Diversification dans le genre Malus

Amandine Cornille

UNIVERSITE PARIS-SUD – UFR des Sciences

Thèse

Pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITE DE PARIS SUD

Par

Amandine CORNILLE

Diversification dans le genre *Malus*

soutenue le 26 octobre 2012, devant le jury d'examen:

Tatiana GIRAUD	Directrice de recherche CNRS, Laboratoire ESE, Orsay, France – Directrice de thèse
Xavier VEKEMANS	Professeur à l'Université des Sciences et Technologies de Lille 1, Lille, France – Rapporteur
Bruno FADY	Directeur de recherche, INRA, Avignon, France - Rapporteur
Sylvain GLEMIN	Chargé de recherche CNRS, ISEM, Montpellier, France - Examinateur
Martin LASCOUX	Professeur à l'Université d'Uppsala, Uppsala, Suède – Président de jury
Christine DILLMAN	Professeur à l'Université de Paris Sud, Paris, France - Examinatrice

Remerciements

La première pensée qui me vient à l'esprit quand je pense aux remerciements est évidemment tournée vers Tatiana et Pierre, mes deux « super-encadrants ». Je n'aurai pas pu rêver meilleurs directeurs de thèse que ce soit d'un point de vue scientifique qu'humain! Je réalise la chance que j'ai de vous avoir eu en encadrants de thèse! Vous avez été formidables! Merci pour ce super sujet de thèse qui a été pour moi une expérience extrêmement stimulante me laissant une liberté et une autonomie scientifique auxquelles je n'aurai jamais pu rêver avant de commencer! Je garderai de super souvenirs de ces trois années!

Tatiana, merci pour ton super encadrement, ta disponibilité quelque soit l'heure, le jour, le moment (pour une thésarde dont la principale difficulté est de gérer son temps ;)). Merci pour ton efficacité impressionnante au travail qui est extrêmement stimulante! Merci pour toutes ces discussions qui m'ont permises de m'éclaircir les idées et les doutes qui me venaient tout au long de ces trois années! Et surtout merci aussi pour ta gentillesse et ton soutien au long de ces trois années! J'espère continuer à travailler à toi, car c'est un VRAI plaisir!!!!!

Pierre, mon premier encadrant avec qui j'ai vécu une « relation à distance », qui s'est passée à merveille!!! Que ce soit sur Orsay ou à Berkeley, tu as toujours été là, à n'importe quel moment de ma thèse, toujours réactif, pour m'aider, pour me soutenir, pour m'encourager. Je me rappellerai nos longues discussions dans le bureau à Orsay (scientifiques ou pas d'ailleurs ;))! J'espère aussi encore longtemps échanger et travailler avec toi!!!!

Pour tout ça, mille mercis à vous deux!!!!

Je tenais à remercier les membres de mon jury d'avoir accepté d'évaluer mon travail: Xavier Vekemans, Bruno Fady, Martin Lascoux, Sylvain Glémin et Christine Dillmann.

Je remercie aussi les différents membres de mes comités de thèse qui m'ont guidés dans mes réflexions: Thierry Robert, Domenica Manicacci, Rémy Petit, Francois Laurens, Jerôme Enjalbert et Maud Tenaillon. En particulier, Maud, je souhaitais te remercier pour les discussions très intéressantes que nous avons sur la domestication ainsi que ton aide sur les scripts Perl! François, je souhaitais te remercier pour nos échanges sur les pommes et les sorties grand public que nous avons faites!

Je souhaite également remercier les collaborateurs avec qui j'ai travaillé et qui m'ont aidés durant ces trois années dans la rédaction des articles: Isabel Roldan Ruiz, Bruno Le Cam, Marina Olonova, Laurence Feugey, Céline Bellard. Et en particulier René Smulders et nos échanges de mails sur des multiples problématiques

Je tenais aussi à remercier Eric Collin pour m'avoir fait découvrir le monde forestier et tous les aspects appliqués de mon travail, de m'avoir permis d'échanger directement avec des

acteurs de la conservation des ressources génétiques. Merci pour ta joie de vivre, ta sympathie et ta bonne humeur! J'espère que nous allons continuer à travailler ensemble encore longtemps!

Je tenais aussi à remercier Nathalie Frascaria-Lacoste pour ta gentillesse, nos discussions et tes précieux conseils pour me guider dans les échanges avec les acteurs de la conservation forestière.

Aurélien Tellier, je te dis merci pour tes conseils et ton aide pour les analyses ABC, avec des conversations au téléphone qui ont parfois durées plusieurs heures! Sans toi, je n'aurai jamais pu aboutir à ces résultats, merci pour ta patience.

Je tenais à remercier Gwendal Restoux, pour tous tes conseils concernant mes analyses mais aussi discussions durant ces trois années. Mon cher collègue de bureau, qui, par dépit, supportait mes questions incessantes concernant les analyses bayésiennes. Merci pour tes relectures de fin de course aussi!

Je tenais aussi à remercier Jacqui Shykoff pour ta présence durant ces trois années de thèse, tes commentaires, tes rappels à moi, tête de linotte, qui oublie souvent les papiers administratifs, et pour ta disponibilité dans les moments de rush de signatures de l'ED!

Je voulais également remercier mes encadrants de Master 1 et 2 qui m'ont permis de découvrir la recherche version « fig-fig wasp interaction » et de me découvrir une passion : Magali Proffit, Finn Kjellberg et Martine Hossaert-McKey.

Je tiens à remercier les financeurs du projet: la Région Ile de France (PICRI), l'IDEEV, la SBF (Société Botanique de France).

Passons au labo. Et oui trois années passées à l'ESE, il y en a à raconter! Des vertes et des pas mûres comme diraient certain! La « dream team »: JT T (avec le T anglais svp), PtiLU (j'ai hâte de voir ta tête quand tu verras que je t'appelle comme ça dans mes remerciement), Alo pour nos conversations longues et rythmées dans notre bureau où tout le monde nous entend rigoler à l'autre bout du couloir, Gwendal, mon super voisin de bureau et d'appartement, qui a toujours été là, quelque soit l'occasion, Juju pour nos super moments passés entre autre avec Mr Hugo, Alex pour nos partages trip piscine en fin de thèse, ca détend. Je tenais aussi à remercier mes compagnons de route: Charles, Jonathan, Céline, Hervé et tous les autres membres de l'ESE que je n'ai pas cité avant!!!!! Merci!

Je tenais à remercier tous mes amis qui m'ont soutenus durant ces trois ans et même avant: les Lillois, les Montpellierains, les Bretons et les Parisiens.

Je voulais dire un ENORMISSIME merci à mes parents qui ont été là et m'ont soutenue durant ces trois années mais aussi depuis que je suis toute jeune, ainsi que ma sœur Lolo.

Enfin, je tiens à remercier mon coach personnel, Youri, qui est, certes arrivé en fin de parcours, mais qui a été un soutien inqualifiable pour cette dernière ligne droite!! Merci merci merci!

Sommaire

Introduction générale

Introduction générale

Comprendre les mécanismes à l'origine de la diversité du vivant est un enjeu majeur de la biologie, avec des implications aussi bien fondamentales qu'appliquées. Le Néo-darwinisme au début du XXème siècle a permis des avancées considérables dans la compréhension de ces mécanismes grâce à l'émergence d'une nouvelle discipline : la génétique des populations. La génétique des populations vise à comprendre et expliquer les processus évolutifs (la dérive génétique, la migration, la mutation et la sélection) qui maintiennent ou génèrent le polymorphisme observé au sein et entre populations (Wakeley, 2008). Depuis une vingtaine d'années, le développement des techniques de biologie moléculaire en parallèle des avancées conceptuelles en génétique des populations ont permis l'acquisition de jeux de données considérables de marqueurs génétiques (séquences, SNP, microsatellites) et l'inférence de paramètres démographiques et historiques issus de modèles simulant l'évolution des populations. Les marqueurs neutres permettent de retracer les histoires démographiques ou généalogiques des espèces, sur toute une gamme d'échelles évolutives, ce qui a révélé des processus de diversification complexes et variés (Alves *et al.*, 2012; Caswell *et al.*, 2008; Delplancke *et al.*, 2011; François *et al.*, 2008; Gladieux *et al.*, 2010b; Joy *et al.*, 2003; Kliman *et al.*, 2000; Li, Stephan, 2006; Li *et al.*, 2011; Tenaillon *et al.*, 2004).

Malgré les avancées théoriques considérables ayant suivi l'avènement du néo-darwinisme et son application à de très nombreux modèles biologiques, les mécanismes à l'origine de la diversification des organismes vivants figurent encore à l'heure actuelle parmi les phénomènes biologiques les moins bien compris. Cela est peut être dû au fait que les processus impliqués se déroulent souvent à des échelles de temps très longs et impliquent différents mécanismes complexes agissant en synergie, dont l'étude pratique est rarement simple. Dans ce contexte, la domestication est un processus de diversification rapide et récent qui offre une opportunité unique de mieux comprendre les phénomènes de diversification. C'est d'ailleurs en s'intéressant à la domestication et à la variabilité intra-espèce des variétés domestiquées que Darwin, en 1859 dans son livre « *On the origin of species* », a mis en évidence un des mécanismes évolutifs à l'origine des espèces, et plus généralement de la diversification : la sélection naturelle. Ainsi, une bonne compréhension

des mécanismes à l'origine de la biodiversité passe par l'étude de processus à diverses échelles évolutives : domestication, diversification intraspécifique et diversification interspécifique (spéciation).

