. petraea*. L'enjeu de ce travail a été de comprendre si les différences de stratégies de reproduction et également de croissance correspondaient à ce qu'on pouvait prédire au vu de leur dynamique écologique.

Chronologiquement, l'étude présentée dans ce chapitre est la dernière que j'ai faite durant ma thèse. Dans le modèle de voisinage présenté dans cette étude, j'avais initialement prévu d'étudier un possible mécanisme empêchant les croisements entre individus apparentés (en relation avec la dépression de consanguinité, Charlesworth & Charlesworth 1987). Sachant que *Q. petraea* a une structure génétique spatiale plus marquée (Streiff *et al.* 1998) et disperse moins son pollen (Jensen *et al.* 2009), j'avais émis l'hypothèse que cette espèce ait pu développer une barrière sexuelle plus forte que *Q. robur* contre les croisements avec des individus apparentés. Malheureusement, j'ai dû abandonner l'étude de l'effet de l'apparentement des arbres adultes sur leur reproduction en lien avec leurs stratégies sexuelles. En effet, lors des analyses, l'ajout de ce paramètre empêchait la convergence vers une situation avec un minimum local. Ce point reste donc à étudier. Initialement j'avais également inclus la hauteur dans le modèle comme un possible paramètre affectant la fécondité mâle mais l'ajout de ce paramètre n'a pas amélioré significativement le modèle (qui comprenait déjà un possible effet de la circonférence sur les fécondités mâles). Je pense que la forte corrélation entre hauteur et circonférence explique ce dernier résultat. Je disposais également d'une carte avec les profondeurs auxquelles le sol avait été retrouvé saturé en eau (d'après l'étude de Bacilieri *et al.* 1995). J'avais choisi d'intégrer également ces données dans le modèle en envisageant un possible effet de la disponibilité en eau du sol sur la fécondité mâle, à l'instar des données topographiques incluses dans ce chapitre. Mais tout comme pour la hauteur, et bien qu'il n'existe pas de réelle corrélation entre topographie (ou circonférence) et hydromorphie du sol dans cette parcelle, la prise en compte de ces données pour expliquer les croisements observés au travers de la fécondité mâle n'améliore pas significativement le modèle.

PERSPECTIVES DE L'ETUDE

Dans ce chapitre je me suis essentiellement concentrée sur les différences entre espèces considérées au travers des croisements intraspécifiques. Pour aller plus loin dans l'étude de la dynamique de colonisation de ces deux espèces, je présente ici l'étude plus détaillée correspondant à une étape clé : l'installation de *Q. petraea* dans un peuplement de *Q. robur*. Pour cela j'ai utilisé l'estimation des paramètres du modèle présenté dans ce chapitre ainsi que la position géographique et les caractéristiques réelles des arbres mais j'ai modifié l'identité des arbres en créant un peuplement largement *Q. robur* dans lequel quelques individus de *Q. petraea* se seraient installés (Figure 1), et vice versa, pour comparaison. Pour ces simulations, tous les caractères influençant les croisements des deux espèces sont pris en compte (dispersion du pollen, y compris anisotropie, phénologie et

circonférence). De plus, ce sont les valeurs des paramètres estimés dans cette étude qui sont pris en compte pour prédire le taux d'hybridation dans les différentes configurations.

Lorsque *Q. petraea* se retrouve en minorité dans un peuplement de *Q. robur*, le taux d'hybridation sur mère *Q. petraea* est bien plus élevé que dans notre étude (i.e. 13.5% contre 0.1%) mais reste plus faible que pour le *Q. robur* dans la même situation (jusqu'à 28%, Table 1). Le doublement du nombre d'individus initial (de 5 à 10) n'a pas une grande influence sur le taux d'hybridation. En revanche la taille du cluster des individus minoritaires (et donc la densité de ces individus) est importante. En effet, plus les individus sont distants et plus leur taux d'hybridation moyen augmente.

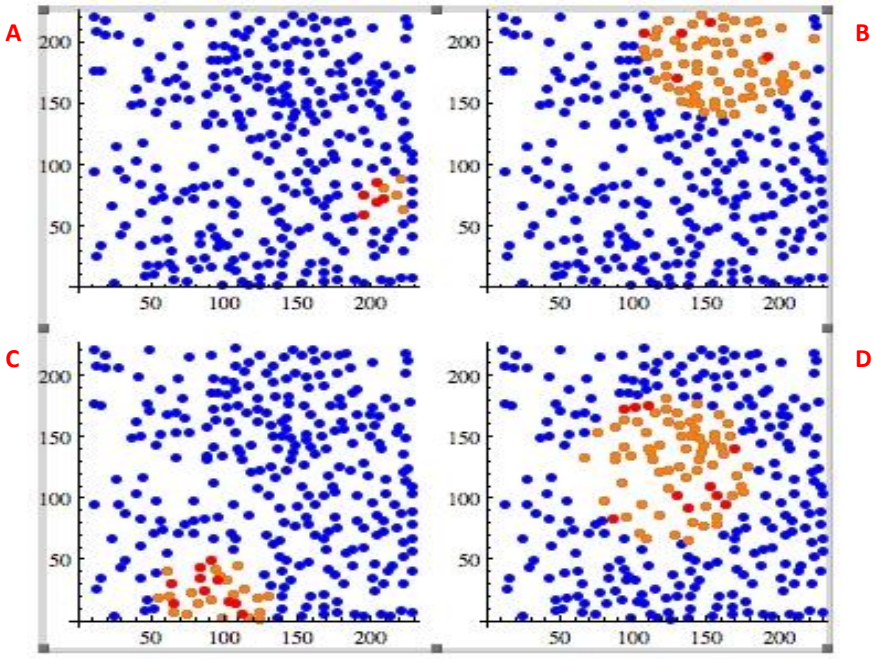


Figure 1 : *Q. robur (en bleu) est ici l'espèce majoritaire. Afin de tester si le nombre et/ou la densité des individus Q. petraea a un effet sur le taux d'hybridation moyen, 100 tirages aléatoires de 5 ou 10 individus dans des rayons différents ont été réalisés. Toutes les positions possibles de Q. petraea sont en rouge et orange. Les points en rouge (= 5 ou 10) sont les arbres Q. petraea tirés au hasard pour lesquels le taux d'hybridation sera prédit à l'aide du modèle. A, B, C, et D sont des configurations où le rayon dans lequel seront tirés aléatoirement les individus de Q. petraea est différent (respectivement 20, 40, 60 and 100 m). Ces configurations ont également été étudiées avec Q. petraea comme espèce majoritaire pour comparaison.*

Rayon du cluster	Nombre d'arbres de l'espèce minoritaire	**Cluster *Q. petraea* dans un peuplement majoritaire de *Q. robur***	**Cluster *Q. robur* dans un peuplement majoritaire de *Q. petraea***
20	5	6.5%	10.9%
40	5	9.3%	17.9%
60	5	10.4%	22.2%
100	5	13.5%	28.6%
40	10	11.9%	25.0%
60	10	10.8%	23.7%
100	10	10.3%	24.2%

Table 1 : *Taux d'hybridation sur mère Q. robur et Q. petraea dans différentes configurations (densité et nombre d'individu) en imaginant qu'aucun pollen n'arrive de l'extérieur. Dans ce tableau sont présentés le taux d'hybridation moyen de 100 tirages aléatoires d'individus pour chaque configuration testée.*

Si les mères *Q. petraea* avaient une barrière à l'hybridation identique à celle des mères *Q. robur* (donc des barrières plus faibles), des taux d'hybridation encore plus importants seraient alors observés (Table 2), compromettant l'installation de *Q. petraea* dans un peuplement déjà établi de *Q. robur :*

Rayon du cluster	Nombre d'arbres de l'espèce minoritaire	**Cluster *Q. petraea* dans un peuplement majoritaire de *Q. robur***	**Cluster *Q. robur* dans un peuplement majoritaire de *Q. petraea***
20	5	54.2%	10.9%
40	5	61.0%	17.9%
60	5	65.0%	22.2%
100	5	69.0%	28.6%
40	10	64.8%	25.0%
60	10	63.7%	23.7%
100	10	63.9%	24.2%

Table 2 : *Taux d'hybridation sur mère Q. robur et Q. petraea dans différentes configurations (densité, nombre d'individu), sans pollen immigrant et si les deux espèces avaient la même barrière à l'hybridation (ici j'ai choisi celle d'une mère Q. robur). Dans ce tableau sont présentés le taux d'hybridation moyen de 100 tirages aléatoires d'individu pour chaque configuration testée.*

L'installation de *Q. petraea* dans un peuplement de *Q. robur* doit donc résulter de très fortes pressions de sélection en faveur des individus *Q. petraea* (croissance, compétition…), comme nous le montrons dans ce chapitre, mais est aussi rendu possible par les très fortes barrières des mères *Q. petraea* contre les croisements avec du pollen de *Q. robur*. Une hypothèse rendant compte de cette observation est que ces barrières asymétriques résultent d'un renforcement asymétrique des barrières. Si tel était le cas, on pourrait réellement parler d'un réel processus d'« isolation ». En effet, c'est le contact avec l'autre

espèce qui entrainerait une pression de sélection importante vers une asymétrie de l'hybridation en renforçant les barrières de l'espèce la plus menacée par l'hybridation, comme cela a été montré récemment pour les drosophiles (Yukilevich 2011). Il serait intéressant de tester cette hypothèse en étudiant les barrières chez des arbres *Q. petraea* en situation d'allopatrie, même si cette situation est peu fréquente (cf. Figure 5 en introduction de cette thèse). Si on trouve des barrières interspécifiques moins fortes dans cette situation, cela viendrait conforter l'hypothèse de renforcement asymétrique.

RÉFÉRENCES

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

Bacilieri R., Ducousso A., Petit R.J. & Kremer A. (1996). Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution*, 50, 900-908.

Charlesworth D. & Charlesworth B. (1987). Inbreeding Depression and its Evolutionary Consequences. *Annu. Rev. Ecol. Syst.*, 18, 237-268.

Jensen J., Larsen A., Nielsen L.R. & Cottrell J. (2009). Hybridization between *Quercus robur* and *Q. petraea* in a mixed oak stand in Denmark. *Ann. Forest Sci.*, 66.

Jones E.W. (1959). Quercus L. *J. Ecol.*, 47, 169-222.

Petit R.J., Bodénès C., Ducousso A., Roussel G. & Kremer A. (2003). Hybridization as a mechanism of invasion in oaks. *New Phytol.*, 161, 151-164.

Steinhoff S. (1993). Results of species hybridization with *Quercus robur* L and *Quercus petraea* (Matt.) Liebl. *Ann. Forest Sci.*, 50, 137s-143s.

Streiff R., Ducousso A. & Kremer A. (1998). Spatial genetic structure and pollen gene flow in a mixed oak stand. *Genet. Sel. Evol.*, 30, S137-S152.

Yukilevich R. (2011). Asymmetrical patterns of speciation uniquely support reinforcement in *Drosophila*. *Evolution*, no-no.

CONCLUSIONS ET PERSPECTIVES

Parcelle mixte de chêne sessile et pédonculé dans la forêt de la Petite Charnie (P26, milieu de pente, 1991)

Mon travail de thèse prolonge celui d'Olivier Lepais, ancien doctorant de l'unité, dont la thèse s'intitulait « Dynamique d'hybridation dans le complexe d'espèces des chênes blancs européens » (Lepais 2008). Son étude sur l'effet de l'abondance relative des espèces sur leur hybridation (Lepais *et al.* 2009) est reconnue comme appartenant à un front de science et est devenu une référence pour les études sur l'hybridation. Elle a permis à la communauté scientifique de redécouvrir « l'effet Hubbs » (Hubbs 1955) au travers d'observations sur les taux d'hybridation à l'échelle des populations. Mon travail de thèse a consisté à poursuivre son travail de deux façons. La première a été de changer d'échelle et d'étudier cet effet au niveau individuel. La seconde a été de modéliser cet effet de manière spatialement explicite en conditions naturelles. De façon générale, mon travail a permis d'apporter de nouveaux éléments concernant l'effet de l'environnement sur l'hybridation de ces deux espèces (abondance relative et absolue des espèces, effet de la spatialisation et de la fragmentation). Il m'a aussi permis d'établir un lien entre les stratégies reproductives et de croissance mises en place par deux espèces proches en lien avec leur dynamique écologique. Enfin, ce travail a été l'occasion d'approfondir la notion d'espèce et de montrer comment le critère d'interfertilité pouvait s'appliquer y compris dans le cas d'un complexe d'espèces dont les limites pouvaient sembler floues.

QUERCUS ROBUR ET *Q. PETRAEA*, DEUX ESPECES ?

Un débat existe quant au statut de ces deux taxons, correspondent-t-ils à une seule et même espèce, vu l'absence de caractère diagnostic et la fréquence des échanges, comme le soutenaient Kleinschmit & Kleinschmit (2000), ou constituent-ils deux espèces avec des barrières à l'hybridation semi-perméables, comme l'affirmaient par exemple Lexer *et al.* (2006), en réponse à Muir & Schlötterer (2005) qui de leur côté doutaient de l'existence même d'hybridation entre ces espèces? L'étude faite dans le **chapitre 1** me semble clarifier ce débat. En effet, cette étude a permis de mettre en évidence que *Q. robur* et *Q. petraea* sont deux groupes différents à tous points de vue (morphologie, génétique, interfertilité, apparentement). Même si elles s'hybrident, ce sont bien deux entités vivant en sympatrie avec des stratégies reproductive et de croissance différentes (étudiées dans le **chapitre 3**). La méthodologie de reconstitution des réseaux de croisements présentée dans le **chapitre 1** est nouvelle ; elle permet de délimiter des espèces même quand celles-ci s'hybrident assez fréquemment, au moins dans les cas où les espèces sont suffisamment polygames. La multiplication d'études comparant in situ différentes méthodes de délimitation d'espèce pourrait aider à mieux comprendre comment la spéciation progresse, en identifiant des cas où certaines méthodes mais pas d'autres permettent d'aboutir à une classification claire des individus. Enfin, cela devrait aider les taxonomistes à prendre des décisions sur le statut taxonomique des entités en présence, si possible en intégrant différentes informations (Padial *et al.* 2010).

L'HYBRIDATION DEPEND DU CONTEXTE

Dans le **chapitre 2**, j'ai montré à quel point l'hybridation dépendait du contexte. Ce travail permet de comprendre comment l'ouverture d'un peuplement forestier, en mettant en contact des individus initialement séparés et en diminuant la densité, devrait mécaniquement augmenter l'hybridation. Disposer d'un modèle prédictif pour l'hybridation (au-delà de la nécessité de la sympatrie) est un progrès important qui mériterait d'être confirmé, soit expérimentalement, en manipulant la composition des peuplements, soit en estimant le taux d'hybridation d'un même peuplement plusieurs années de suite, pour

tester si l'hybridation augmente les années de plus faible production de pollen, ou pour les individus dont la floraison est la plus tardive (et donc recevant à ce titre particulièrement peu de pollen). L'hybridation a longtemps été perçue comme un phénomène rare, or on se rend compte que suite aux simulations présentées dans les perspectives du **chapitre 3**, elle peut être très importante dans certaines configurations, notamment lors de forts déséquilibres de l'effectif des espèces. Il est donc peu pertinent de déterminer une fréquence « moyenne » d'hybridation entre deux espèces sans préciser le contexte.

DE NOUVEAUX ELEMENTS POUR LE MODELE D'INTROGRESSION DE CES DEUX ESPECES

Les différences de stratégie reproductive et de croissance de ces deux espèces étudiées dans le **chapitre 3** confortent le modèle de colonisation des milieux de ces deux espèces. Ainsi, *Q. robur* disperse mieux ses graines (Jones 1959; Petit *et al.* 2003) et son pollen, alors que *Q. petraea* investit davantage son énergie dans la croissance et la compétition. Cependant, les modélisations des croisements intra- et interspécifique de cette thèse ne permettent pas d'évaluer complètement le modèle d'introgression proposé par Petit *et al.* (2003) où *Q. petraea* acquiert des parties du génome de *Quercus robur* par hybridation et puis rétro-croisements ou celui plus général de Currat *et al.* (2008). La phase d'installation de *Q. petraea* dans un peuplement composé en grande majorité de *Q. robur* doit pour cela être considérée, là où l'hybridation est maximale du fait de l'asymétrie d'abondance des espèces (voir perspectives **chapitre 3**). Une fois installé, un arbre (ou un petit groupe d'arbres) de type *Q. petraea* devrait produire un nombre important d'hybrides (voir perspectives **chapitre 2**). Qu'advient-il de ces descendants hybrides? D'après l'étude de Lepais & Gerber (2011), les arbres intermédiaires se reproduisent avec les deux espèces parentales, en fonction de leur abondance. Dans de telles conditions, les hybrides se reproduiront essentiellement avec *Q. robur*. Par contre, en supposant de même que les arbres d'espèce pure se reproduisent aussi bien ou presque avec du pollen d'hybrides qu'avec du pollen conspécifique, les arbres mères *Q. petraea*, peu nombreux, recevront une plus grande proportion de pollen hybride que les arbres mères *Q. robur*. Ainsi des backcross des deux espèces seront formés (backcross *Q. robur* essentiellement issus des mères hybrides F1 et backcross *Q. petraea* essentiellement issus des mères *Q. petraea*). Une modélisation plus fine et plus complète que celle proposée par Currat *et al.* (2008), tenant compte des sexes des individus (père ou mère) et modélisant les abondances relatives de chaque catégorie formée (*Q. petraea* et *Q. robur* purs et hybride F1) selon le modèle de Chan & Levin (2005) devra être réalisée pour mieux comprendre l'asymétrie d'introgression récemment observée chez ces deux espèces. En effet, un signal d'introgression asymétrique a été trouvé à l'aide de SNPs hautement différenciés entre espèces : alors que *Q. petraea* possède des variants privés, l'inverse n'est pas vrai et les marqueurs les plus fréquents chez *Q. robur* se retrouvent en plus faible fréquence chez *Q. petraea* (Guichoux *et al.* 2012, voir **Annexe 3**). Si des backcross des deux espèces sont produits, comment le déséquilibre d'abondance entre les deux espèces peut entrainer une telle asymétrie? Afin de compléter le modèle d'installation de *Q. petraea* dans un peuplement de *Q. robur*, il serait donc important d'étudier plus précisément le système de reproduction des individus F1 avec les espèces pures dans des parcelles mixtes où il existe un grand nombre d'hybrides F1 et avec une résolution suffisante pour délimiter espèces parentales et hybrides. Une étude de modélisation des croisements en limite Nord d'extension de *Q. petraea* ou dans une forêt jeune où *Q. petraea* est en phase d'installation initiale (moins avancée que dans notre parcelle d'étude) pourrait nous apporter des éléments importants sur cette question.

LE DEVENIR DE *Q. ROBUR* DANS UN PEUPLEMENT MIXTE

L'étude du chapitre 3 permet de mieux comprendre la dynamique de colonisation des milieux par *Q. robur* et *Q. petraea* au travers des stratégies reproductive et de croissance de ces deux espèces. *Q. robur* investirait donc préférentiellement son énergie dans la dispersion, coloniserait de nouveaux milieux et *Q. petraea* arriverait en suivant en investissant son énergie dans la croissance et la compétition. De part sa meilleur aptitude à la compétition, on peut s'attendre à ce que l'augmentation de *Q. petraea* se fasse aux dépens de *Q. robur*. Une étude de la régénération d'une parcelle comme celle que nous avons étudiée (Parcelle 26 de la Forêt Domaniale de la Petite Charnie) permettrait d'étudier l'évolution de la composition spécifique et ainsi de tester si *Q. petraea* augmente en abondance aux dépens de *Q. robur*.

LES STRATEGIES REPRODUCTIVES FEMELLE DES DEUX ESPECES

L'étude présentée dans le **chapitre 3** nous a conduit à proposer qu'un seul axe initial de divergence entre deux populations (l'axe appelé *r*/*K*; Macarthur & Wilson 1967) pourrait entrainer une sélection divergente pour de nombreux caractères (*« multifarious divergent selection »*), qu'ils soient relatifs à la fonction mâle ou à la croissance végétative. Or il est clair que plus la sélection divergente porte sur de nombreux caractères, plus elle a des chances d'aboutir à la formation de nouvelles espèces (Nosil 2012). Pour aller plus loin, une étude similaire pourrait être faite sur les stratégies reproductives femelles de ces deux espèces afin d'étudier quelles caractéristiques diffèrent entre espèces en lien avec leur dynamique écologique. Dans notre travail, cela n'a pas été possible car les graines ont été ramassées sur les arbre-mères avant qu'elles ne soient dispersées. Il faudrait pour cela réaliser une étude de parenté complète du jeune peuplement qui s'est installé par régénération naturelle dans notre parcelle d'étude, comme proposé ci-dessus.

TOUJOURS PLUS D'ECOLOGIE …

Une analyse et une interprétation plus « écologique » de la spéciation a aboutit à définir la « spéciation écologique » et a permis d'effectuer de nombreuses avancées (Nosil 2012). Les récents travaux sur les chênes menés au laboratoire (Lepais 2008; Abadie *et al.* 2011) et poursuivis dans cette thèse ont montré à quel point le fonctionnement des barrières à l'hybridation ne pouvait être bien compris que si il était étudié *in situ*, dans des conditions naturelles. L'hybridation apparait ainsi tout autant « écologique » que la spéciation.

REFERENCES

Abadie P., Roussel G., Dencausse B., Bonnet C., Bertocchi E., Louvet J.M., Kremer A. & Garnier-Géré P. (2011). Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *J. Evol. Biol.*, 25, 157-173.

Chan K.M.A. & Levin S.A. (2005). Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution*, 59, 720-729.

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

Guichoux E., Garnier-Géré P., Lagache L., Lang T., Bourry C. & Petit R.J. (2012). Outlier loci highlight the direction of introgression in oaks. *Mol. Ecol.*, *in press (MEC-12-0795.R1)*.

Hubbs C.L. (1955). Hybridization between fish species in nature. *Syst. Zool.*, 4, 1-20.

Jones E.W. (1959). Quercus L. *J. Ecol.*, 47, 169-222.

Kleinschmit J. & Kleinschmit J.G.R. (2000). *Quercus robur* - *Quercus petraea*: a critical review of the species concept. *Glasnik za Sumske Pokuse*, 37, 441-452.