1. Processus de diversification « artificielle » : la domestication

L'étude de systèmes domestiqués a joué un rôle critique dans le développement et l'évaluation de la théorie de l'évolution (Glémin, Bataillon, 2009; Pickersgill, 2007; Purugganan, Fuller, 2009; Zohary, Hopf, 2000). La domestication est considérée comme le stade final d'un processus continu et dynamique qui commence par l'exploitation de taxons sauvages et se poursuit à travers la culture d'individus sélectionnés en environnement naturel, mais non encore différenciés des taxons sauvages. Ce processus se termine par la fixation, à travers la sélection humaine, d'un syndrome de domestication[1] et de différences génétiques distinguant taxons domestiqués et sauvages (Pickersgill, 2007). La connaissance des processus de domestication du stade initial au stade final est une opportunité unique pour comprendre les mécanismes à l'origine de nouvelles espèces, populations ou variétés sur de courtes échelles évolutives.

Il existe de nombreuses études sur la domestication chez les plantes et animaux, montrant des histoires variables de domestication, ce qui a ouvert des débats quant à la nature des processus de domestication. Les principales problématiques qui se posent dans la plupart des modèles biologiques étudiés concernent :

- la vitesse du processus de domestication (rapide ou diffus) ;
- l'origine géographique, le nombre et l'identité de(s) population(s) et espèce(s) contributrice(s) (multiples ou unique) ;
- le mode de diffusion de l'espèce domestiquée à partir de son centre d'origine ;
- l'existence d'un syndrome de domestication ainsi que le moment de son apparition ;
- la détection des gènes impliqués dans ce syndrome ;
- la perte de diversité génétique suite à la domestication (goulet d'étranglement[2]) ;

[1] Transitions morphologiques, phénologiques, physiologiques entre formes cultivées et sauvages associées à une forte sélection par l'homme de traits liés aux conditions de récolte, à la production des graines, ou à la compétition entre graines.

[2] Diminution de la diversité génétique chez le taxon cultivé relativement à son apparenté sauvage due à la sélection humaine et la dérive génétique lors du processus de domestication.

- le rôle des hybridations entre taxons domestiqués et sauvages ;
- l'impact des pratiques de culture sur l'évolution des espèces domestiquées.

Depuis une dizaine d'années, l'accumulation des données génétiques a permis à la fois de compléter les données archéologiques sur ces problématiques, de révéler des particularités du processus de domestication liées à la grande diversité des traits d'histoire de vie des espèces animales et végétales, et aussi de pointer du doigt certaines contradictions temporelles quant à la vitesse du processus de domestication, donnant naissance à un débat concernant cette question.

Vitesse des processus de domestication, le débat : rapide *versus* diffus. Jusque dans les années 2000, la domestication était perçue comme un processus rapide caractérisé par trois étapes principales accompagnant le réchauffement climatique du Pléistocène à l'Holocène : (*i*) un début de culture avec des taxons sauvages suivi d'une domestication rapide donnant naissance au taxon domestiqué, (*ii*) une croissance démographique rapide du taxon domestiqué, et enfin (*iii*) une explosion démographique à l'extérieur du centre d'origine (Allaby et al., 2008; Gross, Olsen, 2009). Dans l'un des lieux les plus étudiés par les biologistes et anthropologues, le Croissant Fertile, l'instantanéité de la domestication semblait se confirmer par l'apparition rapide de restes archéologiques de multiples taxons cultivés majeurs (céréales, légumes secs, lin…). Ce processus de domestication unique et rapide supporte l'idée que les cultures actuelles majeures ont été domestiquées au même moment dans une aire centrale (« *core area* » ou « *golden triangle* » (Fuller et al., 2012)) par le même groupe de fermiers : c'est l'hypothèse du « paquet Néolithique » faisant référence à un groupe de taxons présents dans cette aire centrale (engrain, orge, lentille, pois, pois chiche), dont certains sont actuellement cultivés au niveau mondial. Si cette hypothèse est vraie, les taxons domestiqués devraient former chacun un groupe monophylétique par rapport à leur apparenté sauvage. Inversement, si la domestication a été multiple et dispersée, les taxons domestiqués devraient apparaitre polyphylétiques, au sein des taxons sauvages. De nombreux travaux se sont donc intéressés à tester la monophylie *versus* polyphylie des taxons domestiqués. De nombreux taxons domestiqués du Croissant Fertile, comme le lin (*Linum usitatissimum*), l'orge (*Hordeum vulgare*), l'engrain (*Triticum monococcum*), l'amidonnier (*Triticum turgidum ssp. dicoccon*) et le blé dur (*Triticum durum*),

mais aussi d'Amérique du Sud comme le manioc (*Manihot esculensa*) et la pomme de terre (*Solanum tuberosum*), se sont ainsi révélés monophylétiques (Allaby et al., 2005; Badr et al., 2000; Heun et al., 1997; Olsen, Schaal, 2001; Özkan et al., 2002; Spooner et al., 2005), suggérant une origine unique et donc une invention rapide et localisée de l'agriculture. Cependant, alors que cette hypothèse restait raisonnable il y a encore une dizaine d'années, l'accumulation des données archéologiques a récemment montré que les trois stades du processus de domestication (*i*, *ii*, et *iii*) peuvent être diffus dans le temps. Par exemple, des restes de céréales sauvages ont été découverts dans le Croissant Fertile, datant d'il y a 23000 ans, donc 10000 ans avant la date supposée de la domestication rapide qui suit le début de culture des taxons sauvages (étape *i*) dans cette zone (Weiss et al., 2006). De la même manière, le stade final de domestication par fixation du syndrome de domestication s'est révélé être en réalité un processus très lent chez certains taxons domestiqués comme le blé par exemple (Tanno, Willcox, 2006). Ainsi les données archéologiques montrent la domestication comme une mosaïque de processus lents (Fuller et al., 2012): les trois étapes de domestication se distribueraient sur des échelles de temps bien plus longues que ce que l'on pensait jusqu'alors. Ce modèle est maintenant au centre des débats (Allaby *et al.*, 2008; Clotault *et al.*, 2011; Fuller *et al.*, 2012; Glémin, Bataillon, 2009; Heun *et al.*, 2012). Cette échelle de temps plus diffuse de transition implique que des flux de gènes ont pu avoir lieu entre populations de différentes localités. Ainsi, il est moins probable qu'un taxon domestiqué soit associé à un seul lieu géographique ou que les évènements de domestications multiples soient totalement indépendants, mais qu'ils auraient plutôt impliqué des échanges entre fermiers. D'un point de vue génétique, le modèle diffus de domestication ne va pas nécessairement à l'encontre des études démontrant que certains taxons domestiqués forment des groupes monophylétiques. En effet, la monophylie peut résulter de l'effacement d'évènements de domestication multiples par l'effet de forts et longs goulets d'étranglement sur plusieurs générations (Allaby et al., 2008; Olsen, Gross, 2008), même si cet artefact méthodologique semble exister seulement sous certains modèles de génétique des populations (Ross-Ibarra, Gaut, 2008). De la même manière, les données génétiques nous informent uniquement sur les taxons domestiqués actuels, mais il est possible que certaines lignées aient été perdues par dérive ou par abandon par les fermiers, n'apparaissant pas ainsi dans les études génétiques (Gross, Olsen, 2009). Datant

d'une dizaine d'années, le débat sur la rapidité des processus de domestication est toujours d'actualité (Clotault et al., 2011; Fuller et al., 2012; Gross, Olsen, 2009; Tanno, Willcox, 2006; Zeder et al., 2006). L'utilisation de méthodes, comme *l'approximate Bayesian computation*, permettant de ne plus se limiter aux interprétations trop simplistes basées sur des arbres, mais plutôt de tester des scénarios démographiques et historiques plus précis, a apporté des précisions sur la vitesse des processus de domestication (*e.g.* Clotault et al., 2011)) ainsi que le nombre de populations ayant contribué à l'histoire évolutive des espèces domestiquées (*e.g,.* Molina *et al.*, 2011).

Origine de la domestication et diffusion du taxon domestiqué: Où ? Quoi ? Combien ? Quand ? La connaissance de l'ancêtre sauvage et de sa localisation géographique offre une opportunité unique de comprendre les conséquences génétiques et phénotypiques du processus de domestication (Glémin, Bataillon, 2009).