Lepais O. (2008). Dynamique d'hybridation dans le complexe d'espèces des chênes blancs européens (Ph. D.). In: *Ecole doctorale Sciences et Environnement*. Université Bordeaux 1 Talence, France. .

Lepais O. & Gerber S. (2011). Reproductive patterns shape introgression dynamics and species succession within the European white oak species complex. *Evolution*, 65, 156-170.

Lepais O., Petit R.J., Guichoux E., Lavabre J.E., Alberto F., Kremer A. & Gerber S. (2009). Species relative abundance and direction of introgression in oaks. *Mol. Ecol.*, 18, 2228-2242.

Lexer C., Kremer A. & Petit R.J. (2006). COMMENT: Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Mol. Ecol.*, 15, 2007-2012.

Macarthur R.H. & Wilson E.O. (1967). *The Theory of Island Biodiversity*. Princeton University Press.

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.

Nosil P. (2012). *Ecological speciation*. Oxford University Press, Oxford.

Padial J., Miralles A., De la Riva I. & Vences M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16.

Petit R.J., Bodénès C., Ducousso A., Roussel G. & Kremer A. (2003). Hybridization as a mechanism of invasion in oaks. *New Phytol.*, 161, 151-164.

Source: http://blogs.msdn.com/b/willy-peter_schaub/

ANNEXE 1: TWO HIGHLY VALIDATED MULTIPLEXES (12-PLEX AND 8-PLEX) FOR SPECIES DELIMITATION AND PARENTAGE ANALYSIS IN OAKS (*QUERCUS SPP.)*

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.02983.x

MOLECULAR DIAGNOSTICS AND DNA TAXONOMY

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

E. GUICHOUX*†‡, L. LAGACHE*†, S. WAGNER*†§, P. LÉGER*† and R.J. PETIT*†
*INRA, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France, †University of Bordeaux, UMR Biodiversité Gènes & Communautés 1202, F-33610 Cestas, France, ‡Centre de Recherche Pernod-Ricard, F-94000 Créteil, France, §University of Bonn, Steinmann Institut, D-53115 Bonn, Germany

Abstract

Multiplex PCR is a fast and cost-effective technique allowing increased genotyping throughput of microsatellites. We developed two multiplexes for *Quercus petraea* and *Q. robur*, a 12-plex of EST-SSRs (eSSRs) and an 8-plex of genomic SSRs (gSSRs). We studied the origin of allele calling errors at the human reader and software levels. We showed that the robustness of allele identification can be improved by binning on raw peak sizes prior to genetic data analysis. We checked through simulation the power of these markers for species delimitation and hybrid detection. The resolution achieved with all 20 markers was greatly improved compared to that of previous studies based on a subset of the markers. Preliminary PCR tests suggest that these multiplexes might be useful to study other oak species as well. The strategy used for multiplex microsatellite development (from PCR conditions to the definition of allele calling rules) should be broadly applicable.

Keywords: manual binning, microsatellites, molecular identification, *Quercus spp.*

Received 5 October 2010; revision received 2 December 2010; accepted 10 December 2010

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, 1999; Muir *et al.* 2000; Petit *et al.* 2002, 2004; Barreneche *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, simple sequence repeats (SSRs) have been the markers of choice to study hybridization (Burgarella *et al.* 2009; Viscosi *et al.* 2009; Ortego & Bonal 2010; Penaloza-Ramirez *et al.* 2010) and population genetic structure (Neophytou *et al.* 2010). At the same time, single-nucleotide polymorphisms (SNP) 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), because of 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.*

Correspondence: Rémy J. Petit, INRA-Univ. Bordeaux, UMR BIOGECO, 69 route d'Arcachon, 33612 Cestas, France, Fax: +33557122881; E-mail: petit@pierroton.inra.fr

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 five loci (Dzialuk *et al.* 2005; Lepais *et al.* 2006). Thus, analysing large oak populations at multiple markers remains expensive and time-consuming. In this 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, 1994, 1995; Streiff *et al.* 1998, 1999; Prida *et al.* 2006, 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 10 g of silica gel. 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) and sampled in southwest 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 discs (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 (1 h 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 3-mm tungsten bead was added and the plates were frozen in liquid nitrogen for 2 min before a 1-min disruption step at 30 Hz. DNA quality was estimated on a 1% (w/v) agarose gel stained with GelRed (Biotium, USA). DNA concentration was evaluated on an eight channel Nanodrop spectrophotometer, and concentration of each sample was adjusted to 10 ng/μL on a STARlet 8-channel robot (Hamilton, USA).

Multiplex PCR optimization

Kit-1. Sixty-four eSSRs (Durand *et al.* 2010, Table S2, Supporting Information) derived from expressed sequence tags (ESTs) were first tested on 24 samples from across the European range (12 *Q. petraea* and 12 *Q. robur* trees). They were analysed on a 4000L automatic DNA sequencer (LI-COR Biosciences, USA). Criteria for SSR selection were as follows: 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 because of stutter peaks with MICRO-CHECKER v.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. Final volume was optimized (10 μL) as well as final concentration of Mastermix (0.6 ×), reducing eight times the final cost.

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*

Locus	Primer sequences (5′–3′)	Reference	LG	Dye	[C]	Motif	Size (bp)	*A*	H_o	F_{IS}	F_{ST}
PIE020	GCAGAGGCTCTTCTAAATACAGAACT GGGAGGTTTCTGGGAGAGAT	Durand *et al*. 2010	1	FAM	1.00	AG	97–119	11	0.668	−0.002	0.018
PIE223	TAGAAGCCCAACACGGCTAC AGCAAAACACAAACGCACAA	Durand *et al*. 2010	2	FAM	1.00	GGT	197–221	9	0.749	−0.057	0.108
PIE152	TGTACCTCTTTCCTCTCTCTAAAACT GAATTTCTAAACCACTAGCATTGAC	Durand *et al*. 2010	2	FAM	3.75	TA	230–260	15	0.842	−0.024	0.032
PIE242	TGGAGGGAAAAGAACAATGC TTGCAATCCTCCAAATTTAATG	Durand *et al*. 2010	3	VIC	1.00	TA	102–128	12	0.803	0.045	0.038
PIE102	ACCTTCCATGCTCAAAGATG GCTGGTGATACAAGTGTTTGG	Durand *et al*. 2010	11	VIC	0.50	CT	131–161	9	0.722	−0.047	0.008
PIE243	GGGGTCAGTAGGCAAGTCTTC GAGCTGCATATTTTCCTTAGTCAG	Durand *et al*. 2010	10	VIC	0.25	AG	208–222	6	0.151	0.677	0.070
PIE239	TCAACAAATGGCTCAACAGTG CCCATTTGGTAGCAAAGAGTC	Durand *et al*. 2010	NA	PET	0.63	AT	70–83	11	0.590	−0.082	0.159
PIE227	TACCATGATCTGGGAAGCAAC AAGGGCTTGGTTGGGTTAGT	Durand *et al*. 2010	NA	PET	0.38	TGG	156–177	5	0.546	−0.064	0.207
PIE271	CACACTCACCAACCCTACCC GTGCGGTTGTAGACGGAGAT	Durand *et al*. 2010	2	PET	0.50	TC	180–197	10	0.759	0.019	0.021
PIE267	TCCAACCATCAAGGCCATTAC GTGCGAACAGATCCCTTGTC	Durand *et al*. 2010	3	NED	0.25	AG	80–105	10	0.824	−0.038	0.015
PIE258	TTCTCGATCTCAAAACAAAACCA TTTGATTTGTTTAAGGAAAATTGGA	Durand *et al*. 2010	2	NED	0.75	TC	128–159	19	0.880	0.005	0.039
PIE215	TACGAAATGGAGCTGTTGACC TCTCCTTCTCTTCTGCCATGA	Durand *et al*. 2010	12	NED	0.30	GAG	188–206	6	0.553	0.036	0.125
QrZAG7	CAACTTGGTGTTCGGATCAA GTGCATTTCTTTTATAGCATTCAC	Kampfer *et al*. 1998	2	FAM	0.50	TC	115–153	19	0.874	−0.015	0.025
MsQ13	ACACTCAGACCCACCATTTTTCC TGGCTGCACCTATGGCTCTTAG	Dow *et al*. 1995	6	FAM	0.50	GA	191–221	16	0.785	0.055	0.052
QrZAG112	TTCTTGCTTTGGTGCGCG GTGGTCAGAGACTCGGTAAGTATTC	Kampfer *et al*. 1998	12	VIC	0.40	GA	85–96	12	0.579	−0.005	0.128
QrZAG20	CCATTAAAAGAAGCAGTATTTTGT GCAACACTCAGCCTATATCTAGAA	Kampfer *et al*. 1998	1	VIC	0.15	TC	160–200	19	0.874	−0.015	0.025
QpZAG15	CGATTTGATAATGACACTATGG CATCGACTCATTGTTAAGCAC	Steinkellner *et al*. 1997	9	PET	0.50	AG	108–152	14	0.764	−0.026	0.024
†QpZAG110	GGAGGCTTCCTTCAACCTACTT GATCTCTTGTGTGCTGTATTTTT	Steinkellner *et al*. 1997	8	PET	0.50	AG	206–262	16	0.765	0.009	0.024
QrZAG96	CCCAGTCACATCCACTACTGTCC GGTTGGGAAAAGGAGATCAGA	Kampfer *et al*. 1998	10	NED	0.15	TC	135–194	18	0.628	0.015	0.149
†QrZAG11	CCTTGAACTCGAAGGTGTCC TGGTTGACTAAAGTATGAACTGTTTG	Kampfer *et al*. 1998	10	NED	0.40	TC	238–267	21	0.828	−0.031	0.075

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

PCR mix was composed of 3.5 μL of sterile water, 3 μL of Qiagen Multiplex Buffer (2×), 1 μL of primer premix and 2.5 μL of DNA (10 ng/μL). Concentrations for each primer pair in the primer premix are shown in Table 1. The cycling conditions were as follows: 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 v.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 *Q. 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 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 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 corresponding data was considered as missing.

Results

Multiplex PCR optimization

Among the 27 preselected 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 redesigned: 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 (Figs 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 artefacts. 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 analyse 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 because of 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).

Assignment power

Results of assignment tests on 20 000 simulated genotypes are shown in Fig. 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 (Fig. S3, Supporting Information).

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

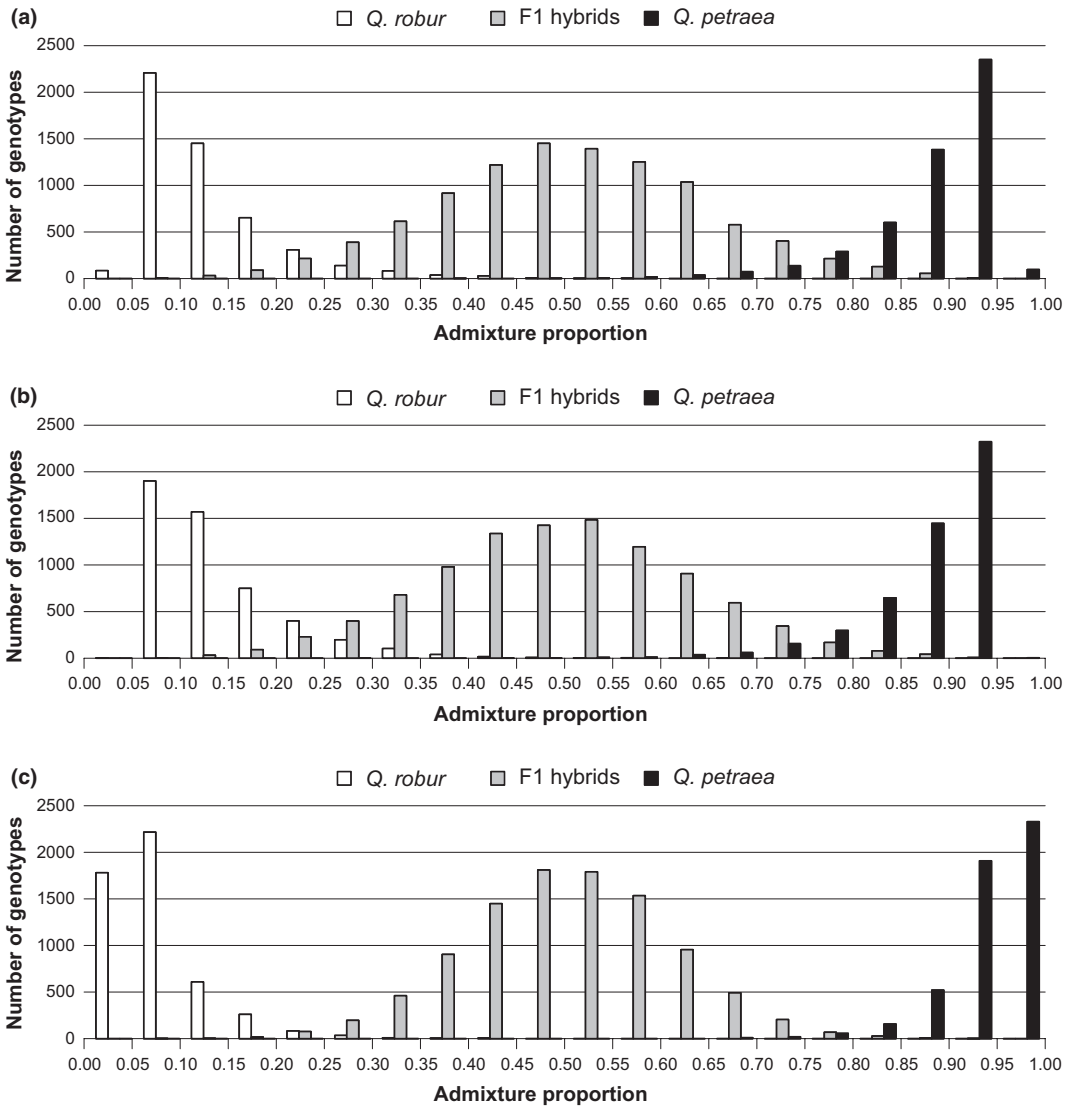


Fig. 1 Assignment of 20 000 simulated genotypes (purebred for both parental species and F1 hybrids). (a) *kit-1* (12-plex). (b) *kit-2* (8-plex). (c) *kit-1* + *kit-2* (12-plex + 8-plex).

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

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%

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: 0.03 bp). However, moderate deviation (>0.1 bp) was observed between sizes measured with each software in 7.8% of genotypes and large deviation (>0.25 bp) was observed in 2.9% of genotypes (maximum deviation: 0.48 bp). 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 data sets 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 because of 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 S1, 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 analyse 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 combination with additional markers, these two multiplexes should 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.

Authors' contributions

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

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).

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, 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.

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, 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.

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, Bodenes 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, Schlterer 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, 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, Bodenes 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.

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, 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.

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.

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.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Disagreement rate measured on 490 samples with *kit-1*. Error rate was measured on 80 samples (6% of the complete dataset).

Table S2 List of 64 EST-SSRs tested to develop *kit-1*. The 12 selected loci are in red.

Fig S1. Multiplex profile with *kit-1*.

Fig S2. Multiplex profile with *kit-2*.

Fig S3. Incorrect assignments for simulated *Q. robur*, *Q. petraea* and F1 hybrids with different intervals used for hybrid assignment.

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.

ANNEXE 2: CURRENT TRENDS IN MICROSATELLITE GENOTYPING

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03014.x

INVITED TECHNICAL REVIEW

Current trends in microsatellite genotyping

E. GUICHOUX,*†‡ L. LAGACHE,*† S. WAGNER,*†§ P. CHAUMEIL,*† P. LÉGER,*† O. LEPAIS,*†¶ C. LEPOITTEVIN,*† T. MALAUSA,** E. REVARDEL,*† F. SALIN*† and R.J. PETIT*†

*INRA, UMR 1202 Biodiversity Genes & Communities, F-33610 Cestas, France, †Univ. Bordeaux, UMR1202 Biodiversity Genes & Communities, Bordeaux, F-33400 Talence, France, ‡Pernod Ricard Research Center, F-94000 Creteil, France, §Univ. Bonn, Steinmann Institut, D-53115 Bonn, Germany, ¶School of Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, UK, **INRA, UMR 1301 IBSV INRA/UNSA/CNRS, F-06903 Sophia-Antipolis, France

Abstract

Microsatellites have been popular molecular markers ever since their advent in the late eighties. Despite growing competition from new genotyping and sequencing techniques, the use of these versatile and cost-effective markers continues to increase, boosted by successive technical advances. First, methods for multiplexing PCR have considerably improved over the last years, thereby decreasing genotyping costs and increasing throughput. Second, next-generation sequencing technologies allow the identification of large numbers of microsatellite loci at reduced cost in non-model species. As a consequence, more stringent selection of loci is possible, thereby further enhancing multiplex quality and efficiency. However, current practices are lagging behind. By surveying recently published population genetic studies relying on simple sequence repeats, we show that more than half of the studies lack appropriate quality controls and do not make use of multiplex PCR. To make the most of the latest technical developments, we outline the need for a well-established strategy including standardized high-throughput bench protocols and specific bioinformatic tools, from primer design to allele calling.

Keywords: binning, high throughput, nextgen sequencing, multiplexing, quality control, SSR

Received 5 October 2010; revision received 24 February 2011; accepted 7 March 2011

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 simple sequence repeats (SSRs) or short tandem repeats (STRs), remain the most popular markers in population genetic studies (Fig. 1). They consist of motifs of one to six nucleotides repeated several times that have a characteristic mutational behaviour (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; Mittal & Dubey 2009; Jones *et al.* 2010). 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; Subirana & Messeguer 2008; Sun *et al.* 2009). Their advantages over single-nucleotide polymorphisms (SNPs), which tend to be used increasingly, 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 because of difficulties for automation and data management, especially when compared to SNPs (Box 1). Hence, the continued use of microsatellites will probably 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,

Correspondence: Rémy J. Petit, Fax: 33 5 57 12 28 81; E-mail: petit@pierroton.inra.fr

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). 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 in a single lane, 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 are 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, 2004). Loci with non-overlapping allele size ranges are labelled with the same fluorescent dye, whereas those with overlapping allele size ranges are labelled with different dyes and resolved individually because of the different characteristic emission spectrum of each dye, hence considerably expanding 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, on both diploid and polyploid species (Jewell *et al.* 2010; Raabova *et al.* 2010), 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 study, 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 primer 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 data sets (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 Data 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 and 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 data sets for similar applications. Even 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

Box 1 SSRs vs. 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 (Fig. 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.

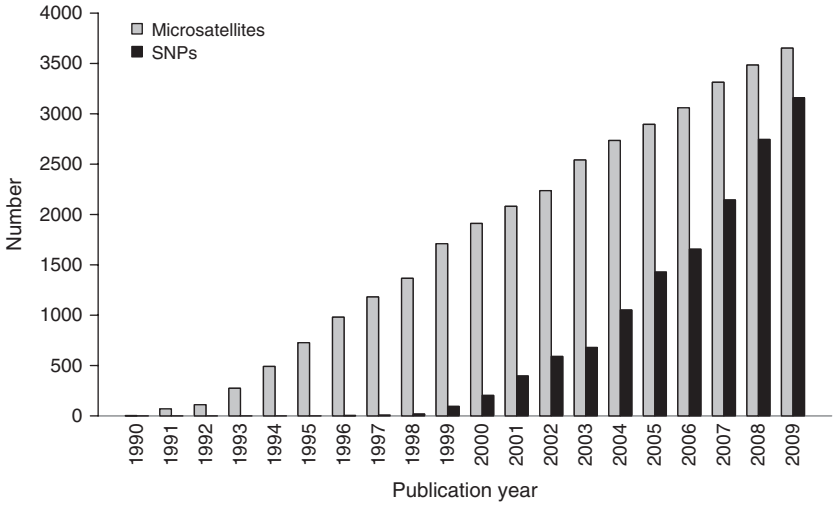


Fig. 1 Evolution of the number of studies relying on SSRs and SNPs since 1990.

Current popularity is not always the best guide to decide which markers to use (Schlötterer 2004). Instead, information 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, <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.

(a) 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 for SSRs in closely related species 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 analysing 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, which is the hallmark of population expansion, 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. Moreover, 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.

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

Application	Relative power of SSRs vs. SNPs	Comments	References
Linkage study, individual identification	2–3	Power proportional to heterozygosity H: $H_{SSR} \sim 2.H_{SNP}$	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 four 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)

(b) 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 when using SSRs (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).

• Variability at highly polymorphic microsatellite markers might not accurately reflect the underlying genomic diversity (Väli *et al.* 2008 but see Ljungqvist *et al.* 2010).

• 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; McCarroll *et al.* 2008).

• With SSRs, there is a need to include common controls among studies and across time. In contrast, 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, making it more difficult to study highly degraded DNA samples, such as faecal and other non-invasive samples, with SSRs than with SNPs (Seddon *et al.* 2005; Morin & McCarthy 2007; 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 probably be high, whereas compound genotyping errors will quickly increase as multiple PCRs are needed to type a single locus.

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 data sets produced.

SSR selection

Source of sequence data

Microsatellite detection requires sequence data. Until recently, the only possibility to identify sequences harbouring 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 harbouring 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 microsatellite development (Abdelkrim *et al.* 2009; Santana *et al.* 2009; Castoe *et al.* 2010; Csencsics *et al.* 2010; Malausa *et al.* 2011). Methodological improvements, such as biotin-based enrichment in SSR motifs, are now being proposed in combination with next-generation sequencing, which should further boost these approaches (Malausa *et al.* 2011). Besides, these approaches generate millions of base pairs of genomic sequence that may be useful for both SSRs-related and SSRs-unrelated research.

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

Organisms studied (%)		Size of repeat units (%)	
Mammals	18	Di-nucleotides	**46**
Other invertebrates	16	Tri-nucleotides	13
Plants	15	Tetra-nucleotides	14
Arthropods	14	Imperfect	26
Amphibian and reptiles	12		
Birds	11	Null alleles check (%)	
Fungi	8		
Fish	6	Yes	**40**
		No	60
Multiplexing (%)			
1–4 markers	15	Error-rate measurement (%)	
5–8 markers	19	Yes	**26**
>8 markers	8	No	74
No	**58**		

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, labelled 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 (Fig. 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.

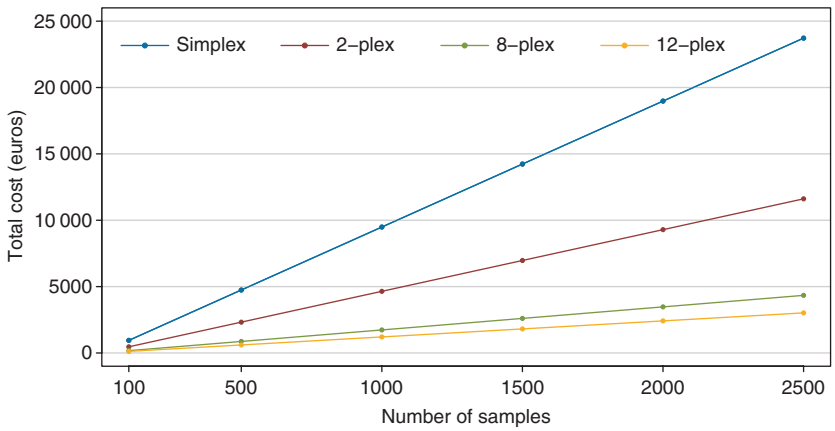


Fig. 2 Overall cost for genotyping 24 SSRs, depending on the multiplex strategy and the number of genotyped samples.

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 5 μL (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 2 μL (Kenta *et al.* 2008). Finally, instead of relying on direct fluorescent labelling of primers, it is possible to use universal tailed primers, one for each fluorescence detection (Missiaggia & Grattapaglia 2006). Such a method allows the same level of marker multiplexing and accuracy in SSR genotyping attained in regular direct-labelled microsatellite fluorescent detection assays, while significantly reducing the costs. This procedure is particularly adapted when many SSRs need to be investigated on relatively few samples.

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 harbouring large and complex genomes, such as conifers, direct approaches might be risky because of 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 the 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). To date, however, most studies (12 of the 15 articles relying on SSR detection with next-generation techniques that we identified, see Data S2, Supporting Information) have relied on whole genome shotgun sequencing, even when genome coverage was low (0.1× in Rasmussen & Noor 2009; 0.02× 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 bioinformatic steps. Software, such as MSAT-COMMANDER (Faircloth 2008) or QDD (Meglécz *et al.* 2010), has 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 two-thirds 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 because 454 technologies delivering >400 bp 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 probably 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 harbouring 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). Choosing perfect motifs should ensure that microsatellite loci follow as much as possible the stepwise mutation model used in coalescent-based methods to infer demographic events (Estoup *et al.* 2001). 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 dinucleotide repeats) rather than on longer ones (≥trinucleotide 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 analysed 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, mononucleotide 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 mononucleotide 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 mononucleotide repeats are of special interest. In contrast, dinucleotide repeat SSRs were most frequently used. Unfortunately, dinucleotide 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) (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 pentanucleotide 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 artefacts (e.g. Schwengel *et al.* 1994). Note also that a few solutions have been proposed to overcome stuttering problems (Box 3).

Number of repeat units. The number of repeats has a critical effect on mutation behaviour 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, SSRs 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, by preserving most of the advantages of SSRs (multiallelic, high diversity) while avoiding some of their drawbacks caused by very high mutation rate (Box 1). For instance, van Asch *et al.* (2010) suggest to select tetranucleotide 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 more than 16 repeats have a higher probability of accumulating interrupted motifs that confound the interpretation of the results.

Primer design

Once the sequences harbouring 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 data sets 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 (Appendix 1).

For multiplexing, primer pairs should have similar annealing temperature range [58–60 °C has been considered to be 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 of 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 (Fig. 5). In particular, SSR loci presenting excessive stuttering, split peaks, null alleles, low heterozygote peak height ratios and other artefacts 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 labelled 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-labelled M13-tail. This technique is economic because the cost of direct fluorescent primer labelling is

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 (Figs 3 and 4):

(1) Low heterozygote peak height ratios (Fig. 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.

(2) Stuttering or shadow bands (Fig. 3c). This corresponds to the amplification of PCR products that differ from the original template by one or a few repeats. This widespread phenomenon complicates the interpretation of electropherograms. Because of 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 labelling can result in slight stuttering because of 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.

(3) Split peaks (Fig. 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-labelled) 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 in the amount of template DNA, down to 10 ng (Lederer *et al.* 2000; Butler 2005b), the decrease in primer concentration, the increase in *Taq* concentration (Fishback *et al.* 1999) or the use of alternative polymerases (Hu 1993; Vallone *et al.* 2008).

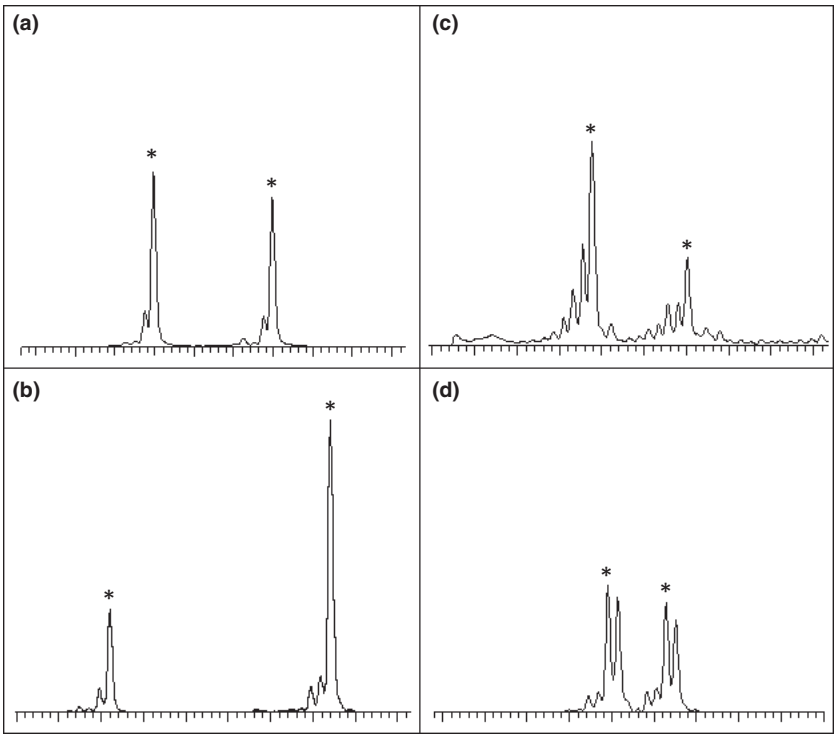


Fig. 3 Illustration of SSR profiles generated on capillary sequencer: correct profile (a), low heterozygote peak height ratios (b), excessive stuttering (c) and split peaks (d). Correct alleles are marked with asterisks.

(4) Null alleles (Figs 4a,b). 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 laboratory, 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). For large-scale population studies, markers should be validated on multiple populations to minimize null allele occurrence (Sinama *et al.* 2011). In the 100 studies that we surveyed, explicit tests of the presence of null alleles were reported in only 40% of the studies.

(5) Primer-dimers, artifactual bands (Fig. 4c) and triallelic patterns (Fig. 4d). These can be caused by the mispriming of primers (Brownie *et al.* 1997; Hill *et al.* 2009). Although the artefacts 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.

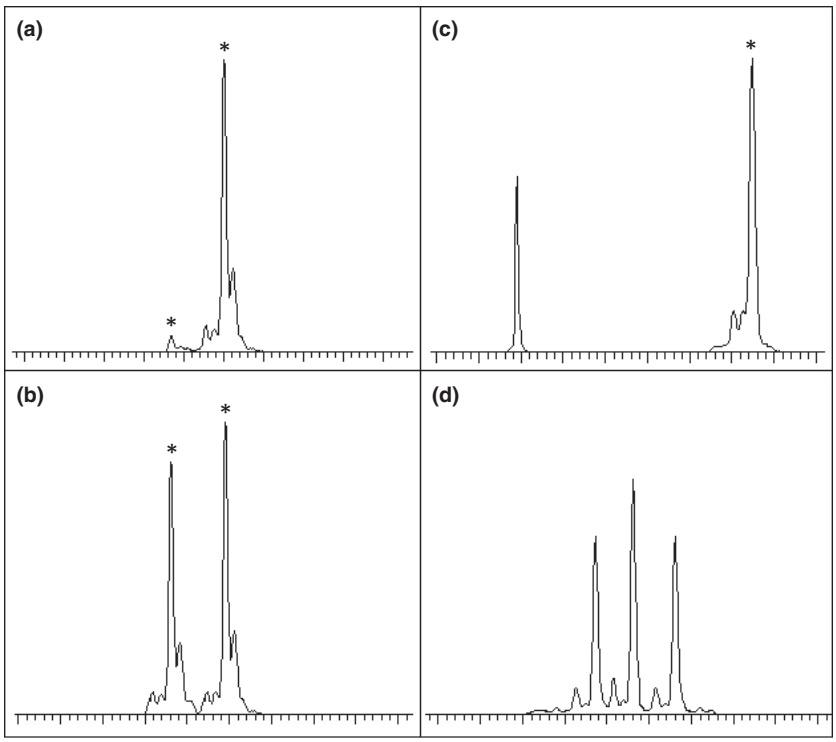


Fig. 4 Illustration of SSR profiles generated on capillary sequencer: weak allele before (a) and after (b) successful primer redesign, artifactual band (c) and triallelic pattern (d). Correct alleles are marked with asterisks.

typically five to ten times higher than the cost of the synthesis of an unlabelled 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 labelled 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 (Sinama *et al.*

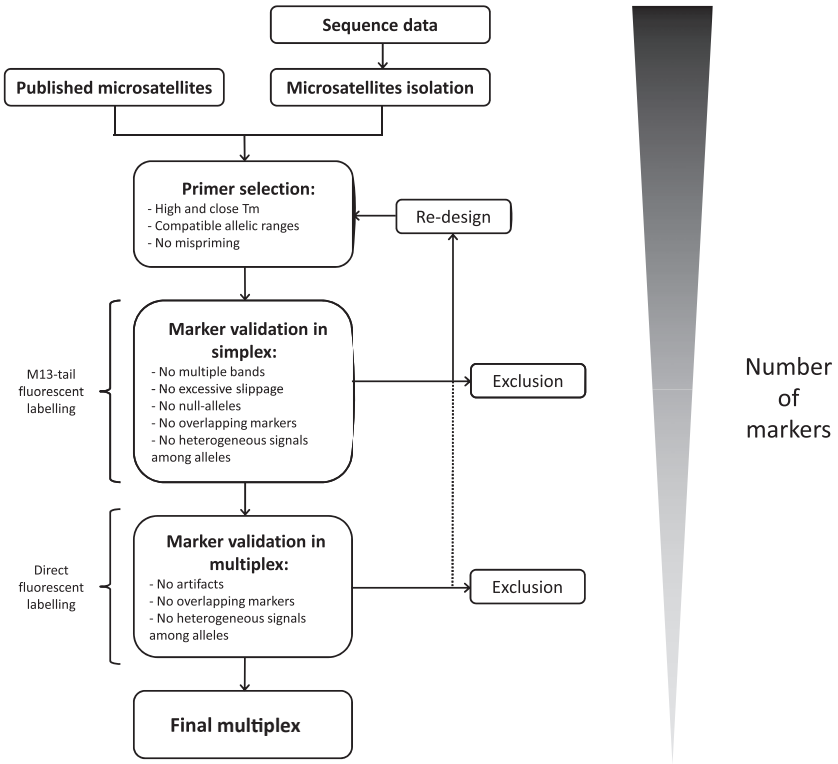


Fig. 5 One possible strategy for the development of multiplex SSRs suitable for high-throughput genotyping.

2011). This will minimize the risks to subsequently discover new alleles differing widely in size and overlapping with the allelic range of other loci labelled 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).

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 2). 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 owing to reaction failure or false positives owing 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 the differences in annealing temperature and maximizes the spacing between markers, the heterozygosity and the number of alleles (Fig. 6).

Multiplex PCR is a sensitive technique. To obtain repeatable results, careful standardization of all steps is needed. In particular, DNA concentration should be standardized (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 artefacts, 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

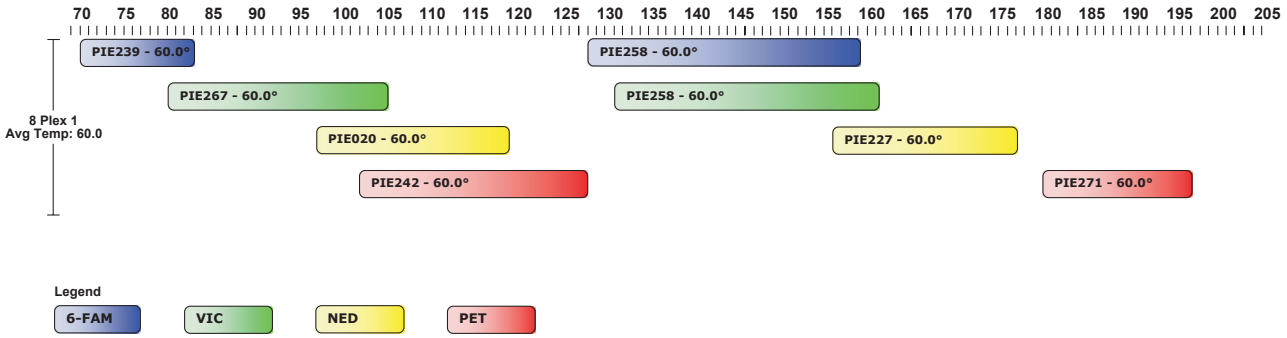


Fig. 6 Example of output obtained with Multiplex Manager software (Holleley & Geerts 2009). This software is used to identify combinations of markers suitable for multiplex reactions. In this example, for each of the eight SSRs, one of the four dyes (6-FAM, VIC, NED and PET) is assigned and the allele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

the Qiagen PCR Multiplex kit. This kit 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. Using different fluorochromes in multiplexes might also produce dye-induced mobility shift, which can lead to allele mis-scoring, with size differences between dyes (for the same allele) up to 3.7 bp (Sutton *et al.* 2011). Hence, strict quality control must be used to limit genotyping errors (see Measuring and reporting error rates section).

To increase the consistency of genetic profiling protocols, testing the quantity and quality of fluorescently labelled primers can be relevant. A simple method to assess primers quality 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 Fig. 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 dinucleotide 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 data sets 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 decreases 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 labour 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 (Morin *et al.* 2010). In one comparative study, 83% of discrepancies between laboratories in scoring dinucleotide alleles were caused by 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 (Appendix 1) and possibly on manual checks, rather than on direct analysis of raw fragment sizes, hence increasing risks of genotyping errors (Amos *et al.* 2007).

Because 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 laboratory, we developed an Excel macro, inspired from FlexiBin (Amos *et al.* 2007), Autobin (http://www4.bordeaux-aquitaine.inra.fr/biogeco/Ressources/Logiciels/Autobin), which automatically analyses raw data generated with commercial software (Appendix 1). The number of samples and loci is automatically detected, alleles in raw sizes are sorted and plotted to detect relevant gaps in size (Fig. 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).

Thousands of data sets 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 (Appendix 1). Binning can also be harmonized using reference genotypes and allelic ladders (Gill *et al.* 2001; LaHood *et al.* 2002; Rathmacher *et al.* 2009).

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

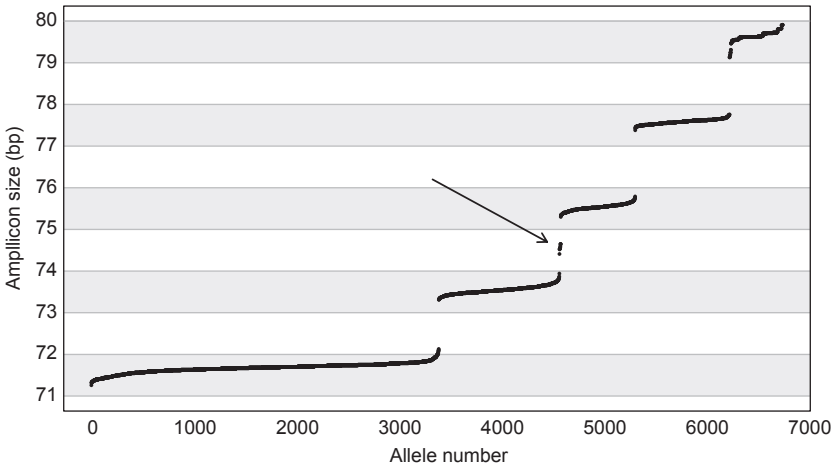


Fig. 7 Size distribution of 6762 alleles for one dinucleotide EST-SSR developed on oaks, achieved with the macro we developed. The arrow indicates the presence of an off-ladder microvariant found in 13 alleles that differs by one base pair from the expected periodicity of 2 bp. Analyses of segregating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

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 into two cases, depending on whether duplicated genotypes include or not a well-characterized control (i.e. concordance checking using standard reference genotypes vs. regenotyping 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 regenotyping 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) (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 regenotyping with reference samples. Nevertheless, from a practical point of view, regenotyping 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 laboratory or in another laboratory) 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.

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 are 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, because 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; Petersen *et al.* 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 performed 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 in 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 data sets 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 with bioinformatic methods aiming at sorting sequences, correcting for sequencing errors and finally summarizing genotype information.

Acknowledgements

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 Stéphanie 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).

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.

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.

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.

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–i56.

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: Biology, Technology, and Genetics of STR Markers*. Elsevier Academic Press, London.

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.

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.

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.

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*, **56**, 931–938.

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

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.

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.

Estoup A, Wilson IJ, Sullivan C, Cornuet J-M, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

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*, **11**, 578–585.

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 fluorophore 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.

Jewell MC, Frere CH, Prentis PJ, Lambrides CJ, Godwin ID (2010) Characterization and multiplexing of EST-SSR primers in *Cynodon* (Poaceae) species. *American Journal of Botany*, **97**, 99–101.

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 AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

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

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.

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.

LaHood ES, Moran P, Olsen J, Grant WS, Park LK (2002) Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Molecular Ecology Notes*, **2**, 187–190.

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*, DOI: 10.1111/j.1755-0998.2011.03002.x.

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 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.

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

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.

Ljungqvist M, Åkesson M, Hansson B (2010) Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on Väli *et al.* (2008). *Molecular Ecology*, **19**, 851–855.

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*, DOI: 10.1111/j.1755-0998.2011.02992.x.

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, Meglécz 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.

McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**, 1166–1174.

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.

Mittal N, Dubey A (2009) Microsatellite markers—A new practice of DNA based markers in molecular genetics. *Pharmacognosy Reviews*, **3**, 235–246.

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

Morin PA, Luikart G, Wayne RK, the SNP workshop group (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, Martien KK, Archer FI *et al.* (2010) Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. *Journal of Heredity*, **101**, 1–10.

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

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.

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.

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.

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.

Petersen J, Poulsen L, Birgens H, Dufva M (2009) Microfludic device for creating ionic strength gradients over DNA microarrays for efficient DNA melting studies and assay development. *PLoS ONE*, **4**, e4808.

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**, 847–859.

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.

Raabova J, Hans G, Risterucci A-M, Jacquemart A-L, Raspé O (2010) Development and multiplexing of microsatellite markers in the polyploid perennial herb, *Menyanthes trifoliata* (Menyanthaceae). *American Journal of Botany*, **97**, 31–33.

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.

Rathmacher G, Niggemann M, Wypukol H *et al.* (2009) Allelic ladders and reference genotypes for a rigorous standardization of poplar microsatellite data. *Trees-Structure and Function*, **23**, 573–583.

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.

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.

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.

Sinama M, Dubut V, Costedoat C *et al.* (2011) Challenges of microsatellite development in Lepidoptera: *Euphydryas aurinia* (Nymphalidae) as a case study. *European Journal of Entomology*, **108**, 261–266.

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.

Subirana JA, Messeguer X (2008) Structural families of genomic microsatellites. *Gene*, **408**, 124–132.

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.

Sun JX, Mullikin JC, Patterson N, Reich DE (2009) Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution*, **26**, 1017–1027.

Sutton JT, Robertson BC, Jamieson IG (2011) Dye shift: a neglected source of genotyping error in molecular ecology. *Molecular Ecology Resources*, **11**, 514–520.

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.

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. *BioTechniques*, **31**, 1320–1324.

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.

Väli Ü, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.

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.

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.

Supporting Information

Additional supporting information may be found in the online version of this article.

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

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

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.

Appendix 1 Non-exhaustive list of software for microsatellites detection and genotyping

Software name	Licence	Functionalities	Type of program	Platforms	Reference
SSR detection and primers design					
AutoDimer	Free	Screening for primer-dimer and hairpins	Visual Basic standalone version or Web application	Platform independent	Vallone & Butler (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)

Appendix 1 Continued

Software name	Licence	Functionalities	Type of program	Platforms	Reference
Pedant	Free	Estimation of maximum likelihood allelic dropout and false allele error rates	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 & Hughes (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)

ANNEXE 3: OUTLIER LOCI HIGHLIGHT THE DIRECTION OF INTROGRESSION IN OAKS

MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/mec.12125

Outlier loci highlight the direction of introgression in oaks

E. GUICHOUX,*†‡ P. GARNIER-GÉRÉ,*†[1] L. LAGACHE,*† T. LANG,*†§ C. BOURY*† and R. J. PETIT*†[1]

*INRA, UMR1202 BIOGECO, Cestas, F-33610, France, †Univ. Bordeaux, UMR1202 BIOGECO, Talence, F-33400, France, ‡Pernod Ricard Research Center, Créteil, F-94000, France, §Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan, 666303, China

Abstract

Loci considered to be under selection are generally avoided in attempts to infer past demographic processes as they do not fit neutral model assumptions. However, opportunities to better reconstruct some aspects of past demography might thus be missed. Here we examined genetic differentiation between two sympatric European oak species with contrasting ecological dynamics (*Quercus robur* and *Quercus petraea*) with both outlier (i.e. loci possibly affected by divergent selection between species or by hitchhiking effects with genomic regions under selection) and nonoutlier loci. We sampled 855 individuals in six mixed forests in France and genotyped them with a set of 262 SNPs enriched with markers showing high interspecific differentiation, resulting in accurate species delimitation. We identified between 13 and 74 interspecific outlier loci, depending on the coalescent simulation models and parameters used. Greater genetic diversity was predicted in *Q. petraea* (a late-successional species) than in *Q. robur* (an early successional species) as introgression should theoretically occur predominantly from the resident species to the invading species. Remarkably, this prediction was verified with outlier loci but not with nonoutlier loci. We suggest that the lower effective interspecific gene flow at loci showing high interspecific divergence has better preserved the signal of past asymmetric introgression towards *Q. petraea* caused by the species' contrasting dynamics. Using markers under selection to reconstruct past demographic processes could therefore have broader potential than generally recognized.