Où : localisation géographique des évènements de domestication ? Le centre d'origine des taxons domestiqués est généralement identifié comme l'aire géographique présentant la plus grande diversité morphologique et génétique chez l'apparenté sauvage (Vavilov, 1926). Les études de génétique qui se sont intéressées à l'origine géographique des espèces cultivées se basent sur des marqueurs neutres permettant de comparer les taxons domestiqués et sauvages, ces derniers présentent des niveaux de variabilités génétiques et morphologiques plus élevés lorsque la domestication est accompagnée d'un goulet d'étranglement. Plusieurs centres de domestication ont été proposés chez les plantes et animaux (Diamond, 2002): le Croissant Fertile, la Chine, la Mésoamérique, les Andes, l'Amazonie, le Sahel, l'Afrique de l'Ouest tropicale et la Nouvelle-Guinée (Figure 1). Dans cette carte classique des centres de domestication, l'Asie Centrale ne figure pas en tant que centre d'origine géographique des taxons domestiqués alors qu'une majorité de fruitiers d'importance économique comme le citronnier, le pommier, le poirier, le cognassier, le néflier, l'amandier, l'abricotier, le cerisier, le pêcher et le prunier furent probablement domestiqués en Asie Centrale, puis ont ensuite rejoint l'Ouest durant l'Antiquité (Janick, 2005), même si aucune donnée génétique ne confirme ces hypothèses sauf rares exceptions

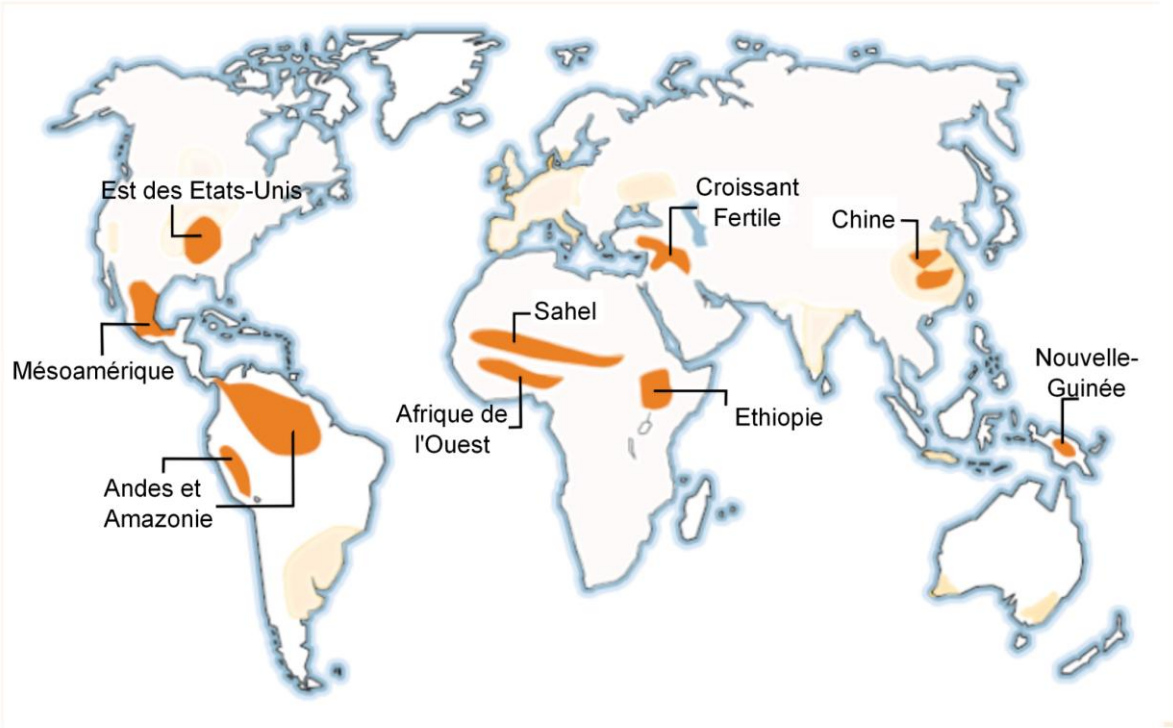


Figure 1. Centres d'origine de domestication des animaux et les plantes (Diamond, 2002)**.**

comme le pommier cultivé (*Malus domestica*) (Velasco *et al.*, 2010). Il est en fait très difficile de déterminer le stade initial de la domestication chez les espèces fruitières car cela exige une combinaison de données archéologiques et génétiques, qui sont pour l'instant peu abondantes et difficiles à acquérir. Jusqu'à très récemment, peu d'études génétiques ont abordé ces questions sur l'origine de la domestication des arbres fruitiers. D'autre part, chez les espèces fruitières, les précisions de datation et de localisation des origines de la domestication via les restes archéologiques sont difficiles : à la fois d'un point de vue anatomique, de par le manque de critères distinctifs entre taxons sauvages et domestiqués (Zohary, Hopf, 2000), et aussi par la conservation difficile des graines ou pépins au cours du temps. Ainsi les découvertes de restes archéologiques sont souvent circonstancielles et anecdotiques, par exemple par la découverte de restes de fruits dans une zone géographique où le taxon sauvage n'est pas présent (Zohary, Hopf, 2000). La localisation des centres de domestication au Proche Orient chez l'olivier et le palmier dattier est liée à l'apparition simultanée des premiers signes de culture et du développement de l'horticulture, en particulier de la technique de greffage, ainsi qu'à la présence d'apparentés morphologiques sauvages dans cette zone géographique (Zohary, Hopf, 2000). Cependant,

rien ne nous empêche de penser que d'autres espèces fruitières, dont les apparentés sauvages sont présents en Asie Centrale, aient pu être domestiquées plus précocement, avant le développement des techniques de greffage il y a 3800 ans.

Quoi et combien : identification des ancêtres sauvages et de leur nombre? Les origines des taxons domestiqués peuvent être uniques (un seul évènement de domestication à partir d'une espèce ou d'une population) ou multiples (plusieurs espèces ou plusieurs populations de la même espèce, domestiquées indépendamment).

La question de l'origine multiple ou unique des taxons domestiqués est encore débattue (Glémin, Bataillon, 2009). L'un des exemples les plus classiques de domestication considéré comme unique est celui du maïs (Matsuoka *et al.*, 2002) : une phylogénie basée sur 99 marqueurs microsatellites incluant la téosinte (*Zea mays ssp. parviglumis)* et le maïs (*Zea mays ssp. mays*) montre que l'ensemble des variétés de maïs cultivées forment un seul groupe monophylétique, dérivé de la téosinte. Cependant, l'inférence d'un évènement unique de domestication peut être un artefact lié aux méthodes basées sur l'étude des arbres phylogénétiques (Allaby *et al.*, 2008). Un exemple illustre particulièrement bien la difficulté de résolution de la question du nombre d'ancêtres à l'origine des taxons domestiqués. Le riz cultivé (*Oryza sativa*) présente deux phénotypes communément appelés *Oryza sativa indica* et *Oryza sativa japonica*. *Oryza sativa* fût ainsi supposée être issue de deux évènements indépendants de domestication à partir de la même population, qui aurait donné ces deux variétés majeures actuelles. Certaines données ont conforté ces hypothèses (Londo et al., 2006), alors que plus récemment d'autres études ont suggéré sur la base de nouveaux marqueurs une domestication unique (Gao, Innan, 2008; Molina *et al.*, 2011).

Le modèle de domestication à partir d'un ancêtre unique a longtemps été accepté par défaut, cependant, le débat reste encore ouvert. En effet, de nombreux taxons domestiqués semblent avoir des origines multiples à partir de populations ou espèces différentes. Les animaux, en particulier le bétail, sont de bons exemples (Bruford *et al.*, 2003). La vache serait issue de deux évènements de domestication indépendants, à partir de l'Auroch (*Bos primigenius primigenius*) en Europe et du zébu en Asie (*B. primigenius namadicus*), amenant l'apparition des deux espèces domestiques actuelles, *Bos taurus et Bos indicus* respectivement (Bruford et al., 2003; Loftus et al., 1994) ; le cheval domestique serait issu de plusieurs domestications indépendantes à partir de différentes populations de la

même espèce sauvage (Vilà et al., 2001; Warmuth et al., 2012) ; la chèvre (*Capra hircus)* serait issue de trois évènements indépendants de domestication en Asie et dans le Croissant Fertile à partir de trois sous-espèces de *C. aegagrus* (Luikart et al., 2001) ; le mouton (*Ovis aries*) serait issu de deux évènements indépendants de domestication, en Asie et au Proche Orient (Bruford *et al.*, 2003). Les plantes aussi présentent des histoires de domestication impliquant de multiples évènements de domestication indépendants. Le coton est un exemple extrême de domestication multiple, avec quatre espèces domestiquées (deux en Amérique, *Gossypium hirsutum* et *G. barbadense*, et deux en Afrique-Asie, *G. arboreum* et *G. herbaceum*) issues de quatre évènements indépendants de domestication à partir d'ancêtres sauvages géographiquement isolés (Wendel, Cronn, 2003). Il existe des exemples également chez les arbres fruitiers : chez l'avocatier (*Persea Americana*), il y aurait eu trois évènements de domestication indépendants en Amérique du Sud à partir d'écotypes locaux (populations sauvages de *Persea Americana*) présentant des distributions au travers du continent sud-américain (Chen *et al.*, 2009a) ; chez l'abricotier (*Prunus armeniaca*), il y aurait eu deux centres de domestication indépendants à partir de populations sauvages en Chine et dans le Proche-Orient (Bourguiba *et al.*, 2012) ; chez le fruitier *Spondias purpurea* de multiples centres de domestication à partir de populations sauvages de la même espèce ont aussi été découverts (Miller, Schaal, 2005). La littérature chez les arbres fruitiers reste cependant beaucoup moins abondante que pour les plantes domestiquées annuelles (céréales et légumes secs).

***Quand* : datation de l'évènement de domestication?** La date des évènements de domestication est difficile à déterminer. Les caractères morphologiques associés à la domestication n'étaient souvent pas encore fixés lors de la domestication (Larson *et al.*, 2012). Ceci a déjà été discuté précédemment dans le cadre de la datation de la domestication des arbres fruitiers, et la possibilité de suggérer l'Asie Centrale comme centre de domestication. Cependant, chez d'autres taxons domestiqués, des analyses morphométriques à partir de restes archéologiques permettent de dater approximativement ces évènements. Par exemple, les chiens seraient apparus il y a 12000 ans en Eurasie (Larson *et al.*, 2012) et la vigne il y a 5000 ans au Proche Orient (Myles *et al.*, 2011). Cependant, rares sont les études basées sur des marqueurs génétiques qui ont fournies des estimations en concordance avec les données archéologiques (Clotault *et al.*, 2011). La principale

explication à ces discordances peut être l'utilisation d'horloges moléculaires inappropriées dans les études de génétique (Zeder *et al.*, 2006).

Origine hybride et diversification post-domestication des taxons domestiqués. Les hybridations[3] jouent un rôle fondamental à la fois dans l'apparition des taxons domestiqués (Mallet, 2007) (domestication par hybridation) et aussi dans leur évolution après la domestication (Arnold, 2004) (flux de gènes post-domestication entre taxons sauvages et domestiqués)**.** Des exemples chez les animaux et les plantes suggèrent qu'il est important de prendre en compte ces hybridations lors de la reconstruction des scénarios de domestication (Arnold, 2004). La domestication par hybridation semble en effet avoir été à l'origine de certains taxons domestiqués. Les variétés de bananes diploïdes (*Musa acuminata*) sont issues d'une hybridation entre deux sous-espèces de *M. acuminata* isolées sur deux îles d'Asie du Sud-est et de Malaisie (Perrier *et al.*, 2011)**.** Le kiwi (*Actinidia deliciosa)* serait issu d'un évènement d'allopolyploïdisation (Atkinson *et al.*, 1997). Le blé tendre (*Triticum aestivum*) dérive d'un croisement entre l'amidonnier *T. dicoccum* et l'égilope *Aegilops tauschii*. Les hybridations post-domestication répétées semblent aussi fréquentes et, même si leurs conséquences d'un point de vue évolutif sont encore mal connues, dans certains cas, elles ont sûrement joué un rôle dans le développement de la grande diversité de races et variétés existantes (Bruford *et al.*, 2003). Chez le cochon et le mouton, des hybridations entre populations sauvages et cultivées ont été à la base de nouvelles races domestiquées (Arnold, 2004). Chez les maïs, des flux de gènes de *Zea mays* ssp. *mexicana,* une sous-espèce de téosinte des hauts plateaux mexicains, vers certaines variétés locales des hauts plateaux mexicains (Matsuoka *et al.*, 2002; van Heerwaarden *et al.*, 2011), ont été suggérés comme étant à l'origine de l'adaptation des variétés locales aux conditions extrêmes de survie et de maturation des hauts plateaux mexicains (Arnold, 2004). Chez la vigne (*V. v. vinifera)*, les cultivars d'Europe de l'Ouest, originaires des populations sauvages de *V. vinifera sylvestris* du Proche Orient, ont subi des introgressions par les populations sauvages de *V. vinifera sylvestris* d'Europe de l'Ouest lors de l'introduction de la vigne en Europe il y a 2800 ans.

[3] Processus par lequel se forme une descendance issue d'individus de populations ou espèces différentes qui sont distinguables sur la base d'un ou plusieurs caractères héritables.

Ainsi les flux de gènes des taxons sauvages vers les taxons domestiqués semblent avoir joué un rôle dans la diversification post-domestication de plusieurs espèces domestiquées.

Perte de diversité génétique durant la domestication (goulet d'étranglement) et conséquences des hybridations post-domestication entre apparentés sauvages et cultivés. La domestication peut avoir pour conséquence la diminution de la diversité génétique chez le taxon domestiqué relativement à son apparenté sauvage, du fait de la sélection par l'homme et de la dérive génétique (Gross, Olsen, 2009), un phénomène appelé goulet d'étranglement. Chez les plantes annuelles, des pertes considérables de diversité génétique ont été documentées (60-90%), comme chez le tournesol (*Helianthus annuus*) (Liu, Burke, 2006), le maïs (*Zea mays ssp mays*) (Tenaillon *et al.*, 2004), le soja (Hyten *et al.*, 2006) et le blé (Haudry *et al.*, 2007). Les barrières reproductives entre taxons sauvages et domestiqués sont souvent très faibles, voire inexistantes, de par leur divergence récente, impliquant ainsi des hybridations interspécifiques fréquentes (Ellstrand *et al.*, 1999). Elles ont pour conséquences l'effacement des goulets d'étranglement chez le taxon domestiqué et la baisse de l'apparentement génétique au(x) progéniteur(s) ancestral(ux) si les flux de gènes proviennent d'autres espèces sauvages que le(s) contributeur(s) initial(ux), générant des niveaux de diversité et structures génétiques très difficiles à interpréter. Ce type de situation se retrouve chez des animaux tels que les chiens (*Canis lupus familiaris*) (Larson *et al.*, 2012), le mouton d'Eurasie de l'Ouest (*Ovis orientalis aries)* (Chessa *et al.*, 2009), le cheval (Warmuth *et al.*, 2012) mais aussi chez des arbres fruitiers tels que la vigne cultivée (*Vitis vinifera vinifera)* (Myles *et al.*, 2011) ou l'amandier (*Prunus amygdalus*), (Delplancke *et al.*, 2011) ainsi que des céréales comme le maïs (Hufford *et al.*, 2012; van Heerwaarden *et al.*, 2011). Les flux de gènes du taxon domestiqué vers les apparentés sauvages se sont aussi révélés fréquents (Ellstrand *et al.*, 1999) mais leurs conséquences d'un point de vue évolutif sur les apparentés sauvages restent encore peu connus.

2. Processus de diversification naturelle : divergences interspécifiques et intraspécifiques avec flux de gènes

Diversifications interspécifiques et intraspécifiques : avec ou sans flux de gènes? La littérature scientifique s'intéressant aux processus de divergence entre espèces est très

riche (Network, 2011). Les évènements de spéciation ont été historiquement catégorisés en quatre grands modes selon la géographie : allopatrie, péripatrie, sympatrie et parapatrie (Coyne, Orr, 2004). Cependant, ces classifications basées sur la géographie ont récemment été remises en question (Fitzpatrick *et al.*, 2009; Hey, 2006). Les problématiques se focalisent maintenant sur l'importance des flux de gènes durant la spéciation (Mallet *et al.*, 2009), en particulier pour les plus contestés des modes de spéciation : sympatrique et parapatrique, impliquant des flux de gènes durant la divergence. La divergence entre espèces avec flux de gènes apparaît possible si la sélection est assez forte, en particulier dans des régions génomiques incluant des gènes impliqués dans le choix du partenaires et/ou dans les traits morphologiques impliqués dans la réponse à la sélection sexuelle (Feder *et al.*, 2012; Hey, 2006). Deux principales voies d'études se sont formées, la première concerne la détection de l'existence de flux de gènes au moment de la divergence, voire leurs estimations au travers de méthodes d'inférences démographiques entre espèces en divergence (Bertorelle *et al.*, 2010; Pinho, Hey, 2010). La deuxième concerne la détection d' « îles » ou de « continents » génomiques plus différenciés que le reste du génome entre espèces sœurs, ce qui témoignerait de l'existence de régions génomiques sous forte sélection divergente lors de la divergence entre populations en présence de flux de gènes (Feder *et al.*, 2012; Michel *et al.*, 2010). Nous nous intéresserons ici uniquement à la première partie, l'estimation du degré de flux de gènes lors de la divergence en l'absence d'isolement géographique, qui a fait l'objet d'une partie de ma thèse. De récentes études ont permis d'estimer la part de flux de gènes dans les processus de divergence interspécifiques par l'utilisation de marqueurs neutres et d'inférences Bayésiennes. De nombreux organismes ont été étudiés : des mammifères (Geraldes *et al.*, 2008), des amphibiens (Nadachowska, Babik, 2009; Nosil, 2008), des plantes (Li *et al.*, 2011; Städler *et al.*, 2005; Zheng, Ge, 2010; Zhou *et al.*, 2007), des oiseaux (Carling et al., 2010), des champignons (Gladieux *et al.*, 2010a) ou encore des insectes (Hey, Nielsen, 2004). Les études révèlent des niveaux variables de flux de gènes entre populations divergentes, souvent asymétriques, avec certains cas avérés de spéciation avec des flux de gènes, comme par exemple entre espèces/populations de papillons (Kronforst *et al.*, 2006; Salazar *et al.*, 2008), de salamandres (Nadachowska, Babik, 2009; Niemiller *et al.*, 2008), de riz (Zheng, Ge, 2010) et de tomates (Städler *et al.*, 2008). Même si l'estimation et la datation des flux gènes

restent encore un gros challenge, les nouvelles méthodes basées sur la coalescence et la comparaison de scénarios alternatifs par ABC (*approximate Bayesian computation*) sont prometteuses dans l'identification des nouveaux cas de spéciation avec flux de gènes (Smadja, Butlin, 2011).