Keywords: asymmetric introgression, divergent selection, gene flow barriers, genetic assignment, outlier loci, *Quercus petraea*, *Quercus robur*, single nucleotide polymorphism

Received 15 March 2012; revision received 29 September 2012; accepted 4 October 2012

Introduction

In population genetics analyses, loci considered to be under selection are typically discarded in attempts to infer past demographic processes (Beaumont 2005; Helyar *et al.* 2011). The rationale for removing these loci from such analyses is that locus-specific effects caused by selection will bias population genetic inferences that traditionally assume selective neutrality (e.g. Wright 1931; Hudson 1990; Wakeley & Hey 1997; Nielsen & Wakeley 2001). Yet, markers known to be under selection have been used to estimate dispersal in two specific situations: genetic clines along environmental gradients and 'tension' hybrid zones (e.g. Barton & Hewitt 1981; Mallet *et al.* 1990; Szymura & Barton 1991; Lenormand *et al.* 1998). More studies are needed to evaluate the potential utility of genes under selection to reconstruct historical patterns of gene flow in other situations.

In an island model of migration at equilibrium, there is an inverse nonlinear relationship between genetic

Correspondence: Pauline Garnier-Géré, Fax: 33 5 57 12 28 81; E-mail: pauline@pierroton.inra.fr
Rémy J. Petit, Fax: 33 5 57 12 28 81; E-mail: petit@pierroton.inra.fr
[1]These authors contributed equally to this work.

structure and gene flow (Wright 1931). Thus, when gene flow is high, differences in allelic frequencies among populations become very low and genotyping or sampling errors become relatively more important (Waples 1998). In contrast, when gene flow is low, differences in allele frequencies among populations should be large, thereby facilitating the characterization of genetic structure. As selection can reduce effective gene flow and increase divergence (Bengtsson 1985), loci influenced by selection could provide more precise indications of genetic structure than others (Nosil *et al.* 2009). Such loci could be particularly helpful for assessing relative differences in levels of gene flow, especially in high gene flow species (see Appendix S1, Supporting information for a numerical example).

Targeting loci under divergent selection or tightly linked with them could be particularly relevant for reconstructing the main direction of gene flow. If gene flow is asymmetric between two populations, we expect that the population receiving more immigrants will be more variable and harbour more private alleles than the other population (e.g. Quintana-Murci *et al.* 2008; Marsden *et al.* 2011). However, if overall gene flow is high, differences in levels of diversity or in allele frequencies among populations might be slight and error-prone (Waples 1998; Neigel 2002). In contrast, the signature of asymmetric gene flow should be strong at loci under divergent selection.

A prerequisite for testing the potential of selected loci for such purposes is to accurately identify them. Loci showing high allelic-frequency divergence, which are possibly affected by selection in the corresponding genomic region, are typically detected with F_{ST}-based outlier methods (Beaumont & Nichols 1996; Beaumont 2005; Foll & Gaggiotti 2006; Excoffier *et al.* 2009). These methods can identify relatively highly differentiated markers (so-called outlier loci) in comparison to expected levels under neutrality inferred from coalescent simulations (Luikart *et al.* 2003; Li *et al.* 2012). They are increasingly used to study nonmodel species and speciation processes (Butlin 2008; Nosil *et al.* 2009; Garvin *et al.* 2010; Helyar *et al.* 2011).

In this study, we decided to focus on gene flow between closely related plant species rather than between conspecific populations, as divergent selection should be high in this case (Nosil *et al.* 2009). Moreover, interspecific gene flow is often asymmetric in plants (Arnold 1997; Abbott *et al.* 2003). This asymmetry can be caused by differences in fertilization success and offspring survival (Tiffin *et al.* 2001; Lowry *et al.* 2008), differences in abundance at the time of mating (Lepais *et al.* 2009) or differences in population dynamics (Currat *et al.* 2008). We selected a pair of partly interfertile white oak species, pedunculate oak (*Quercus robur*) and sessile oak (*Quercus petraea*), which are widely distributed over Europe and have overlapping distribution ranges (the range of *Q. petraea* being largely included within that of *Q. robur*). These two species are patchily distributed as a function of the environment, resulting in numerous contact zones where hybridization can take place, forming so-called mosaic hybrid zones (Streiff *et al.* 1999; Jensen *et al.* 2009). Despite evidence of hybridization and introgression, *Q. robur* and *Q. petraea* remain ecologically and morphologically differentiated (Kremer *et al.* 2002) and have strong postpollination prezygotic sexual barriers, as revealed by a recent large-scale interspecific crossing study (Abadie *et al.* 2012).

Another important prerequisite for our study was to accurately delimit these two closely related interfertile oak species, which has been a long-lasting goal for botanists and geneticists (Cousens 1963; Carlisle & Brown 1965; Bodénès *et al.* 1997; Muir *et al.* 2000; Coart *et al.* 2002; Kremer *et al.* 2002; Scotti-Saintagne *et al.* 2004; Kelleher *et al.* 2005). Encouraging results have been obtained recently by selecting some of the most discriminating microsatellites identified to date (Guichoux *et al.* 2011). However, greater discriminatory power might be obtained by focusing on F_{ST}-based outlier loci showing high interspecific divergence. The objective is then to use these loci to test hypotheses regarding past demographic events that emerge from considerations of oaks' life histories.

Quercus petraea, a shade-tolerant species, must typically follow the more pioneering oak species, *Q. robur*, during forest successions, as it probably did during the postglacial recolonization of Europe (Petit *et al.* 2003). Thus, there is a phase where immigrant *Q. petraea* trees have to establish in areas already dominated by *Q. robur*. Under such conditions, introgression is expected to be strongly asymmetric towards the late invader, according to neutral models of colonization dynamics (Currat *et al.* 2008). On one hand, alleles from the resident species (*Q. robur*) that leak into the genome of the colonizing species (*Q. petraea*) can rapidly increase in frequency at the time of expansion, resulting in high introgression in the expanding species (*Q. petraea*). On the other hand, less introgression is expected towards *Q. robur*, the resident species, which is already at carrying capacity. Asymmetric introgression would also be consistent with the finding that *Q. robur* female flowers are more easily fertilized by *Q. petraea* pollen than the converse in artificial crosses (Steinhoff 1993). Consequently, late-successional *Q. petraea* should have greater genetic diversity than the early successional *Q. robur*, at least if there are similar initial levels of diversity in the two species. However, if interspecific genetic exchanges are not exceedingly rare,

as might be inferred from previous studies of the species (Streiff *et al.* 1999; Jensen *et al.* 2009; Lepais *et al.* 2009; Lagache *et al.* 2012), the asymmetry signal might be weak or absent. Under such conditions, highly divergent genes that have experienced reduced effective interspecific gene flow might be particularly useful for detecting the signal of ancient asymmetric introgression.

The objective of this work was to use outlier loci to test if the direction of introgression matches predictions from the demo-genetic models described above, thus demonstrating their utility to study demographic processes. The two prerequisites of this study were to accurately identify *Q. robur* and *Q. petraea* purebreds (and remove admixed individuals) and to identify outlier loci. For these purposes, we applied a model-based outlier detection method to a set of single nucleotide polymorphisms (SNPs) enriched with markers showing high differentiation between species in a discovery panel. We compared the ability of outlier SNPs and non-outlier SNPs to delimitate species using existing methods. We then tested for differences in the genetic diversity and structure of the two species using both types of markers, to check if they are consistent with a signature of ancient asymmetric introgression.

Materials and methods

Material

We sampled 855 oak trees in six mixed stands of *Quercus robur* and *Quercus petraea* in northern France (Petite Charnie, Vitrimont, Charmes, Lure, Cuve, Mondon, see Appendix S2, Supporting information for the populations' geographic locations and sample sizes, and Appendix S3, Supporting information for the species' distributions in Europe). One stand (Petite Charnie) includes 278 adult trees and 380 offspring (in 51 half-sib families, see Guichoux *et al.* 2011). Leaves or buds were sampled and stored immediately at −20 °C or in silica gel.

DNA isolation

DNA was isolated from leaves or buds using an Invisorb DNA plant HTS 96 kit (Invitek, Berlin, Germany), following the manufacturer's instructions, except for the lysis step (1 h at 65 °C). DNA quality was estimated by separating the samples in 1% (w/v) agarose gel then staining with bromophenol blue. The DNA concentration in the samples was evaluated using an Infinite 200 microplate reader (Tecan, Männedorf, Switzerland) in conjunction with a Quant-it dsDNA Broad-Range Assay kit (Invitrogen, Carlsbad, CA, USA). The concentration of each sample was then adjusted to 50 ng/μL by a STARlet 8-channel robot (Hamilton, Reno, NV, USA).

SNP choice

Most of the SNPs were chosen from a larger set of 9080 SNPs that had been previously validated by allelic resequencing of 584 gene fragments within the framework of the EVOLTREE network of excellence activities (http://www.evoltree.eu/; SNPs available via the *Quercus* Portal at https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=snp). We selected a subset of 346 validated polymorphic SNPs (Phred Score > 30) from the resequencing study, by applying both technical and biological criteria, using an automatic pipeline developed for these data. In particular, we enriched the list with SNPs expected to better differentiate the species (either from high interspecific differentiation estimates in a small panel of individuals, or from their location in genes putatively involved in drought stress tolerance, a trait that differentiates the two species; see Appendices S4 and S5 for details on functional categories). The aim was also to maximize the number of genes by targeting few SNPs per gene. In addition, 32 SNPs from functional and expressional candidate genes not included in the previous resequencing study were identified by *in silico* analysis (Appendices S4 and S5). In the final list of 384 SNPs (Appendix S5, Supporting information), all markers except those derived *in silico* met stringent technical criteria (successful amplification for at least 2/3 of the sampled individuals in each species, Illumina scores above 0.6, and at least 60-bp spacing between SNPs within genes). These SNPs represent 227 different genes (15 genes for the 32 *in silico* SNPs) with on average 1.7 SNPs per gene.

SNP genotyping

The required SNP format for online submission to the Illumina Assay Design Tool (ADT; Illumina Inc., San Diego, CA, USA) was prepared with a Perl script adapted from Lepoittevin *et al.* (2010), which predicts design feasibility. SNP genotyping was performed with the 384-plex GoldenGate assay (Illumina Inc.) based on the VeraCode technology. We followed the manufacturer's instructions, using 250 ng of DNA as starting quantity for each sample. Three negative controls were added to each batch of the five 96-well plates. The acquired data were analysed (i.e. SNPs were clustered for genotypic class calls) using BeadStudio (Illumina Inc.) according to recommended procedures (Close *et al.* 2009; Lepoittevin *et al.* 2010), except that we initially retained SNPs lacking one homozygote cluster and those showing cluster compression, that is, members of

genotypic classes that were closer to each other than expected on a normalized 0–1 scale in cluster plots. In the Petite Charnie stand, SNP data were validated using parent/progeny relationships determined using microsatellite (SSR) data (Guichoux *et al.* 2011). This allowed *a posteriori* validation of all SNPs, even in cases of cluster compression. Monomorphic loci and loci in total linkage disequilibrium with another locus were discarded from subsequent analyses.

Assignment methods for accurate species delimitation

We used the Bayesian clustering algorithm implemented in STRUCTURE 2.3.3 (Pritchard *et al.* 2000) to classify individual SNP genotypes and compared the results with those for SSR genotypes previously reported (Guichoux *et al.* 2011). After a burn-in of 50 000 steps followed by 50 000 Markov chain Monte Carlo repetitions, we calculated average assignment scores over 10 runs with K (number of groups) set to two, corresponding to the two species. A key step in any such analysis is to choose appropriate threshold values for the assignment scores to identify purebred individuals efficiently (Vähä & Primmer 2006). Purebreds have expected admixture levels of 0 and 1, F1 hybrids of 0.5, and backcrosses of 0.25 and 0.75. Thus, threshold values of 0.125 and 0.875 are optimal for distinguishing between purebreds and first-generation backcrosses, which was deemed sufficient for this study, even though later-generation backcrosses certainly occur in this system. To confirm the relevance of these thresholds under the simplifying assumption that the examined populations consist solely of purebreds, F1s and first-generation backcrosses, we simulated with HYBRIDLAB 1.0 (Nielsen *et al.* 2006) 1000 genotypes for each of the following categories: purebreds (2), F1s and first-generation backcrosses (2). Allelic frequencies of purebreds were used as reference and observed assignment scores were compared to theoretical expectations.

We also tested the repeatability of the assignment scores by performing further clustering analyses using only half of the validated SNPs, grouped into two independent subsets (designated A and B) randomly drawn from the complete set. Finally, we tested the ability of varying numbers of SNPs to assign purebred individuals. For this purpose, we created SNP subsets (2, 4, 8, 16, 32, 64, 128, 256 and all SNPs), with each subset comprising the SNPs with the highest possible interspecific F_{ST}. We then compared the STRUCTURE clustering results for each of these subsets on the basis of a performance index, defined as the efficiency multiplied by the accuracy, as in Vähä & Primmer (2006). Efficiency is 'the proportion of individuals in a category that are correctly identified' (e.g. *Q. robur* identification efficiency = the number of individuals in the *Q. robur* group that are correctly assigned divided by the total number of *Q. robur* individuals, including those falsely assigned to other groups). Accuracy is 'the proportion of an identified group that truly belongs to that category' (e.g. *Q. robur* identification accuracy = the number of individuals in the *Q. robur* group that are correctly assigned divided by the total number of individuals in the *Q. robur* group, including those falsely assigned to the group).

Diversity analyses and outlier detection method

For each species, allelic frequencies, genotypic frequencies, expected heterozygosity (H_e; Nei 1973) and inbreeding coefficients (F_{IS}; Weir & Cockerham 1984) were estimated for each SNP and their average across loci was computed. Only individuals with multilocus genotypes having <10% of missing data were included. Intra- and interspecific F_{ST} estimates (Wright 1951) were computed using ARLEQUIN 3.5.1.2 (Excoffier *et al.* 2009).

The main objective was to contrast diversity patterns between outlier loci and nonoutlier loci. We searched for outlier loci, that is, loci showing higher levels of interspecific genetic differentiation than expected under neutrality, by using the coalescent simulation module implemented in ARLEQUIN, which extends the Beaumont & Nichols method (1996) to a finite number of demes in the symmetrical island migration model and to a variable mutation rate across loci (Excoffier *et al.* 2009).

A main issue was to choose a mean targeted F_{ST} value for the simulations (hereafter called reference F_{ST} value). For that, the ideal would be to have randomly chosen SNPs available across the genome, preferably far away from the influence of coding regions, so that they could be considered to be mostly affected by demographic effects. Unfortunately, such markers are usually not available in nonmodel species. Therefore, the mean F_{ST} is often used as initial reference value, assuming no selection effects overall when using a large number of random markers. In our case, the markers included a large proportion of highly differentiated SNPs and SNPs from candidate genes of ecologically divergent traits among species. Given this choice, using the mean F_{ST} value as reference would assuredly overestimate the 5% quantile of the simulated distribution (Helyar *et al.* 2011). The number of outliers detected with such a reference would therefore be underestimated, which would be very conservative.

To account for the uncertainty in the reference F_{ST} value in oaks, we followed two different approaches for outlier detection: one using a reference F_{ST} value of 0.04, which is based on a multilocus scan from different markers in the same species (Scotti-Saintagne *et al.*

2004), and the very conservative approach described previously, which uses the observed mean F_{ST} value (0.22) as reference. We also derived the neutral envelope differently to the default ARLEQUIN option to better account for our particular case study: first by choosing a trial subset of SNPs with a mean F_{ST} value equal to the reference value; second by adjusting this reference value so that the bias in the mean simulated F_{ST} value for two demes only is accounted for (see Slatkin 1991); finally by retaining only genealogies with one mutation to model SNPs (See Appendix S4, Supporting information for more details on how we ran the outlier tests). We further explored the robustness of outlier detection in our data in more complex situations (Excoffier *et al.* 2009), by testing a hierarchical model with the two species demes each composed of six populations. We also tested for the presence of outliers within each species using as reference intraspecific F_{ST} values the mean observed values (0.012 for *Q. robur* and 0.013 for *Q. petraea*, based on data for all 262 SNPs, see Table 1). In all cases, the null F_{ST} distribution was built as a function of H_{WP}, the mean within-deme heterozygosity value, and observed values were tested as outliers in comparison with the 95th percentile of the simulated distributions.

Graphical comparison of genotype likelihoods

The data set was analysed with the genotype-likelihood approach of Paetkau *et al.* (1995) and Waser & Strobeck (1998), which allows direct, convenient visualization of genetic differences between individuals of two groups. We plotted two likelihoods for each genotype corresponding to their probabilities of generation based on *Q. robur* and *Q. petraea* allelic frequencies, respectively, in the form of biplots. To compute these likelihoods, allele frequencies at each locus in each 'pure' species are first computed. Then, the genotypic likelihood at each locus 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 are multiplied across loci assuming that they are independent (Paetkau *et al.* 2004) to yield an overall likelihood. As genotype likelihoods are products across loci, their values are geometrically affected by the number of SNPs included in the computation, so only individuals with nearly complete multilocus genotypes were considered. We compared results obtained using four sets of loci (12 SSRs, all SNPs, nonoutlier SNPs and outlier SNPs) with species and admixed categories previously defined on the basis of all validated SNPs. Genotype likelihoods were computed with GENALEX 6.4 (Peakall & Smouse 2006). For each category, we also plotted the coordinates of the mean likelihood value of all individuals belonging to that category.

Table 1 Genetic parameters for the two oak species (*Quercus robur*, *Quercus petraea*) based on all SNPs

Group	*N*	H_e	F_{IS}	Intraspecific F_{ST}
Q. robur	436	0.221	−0.004	0.012
Q. petraea	329	0.217	0.001	0.013

N, sample size; H_e, mean expected heterozygosity across individuals; F_{IS}, mean value across individuals (within populations and then across them) of fixation indices.

Results

SNP genotyping

A total of 855 individuals of the two species (*Quercus petraea* and *Quercus robur*) were genotyped at 384 SNPs. After all validation steps, 262 out of 384 SNPs were retained for further analyses (68%). We excluded in particular six SNPs that were in complete linkage disequilibrium with another locus. The parent-pair analyses further led to the exclusion of 24 SNPs that did not segregate according to Mendelian expectations, including 11 SNPs that had compressed clusters (16% of this category) and 13 SNPs that had uncompressed clusters (6% of this category; Appendix S7, Supporting information). The retained SNPs all had inconsistency rates lower than 5% in parent-pair analyses. Overall, this validation procedure increased the final success rate by 13% compared with recommended procedures (Close *et al.* 2009; Lepoittevin *et al.* 2010) and decreased the error rate of the selected SNPs.

Species assignment

With the chosen clustering thresholds (0.125 and 0.875, see Materials and methods), we assigned each individual to one of the following categories: (i) 'purebreds' (which should include mostly 'pure' *Q. robur* or *Q. petraea* individuals and, if present, some second and later-generation backcrosses) and (ii) 'admixed individuals'. Assignment results based on the 262 retained loci revealed a low proportion of admixed trees (9%), about half the estimate based on 12 SSRs (17%, see Appendix S8, Supporting information). The stability of assignment values was very high for purebreds when comparing the two subsets of 131 SNPs (97% correspondence, see Fig. 1). Assignment scores for purebreds obtained from the SNP analysis were also very similar to those obtained using SSRs (95% correspondence), despite the lower number of SSR loci (12). In contrast,

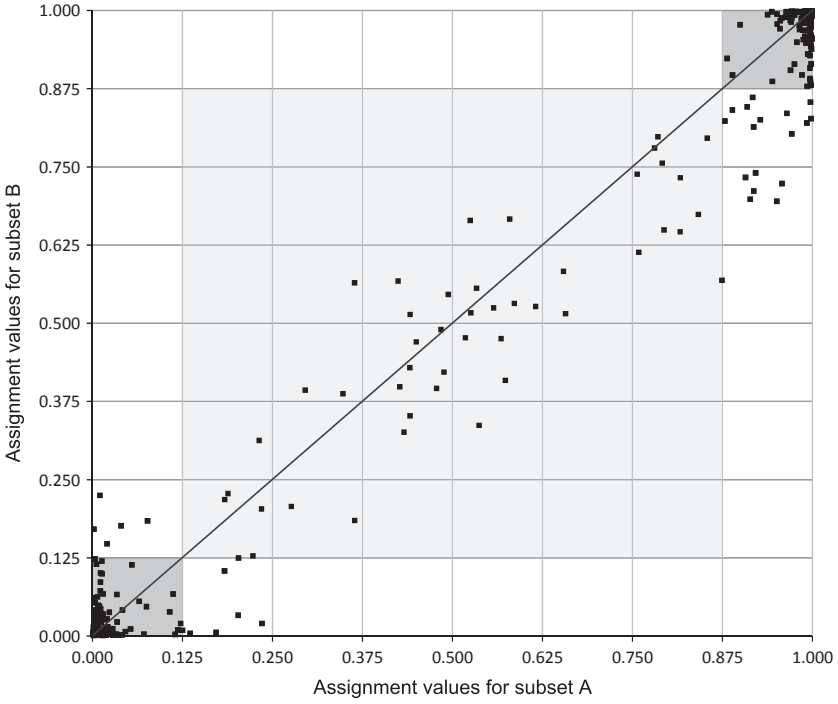


Fig. 1 Correspondence between assignment scores for each individual genotyped with two randomly drawn subsets (A and B) of 131 SNPs from all 262 SNPs. Points close to the diagonal represent individuals with repeatable assignments based on the two subsets. The thresholds for the categories are provided in the text. On the basis of these thresholds, purebred individuals with repeatable assignment scores are found inside the two small grey squares.

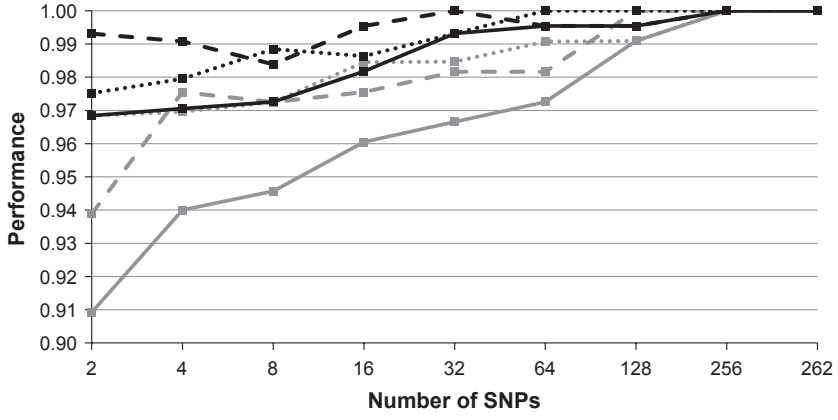


Fig. 2 Efficiency, accuracy and overall performance of assignments for *Quercus robur* (black line) and *Quercus petraea* (grey line) individuals. Efficiency (dashed line), accuracy (dotted lines) and performance (full lines) as functions of the number of SNPs ordered by decreasing interspecific F_{ST} values.

assignments for the admixed category were less stable across the two subsets of 131 SNPs (66% correspondence, Fig. 1), indicating that assignment is less precise in this group. When using few SNPs showing the highest interspecific F_{ST}, assignment performance (Vähä & Primmer 2006) remained high for both species (Fig. 2). Interestingly, the performance for assigning *Q. robur* individuals was always higher than for assigning *Q. petraea* individuals, regardless of the number of SNPs used, due to a better efficiency and accuracy (Fig. 2). Therefore, *Q. robur* individuals require genotyping at fewer SNPs than *Q. petraea* for equally robust assignment.

We also compared assignment values of simulated genotypes with expectations. The results show that all genotypic classes were clearly separated with few incorrect assignments (see Appendix S9, Supporting information).

Genetic structure and outlier detection

The mean expected heterozygosity (H_e) across loci was similar for the two species (0.221 for *Q. robur* and 0.217 for *Q. petraea*, see Table 1). Mean F_{IS} values across loci were very close to zero and did not differ significantly between the species ($P = 0.7$, see Table 1). Within each species, a large number of loci (>90%) were at Hardy–Weinberg equilibrium and all loci included in the analyses were at linkage equilibrium, as required by the initial assumptions of both the STRUCTURE clustering and

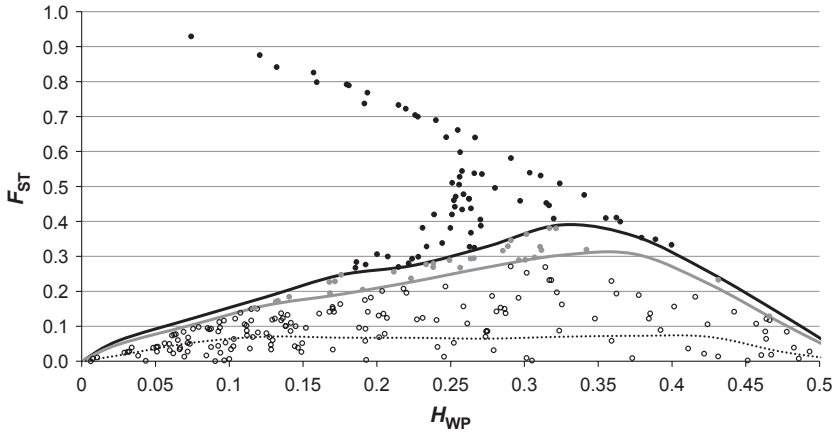


Fig. 3 Distribution of interspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), calculated using a hierarchical island model (two demes and six populations within each deme, reference F_{ST} = 0.04). Outlier and nonoutlier loci are represented by filled and open circles, respectively. Loci represented by grey circles were excluded from the nonoutlier category. Black line, 95th quantile distribution; grey line, 90th quantile distribution; dotted line, 50th quantile distribution.

the genotype-likelihood descriptive methods. As expected given our choice of SNPs, the mean interspecific F_{ST} across loci was much higher (0.22) than previously published estimates (Scotti-Saintagne *et al.* 2004), with some SNPs showing very high values (up to 0.93, see Fig. 3).

The distribution of observed F_{ST} values as a function of mean within-species diversity has a remarkable croissant shape (Fig. 3). We interpret this as a mathematical artefact caused by the fact that F_{ST} at each diallelic locus is constrained to vary within some limits that depend on the minor allele frequency (and thus on diversity) and on the number of populations (e.g. Petit *et al.* 1995; Hedrick 2005). With a reference F_{ST} value of 0.04, the proportions of outlier loci detected (i.e. located above the 95th percentile of the simulated distribution) when using the two demes only or the hierarchical island model were similar (74 and 68 outliers out of 262, respectively). We focus on the latter model in the following as it was slightly more conservative (Fig. 3 and Appendix S5, Supporting information). Moreover, 28 loci located between the 90th and 95th percentiles were excluded from the nonoutlier category, as suggested by Nosil *et al.* (2009). At the intraspecific level, the proportion of outliers detected were 5% (13) outliers in *Q. petraea* and 4% (10) in *Q. robur*, see Appendices S11 and S12. These values are very close to the type I error rate (5%). Eight of these intraspecific outliers were also interspecific outliers and were excluded from subsequent analyses to facilitate interpretations. Thus, a total of 60 interspecific outliers (24%) and 166 nonoutliers (65%) were finally considered. Mean estimates of interspecific F_{ST} were 0.093 for nonoutlier loci, 0.504 for outlier loci and 0.210 across all loci. The levels of genetic differentiation computed among populations within each species did not differ significantly between interspecific outliers and nonoutliers (see intraspecific F_{ST} values in Table 2). As expected, using an initial reference F_{ST} value of 0.22 instead of 0.04 resulted in a much lower proportion of

Table 2 Comparison of genetic diversity and inter- and intraspecific differentiation at outlier and nonoutlier loci, with two different reference F_{ST} values (0.04 and 0.22)

				H_e					Intraspecific F_{ST}		
Reference F_{ST}	Type of loci	N	F_{ST}	Total	Mean within species	*Quercus robur*	*Quercus petraea*	P†	*Quercus robur*	*Quercus petraea*	P
0.04	Outliers	60‡	0.504	0.511	0.255	0.163	0.347	***	0.010	0.010	ns
	Nonoutliers	166	0.093	0.401	0.201	0.228	0.173	**	0.013	0.015	ns
	P		***	***		*	***		ns	ns	
0.22	Outliers	13	0.756	0.393	0.197	0.094	0.299	***	0.003	0.012	ns
	Nonoutliers	247	0.191	0.440	0.220	0.228	0.213	ns	0.013	0.013	ns
	P		***	ns		***	***		***	ns	

N, sample size; H_e, mean expected heterozygosity across loci.

†The significance of differences, obtained from Student t-tests, in values between the species (ns, not significant; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

‡For calculating intraspecific F_{ST} values the eight intraspecific outliers were considered, to enable comparison with nonoutliers.

outlier loci (13 interspecific outliers only, out of 262 SNPs).

Genotype likelihoods and diversity patterns at outlier and nonoutlier loci

Log-likelihoods of genotypes were plotted to visualize their similarity to either *Q. robur* (*x*-axis) or *Q. petraea* (*y*-axis; Fig. 4A). Using observed genotypes at the 12 SSRs, several admixed trees could not be distinguished from purebreds, and even some purebreds (indicated by red or blue circles) were not clearly separated on their respective sides of the diagonal. In contrast, the corresponding biplots showed that admixed trees were correctly differentiated from purebreds when the complete SNP data set was used (Fig. 4B).

The total expected heterozygosity (H_t) was significantly higher at outliers than at nonoutliers (0.511 vs. 0.401, $P < 0.001$, see Table 2). At outlier loci (60 SNPs), the mean expected diversity H_e was higher for *Q. petraea* than for *Q. robur* (0.347 vs. 0.163, $P < 0.001$). In contrast, at nonoutlier loci (166 SNPs), H_e was lower for *Q. petraea* than for *Q. robur* (0.173 vs. 0.228, $P < 0.01$). Finally, H_e values calculated in *Q. petraea* and *Q. robur* using data for all SNPs were not significantly different (0.217 and 0.221, $P = 0.85$). Similar results were observed in the conservative approach with a reference F_{ST} value of 0.22: at outlier loci (13 SNPs), H_e was higher for *Q. petraea* than for *Q. robur* (0.299 vs. 0.094, $P < 0.001$). In contrast, at nonoutlier loci (247 SNPs), H_e was slightly lower for *Q. petraea* than for *Q. robur* (0.213 vs. 0.228, $P = 0.35$, Table 2). The same patterns (higher genetic diversity in *Q. petraea* than in *Q. robur* at

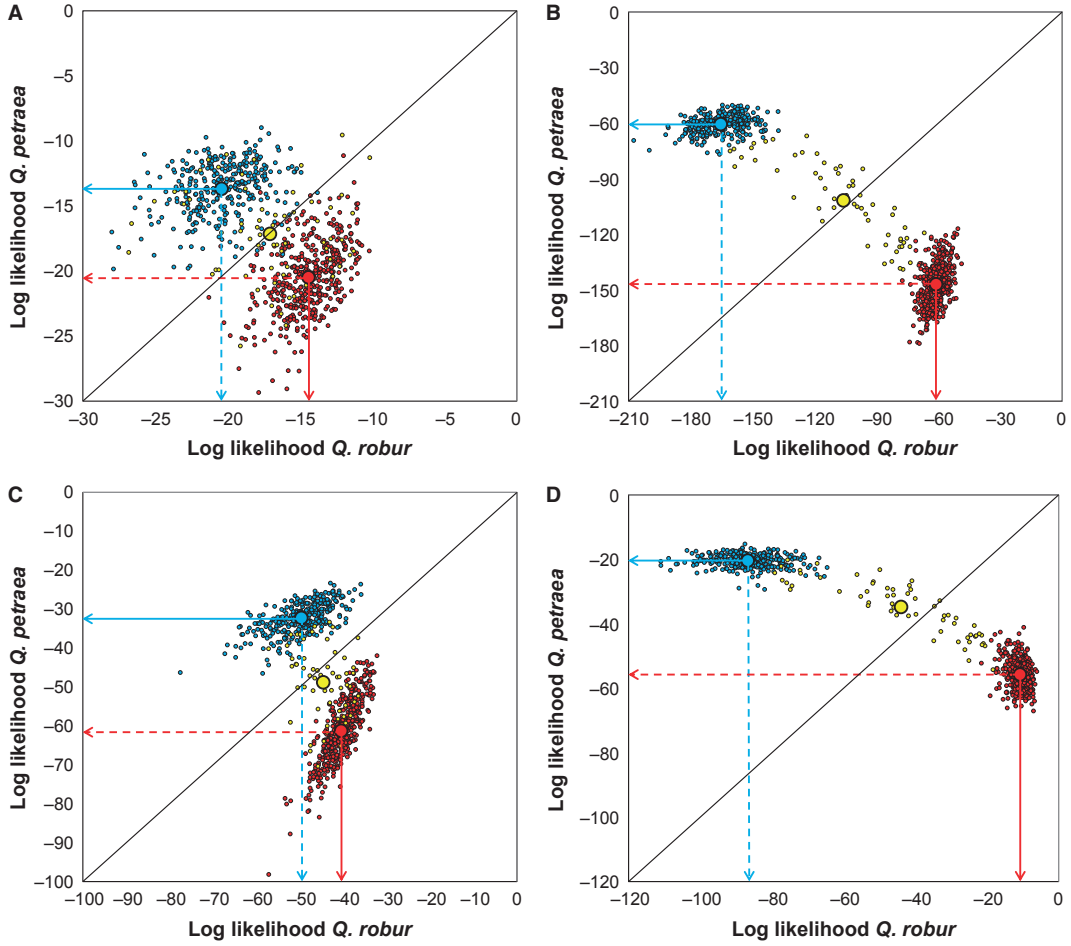


Fig. 4 Biplots of log-likelihoods of assignment to *Quercus petraea* and *Quercus robur* across all individuals. Three groups are distinguished—*Q. petraea* (blue), *Q. robur* (red) and admixed individuals (yellow)—with different categories of markers (A, 12 SSRs; B, all SNPs; C, 166 nonoutlier loci; D, 60 outlier loci). Mean values for each group are indicated by larger circles. Full arrows indicate mean log-likelihoods of conspecific identity and dotted arrows mean log-likelihoods of allospecific identity. The diagonal line helps identify asymmetries in assignment probabilities between species.

outliers but not at nonoutliers) were observed when using two different functional subsets of loci (loci involved in drought stress vs. loci involved in other functions, see Appendix S13, Supporting information).

Similar insights are obtained when considering mean log-likelihood of conspecific identity (full lines in Fig. 4C, D). Even stronger differences between species were detected when comparing mean log-likelihood values of allospecific identity at outlier vs. nonoutlier loci (dotted lines in Fig. 4C, D). For outlier loci, the values differ between the two species by 30 orders of magnitude, implying that *Q. petraea* genotypes are less well affected to *Q. robur* (−86) than *vice versa* (−56). In contrast, for nonoutlier loci, the corresponding difference between mean values of allospecific identity is weaker and in the opposite direction (Fig. 2C). These findings can be related to the counts of private or quasi-private alleles (defined here as alleles present at frequencies higher than 0.5 in one species that are either absent or present at frequencies lower than 0.01 in the other species). There is a higher number of quasi-private alleles in *Q. petraea* (15) than in *Q. robur* (2) among the 60 outlier loci ($P < 0.05$, Fisher test), with the same trend, though not significant, for private alleles. No such trend was found at nonoutlier loci. All the patterns and trends described above were consistent with those found when considering the much smaller number of outliers detected with the very conservative approach or when using different functional subsets of loci.

Discussion

Our increasing ability to isolate large numbers of loci raises questions about the ideal loci to use for reconstructing population structure and demographic history. Loci likely to be affected by the direct or indirect effects of selection are generally excluded to avoid bias when inferring demographic processes, except in the case of spatial gradients and when using loci with known selection intensities. In contrast, interspecific outlier loci were used in our study to better understand past introgression dynamics, considered to reflect an episode of the species' past demography.

Detection and interpretation of outlier and nonoutlier loci

As a prerequisite for exploring the potential interest of outlier loci for characterizing demographic processes, they have to be identified accurately. Using a reference F_{ST} value from the literature, and excluding SNPs that were intraspecific outliers, we detected 60 outlier SNPs (23% of the total, far above the type I error threshold of 5%), with interspecific F_{ST} values ranging from 0.27 to 0.93, that is, on average five times higher than at nonoutliers. This high rate of outliers is consistent with our strategy to deliberately enrich the panel of SNPs genotyped with markers likely to show high divergence between species. Whereas the number of outliers inferred from the reference value could be over-estimated, the number of outliers that were detected directly from the enriched panel (13, that is, 5% of the total) is surely under-estimated, and the reality probably falls between these limits, considering also the large sample size used (around 800 gametes for each species at each locus). In any case, using the smaller set of outliers did not change the diversity patterns and trends observed in comparison with nonoutliers, confirming the robustness of our interpretation. This interpretation was further supported by the fact that the same patterns of higher genetic diversity in *Quercus petraea* than in *Quercus robur* at outliers, but not at nonoutliers, were observed when using two different functional subsets of loci.

Strong outlier patterns have been classically interpreted as being caused by divergent selection affecting the loci themselves or genes strongly linked with them (Storz 2005). Moreover, our selection of SNPs showing a priori high levels of interspecific differentiation or located within candidate genes of ecologically divergent traits between these two oak species might have increased the chance that some outliers are of adaptive significance. Yet, association genetics and functional studies are ultimately required to confirm that particular loci are directly involved in species divergent trait variation. Indeed, alternative explanations for strong genetic divergence at some loci exist and are difficult to rule out (see e.g. Klopfstein *et al.* 2006; Excoffier & Ray 2008; Bierne *et al.* 2011). Problems of interpretations can also arise for nonoutliers (see e.g. Le Corre & Kremer 2003; Latta 2004; Charlesworth 2006; Kremer & Le Corre 2011). Despite these difficulties, the contrast between loci having different levels of divergence should remain informative as long as the average effective gene flow between species is greater at nonoutliers than at outliers.

Species delimitation and SNP discriminatory power

Another prerequisite for our study was to correctly identify individuals belonging to each oak species (i.e. purebreds). Results of the STRUCTURE clustering analysis with 262 SNPs showed that the assignments of individuals to species largely outperformed those from previous studies of the same species based on small sets of SSR loci (Muir *et al.* 2000; Jensen *et al.* 2009; Lepais *et al.* 2009). Validation of assignment performance requires the use of independent samples (Waples 2010). We

therefore confirmed the repeatability of the results for purebreds using independent SNP data sets. We also found that the proportion of admixed trees is prone to overestimation when few loci are used, as previously noted (Vähä & Primmer 2006). Due to the greater abundance of purebred than admixed trees (hybrids *sensu lato*), more purebred trees are likely to be misassigned as hybrids than the converse. Consequently, reducing the precision of assignment by using less loci would artificially increase the proportion of the admixed category, as might have happened in previous studies based on smaller sets of loci or less powerful markers (e.g. Jensen *et al.* 2009; Lepais *et al.* 2009). Moreover, using only the two loci with the highest interspecific F_{ST} (mean = 0.9), assignment performance reached 97% for *Q. robur* and 91% for *Q. petraea*. In contrast, as many as 49 SNP loci with the lowest differentiation (mean interspecific F_{ST} = 0.02) were required to reach similar performance, confirming that locus selection is critical for species delimitation. Overall, our results illustrate the great value of SNPs for assigning individual genotypes: their lower allelic diversity compared with other loci, especially SSRs (Rosenberg *et al.* 2003), can be compensated for by using more loci or selecting outlier loci (Liu *et al.* 2005).

Signals of asymmetric introgression between species

Assignment results based on all SNPs highlight a genetic asymmetry between the two species, *Q. robur* trees being more easily assigned to the purebred category than *Q. petraea* trees. This can be related to the fact that, at outlier loci, *Q. petraea* has higher genetic diversity than *Q. robur*. Altogether, the results based on outlier loci fit well with our expectations for the introgression dynamics between these two species: past asymmetric introgression towards *Q. petraea* should have increased its diversity and decrease the number of private alleles in *Q. robur*. These findings, based solely on data from purebreds and using trees sampled in different populations, point to a relatively ancient and general trend towards asymmetric introgression in the predicted direction (Currat *et al.* 2008).

In the oak colonization model proposed by Petit *et al.* (1997, 2003) to account for shared chloroplast DNA variation across species, a hybrid phase is hypothesized to occur at the time of establishment of *Q. petraea* invading stands already occupied by *Q. robur* through pollen dispersal. Such populations then evolve to yield backcrosses and eventually typical *Q. petraea* trees within a few generations. Thus, there is a stage in the colonization process where the diversity of *Q. petraea* populations would be maximal. Following this, reproductive isolation would rapidly reemerge (see e.g. Gilman & Behm 2011). As loci under divergent selection should be less likely to experience subsequent genetic exchanges between species than other genes, they should retain the initial introgression signal and correspondingly increased genetic diversity in *Q. petraea* most strongly.

At nonoutlier loci, genetic diversity is instead slightly greater in *Q. robur* than in *Q. petraea*, significantly so for the approach using the 0.04 reference F_{ST} value. As nonoutlier loci should behave most closely to neutral expectations, this observation could indicate that *Q. robur* may have a larger effective population size than *Q. petraea*, in line with its greater distribution range and greater dispersal ability through both pollen and seeds (Petit *et al.* 2003). While the latter inference should be confirmed using other methods such as Isolation with Migration coalescent modelling (e.g. Nielsen & Wakeley 2001; Hey & Nielsen 2004), it illustrates the potential benefits of relying on the two different groups of loci to reconstruct particular demographic features of hybridizing species. In fact, in these oaks, the uninformed use of only one class of markers (e.g. only those that are presumably neutral or only those likely to be under divergent selection) would result in opposite conclusions regarding the genetic diversity maintained by each species and the direction of introgression, highlighting the value of jointly considering and comparing results obtained with both groups of markers (Nosil *et al.* 2009). In our study, we were interested in the statistical signals emerging across different sets of loci, not in the behaviour of individual loci. The outlier group potentially includes genes affected by divergent selection, but the approach does not rely on every single locus in that group being actually under divergent selection. Similarly, the approach does not depend on each nonoutlier locus behaving in a strictly neutral manner. Studies aiming at inferring demographic history driven by drift, bottlenecks, gene flow and inbreeding effects are typically based on genome-wide effects of a large number of markers (Luikart *et al.* 2003), whereas selection studies generally focus on particular genes and their locus-specific effects. The approach used here is an original combination of both methods.

Conclusions

We have shown that outlier loci retain signatures of past asymmetric introgression events presumably caused by differences in colonization history, a signature that is missing at other loci. Our approach takes advantage of the nonlinear dependence of genetic structure on levels of gene flow and of the fact that divergent selection can reduce effective gene flow (Bengtsson 1985; Barton & Bengtsson 1986; Nosil *et al.* 2009). Small differences in

gene flow that are usually hard or impossible to detect might become apparent by using loci under divergent selection. While this study dealt with interspecific genetic exchanges, the principles are general. This suggests that loci experiencing reduced effective gene flow due to selection could help reconstruct other aspects of species' demographic histories, providing insights complementary to those obtained with loci evolving closer to neutral predictions. An important perspective is to incorporate variable selection coefficients in model-based approaches aiming at inferring past demographic processes to take advantage of the demographic signal available in the different categories of loci.