La différenciation des populations au sein d'une espèce représente un processus de diversification à part entière, surtout lorsque l'adaptation locale entre en jeu, et est pourtant rarement prise en compte lorsqu'on parle de biodiversité. La différenciation des populations est la résultante des mêmes forces évolutives (mutation, dérive génétique, migration et sélection naturelle) que la spéciation (Coyne, Orr, 2004). Les niveaux de différenciation peuvent d'ailleurs être plus élevés entre populations qu'entre espèces (Lexer, Widmer, 2008). D'ailleurs, les mêmes problématiques qu'à l'échelle spécifique se posent à l'échelle populationnelle (la barrière espèce/population est d'ailleurs souvent difficile à définir chez certains modèles, comme les arbres par exemple) concernant l'importance des flux de gènes face à la sélection pour des adaptations locales et la différenciation des populations (Savolainen *et al.*, 2007). L'étude des processus historiques responsables de l'organisation spatiale de la diversité génétique entre espèces phylogénétiquement proches et au sein des espèces, est appelée phylogéographie (Avise, 2000; Avise, 2009). Cette discipline a permis de mieux comprendre les mécanismes évolutifs impliqués dans les processus de diversification intraspécifique en inférant les barrières aux flux de gènes et en explorant les conséquences des fragmentation-recolonisation en terme de différenciation génétique (Petit *et al.*, 2002; Taberlet *et al.*, 1998). L'un des thèmes les plus abordés en phylogéographie est l'étude de la conséquence des oscillations climatiques durant la dernière grande glaciation (23000-11000 ans) sur la distribution de la diversité génétique des populations et de leurs différenciations (principalement en Europe et Amérique du Nord). Cette glaciation a eu pour conséquence la séparation de populations au sein de refuges glaciaires[4]; puis, suite au réchauffement du début de l'Holocène (10000 ans), ces populations ont recolonisé des territoires vers le Nord, créant probablement des zones hybrides entre lignées évolutives ayant divergé. Ces processus de migration/extinction/recolonisation de populations sur une échelle de temps courte (30000 ans) représentent ainsi un modèle idéal pour l'étude des processus de

[4] Site/localisation géographique de populations reliques ou isolées d'une espèce ayant subit des contractions de son aire géographique suite aux glaciations.

divergence de populations avec contact secondaire. De nombreuses études se sont intéressées à ces processus de recolonisation pour des espèces tempérées européennes (Heuertz *et al.*, 2006; Heuertz *et al.*, 2004; Pauwels *et al.*, 2012; Petit *et al.*, 2002; Vercken *et al.*, 2010). Des refuges glaciaires (Péninsule Ibérique, Italie, Balkans et Carpates) ainsi que des zones de sutures localisées en Europe (Hewitt, 2004; Schmitt, 2007; Taberlet *et al.*, 1998) semblent partagés entre espèces tempérées. Cependant, les questions concernant l'importance et la quantification des flux de gènes entre lignées évolutives dans ces processus de diversification restent limitées aux interprétations issues des structures phylogéographiques observées. Les méthodes d'inférences démographiques permettant de tester des scénarios alternatifs avec ou sans flux de gènes sont encore peu utilisées dans ce domaine (Bertorelle *et al.*, 2010; François *et al.*, 2008).

Capacité à disperser des espèces. La dispersion du pollen et des graines sont responsables des flux de gènes entre populations et individus, et par conséquent ont un rôle majeur dans les processus de diversification. En effet, la probabilité d'observer une divergence entre espèces au sein d'une région diminue avec l'intensité des flux de gènes intraspécifiques. Cet attendu théorique a été démontré chez un grand nombre de modèles : chauves-souris, mammifères carnivores, oiseaux, plantes à fleurs, lézards et escargots (Kisel, Barraclough, 2010). Parmi les diverses méthodes d'estimation des capacités de dispersion, l'estimation indirecte à travers la caractérisation de la structuration génétique spatiale intraspécifique (SGS), ainsi que de la détection d'isolement par la distance, permet de quantifier, voire même de comparer entre populations et entre espèces, les capacités de colonisation et de dispersion (Vekemans, Hardy, 2004). Ces données sont essentielles pour la gestion des populations naturelles mais aussi pour prédire les changements de distribution à long terme des espèces face aux changements climatiques ou lors d'invasions biologiques (Petit *et al.*, 2004a)

Approches statistiques d'inférences de paramètres démographiques dans les modèles de divergence avec flux de gènes. L'utilisation de marqueurs génétiques (séquences, SNP, microsatellites) et le développement de méthodes analytiques ont significativement amélioré l'étude des populations en divergence (Pinho, Hey, 2010). Une nouvelle discipline visant à retracer l'histoire démographique de populations

phylogénétiquement proches à partir du polymorphisme génétique neutre distribué aléatoirement sur le génome est alors née : la « génétique des populations en divergence » (Kliman et al., 2000). L'utilisation de modèles de coalescence à travers des généalogies de gènes permet d'estimer des paramètres classiques de génétique des populations tels que les tailles efficaces, les taux de migration et le temps depuis la divergence.

Certains des programmes actuellement les plus utilisés permettant l'étude de la différenciation génétique et de l'introgression entre les populations, première étape de l'étude de population en divergence, sont basées sur des inférences Bayésiennes et des simulations des Chaines de Markov Monte Carlo (MCMC) comme implémentées dans le logiciel *STRUCTURE* (Pritchard *et al.*, 2000). Cependant, ce logiciel ne permet pas d'inférer des paramètres démographiques (tailles efficaces, taux de migration et temps depuis la divergence), essentiels pour la compréhension des processus de divergence avec flux de gènes. Ainsi, des modèles basés sur la théorie de la coalescence se sont développés, permettant d'estimer ces paramètres démographiques simultanément (Nielsen, Wakeley, 2001). Ces modèles supposent une divergence de populations avec flux de gènes. Plusieurs logiciels disponibles à ce jour permettent d'utiliser de manière accessible ces modèles et d'inférer ainsi les paramètres démographiques associés à une divergence entre deux (*IMa*) voire plusieurs populations (*IMa2*) (Hey, Nielsen, 2004; Hey, Nielsen, 2007). Le calcul de la vraisemblance est cependant très complexe, restreignant l'utilisation de ces logiciels à des scénarios évolutifs et modèles d'évolution moléculaires simples (Csilléry *et al.*, 2010) ou à des jeux de données restreints (Wang, Hey, 2011). D'autre part, malgré les avancées des capacités informatiques, ces méthodes ne peuvent plus assurer des temps de simulation raisonnables face au nombre et à la complexité des modèles démographiques devant être comparés. Une nouvelle approche qui approxime la vraisemblance est née, l'*approximate Bayesian computation* (ABC), associant les avantages du modèle Bayésien, mais contournant le problème du calcul de la vraisemblance (Csilléry *et al.*, 2010). L'ABC permet de comparer des modèles évolutifs plus complexes et réalistes que le modèle « isolement avec migration » entre deux ou plusieurs populations (Hey, Nielsen, 2004; Hey, Nielsen, 2007; Nielsen, Wakeley, 2001). Son utilisation s'est ainsi accrue de manière considérable depuis quelques années afin de résoudre de multiples problématiques (Patin *et al.*, 2009; Row *et al.*, 2011; Tellier *et al.*, 2011; Wegmann, Excoffier, 2010), en particulier la comparaison de

modèles avec et sans flux de gènes, permettant ainsi d'avoir une meilleure compréhension du processus de spéciation et plus généralement de diversification (Encadré 1) (Smadja, Butlin, 2011).

Encadré 1: Approximate Bayesian computation (ABC) et estimation de paramètres

L'ABC est une approche d'inférence Bayésienne basée sur des statistiques résumées contournant ainsi l'estimation de la vraisemblance, qui est difficile à estimer dans le cas de scénarios évolutifs complexes. Cette méthode présente aussi les avantages des approches Bayésiennes prenant en compte des informations *priors* (distributions de probabilité sur la valeur d'un paramètre avant que les données soient examinées) et permettant l'inférence de distributions postérieures sur les paramètres des modèles (distribution conditionnelle d'un paramètre sachant les données). Elle est basée sur la simulation de millions de généalogies supposant différentes valeurs de paramètres sous différents modèles. Les simulations qui présentent les valeurs de statistiques résumées les plus proches de celles du jeu de données observé sont retenues pour estimer les distributions postérieures. Les principales étapes de l'ABC sont la définition des modèles à tester, le choix du modèle, puis l'inférence des valeurs de paramètres.