Acknowledgements

We thank E. Dreyer and O. Brendel for material and information on the oak stands from eastern France, A. Ducousso and J.-M. Louvet for information on the oak stands from Petite Charnie and S. Wagner for the sampling, and all the persons who contributed to the preliminary candidate gene lists for the oak allelic resequencing project (P. Abadie, C. Bodénès, C. Burban, T. Decourcelle, J. Derory, M.-L. Desprez-Loustau, A. Kremer, G. Le Provost, C. Plomion and C. Robin). We are grateful to J.-M. Frigerio and C. Lepoittevin for their help in developing the new version of the SNP2Illumina perl script, to S. Ueno and I. Lesur for bioinformatic support, to F. Alberto for providing data on previously validated SNPs, to M. Navascués for sharing with us his ideas on the identification of informative loci and to the referees for helpful advices. Genotyping was performed in the Genome-Transcriptome facility of the Functional Genomic Centre of Bordeaux. E.G. was employed during his PhD by the Pernod Ricard Research Center and then as postdoc by INRA (OAKTRACK project, ANR-10-EMMA-0016). T.L. received successive postdoctoral fellowship grants (the first for 18 months and the second for 6 months) from the TRANSBIODIV and LinkTree projects. L.L. was funded by a PhD grant from the EVOLTREE network of Excellence and LinkTree project.

References

Abadie P, Roussel G, Dencausse B *et al.* (2012) Strength, diversity and plasticity of postmating reproductive barriers between two hybridizing oak species (*Quercus robur* L. and *Quercus petraea* (Matt) Liebl.). *Journal of Evolutionary Biology*, **25**, 157–173.

Abbott RJ, James JK, Milne RI, Gillies ACM (2003) Plant introductions, hybridization and gene flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **358**, 1123–1132.

Arnold M (1997) *Natural Hybridization and Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, UK.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.

Barton NH, Hewitt GM (1981) The genetic basis of hybrid inviability in the grasshopper *Podisma pedestris*. *Heredity*, **47**, 367–383.

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 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* (eds. Greenwood JJ, Harvey PH, and Slatkin M), pp. 31–42. Cambridge University Press, Cambridge, UK.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

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.

Butlin RK (2008) Population genomics and speciation. *Genetica*, **138**, 409–418.

Carlisle A, Brown AHF (1965) The assessment of the taxonomic status of mixed oak (*Quercus spp.*) populations. *Watsonia*, **6**, 120–127.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, **2**, e64.

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, Van Bockstaele E, Lootens P, Roldan-Ruiz I (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.

Cousens JE (1963) Variation of some diagnostic characters of the sessile and pedonculate oaks and their hybrids in Scotland. *Watsonia*, **5**, 273–286.

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

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

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

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.

Gilman RT, Behm JE (2011) Hybridization, species collapse, and species reemergence after disturbance to premating mechanisms of reproductive isolation. *Evolution*, **65**, 2592–2605.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). *Molecular Ecology Resources*, **11**, 578–585.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

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.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

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**, 706.

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.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kremer A, Le Corre V (2011) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, **108**, 375–385.

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.

Lagache L, Klein EK, Guichoux E, Petit RJ (2012) Fine-scale environmental control of hybridization in oaks. *Molecular Ecology*, doi: 10.1111/mec.12121.

Latta RG (2004) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, **161**, 51–58.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Lenormand T, Guillemaud T, Bourguet D, Raymond M (1998) Evaluating gene flow using selected markers: a case study. *Genetics*, **149**, 1383–1392.

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.

Li JR, Li HP, Jakobsson M, Li S, Sjodin P, Lascoux M (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, **21**, 28–44.

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.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **363**, 3009–3021.

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.

Mallet J, Barton N, Gerardo LM, Jose SC, Manuel MM, Eeley H (1990) Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics*, **124**, 921–936.

Marsden CD, Lee Y, Nieman CC *et al.* (2011) Asymmetric introgression between the M and S forms of the malaria vector, *Anopheles gambiae*, maintains divergence despite extensive hybridization. *Molecular Ecology*, **20**, 4983–4994.

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

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, **70**, 3321–3323.

Neigel JE (2002) Is F_{ST} obsolete? *Conservation Genetics*, **3**, 167–173.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

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.

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

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.

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, Bahrman N, Baradat P (1995) Comparison of genetic differentiation in maritime pine (*Pinus pinaster* Ait) estimated using isozyme, total protein and terpenic loci. *Heredity*, **75**, 382–389.

Petit RJ, Pineau E, Demesure B, Bacilieri R, Ducousso A, Kremer A (1997) Chloroplast DNA footprints of postglacial recolonization by oaks. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9996–10001.

Petit RJ, Bodénès C, Ducousso A, Roussel G, Kremer A (2003) Hybridization as a mechanism of invasion in oaks. *New Phytologist*, **161**, 151–164.

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

Quintana-Murci L, Quach H, Harmant C *et al.* (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences*, **105**, 1596–1601.

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.

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.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.

Steinhoff S (1993) Results of species hybridization with *Quercus robur* L. and *Quercus petraea* (Matt) Liebl. *Annals of Forest Science*, **50**, 137s–143s.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Streiff R, Ducousso A, Lexer C, Steinkellner H, Gloessl J, Kremer A (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.

Szymura JM, Barton NH (1991) The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, **45**, 237–261.

Tiffin P, Olson MS, Moyle LC (2001) Asymmetrical crossing barriers in angiosperms. *Proceedings of the Royal Society of London B: Biological Sciences*, **268**, 861–867.

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.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

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.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

E.G., P.G.-G. and R.J.P. designed the experiment when E.G. was a PhD student under the supervision of R.J.P. P.G.-G. coordinated the resequencing study that provided the main source of SNPs (95%) for this study. T.L. and P.G.-G. developed the bioinformatics tools used to identify and validate these SNPs. L.L., E.G. and P.G.-G. conceived the Illumina genotyping assay. L.L. and C.B. performed SNP genotyping and provided multilocus genotypes. E.G. analysed the data with the help from P.G.-G. (outlier detection) and R.J.P. (species delimitation). E.G., P.G.-G. and R.J.P. wrote the paper. All authors have checked and approved the final version of the manuscript.

Data accessibility

Summary SNP data are available from Appendix S5 (Supporting information).

SNP genotypes are available from Dryad: doi:10.5061/dryad.3g140.

Locus name, sequence, target functional trait, gene annotation and contig reference are available from the *Quercus* Portal: https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=snp.

Contig assembly, contig blast and data mining are available from the GENOTOUL Forest Trees Contig Browser Oak: http://genotoul-contigbrowser.toulouse.inra.fr:9092/Quercus_robur/index.html.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Impact of selection on indirect measures of gene flow.

Appendix S2 Sampling sites in France.

Appendix S3 Distribution of *Quercus petraea* (blue) and *Quercus robur* (red) across Europe (Ducousso & Bordacs, with permission).

Appendix S4 Detailed SNP selection and outlier detection methods.

Appendix S5 Detailed characteristics of the 384 SNPs selected for inclusion in the Illumina array (.xls file).

Appendix S6 Distribution of interspecific F_{ST} values as a function of the mean values of within-demes heterozygosity (H_{WP}), simulated with the FDIST method implemented in ARLEQUIN 3.5.1.2 (Excoffier *et al.* 2009).

Appendix S7 Characteristics of the 384 SNPs used.

Appendix S8 Barplot of the number of individuals (among 855) assigned to different categories from either 12 SSRs (light colours) or 262 SNPs (dark colours).

Appendix S9 Assignment score of 5000 simulated genotypes, i.e. 1000 purebreds *Quercus robur*, 1000 purebreds *Quercus petraea* and 3000 admixed individuals (1000 F1s, 1000 backcrosses *Q. robur* and 1000 backcrosses *Q. petraea*) from 262 SNPs.

Appendix S10 Distribution of observed interspecific F_{ST} values (calculated between purebreds over 262 SNPs).

Appendix S11 Distribution of intraspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), using a finite island model with six demes for *Quercus robur*.

Appendix S12 Distribution of intraspecific F_{ST} values for all 262 SNPs as a function of their mean within-species heterozygosity value (H_{WP}), using a finite island model with six demes for *Quercus petraea*.

Appendix S13 Genetic diversity and inter- and intraspecific differentiation at outlier vs. non-outlier loci for the subset of loci involved in drought stress (A) and for the subset of loci not involved in drought stress (B).

ANNEXE 4: GENETIC DIVERSITY INCREASES INSECT HERBIVORY ON OAK SAPLINGS

OPEN ACCESS Freely available online

PLOS ONE

Genetic Diversity Increases Insect Herbivory on Oak Saplings

Bastien Castagneyrol[1,2]*, Lélia Lagache[1,2], Brice Giffard[1,2], Antoine Kremer[1,2], Hervé Jactel[1,2]

1 University Bordeaux, BIOGECO, UMR1202, Talence, France, **2** INRA, BIOGECO, UMR1202, Cestas, France

Abstract

A growing body of evidence from community genetics studies suggests that ecosystem functions supported by plant species richness can also be provided by genetic diversity within plant species. This is not yet true for the diversity-resistance relationship as it is still unclear whether damage by insect herbivores responds to genetic diversity in host plant populations. We developed a manipulative field experiment based on a synthetic community approach, with 15 mixtures of one to four oak (*Quercus robur*) half-sib families. We quantified genetic diversity at the plot level by genotyping all oak saplings and assessed overall damage caused by ectophagous and endophagous herbivores along a gradient of increasing genetic diversity. Damage due to ectophagous herbivores increased with the genetic diversity in oak sapling populations as a result of higher levels of damage in mixtures than in monocultures for all families (complementarity effect) rather than because of the presence of more susceptible oak genotypes in mixtures (selection effect). Assemblages of different oak genotypes would benefit polyphagous herbivores via improved host patch location, spill over among neighbouring saplings and diet mixing. By contrast, genetic diversity was a poor predictor of the abundance of endophagous herbivores, which increased with individual sapling apparency. Plant genetic diversity may not provide sufficient functional contrast to prevent tree sapling colonization by specialist herbivores while enhancing the foraging of generalist herbivores. Long term studies are nevertheless required to test whether the effect of genetic diversity on herbivory change with the ontogeny of trees and local adaptation of specialist herbivores.

Citation: Castagneyrol B, Lagache L, Giffard B, Kremer A, Jactel H (2012) Genetic Diversity Increases Insect Herbivory on Oak Saplings. PLoS ONE 7(8): e44247. doi:10.1371/journal.pone.0044247

Editor: Justin Wright, Duke University, United States of America

Received May 8, 2012; **Accepted** July 31, 2012; **Published** August 28, 2012

Funding: This study was founded by the EVOLTREE Network of Excellence (EU FP7 project 016322). BC was funded by a grant for the BACCARA project, which received funding from the European Commission's Seventh Framework Programme (FP7/2007–2013), under grant agreement no. 226299. LL received funding from the LinkTree project (ANR BIODIVERSA) and the EU Network of Excellence EvolTree. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bastien.castagneyrol@gmail.com

Introduction

Over the last decades, the role that biodiversity plays in ecosystem functioning has emerged as a key issue in ecology [1,2,3]. Although a majority of studies have focussed on the effect of plant diversity on primary production [4], a growing attention is being paid on other ecosystem services provided by biodiversity such as pest regulation.

The diversity – resistance hypothesis states that species rich plant communities suffer less feeding damage by herbivores than plant monocultures [5,6,7]. However two opposite effects of plant diversity on herbivory have been observed [8]. A given focal plant species can experience more damage when associated with other plant species that are more attractive or palatable for herbivores [8,9]. This pattern is known as associational susceptibility and seems to mainly involve generalist herbivore species [7]. Conversely, a focal plant species can have less herbivore damage (*i.e.* associational resistance) when the presence of non conspecific neighbours (i) reduces host plants concentration and the probability to be located by specialist herbivores [10]; (ii) provides physical or chemical barriers to host colonisation [11,12,13] and (iii) increases the abundance, the diversity and/or the efficiency of natural enemies [14,15,16,17]. Several meta-analyses have shown that associational resistance is more frequent than associational susceptibility but the balance between these two mechanisms is likely to depend on the identity of host plant species, herbivore feeding guilds or the way herbivory is assessed (abundance of herbivores *vs* biomass removed) [5,7,8,18].

Intraspecific diversity (*i.e.* genetic diversity) is a key component of biodiversity. Recent research in the field of community genetics has shown that host plant genotype is one of the ecological filters shaping the structure of insect species assemblages [19,20,21] and that insect species diversity increases with the genetic diversity in host plant populations [22,23,24]. The question of the effects of genetic diversity on ecosystem functioning has also attracted considerable interest in recent years. It has been shown that most of the ecosystem functions provided by species diversity are also supported by genetic diversity, including plant productivity [25,26] nutrient cycling [27], temporal stability [28,29,30] and resistance to invasion [22]. Despite the similarity between the effects of plant species and plant intraspecific diversity on ecosystem properties, the mechanisms underlying the biodiversity-ecosystem functioning relationship may be different at the two scales. For example, Cook-Patton *et al.* [25] showed that the increase in arthropod species richness with plant genetic diversity is mediated by arthropod abundance while resource specialisation is the main factor explaining the increase in arthropod species richness with plant species richness. It is therefore necessary to verify whether the

ecological processes leading to associational resistance or susceptibility in plant species assemblages also apply to assemblages of plant genotypes.

Because herbivore species richness and abundance generally increase with genetic diversity in host plant populations [22,25,26], associational susceptibility may be more likely to occur than associational resistance in mixtures of host plant genotypes. In addition, generalist insect herbivores (such as grasshoppers and many leaf chewers) are known to develop better on plant species mixtures due to food resources complementation or toxins dilution, a phenomenon known as diet mixing [31,32] that has been reported for mixtures of plant genotypes [33,34]. Generalist herbivores are then expected to cause higher damage in genotype mixtures. Recent studies have shown positive [23,35] or neutral [36] effects of host plant genetic diversity on the abundance of specialist herbivores. It is therefore still uncertain whether plant genetic diversity might have different effects on herbivores with different diet breadth or feeding behaviour [37,38].

With a few exceptions [27,39,40,41], studies on the functional consequences of genetic diversity for ecosystem functioning have focused on hybrids [20,42,43] or clones [23,25,35,44,45,46]. Because these studies were designed so as to increase the contrast between plant genotypes, they may not be relevant to more complex processes occurring in more natural conditions [46]. We present here one of the first attempt to assess the effect of casual intraspecific plant diversity on natural insect herbivory. Using an experimental plantation of pedunculate oak saplings, we tested the following hypotheses: (i) the genetic diversity of young trees tends to increase insect herbivory (*i.e.* associational susceptibility) and (ii) the magnitude of the effect depends on host specialization of insect herbivores, being higher for more generalist species. To test these hypotheses we designed a common garden experiment with 90 synthetic mixtures of oak saplings composed of one to four half-sib families. We genotyped all saplings and evaluated the amount of damage caused by ectophagous insect herbivores (less specialized) and endophagous leaf miners (more specialized) on each individual sapling. We assessed the level of genetic diversity in each mixture and estimated the correlation between diversity and insect herbivory.

Materials and Methods

No specific permits were required for the described field studies. The site on which the experimental common garden was established is owned by our institute (INRA) and is no subjected to any protection scheme. This work did not involve any endangered or protected species or area.

Experimental design

In autumn 2007, we collected acorns from the canopy of four mature pedunculate oaks (*Quercus robur*), referred to hereafter as 'mother trees', sampled at random within a 10 km radius at a site 40 km south of Bordeaux (44°440 N, 00°460 W). In March 2008, we sowed the acorns at the nursery of the forest research centre of the French National Institute for Agricultural Research (INRA), to produce four half-sib families of oak seedlings. The seedlings were grown in individual pots containing peat and were treated with fungicide and insecticide during the first growing season (*i.e.* 2008), to prevent damage before planting. In March 2009, the seedlings were transplanted to a clearing surrounded by pine trees (*Pinus pinaster*) and broadleaved species (*Quercus robur, Quercus rubra* and *Betula pendula*).

Six different blocks were established, with 15 plots in every block, each plot corresponding to one of the 15 possible combinations of one ($n = 4$ plots, *i.e.* one per family), 2 ($n = 6$), 3 ($n = 4$) and 4 ($n = 1$) families per plot. Each plot contained four rows of three seedlings; the seedlings were 0.2 m apart and the plot area was 0.24 m^2 (0.60×0.40). Within each plot, oak families were planted at equal density in a regular alternate pattern, such that seedlings from the same family were never adjacent in mixed plots. The plots were separated by a distance of 3 m and were randomly distributed within the blocks. Blocks were 14 m ×6 m in size and were located 4 m apart (Figure S1).

The experimental site was fenced to prevent grazing by mammalian herbivores. The herbaceous plants growing between plots were removed by mowing, twice yearly. Pine bark chips were spread on the soil of each plot to control the vegetation and limit evaporation. Plots were watered during the summer of 2009, to minimise seedling mortality. In August 2011, 25 out of the 1080 planted seedlings were dead (*i.e.* 1055 survived).

Herbivory assessment

Insect herbivory was assessed by the visual inspection of 20 leaves on each four-year-old sapling, in August 2011. Five leaves were sampled at the tip and five at the base of two branches randomly chosen at the top and two branches randomly chosen towards the bottom of the sapling. We also recorded the total height of each sapling during this herbivory assessment.

Herbivore damage on oak leaves was assigned to four different trophic guilds: *chewers* (mostly adult Curculionidae or Chrysomelidae and Lepidoptera caterpillars), *skeletonisers* (adult grasshoppers and Tenthredinoidea larvae), *rollers* (mostly Lepidoptera larvae) and *miners* (mostly Microlepidoptera larvae). No gall makers were observed. The percentage leaf area affected was visually estimated for each leaf and each guild using six classes (0%, 1–5%, 6–15%, 16–25%, 26–50%, 51–75% and >76%) and then averaged per sapling.

Damage due to skeletonisers and leaf-rollers were very rare. We therefore pooled these two guilds with the chewers and classified the damage inflicted as being due to 'ectophagous insects'. Previous work by Giffard *et al.* [47] in the same study area showed that most of ectophagous insect herbivores found feeding on *Q. robur* are polyphagous species able to consume plant tissues from different genera and families and may be then considered as generalists (see [47] for a list of the commonest species). Leaf miners are different from the other insect herbivores found on oak saplings in that they are endophagous and much more specialized (they develop on a narrow spectrum of species within the Fagaceae family). Damage by leaf miners was quite frequent but minor in term of leaf area impacted. In addition, the leaf surface affected by a mine is dependent on the timing of assessment, while the presence or absence of a mine is not. We therefore used the density of mines per sapling (number of mines/20 leaves) to quantify damage due to these specialist insects.

Genotyping of oak saplings

All oak saplings and the four mother trees were genotyped with 12 microsatellite markers (see Guichoux *et al.* [48] for details), using one leaf per sapling and per mother tree collected in August 2010. Leaves were dried and stored separately before DNA extraction and gene amplification. We isolated DNA from five leaf discs, each 5 mm in diameter, from each sample with the Invisorb DNA plant HTS 96 kit (Invitek, Berlin, Germany). We used the 12plex SSR (Single Sequence Repeats) kit developed by Guichoux *et al.* [48] for genotyping. We scored SSR profiles, using real allele sizes and alleles were binned with the Microsoft Excel macro AUTOBIN program (available from http://www4.bordeaux-

aquitaine.inra.fr/biogeco/Ressources/Logiciels/Autobin) developed by Guichoux *et al.* [49].

Among the 1059 surviving individuals in 2010, 1032 were successfully genotyped. The mean proportion of loci succesfully typed was 99.7%. The mean number of alleles per locus was 11 (range: 6–19). More detailed information about genetic structure of the oak seedlings population is provided in Table S1. Seventeen offspring (1.7%) were excluded from the analysis because their genotype at multiple loci did not match that of any mother tree. 134 offsprings showed only one mismatch with the corresponding mother tree. These offsprings were used to identify loci with genotyping errors before correction. Error rates based on these comparisons were low for 10 markers (<2%), high for one single-nucleotide marker (1.92% for the PIE258 marker) and high for another marker (10.49% for the PIE020 marker). Comparisons of the genotypes of mother trees and offspring revealed that manual binning was incorrect for the single-nucleotide marker and a null allele in the offspring of one mother tree, for the PIE020 marker. Single-nucleotide errors were corrected for further analysis and manual binning was repeated for the PIE258 marker. The PIE020 marker was removed from the data set. We finally retained 11 markers for the genotyping of 1016 offsprings plus the four mother trees.

Estimation of genetic diversity

We initially used the number of maternal lineages per plot as a measure of genetic diversity. However, as a given mother tree could have been pollinated by several father trees, the offspring may be half-sibs or full-sibs and the proportion of the two types of saplings could vary within families and within sapling assemblages. The number of maternal lineages per plot may therefore underestimate genetic diversity and be poorly correlated with variation in insect damage.

We then determined SSR genotypes, to calculate the genetic relatedness between oak saplings, thereby improving estimates of genetic diversity per plot and switching from an almost categorical (1, 2, 3 or 4 maternal lineages per plot) to a more continuous (90 individual scores of genetic diversity) variable. Hereafter, *genetic diversity* (GD) refers to the number of maternal lineages per plot, whereas *genetic relatedness* (GR) refers to the mean between-saplings relatedness per plot.

Genetic relatedness was calculated with CoAncestry software [50]. We used the DyadML estimator (a dyadic likelihood estimator described in [51]) because the simulated values of relatedness it provided were the closest to expected values (*i.e.* 0.5 for full sibs, 0.25 for half sibs and 0 for unrelated saplings). GR was calculated for all pairs of individuals ($n = ½(1016 \times (1016–1)) = 515{,}620$ pairs) and we used these values to calculate a mean genetic relatedness for each plot. Mean genetic relatedness significantly differed between plots with different numbers of maternal lineages (Kruskal-Wallis test: $K_{calc} = 845.55$, $df = 3$, $p<0.001$), decreasing with increasing number of lineages (Figure S2). However, genetic relatedness also varied considerably within each level of genetic diversity, thus supporting the use of the two indices. As they were highly correlated ($r = –0.80$, Figure S2), GD and GR were introduced separately in further models.

Owing to missing genotypes (dead saplings, unamplified DNA, mismatch between observed genotype and mother tree), mean relatedness was averaged across a variable number of individual saplings per plot (9 to 12). For the sake of consistency, missing genotypes were also removed before the analysis of insect damage data. The final dataset contains 1002 individuals (6 blocks ×15 plots ×12 trees –25 dead saplings –53 unamplified or mismatched genotypes).