Les modèles

L'un des points forts de l'ABC est la comparaison de scénarios démographiques et historiques. Les premières questions à se poser sont donc : quels modèles doit-on tester et quelle sera leur complexité? Ces modèles proposés sont souvent construits sur la base de l'expérience et des connaissances préliminaires existant sur les modèles biologiques étudiés, et des données historiques. Cette étape passe par la conceptualisation mathématique de ces connaissances préliminaires par la définition des distributions *priors* pour chaque paramètre des modèles. Un grand nombre de modèles peuvent être proposés ; il faut cependant essayer de limiter le nombre de modèles testés ainsi que leur complexité afin de trouver l'explication la plus parcimonieuse.

Inférence et choix des modèles

Une fois les modèles à tester établis, l'étape suivante consiste, sous chaque modèle évolutif, à échantillonner de multiples fois une valeur de paramètre, ϑi, dans une distribution *prior* afin de simuler un jeu de données *yi*. Ensuite à partir des données simulées, il s'agit de

calculer la valeur de statistique résumée associée au jeu de données simulé *S(yi)*, et la comparer à la valeur de la statistique résumée dans le jeu de données observé *S(yo)*, en utilisant une mesure de distance (par exemple euclidienne). Si la distance entre *S(yi)* et *S(yo)* est inférieure à un seuil, appelé tolérance *ε*, la valeur de paramètre est acceptée. Lors de mes simulations, 0,1% des valeurs les plus proches étaient retenues, soit les 1000 meilleures simulations. Ces valeurs sont représentées par des points orange sur la figure a. Elles sont ensuite ajustées par régression linéaire localement pondérée (Leuenberger, Wegmann, 2009) (ligne verte), donnant plus de poids aux simulations produisant des statistiques résumées proches des valeurs observées. Après réajustement, la nouvelle valeur de paramètre (histogramme vert) forme l'échantillon de la distribution postérieure. Cette étape est réalisée pour chacun des modèles. Ensuite, la comparaison des modèles se fait par l'estimation des probabilités postérieures ainsi que les *Bayes factor* de chacun des modèles, pour ainsi choisir le modèle le plus probable. Une fois le modèle choisi, les paramètres sont estimés à partir de ce modèle.

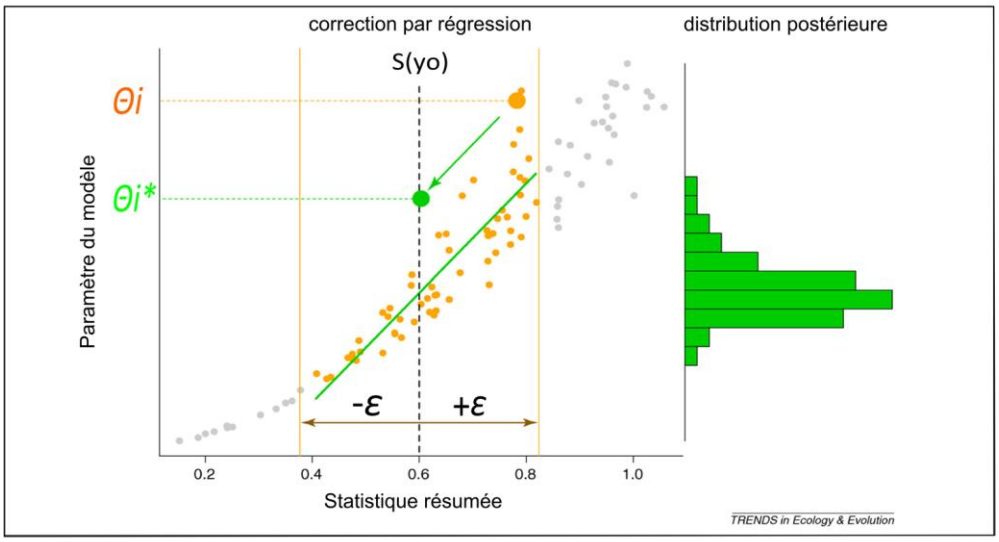
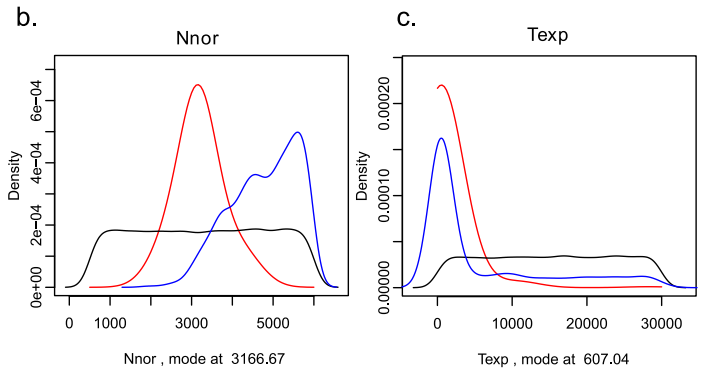


Figure a. Principe de l'ajustement par régression linéaire lors de l'utilisation de l'ABC. *Θi* : valeur de paramètre, *Θi**: valeur de paramètre après réajustement*, yi :* jeu de données simulé, *S(yi):* statistique résumée associée au jeu de données simulé, *S(yo) :* statistique résumée dans le jeu de données observé, *-ε* : tolérance, ligne verte : régression linéaire (Figure tirée de Csilléry *et al.*, 2010).

Estimation des paramètres : fiabilité ?

Dans la littérature, les estimations des paramètres sont souvent présentées sous forme de courbes avec les distributions *priors* (noires dans les figures b et c) et postérieures (rouges

dans les figures b et c). Cependant, les courbes de rejet sont rarement présentées, et sont pourtant très informatives sur la capacité des modèles à prédire et estimer les paramètres. Les courbes de rejet (bleues dans les figures b et c) correspondent à la distribution des valeurs de paramètres dont les valeurs *So* ont été acceptées sur le critère du seuil de tolérance ϵ, mais avant le réajustement par régression.



Si la courbe de rejet est en concordance avec la distribution postérieure des paramètres (en rouge, figure c) alors, les estimations, de *visu*, peuvent être considérées comme fiables (figure c, paramètre estimé *Texp*). Ces résultats validés *de visu* sont souvent ensuite aussi validés lors du *model checking* qui consiste à tester statistiquement si les statistiques simulées sous le modèle choisi sont non significativement différentes des statistiques résumées observées, afin de valider les estimations de paramètres. Dans certains cas cependant la courbe de rejet est décalée par rapport à la courbe postérieure (Figure b) : l'ajustement par régression a amené une estimation trop biaisée des paramètres traduisant un problème d'ajustement du modèle aux données observées. Cette estimation au travers de la régression engendre des problèmes d'estimation de paramètres pour le modèle donné. Ainsi l'observation des courbes de rejet est souvent un bon moyen de savoir si les estimations de paramètres sont fiables ou non, sans passer par le *model checking* directement.

3. Les arbres, un modèle particulier d'étude des processus de diversification

Les arbres sont des plantes vasculaires ligneuses pérennes ne formant pas un groupe monophylétique mais partageant des spécificités communes d'un point de vue évolutif : des faibles taux de spéciation malgré une grande diversité génétique, une grande différenciation au niveau de leurs traits adaptatifs malgré de forts flux de gènes, une grande capacité à maintenir l'intégrité génétique des espèces malgré des hybridations interspécifiques très fréquentes, ainsi que des traits d'histoire de vie qui les distinguent des autres plantes, tels qu'une longue durée de vie et un long temps de génération (Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). On peut ainsi s'attendre à des mécanismes de diversification originaux à diverses échelles évolutives (domestication, diversification intraspécifique, spéciation) chez ces modèles. Pourtant, la littérature concernant ces problématiques reste encore limitée comparativement à d'autres modèles de plantes comme les espèces annuelles, malgré leur importance dans l'économie actuelle et les retombées de ces études en termes d'amélioration variétale et de conservation des ressources génétiques, sans oublier bien sûr l'intérêt académique de ces questions.

Propagation végétative, temps de génération et méthode de sélection chez les arbres fruitiers : impact sur le processus de domestication. La combinaison du greffage et des méthodes de sélection a sûrement dû jouer un rôle primordial dans l'histoire de la domestication des arbres fruitiers.