Statistical analyses

Response variables (*i.e.* herbivory by ectophagous insects, abundance of leaf miners and sapling height) were analysed using each individual tree as a replicate to make it possible to test for the possible effects of interactions between mother tree identity (MT) and GD or GR. We accounted for spatial replication by nesting the 'population effect' (*i.e.* 1 population = 1 plot = 12 saplings) within the block effect, both factors being treated as random effects in all mixed models, in order to specify that individual observations were correlated within blocks and within plots.

We first tested the MT effect on response variables in monocultures alone, to avoid confounding factors. Mother tree identity was part of the experimental design and we were interested in its influence on the mean of herbivory by ectophagous insects, abundance of leaf miners and tree height. We therefore treated this factor as a fixed effect because there were not enough levels on which to base an estimate of the variance of the total population (only four different mother trees).

In order to determine potential genetic effects on sapling height, we first performed two sets of linear mixed models with MT and GD or GR, separately, and their interactions as fixed effects. We then carried out linear mixed models to test the effect of sapling height, MT, GD or GR, and their interactions on herbivory by ectophagous insects (% leaf area damaged) and by endophagous insects (abundance of leaf miners), separately. Prior to analyses, continuous explanatory variables (GD, GR and sapling height) were centred (*i.e.* subtracting the sample mean from all observations) and reduced (*i.e.* dividing centred variables by their sample standard deviation) in order to make model coefficients comparable within and between models [52] and to allow estimating the magnitude of effects. Centring variables also makes main effects biologically interpretable even when involved in interactions [52].

In all mixed models, we applied a model simplification procedure and reduced each maximal mixed model by removing non significant interaction terms, starting with the highest order interaction, to finally retain the least parameterized models including only simple terms and significant interaction terms.

Test statistics for fixed effects were based on F values for linear mixed models (herbivory by ectophagous insects and sapling height) and on χ^2 values (loglikelihood ratio tests with one degree of freedom) for generalised linear mixed models performed on the abundance of leaf miners. Log-likelihood R^2 values were calculated to estimate the amount of variance explained by each independent variable [53].

Data for sapling height and damage due to ectophagous insects were analysed with linear mixed models with the *lme* procedure [54] in R [55]. Tree height was square-transformed and percentage data were transformed with the *logit* function [56] to meet the assumptions of these tests, which were checked by graphical analyses and Shapiro-Wilk tests on model residuals. The abundance of leaf miners per tree was expressed as counts, which were analysed with generalized linear mixed models by specifying a Poisson error structure, with the *lmer* procedure in the lme4 package in R [57].

We used the method developed by Loreau and Hector [58] and adapted by Unsicker *et al.* [1] to quantify the net genetic diversity effect on herbivore damage. We first calculated the observed relative forage of the family i (RF_{Oi}) as the ratio of the damage observed on each family i (i from 1 to 4) in a mixture (C_i) to that

observed on this family in monoculture (M_i) [1]:

$$RF_{Oi} = C_i / M_i \quad (1)$$

The expected relative forage of the family i (RF_{Ei}) under the null hypothesis (*i.e.* no effect of genetic diversity on damage) was simply its proportion in the mixture, *i.e.* $1/n$ where n is the number of families in the mixture [1,58].

The deviation of the observed relative damage in a mixture from the relative damage expected in the corresponding monoculture was thus:

$$DRC_i = RF_{Oi} - RF_{Ei} = C_i/M_i - (1/n) \quad (2)$$

The total observed damage in the mixture was calculated as:

$$F_O = \Sigma_i RF_{Oi} \quad (3)$$

The total expected damage in the mixture was calculated as:

$$F_E = \Sigma_i RF_{Ei} \quad (4)$$

A positive NGDE indicates associational susceptibility (higher level of damage observed in mixtures than expected from mean damage levels in the corresponding monocultures), whereas a negative NGDE indicates associational resistance (lower level of damage observed in mixtures than expected from mean damage levels in the corresponding monocultures).

The NGDE can be further divided into two additive components: a complementarity effect (CE) and a selection effect (SE) [1,58].

$$NGDE = F_O - F_E \quad (5)$$

The CE is assessed by calculating the mean ΔRC_i over all families at the plot level:

$$CE = n \times \overline{\Delta RC} \times \overline{M} \quad (6)$$

CE measures the change in mean relative forage of the species. CE is positive when the mean relative forage increases *i.e.* when oak families are, on average, consumed more in mixtures than it would be expected from their consumption in monocultures.

The calculation of SE takes into account the covariance between ΔRC_i and M_i:

$$SE = n' \mathrm{cov}(DRC, M) \quad (7)$$

SE values are used to determine whether there is a relationship between consumption in the monoculture and relative forage in mixtures. SE is positive when plant species that are consumed in larger amounts in monocultures (less resistant) also have higher relative forage values in mixtures, thus making a greater contribution to total plot damage.

NGDE, CE and SE were calculated for all levels and combinations of mixtures within each block, giving a total of 66 comparisons between observed and expected values. The significance of each effect (NGDE, CE, SE) was determined by one sided t-tests [58]. We first tested grand mean values across all mixtures against zero, to determine whether they differed significantly from the weighted average of the response variable in monocultures. We also assessed the significance of the NGDE, CE and SE against zero for each level of genetic diversity. We used analyses of variance to assess change in these three effects along the gradient of GD [58].

Results

Effects of genetic diversity and relatedness on sapling height

Mean sapling height significantly differed between oak families ($F_{3,909} = 2.88$, $p = 0.035$) but we observed no significant effect of genetic diversity (GD: $F_{1,83} = 0.02$, $p = 0.880$) or genetic relatedness (GR: $F_{1,83} < 0.01$, $p = 0.984$) on sapling height.

Effects of genetic diversity and relatedness on insect herbivory

Damage due to ectophagous insects was significantly affected by MT, GD, GR and sapling height (H), but not by interactions between these factors (Table 1). Significant differences in damage levels between families were observed in monocultures (on average 5.5 and 8.3% of leaf area was removed in the more and the less resistant families, respectively), suggesting a genetic control of oak saplings resistance to ectophagous insects ($F_{15,244} = 4.81$; $P = 0.015$; $R^2 = 0.04$, Figure 1A). Damage also increased significantly with the GD of saplings (Figure 2B) and decreased significantly with increasing GR, regardless of the family considered (Figure 2C), indicating that the presence of more genetically diverse neighbours increased the risk of damage and that this risk increased with the diversity of conspecific neighbours. However, the magnitude of this effect was low, leaf area removed being on average 6.9% in monocultures and 7.9% in 4-families mixtures. Damage by ectophagous insects also increased significantly with sapling height (Table 1, Figure 2A). The effects of GD, GR and H on damage by ectophagous herbivores were comparable in terms of magnitude, as shown by standardized model coefficients (Table 1). The effects of GD and GR on herbivory seemed to be direct rather than mediated by the genetic control of sapling height as (i) GD and GR had no effect on height and (ii) MT×H ($F_{3,905} = 1.56$; *p-value* $= 0.198$), GD×H ($F_{1,901} = 0.08$; *p-value* $= 0.775$) and GR×H ($F_{1,907} = 2.92$; *p-value* $= 0.088$) interactions had no significant effect on damage (Table 1).

For endophagous herbivores (*i.e.* leaf miners), sapling height emerged as the main factor determining their abundance on individual saplings (Table 1, Figure 2D). GD×H and GR×H interactions also significantly affected the abundance of leaf miners (Table 1) while H×MT did not ($\chi^2 = 1.05$; *p-value* $= 0.789$). The coefficient estimate of GD×H interaction term was negative (Table 1) which means that the strength of the effect of sapling height on abundance of leaf miners decreased when increasing GD. The opposite was true for GR×H (Table 1), which is consistent with the negative covariation between GD and GR.

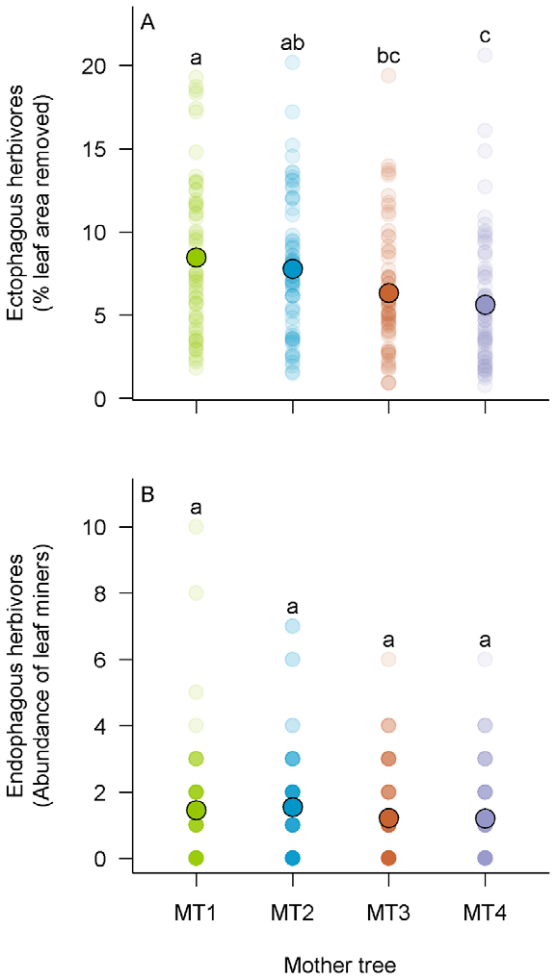


Figure 1. Effect of mother tree identity on insect herbivores in monocultures. (A) Effect of mother tree identity on damage (% leaf area removed) due to ectophagous herbivores. (B) Effect of mother tree identity on the abundance of endophagous insect herbivores. Semi transparent coloured circles represent individual saplings. Darkest circles represent overlapping datapoints. Solid black circled dots indicate the mean values in monocultures for all saplings and all blocks. Same letter above two lines of dots indicates that the corresponding means were not significantly different (LMM and GLMM on monoculture plots).
doi:10.1371/journal.pone.0044247.g001

However, standardized coefficients of regression of both GD×H and GR×H were low compared to the coefficient of regression for H. The simple effects of MT (Figure 1B), GD (Figure 2E) and GR (Figure 2F) on leaf miner abundance were not significant (Table 1).

Net genetic diversity effect

The net genetic diversity effect (NGDE) on herbivory by ectophagous insects was overall significantly positive (Table 2), indicating a higher level of damage in mixtures than expected from monocultures (*i.e.* associational susceptibility). Both complementarity and selection effects (CE and SE) were significantly different from zero (Table 2) but had opposite signs (Figure 3): mean CE was positive and more than three times higher than mean SE, which was negative. The resulting positive NGDE was therefore principally due to the positive complementarity effect. Mean NGDE and CE were consistently positive at each level of genetic diversity, and SE was significantly negative at all but the higher level of GD (Figure 3, Table 2). A negative Selection Effect indicates a negative covariation between damage in monocultures and the deviation between observed and expected relative damage in mixtures (Figure S3). For families with lower levels of damage in monocultures (*i.e.* intrinsically more resistant), herbivory in mixtures was much higher than expected (Figure S3).

There was no significant NGDE or CE on the abundance of endophagous herbivores (Table 2). By contrast, SE was significant and negative at all levels of genetic diversity (Figure 3, Table 2). Thus, for the families showing a tendency for higher infestation in monocultures (more susceptible), relative infestation levels were lower than expected in mixtures, the opposite being true for less susceptible families. CE and SE were of similar magnitude but of opposite signs, accounting for the null NGDE.

NGDE and CE were not significant for sapling height, for either the grand mean, or for any of the levels of genetic diversity, with exception of the 4-families mixtures (Table 2). SE was consistently and significantly negative (but for GD = 4).

Discussion

Based on a large number of samples and a manipulative experiment this study shows for the first time that genetic diversity can trigger associational susceptibility to insect herbivory [8,9] in tree saplings. This process describes an increase in insect herbivory with increasing genetic diversity in host population.

The relative importance of genetic diversity *vs.* other ecological factors as drivers of ecosystem processes is a central issue for community genetics [46,59,60]. In our study we found that the effects of genetic diversity or relatedness on insect herbivory were overall significant but low in terms of magnitude. These results are consistent with the small effects of tree genetic diversity on structuring the insect community associated with pedunculate oak, as recently reported by Tack *et al.* [39,40]. In addition, we showed that sapling height was as important as genetic diversity for predicting generalist herbivore damage and the best predictor of endophagous herbivores abundance. These findings suggest that the influence of host tree genetic diversity on insect herbivores may originate in the variance of particular functional traits.

If genetically based differences in tree susceptibility to herbivores is now well documented [39,61,62,63,64], the effect of genetic diversity on insect damage has rarely been investigated and most often on crops or herbaceous plants [23,36,65]. Recently, Tack *et al.* [39,40] studied the effects of genotype identity and diversity on the structure of endophagous insect communities on *Quercus robur*, but they did not measure corresponding herbivory. We are aware of only two studies that investigated the relationship between tree genetic diversity and insect damage. They reported a trend towards higher levels of pest damage in monocultures than in mixtures of willow clones [66,67] but they focused on only two specialized leaf beetles. The consequences of tree genetic diversity on total herbivory remain largely unknown. A similar concern is currently emerging about the effect of plant species diversity on insect herbivory. The diversity-resistance relationship has been clearly demonstrated by meta-analyses focusing on individual species-species interactions [7,68,69]. However, the diet breadth of insect herbivores emerged as a key factor accounting for differences in insect response to plant diversity. Herbivory by oligophagous species is often reduced in mixed-species forests in comparison to monospecific forests, whereas the response of polyphagous insect species is more variable [7]. Several examples of such opposite patterns have recently been reported, with higher levels of damage caused by polyphagous insects [70,71] and lower abundance of oligophagous insects [72] in more diverse plant

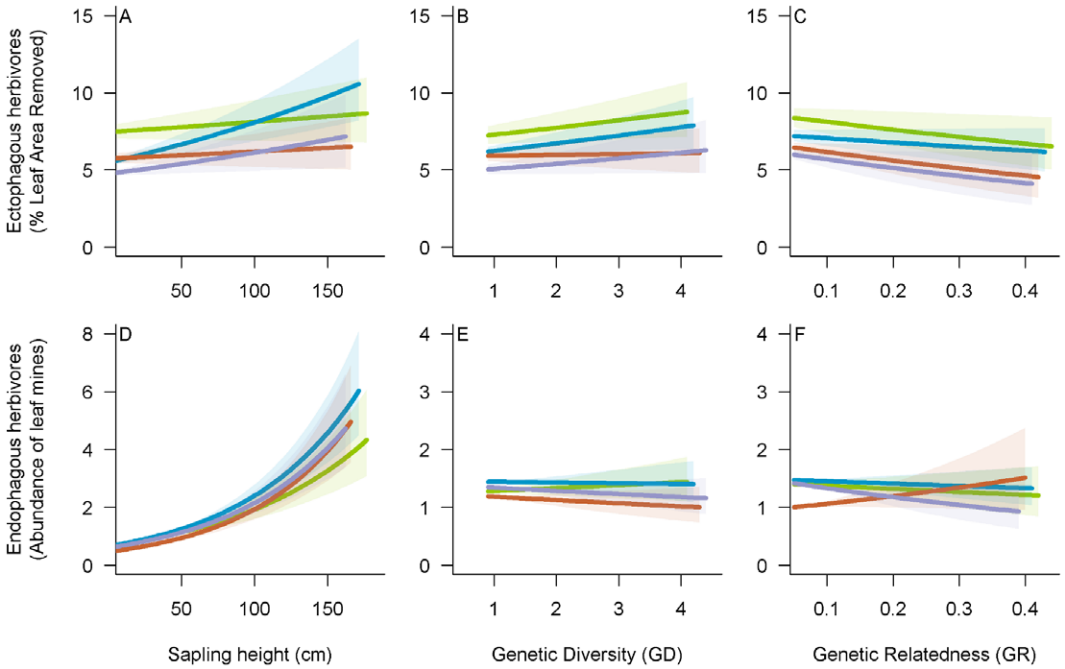


Figure 2. Effects of sapling height, genetic diversity and genetic relatedness on ectophagous and endophagous insects. Effects of sapling height (A, D), genetic diversity (B, E) and genetic relatedness (C, F) on damage due to ectophagous insects (A, B, C) and on the abundance of endophagous insects (D, E, F). The different colours indicate regression lines for different families (MT1: green, MT2: blue, MT3: red, MT4, purple). The shaded areas indicate the corresponding 95% confidence intervals.
doi:10.1371/journal.pone.0044247.g002

communities. The effect of plant diversity on total insect herbivory may then primarily depend on the share of generalist and specialist herbivore species. Here we tentatively addressed this issue by considering two guilds of herbivores of contrasting diet breadth. Oak leaf-miners are oligophagous species that develop a narrow range of species within the *Quercus* and the *Castanea* genera while all the ectophagous insects we observed in the field were polyphagous species able to feed on host plants belonging to different families (see [47] for the list of insect herbivore species found on oak trees in the study area). As endophagous herbivores, leaf-miners have an intimate relationship with their host and are expected to be more dependent on host genotype than ectophagous insects that can move freely and exploit several hosts during their development [20,46].

Table 1. Summary of the results of linear mixed models assessing the effect of sapling height (H), mother tree identity (MT), genetic diversity (GD) and genetic relatedness (GR) between oak saplings and their interactions on herbivory by ectophagous insects and on abundance of endophagous insects (leaf-miners).

		Ectophagous insects					Endophagous insects			
		df[a]	Coefficients of regression (± SE)	F-value	p-value	Log-likelihood R^2	Coefficients of regression (± SE)	χ^2	p-value	Log-likelihood R^{2b}
Genetic diversity	H	1, 908	0.05 ± 0.02	6.51	**0.011**	0.006	0.31 ± 0.052	127.53	**<0.001**	
	MT	3, 908		18.39	**<0.001**	0.052		7.65	0.054	
	GD	1, 83	0.06 ± 0.03	4.45	**0.038**	0.004	−0.001 ± 0.080	<0.001	0.995	
	H × GD	-					−0.003 ± 0.088	12.50	**0.014**	0.004
Genetic relatedness	H	1, 908	0.05 ± 0.02	6.57	**0.011**	0.007	0.31 ± 0.05	128.48	**<0.001**	
	MT	3, 908		19.22	**<0.001**	0.054		7.51	0.057	
	GR	1, 83	−0.06 ± 0.03	4.49	**0.037**	0.004	−0.04 ± 0.08	1.08	0.300	
	H × GR	-					0.08 ± 0.27	15.49	**0.004**	0.007

Results are given from LMM and Poisson GLMM for ectophagous and endophagous herbivores respectively.
[a]*df* degrees of freedom (numerator, denominator).
[b]Log-likelihood R^2 were not estimated in case of significant H × GR and H × GD interactions.
doi:10.1371/journal.pone.0044247.t001

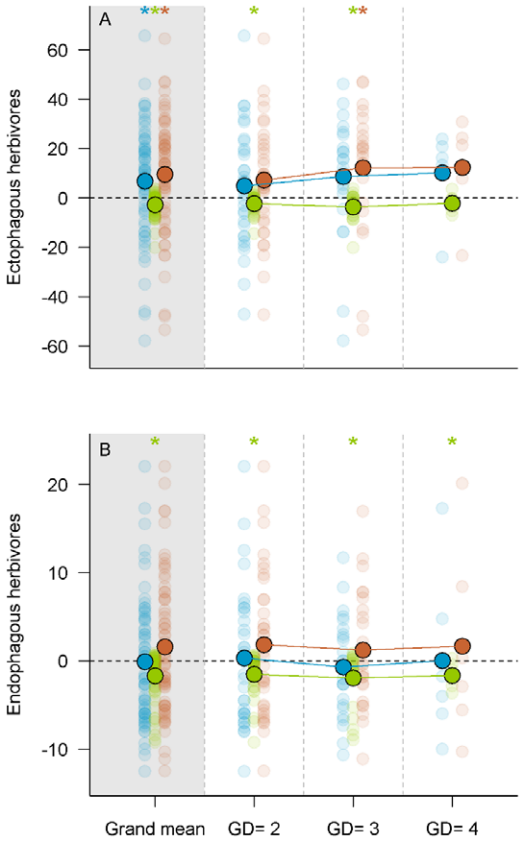


Figure 3. Non-additive effect of genetic diversity insect herbivores. (A) Test of the non-additive effect of genetic diversity on ectophagous insects. (B) Test of the non-additive effect of genetic diversity on endophagous insects. Semi transparent circles represent individual values per plot for net genetic diversity effect (NGDE, blue), complementarity effect (CE, red) and selection effect (SE, green). Solid black circled dots are the averaged values for all plots (grand mean) and each level of genetic diversity (GD). The '*' symbol are for means value significantly different from zero.
doi:10.1371/journal.pone.0044247.g003

Response of ectophagous herbivores

In the present study, we observed that the four oak families displayed different levels of resistance to generalist insect herbivores but also differed significantly in sapling height. Herbivory by generalist insects increased with sapling height. However there was no significant interaction between the effect of sapling height and genetic identity on damage by ectophagous herbivores. Moreover the genetic diversity and relatedness had only weak effects on sapling height whereas they significantly affected ectophagous insect herbivory. So the observed increase in damage caused by these herbivores in genetically diverse oak sapling mixtures was not mediated by differences in height. Yet, four ecological mechanisms may account for the observed relationship between genetic diversity and herbivory by ectophagous insects.

(i) Herbivore abundance. Herbivore abundance has been reported to increase with genetic diversity [25]. Associational susceptibility may then have been driven by an increase in abundance of generalist herbivores in genotype mixtures. However, as we did not sampled insects, we cannot validate this hypothesis. In addition, the relationship between herbivore density and herbivory damage remains unclear and we are not aware of any study that convincingly demonstrated an increase in herbivory with the abundance of herbivores.

Table 2. Summary of *t* values from *t*-tests for net genetic diversity effect (NGDE), complementarity effect (CE) and selection effect (SE) on damage of ectophagous and abundance of endophagous (leaf-miners) insects, and on sapling height, for all mixtures (grand mean) and for each level of genetic diversity (GD).