Pour les espèces de plantes pérennes, la domestication passe souvent par un changement de la reproduction sexuée (milieu naturel) vers la reproduction végétative (milieu cultivé), par bouture, par surgeon (rejeton qui pousse au pied d'un arbre), ou par greffage (Zohary, Hopf, 2000). Ce contraste entre les taxons domestiqués et leurs apparentés sauvages joue probablement un rôle dans l'évolution des espèces pérennes végétales cultivées. La reproduction végétative permet, chez des espèces végétales allogames (strictes ou non), de maintenir des génotypes élites d'intérêt agronomique à la génération suivante et d'éviter ainsi les désavantages de la ségrégation mendélienne inhérents à la reproduction sexuée (McKey *et al.*, 2010). La capacité naturelle à se propager

végétativement de certaines espèces sauvages pérennes leur a ainsi permis d'être domestiquées tôt dans l'histoire de l'humanité, tels que le figuier, l'olivier ou la vigne, alors que des espèces comme le pommier, poirier ou cerisier ont été cultivées plus tardivement (Zohary, Hopf, 2000). La propagation végétative des arbres cultivés implique peu de recombinaison et de cycles sexués, amenant les taxons cultivés à être très proches génétiquement du pool génétique de leurs apparentés sauvages. De plus, les arbres cultivés présentent des temps de génération beaucoup plus longs que les espèces annuelles. Ainsi, la différenciation génétique par rapport à l'apparenté sauvage devrait être beaucoup plus lente (en termes d'années) que chez les espèces annuelles. L'observation de syndromes de domestication peu marqués chez les arbres fruitiers pourrait ainsi s'expliquer par la limitation du nombre de générations sur laquelle la sélection a pu agir (Miller, Gross, 2011).

Chez les arbres fruitiers, l'amélioration variétale est souvent basée sur des croisements aléatoires entre génotypes intéressants, dont les phénotypes intéressants sont ensuite maintenus par greffage. La question de l'impact de ces pratiques sur la structuration génétique et l'évolution des espèces pérennes reste encore très peu étudiée (McKey *et al.*, 2010), en particulier chez les arbres fruitiers.

Les arbres : impact des flux de gènes intraspécifiques et interspécifiques et des larges tailles efficaces sur les mécanismes de diversification artificielle et naturelle. L'évolution des arbres ne peut être comprise sans considérer leurs caractéristiques principales (Petit, Hampe, 2006). Or, l'une d'entre elles est leur forte capacité à échanger des gènes au travers des hybridations interspécifiques et intraspécifiques.

Les flux de gènes intraspécifiques sont en général plus importants chez les arbres que chez les herbacées à système de reproduction similaire (Petit, Hampe, 2006). Beaucoup d'études ont en effet montré une faible structuration génétique spatiale et de fortes capacités de dispersion chez de nombreuses espèces d'arbres (Lascoux *et al.*, 2004; Petit *et al.*, 2004a; Petit, Hampe, 2006; Vekemans, Hardy, 2004), en particulier chez les espèces tropicales ayant la caractéristique d'être en très faible densité dans les forêts. La combinaison de ces traits d'histoire de vie, associée à de larges tailles efficaces de population, implique des effets de dérive très peu prononcés, limitant la divergence entre populations ou espèces et favorisant la rétention de polymorphisme ancestral entre espèces

phylogénétiquement apparentées. Par exemple, des espèces de pins ayant divergé il y a 10 millions d'années ne diffèrent que de 5% au niveau des sites synonymes de séquences nucléaires (Savolainen, Pyhäjärvi, 2007). D'autre part, les arbres apparaissent plus enclins aux croisements interspécifiques que d'autres plantes (Petit, Hampe, 2006). Ces hybridations interspécifiques par contacts secondaires combinées au polymorphisme ancestral ont pour conséquence une évolution réticulée des lignées, limitant l'étude des mécanismes de diversification chez les arbres au travers des approches phylogénétiques classiques, ne prenant pas en compte cette rétention de polymorphisme ancestral. Ainsi, l'étude des mécanismes de diversification chez les arbres par les nouvelles méthodes d'analyses d'inférence en génétique des populations, intégrant ces effets de rétention de polymorphisme ancestral et des modèles évolutifs complexes, permettra de répondre à des questions encore peu résolues chez ces modèles, concernant l'importance des flux de gènes lors des processus de diversification.

4. Le modèle pommier, le genre *Malus*

Le genre *Malu*s représente un modèle idéal pour tester des hypothèses sur les mécanismes de diversification chez les arbres, en particulier chez les arbres fruitiers. Ce genre regroupe l'ensemble des espèces de pommiers distribuées à travers les régions tempérées dont les délimitations morphologiques sont très floues et le nombre encore controversé (dix à 100 morpho-espèces selon les auteurs). L'incertitude sur le nombre d'espèces de *Malus* résulte d'un manque de caractères morphologiques diagnostics, de chevauchements géographiques et d'hybridations interspécifiques fréquentes (Juniper, Mabberley, 2006). Parmi l'ensemble des espèces du genre, la seule espèce cultivée, *Malus domestica*, est utilisée pour la production commerciale de pommes. Le genre *Malus* serait originaire du Sud de la Chine, des provinces du Guizhou, du Sichuan et du Yunnan. Ces trois régions présentent une très grande richesse spécifique, comportant potentiellement plus de 3/5 des espèces de *Malus* existant dans le monde. Il semble probable que le genre soit apparu durant l'Eocène (55,5-33,7 millions d'années), période d'apparition de plusieurs genres dans la famille des Rosaceae (Juniper, Mabberley, 2006). Une espèce ancestrale présentant huit chromosomes aurait subi une auto-polyploïdisation (≈50 millions d'années) aboutissant à un caryotype à 18 chromosomes, dont un aurait été perdu, aboutissant à un

caryotype à 17 chromosomes, caractéristique de la sous-tribu des Pyreae des Rosaceae, qui comprend l'ensemble des espèces de pommiers et de poiriers (Giovannoni, 2010). Le scénario de dispersion et de diversification du genre est cependant encore très mal connu. Le genre aurait connu une radiation spécifique durant les périodes préglaciaires, interglaciaires et postglaciaires, avec des espèces colonisant le reste de l'Asie, de l'Europe et de l'Amérique (Juniper 2006). Des phylogénies basées sur des séquences nucléaires et chloroplastiques ont permis d'identifier les espèces basales du genre, mais la phylogénie manquait de résolution pour le groupe incluant le pommier domestiqué et plusieurs espèces sauvages (Harris *et al.*, 2002; Robinson *et al.*, 2001), en particulier pour les quatre espèces sauvages supposées être des contributrices potentielles au génome du pommier domestiqué (Juniper, Mabberley, 2006): *M. sylvestris* (Europe), *M. baccata* (Sibérie), *M. sieversii* (Asie Centrale) *et M. orientalis* (Caucase) (Figure 2).

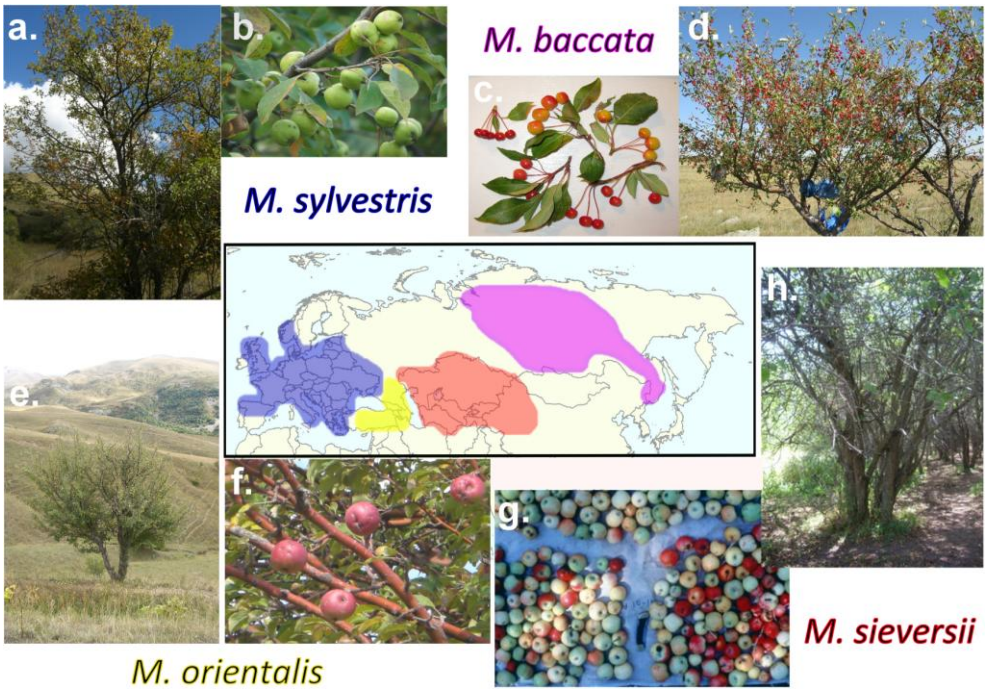


Figure 2. Distributions géographiques des quatre espèces de pommiers sauvages étudiées (*Malus domestica* n'est pas présente en milieu naturel). *Malus sylvestris* (a) est distribuée à travers l'Europe de l'Ouest et de l'Est (couleur bleue sur la carte) et présente des pommes petites (Ø : 2-5 cm), vertes et acides (b). *Malus baccata* (d) se distribue à travers la Russie du Sud-Est ainsi qu'en Sibérie (violet), et présente de petites pommes type « cerise » (Ø < 1 cm) (c). *Malus orientalis* (e) se distribue dans le Caucase, avec certains individus repérés en Russie (jaune), avec des pommes de tailles similaires à celles de *Malus sylvestris* vertes/rougeâtres et acides (f). *Malus sieversii* (h) se distribue en Asie Centrale (Kazakhstan, Kirghizistan, Tadjikistan et Xinjiang en Chine), avec des pommes présentant des diversités de couleurs et de morphologies exceptionnelles (g) (photos : USDA, Forsline, 2004).