		Grand mean	GD = 2	GD = 3	GD = 4
	df	65	35	23	5
Ectophagous insects	NGDE	**2.44***	1.30	1.79	1.43
	CE	**3.37****	1.95	**2.39***	1.57
	SE	**−5.51*****	**−4.07*****	**−3.51****	−1.38
Endophagous insects	NGDE	−0.08	0.28	−0.61	0.01
	CE	1.70	1.40	0.89	0.37
	SE	**−5.46*****	**−3.96*****	**−3.16****	**−2.69***
Sapling height	NGDE	0.57	1.01	−1.12	2.76 *
	CE	1.05	1.34	−0.77	2.90 *
	SE	**−4.31*****	**−3.51****	**−2.46***	−0.91

Significant t-values are in bold: (***) P-value <0.001, (**) 0.001< P-value <0.01, (*) 0.01< P-value <0.05.
doi:10.1371/journal.pone.0044247.t002

(ii) Mixing diet. Herbivory by ectophagous insects increased in genotype mixtures because of a higher consumption of the four oak families, regardless their intrinsic susceptibility in monocultures, as evidenced by a significant and positive complementarity effect. This is consistent with the observation that generalist insect herbivores can increase their fitness by feeding on different host plants [1,2]. Different plant genotypes may provide insects with feeding resources of different qualities [41]. Mixtures of genotypes are therefore likely to improve diet mixing, which is known to benefit generalist herbivores [1,33,34]. It has been also proposed that feeding on different host plants results in the dilution of toxic compounds present in the plant tissues, allowing a more balanced input of nutrients [31,32]. It should be of great interest now to investigate leaf chemistry and check whether the blends of secondary metabolites involved in plant defence can explain herbivory patterns in genotypes mixtures and possibly changes with the genetic diversity of mixtures.

(iii) Spill over. The higher damage by herbivorous insects in plant species mixtures (*i.e.* associational susceptibility) has been initially attributed to a spill over of generalist herbivores from their preferred host plants to nearby suitable but less suitable host plants [9]. Despite the fact we did not monitor the temporal dynamic of ectophagous insects on individual oak saplings, the negative selection effect we report, though low, may account for such a spill over. Indeed, a negative selection reveals the existence of negative covariance between observed damage in mixtures and observed damage in monocultures: the increase in damage with genetic diversity was higher for families that suffered less damage when growing in monocultures. This is consistent with the hypothesis of greater colonisation through contagion, with the transfer of insects from more to less palatable families, in sapling mixtures. A similar behaviour was recently reported by Utsumi *et al.* [35] who observed a shift of insect herbivores from more to less preferred host genotypes in mixtures of annual plants.

(iv) Host location. The way insect herbivores perceive their host plants may change with their genetic diversity as recently suggested by Crawford *et al.* [23] who showed a non additive increase in gall abundance on patches of *Solidago altissima* with higher genetic diversity. At the plot scale, associational susceptibility may be explained by a better patch detection by foraging herbivores. For example, the mixture of sapling genotypes may have increased the probability of incorporating tall saplings that could be easier to detect and colonise, as suggested by our observation of a significant effect of sapling height on insect damage. Consistent with this hypothesis, the difference in height between taller (75^{th} percentile of heights distribution) and medium-sized (median of heights distribution) saplings tended to increase with genetic diversity at the plot level (Figure S4). In addition to visual cues that shape plant "physical" apparency, host plant location by insect herbivores is most often mediated by olfactory cues [73]. Host-plant recognition depends on ratios of plant volatiles and not just on detection of the presence or absence of particular compounds [74]. Insects use blends of volatile compounds to distinguish between host and non-host plant species. It has recently been suggested that there is redundancy in the composition of host odour blends, with some components being substitutable to others [75]. It is therefore possible that a mix of host plant genotypes is more likely to produce the right combination of attractants than a monoculture of a single plant genotype [76].

However, as we did not sample insect herbivores, it is difficult to determine which one of these mechanisms is the more likely or if they operate synergistically. For example, for a single herbivore, diet mixing might actually lead to less herbivory if that individual is able to acquire more nutrients with a variety of host genotype consumed. But on the other hand, if mixed diets are more preferable, and if mixed diets are associated with mixed host finding cues, it might attract more individuals and lead to greater overall herbivore damage. As a result, the abundance of herbivores may have ultimately been the primary driver of associational susceptibility.

Response of endophagous herbivores

None of the genetic attributes (identity, diversity or relatedness) had a significant effect on the abundance of specialist herbivores (leaf-miners). This finding is consistent with previous studies showing that genotype [39,40] and genetic diversity [39] are poor predictors of the diversity of specialist herbivores (leaf-miners and gall-makers) feeding on oaks. Instead, sapling height emerged as the key determinant of leaf miner abundance. Consistently, the maternal lineages that produced the tallest saplings (MT2 and MT1) were also more infested by leaf miners (although not significantly) than those in which saplings were significantly smaller suggesting that genetically based differences in sapling height may drive differences in the abundance of leaf miners. The negative selection effect on abundance of leaf-miners indicates that the oak families that tended to be less infested in monocultures also tended to be more often colonised by leaf miners in mixtures than in monocultures. As the larval stages of leaf miners cannot relocate after oviposition and cannot shift from one host plant to the next in order to find new (spillover) or complementary (mixed diet) feeding resources, the distribution of leaf miners between and within plots thus reflects the choice of oviposition site by females. As proposed for ectophagous herbivores, a possible explanation of the negative selection effect is that mixing genotypes resulted in a greater probability of including taller and then more attractive saplings to endophagous insects (Figure S4). One cannot exclude that the combination of relevant attractants was also more likely to occur in more diverse genotypes mixtures.

The mean abundance and species richness of leaf miners were very low in our experiment, with abundance scores of 0 to 10 mines in 20 leaves per sapling and 85% of total abundance represented by only two of the nine observed species (*Phyllonorycter* sp. and *Stigmella* sp.). Separate analyses of each species of leaf miner would have generated too many zero counts, so we decided to pool data into a single category of "insect specialists". However, Tack and Roslin [39] showed that *Phyllonorycter* sp. and *Stigmella* sp. responded differently to genetic and environmental treatments. Considering the abundance of several leaf-miner species together may therefore have prevented the detection of species-specific abundance patterns. Further investigation, after the oak saplings have been colonised by a larger number of insect species, are required for more detailed comparisons of the responses of generalist and specialist herbivores to genetic diversity.

Conclusion

Unlike plant species richness, plant genetic diversity may not provide sufficient functional contrast to prevent host colonization by specialist herbivores while enhancing the foraging of generalist herbivores. Overall, we observed a significant effect of tree genetic diversity on generalist herbivores but not on specialists. Increasing genetic diversity resulted in higher damage by generalist herbivores because of (i) a general increase in leaf consumption in more diverse genotype mixtures (*i.e.* positive complementarity effect) and (ii) an increased damage exposure of individuals from more resistant genotypes in the vicinity of individuals from more susceptible genotypes (*i.e.* negative selection effect). To date many studies have shown that herbivore diversity increases with host plant genetic diversity [22,23,24,25,26]. There is a need now to reconcile the two approaches and investigate the relationship between the diversity of herbivores and the resulting herbivory along gradients of plant genetic diversity. In addition, when saplings develop into young trees, they may be more easily located and infested by more specialized herbivores, therefore benefiting from being part of a mixed-genotype community [66,67]. We would therefore advocate long-term monitoring of the dynamics of sapling colonisation by insects with various degrees of host plant range limitation, to determine whether the magnitude and direction of the effect of genetic diversity on associational herbivory change with the ontogeny of focal tree species.

Supporting Information

Figure S1 Experimental design. Each colored square represents an individual oak sapling.
(DOCX)

Figure S2 Effect of the number of half-sib families per plot (Genetic Diversity, GD) on the mean genetic relatedness among oak seedlings within plots (Genetic relatedness, GR). Open circles represent individual plots ($n = 90$); filed circles represent mean genetic relatedness per level of genetic diversity.
(DOCX)

Figure S3 Negative selection effect of genetic diversity on exophagous herbivores.
(DOCX)

Figure S4 Effects of genetic diversity on oak height heterogeneity within plots. Each dot represents the mean difference ($\pm$ SE) between the 75^{th} percentile and the median (50^{th}

percentile) of sapling heights distribution within plots. This difference indicates how far taller trees were apparent to herbivores within plots.
(DOCX)

Table S1 Summary of genetic data describing the genetic structure of the population used to constructed experimental plots.
(DOCX)

Acknowledgments

We thank the INRA experimental units at Pierroton for their assistance in the establishment of the common garden experiments. We are also grateful to Audrey Lefrançois and Rodolphe Lauverjon for their help in the field. Genotyping was carried out at the genome/transcriptome facility at Pierroton.

Author Contributions

Conceived and designed the experiments: HJ BC. Performed the experiments: BC LL. Analyzed the data: BC LL BG AK HJ. Contributed reagents/materials/analysis tools: BC LL BG. Wrote the paper: BC HJ. Obtained fundings: HJ AK.

References

1. Unsicker S, Oswald A, Köhler G, Weisser W (2008) Complementarity effects through dietary mixing enhance the performance of a generalist insect herbivore. Oecologia 156: 313–324.
2. Bertheau C, Brockerhoff EG, Roux-Morabito G, Lieutier F, Jactel H (2010) Novel insect-tree associations resulting from accidental and intentional biological 'invasions': a meta-analysis of effects on insect fitness. Ecology Letters 13: 506–515.
3. Hooper DU, Chapin FS, Ewel JJ, Hector A, Inchausti P, et al. (2005) Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. Ecological Monographs 75: 3–35.
4. Cardinale BJ, Matulich KL, Hooper DU, Byrnes JE, Duffy E, et al. (2011) The functional role of producer diversity in ecosystems. American Journal of Botany 98: 572–592.
5. Andow DA (1991) Vegetational Diversity and Arthropod Population Response. Annual Review of Entomology 36: 561–586.
6. Balvanera P, Pfisterer AB, Buchmann N, He JS, Nakashizuka T, et al. (2006) Quantifying the evidence for biodiversity effects on ecosystem functioning and services. Ecology Letters 9: 1146–1156.
7. Jactel H, Brockerhoff EG (2007) Tree diversity reduces herbivory by forest insects. Ecology Letters 10: 835–848.
8. Barbosa P, Hines J, Kaplan I, Martinson H, Szczepaniec A, et al. (2009) Associational Resistance and Associational Susceptibility: Having Right or Wrong Neighbors. Annual Review of Ecology Evolution and Systematics 40: 1–20.
9. White JA, Whitham TG (2000) Associational susceptibility of cottonwood to a box elder herbivore. Ecology 81: 1795–1803.
10. Root RB (1973) Organization of a Plant-Arthropod Association in Simple and Diverse Habitats: The Fauna of Collards (*Brassica Oleracea*). Ecological Monographs 43: 95–124.
11. Finch S, Collier RH (2000) Host-plant selection by insects: a theory based on 'appropriate/inappropriate landings' by pest insects of cruciferous plants. Entomologia Experimentalis Et Applicata 96: 91–102.
12. Floater GJ, Zalucki MP (2000) Habitat structure and egg distributions in the processionary caterpillar *Ochrogaster lunifer*: lessons for conservation and pest management. Journal of Applied Ecology 37: 87–99.
13. Jactel H, Petit J, Desprez-Loustau M-L, Delzon S, Piou D, et al. (2012) Drought effects on damage by forest insects and pathogens: a meta-analysis. Global Change Biology 18: 267–276.
14. Hambäck PA, Ågren J, Ericson L (2000) Associational resistance: insect damage to purple loosestrife reduced in thickets of sweet gale. Ecology 81: 1784–1794.
15. Jactel H, Menassieu P, Vetillard F, Gaulier A, Samalens JC, et al. (2006) Tree species diversity reduces the invasibility of maritime pine stands by the bast scale, *Matsucoccus feytaudi* (Homoptera: Margarodidae). Canadian Journal of Forest Research 36: 314–323.
16. Riihimaki J, Kaitaniemi P, Koricheva J, Vehvilainen H (2005) Testing the enemies hypothesis in forest stands: the important role of tree species composition. Oecologia 142: 90–97.
17. Castagneyrol B, Jactel H (2012) Unravelling plant-animals diversity relationships: a meta-regression analysis. Ecology In press.
18. Vehvilainen H, Koricheva J, Ruohomaki K (2007) Tree species diversity influences herbivore abundance and damage: meta-analysis of long-term forest experiments. Oecologia 152: 287–298.
19. Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, et al. (2006) A framework for community and ecosystem genetics: from genes to ecosystems. Nature Reviews Genetics 7: 510–523.
20. Wimp GM, Wooley S, Bangert RK, Young WP, Martinsen GD, et al. (2007) Plant genetics predicts intra-annual variation in phytochemistry and arthropod community structure. Molecular Ecology 16: 5057–5069.
21. Bangert RK, Allan GJ, Turek RJ, Wimp GM, Meneses N, et al. (2006) From genes to geography: a genetic similarity rule for arthropod community structure at multiple geographic scales. Molecular Ecology 15: 4215–4228.
22. Crutsinger GM, Reynolds WN, Classen AT, Sanders NJ (2008) Disparate effects of plant genotypic diversity on foliage and litter arthropod communities. Oecologia 158: 65–75.
23. Crawford KM, Crutsinger GM, Sanders NJ (2007) Host-plant genotypic diversity mediates the distribution of an ecosystem engineer. Ecology 88: 2114–2120.
24. Johnson MTJ, Lajeunesse MJ, Agrawal AA (2006) Additive and interactive effects of plant genotypic diversity on arthropod communities and plant fitness. Ecology Letters 9: 24–34.
25. Cook-Patton SC, McArt SH, Parachnowitsch AL, Thaler JS, Agrawal AA (2011) A direct comparison of the consequences of plant genotypic and species diversity on communities and ecosystem function. Ecology 92: 915–923.
26. Crutsinger GM, Collins MD, Fordyce JA, Gompert Z, Nice CC, et al. (2006) Plant genotypic diversity predicts community structure and governs an ecosystem process. Science 313: 966–968.
27. Madritch MD, Hunter MD (2002) Phenotypic diversity influences ecosystem functioning in an oak sandhills community. Ecology 83: 2084–2090.
28. Booth RE, Grime JP (2003) Effects of genetic impoverishment on plant community diversity. Journal of Ecology 91: 721–730.
29. Gamfeldt L, Kallstrom B (2007) Increasing intraspecific diversity increases predictability in population survival in the face of perturbations. Oikos 116: 700–705.
30. Reusch TBH, Ehlers A, Hammerli A, Worm B (2005) Ecosystem recovery after climatic extremes enhanced by genotypic diversity. Proceedings of the National Academy of Sciences of the United States of America 102: 2826–2831.
31. Karban R, Karban C, Huntzinger M, Pearse I, Crutsinger G (2010) Diet mixing enhances the performance of a generalist caterpillar, *Platyprepia virginalis*. Ecological Entomology 35: 92–99.
32. Bernays EA, Bright KL, Gonzalez N, Angel J (1994) Dietary Mixing in a Generalist Herbivore: Tests of Two Hypotheses. Ecology 75: 1997–2006.
33. Kotowska AM, Cahill JF Jr., Keddie BA (2010) Plant genetic diversity yields increased plant productivity and herbivore performance. Journal of Ecology 98: 237–245.
34. Mody K, Unsicker SB, Linsenmair KE (2007) Fitness related diet-mixing by intraspecific host-plant-switching of specialist insect herbivores. Ecology 88: 1012–1020.
35. Utsumi S, Ando Y, Craig TP, Ohgushi T (2011) Plant genotypic diversity increases population size of a herbivorous insect. Proceedings of the Royal Society B: Biological Sciences: Pages?
36. Genung M, Crutsinger G, Bailey J, Schweitzer J, Sanders N (2012) Aphid and ladybird beetle abundance depend on the interaction of spatial effects and genotypic diversity. Oecologia: Springer Berlin/Heidelberg. 167–174.
37. Kaplan I, Denno RF (2007) Interspecific interactions in phytophagous insects revisited: a quantitative assessment of competition theory. Ecology Letters 10: 977–994.
38. Ali JG, Agrawal AA (2012) Specialist versus generalist insect herbivores and plant defense. Trends in plant science *In press*.
39. Tack AJM, Roslin T (2011) The relative importance of host-plant genetic diversity in structuring the associated herbivore community. Ecology 92: 1594–1604.
40. Tack AJM, Ovaskainen O, Pulkkinen P, Roslin T (2010) Spatial location dominates over host plant genotype in structuring an herbivore community. Ecology 91: 2660–2672.
41. Madritch MD, Hunter MD (2005) Phenotypic variation in oak litter influences short- and long-term nutrient cycling through litter chemistry. Soil Biology and Biochemistry 37: 319–327.
42. Dungey HS, Potts BM, Whitham TG, Li HF (2000) Plant genetics affects arthropod community richness and composition: Evidence from a synthetic eucalypt hybrid population. Evolution 54: 1938–1946.
43. Hochwender CG, Fritz RS (2004) Plant genetic differences influence herbivore community structure: evidence from a hybrid willow system. Oecologia 138: 547–557.
44. Barbour RC, O'Reilly-Wapstra JM, De Little DW, Jordan GJ, Steane DA, et al. (2009) A geographic mosaic of genetic variation within a foundation tree species and its community-level consequences. Ecology 90: 1762–1772.

45. Johnson MTJ, Agrawal AA (2005) Plant genotype and environment interact to shape a diverse arthropod community on evening primrose (*Oenothera biennis*). Ecology 86: 874–885.
46. Tack AJM, Johnson MTJ, Roslin T (2012) Sizing up community genetics: it's a matter of scale. Oikos 121: 481–488.
47. Giffard B, Corcket E, Barbaro L, Jactel H (2012) Bird predation enhances tree seedling resistance to insect herbivores in contrasting forest habitats. Oecologia 168: 415–424.
48. Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp.). Molecular Ecology Resources 11: 578–585.
49. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, et al. (2011) Current trends in microsatellite genotyping. Molecular Ecology Resources 11: 591–611.
50. Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Molecular Ecology Resources 11: 141–145.
51. Milligan BG (2003) Maximum-Likelihood Estimation of Relatedness. Genetics 163: 1153–1167.
52. Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Methods in Ecology and Evolution 1: 103–113.
53. Kramer M. R2 Statistics for mixed models (2005) Kansas state university, Manhattan.
54. Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Development Core Team (2011) nlme: Linear and Nonlinear Mixed Effects Models, R package version 3.1–102.
55. R Development Core Team (2010) R: A language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria.
56. Warton DI, Hui FKC (2010) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92: 3–10.
57. Bates D, Maechler M, Bolker B (2011) lme4: Linear mixed-effects models using S4 classes. R package version 0.999375–42 ed.
58. Loreau M, Hector A (2001) Partitioning selection and complementarity in biodiversity experiments. Nature 412: 72–76.
59. Hughes AR, Inouye BD, Johnson MTJ, Underwood N, Vellend M (2008) Ecological consequences of genetic diversity. Ecology Letters 11: 609–623.
60. Hersch-Green EI, Turley NE, Johnson MTJ (2011) Community genetics: what have we accomplished and where should we be going? Philosophical Transactions of the Royal Society B: Biological Sciences 366: 1453–1460.
61. Tikkanen O-P, Rousi M, Ylioja T, Roininen H (2003) No negative correlation between growth and resistance to multiple herbivory in a deciduous tree, *Betula pendula*. Forest Ecology and Management 177: 587–592.
62. Smith DS, Bailey JK, Shuster SM, Whitham TG (2011) A geographic mosaic of trophic interactions and selection: trees, aphids and birds. Journal of Evolutionary Biology 24: 422–429.
63. Silfver T, Roininen H, Oksanen E, Rousi M (2009) Genetic and environmental determinants of silver birch growth and herbivore resistance. Forest Ecology and Management 257: 2145–2149.
64. Ito M, Ozaki K (2005) Response of a gall wasp community to genetic variation in the host plant *Quercus crispula*: a test using half-sib families. Acta Oecologica 27: 17–24.
65. Hajjar R, Jarvis DI, Gemmill-Herren B (2008) The utility of crop genetic diversity in maintaining ecosystem services. Agriculture, Ecosystems & Environment 123: 261–270.
66. Peacock L, Hunter T, Turner H, Brain P (2001) Does host genotype diversity affect the distribution of insect and disease damage in willow cropping systems? Journal of Applied Ecology 38: 1070–1081.
67. Peacock L, Herrick S (2000) Responses of the willow beetle *Phratora vulgatissima* to genetically and spatially diverse *Salix spp*. plantations. Journal of Applied Ecology 37: 821–831.
68. Letourneau DK, Armbrecht I, Rivera BS, Lerma JM, Carmona EJ, et al. (2011) Does plant diversity benefit agroecosystems? A synthetic review. Ecological Applications 21: 9–21.
69. Tonhasca A, Byrne DN (1994) The effects of crop diversification on herbivorous insects: a meta-analysis approach. Ecological Entomology 19: 239–244.
70. Schuldt A, Baruffol M, Böhnke M, Bruelheide H, Härdtle W, et al. (2010) Tree diversity promotes insect herbivory in subtropical forests of south-east China. Journal of Ecology 98: 917–926.
71. Sobek S, Scherber C, Steffan-Dewenter I, Tscharntke T (2009) Sapling herbivory, invertebrate herbivores and predators across a natural tree diversity gradient in Germany's largest connected deciduous forest. Oecologia 160: 279–288.
72. Otway SJ, Hector A, Lawton JH (2005) Resource dilution effects on specialist insect herbivores in a grassland biodiversity experiment. Journal of Animal Ecology 74: 234–240.
73. Visser JH (1986) Host Odor Perception in Phytophagous Insects. Annual Review of Entomology 31: 121–144.
74. Bruce TJA, Wadhams LJ, Woodcock CM (2005) Insect host location: a volatile situation. Trends in Plant Science 10: 269–274.
75. Bruce TJA, Pickett JA (2011) Perception of plant volatile blends by herbivorous insects: Finding the right mix. Phytochemistry 72: 1605–1611.
76. Glinwood R, Ahmed E, Qvarfordt E, Ninkovic V, Pettersson J (2009) Airborne interactions between undamaged plants of different cultivars affect insect herbivores and natural enemies. Arthropod-Plant Interactions 3: 215–224.

http://lapinobservateur.over-blog.com