Ces quatre espèces présentent des distributions géographiques et des écologies différentes, alors qu'étant morphologiquement proches du pommier cultivé (*M. domestica*). Un scénario de domestication probable est le suivant : le pommier cultivé (*M. domestica*) serait initialement apparu à partir de *M. sieversii* en Asie centrale dans les montagnes du Tian Shan il y a 8000 ans environ, et il aurait ensuite été transporté vers l'Ouest le long de la route de la Soie, où il aurait pu s'hybrider avec *M. orientalis* et *M. baccata*, puis avec *M. sylvestris* suite à son introduction en Europe par les Romains il y a 3000 ans (Forsline *et al.*, 2002). Certains auteurs, comme Barrie Juniper, contestent ce scenario et considèrent que *M. sieversii* est le seul contributeur au génome de *M. domestica* (Juniper, Mabberley, 2006). Les données génétiques ne permettaient pas de trancher sans ambiguïté en faveur de l'un ou l'autre des scénarios au début de ma thèse.

Ce complexe de cinq espèces, toutes diploïdes et phylogénétiquement proches, incluant une espèce domestiquée, constitue un modèle idéal pour étudier des processus de diversification à des échelles évolutives différentes.

5. Buts de la thèse

Cette thèse vise à comprendre les mécanismes généraux impliqués lors de diversification interspécifiques et intraspécifiques dans le genre *Malus,* qui regroupe l'ensemble des espèces de pommiers à travers le monde. En particulier, je me suis focalisée sur l'unique espèce domestiquée (*Malus domestica*) et quatre espèces sauvages ayant été suggérées comme de possibles espèces contributrices au génome du pommier cultivé. Je me suis intéressée aux mécanismes évolutifs neutres impliqués dans la diversification de ces espèces à différentes échelles évolutives : a) histoire de la domestication du pommier cultivé (*M. domestica*) ; b) diversification intraspécifique chez un contributeur sauvage (*M. sylvestris*) à l'espèce cultivée; c) histoires des spéciations entre quatre espèces sauvages apparentées et une espèce domestiquée; d) estimation de la structure génétique spatiale et des capacités de dispersion de trois espèces sauvages contributrices au génome du pommier cultivé (*M. orientalis*, *M. sylvestris* et *M. sieversii*), ainsi que des niveaux d'introgressions du pommier cultivé vers chacune de ces espèces dans leur région d'origine, afin d'estimer l'importance des hybridations entre espèces d'arbres cultivées et sauvages. Pour répondre à ces questions, j'ai utilisé un total de 1549 individus de chacune des cinq espèces, échantillonnés

à travers l'Eurasie et génotypés à l'aide de 26 marqueurs microsatellites. A cela, j'ai ajouté 13 loci nucléaires amplifiés pour huit à dix individus selon les espèces afin d'étudier l'histoire des spéciations dans le genre *Malus* (Manuscrit C). Enfin, j'ai participé à une étude concernant la coévolution d'un des champignons pathogène du pommier cultivé, la tavelure du pommier *(Venturia inaequalis),* dont les résultats sont présentés en Annexe 2. Ma thèse s'articule donc en quatre parties principales:

A. Domestication du pommier cultivé (*Malus domestica)*: comprendre l'histoire de sa domestication, déterminer son origine et comprendre quelles sont les espèces sauvages ayant contribué à son génome, et en quelles proportions s'il y en a eu plus d'une. ***Manuscrit A**: **New Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties.*** *PloS Genet. 2012. 8(5) e1002703*

B. Phylogéographie du contributeur européen (*Malus sylvestris*) au génome du pommier cultivé: comprendre l'origine de la structuration des populations d'une des espèces contributrices au génome du pommier cultivé. **Manuscrit B:** ***Post-glacial recolonization history of the European crabapple (*Malus sylvestris *Mill.), a wild contributor to the domesticated apple.*** Resubmitted to Molecular Ecology.

C. Histoires de spéciations entre quatre espèces de pommiers sauvages en Eurasie et du pommier cultivé : reconstruction des histoires de spéciations récentes au sein de ce genre afin de comprendre si les espèces dans le genre *Malus* apparaissent et se maintiennent avec ou sans flux gènes. **Manuscrit C:** ***Speciation histories of four wild apple species in Eurasia and of the cultivated apple*.** In prep.

D. Hybridations interspécifiques du pommier cultivé vers les trois espèces contributrices sauvages et estimation de la structuration génétique spatiale chez l'espèce asiatique et l'espèce caucasienne. **Manuscrit D:** ***Crop-to-wild gene flow and spatial genetic structure in the closest wild relatives of the cultivated apple.*** Submitted to Evolutionnary Applications.

E. Annexe 2: Le Van *et al*. **Manuscrit E: *Evolution of pathogenicity traits in the apple scab fungal pathogen in response to the domestication of its host.*** Evolutionary Applications. 2012: In press.

Manuscript A: Domestication du pommier cultivé

(*Malus domestica*)

Manuscript A : New Insight into the History of Domesticated Apple : Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. PloS Genet. 2012. 8(5) e1002703

Amandine Cornille [1,2†], Pierre Gladieux[1,2], Marinus J. M. Smulders[3], Isabel Roldán-Ruiz[4], François Laurens[5,6,7], Bruno Le Cam[5,6,7], Anush Nersesyan[8], Joanne Clavel[1,2], Marina Olonova[9], Laurence Feugey[5,6,7], Ivan Gabrielyan[8], Xiu-Guo Zhang[10], Maud I. Tenaillon[11], Tatiana Giraud[1,2]

† Corresponding author: amandine.cornille@gmail.com

1. CNRS, Laboratoire Ecologie Systématique et Evolution - UMR8079, Bâtiment 360, 91405 Orsay, France; Univ. Paris Sud, 91405 Orsay, France; 2. AgroParisTech, 91405 Orsay, France; 3. Plant Research International, Wageningen UR Plant Breeding, PO Box 16, 6700 AA Wageningen, The Netherland; 4. ILVO, Plant – Growth and Development, Caritasstraat 21, 9090 Melle, Belgium; 5. INRA, IRHS, PRES UNAM, SFR QUASAV, Rue G. Morel F-49071 Beaucouzé, France ; 6. Université d'Angers, IRHS, PRES UNAM, SFR QUASAV, Blvd Lavoisier F-49071 Angers, France ; 7. Agrocampus Ouest, IRHS, PRES UNAM, SFR QUASAV, Rue Le Nôtre F-40045 Angers, France ; 8. Institute of Botany, Armenian National Academy of Sciences, Department of Plant Taxonomy, 375063 Yerevan, Armenia; 9. Biological Institution, Tomsk State University, 36 Lenin av., Tomsk, 634050 Russia; 10. Department of Plant Pathology, Shandong Agricultural University, Taiwan, China; 11. CNRS, UMR de Génétique Végétale, INRA/CNRS/Univ Paris-Sud, F-91190 Gif-sur-Yvette, France.

ABSTRACT

Apple is the most common and culturally important fruit crop of temperate areas. The elucidation of its origin and domestication history is therefore of great interest. The wild Central Asian species *Malus sieversii* has previously been identified as the main contributor to the genome of the cultivated apple (*Malus domestica*), on the basis of morphological, molecular and historical evidence. The possible contribution of other wild species present along the Silk Route running from Asia to Western Europe remains a matter of debate, particularly with respect to the contribution of the European wild apple. We used microsatellite markers and an unprecedented large sampling of five *Malus* species throughout Eurasia (839 accessions from China to Spain) to show that multiple species have contributed to the genetic makeup of domesticated apples. The wild European crabapple *M. sylvestris*, in particular, was a major secondary contributor. Bidirectional gene flow between the domesticated apple and the European crabapple resulted in the current *M. domestica* being genetically more closely related to this species than to its Central Asian progenitor, *M. sieversii*. We found no evidence of a domestication bottleneck or clonal population structure in apples, despite the use of vegetative propagation by grafting. We show that the evolution of domesticated apples occurred over a long time period and involved more than one wild species. Our results support the view that self-incompatibility, a long life span and cultural practices, such as selection from open-pollinated seeds, have facilitated introgression from wild relatives and the maintenance of genetic variation during domestication. This combination of processes may account for the diversification of several long-lived perennial crops, yielding domestication patterns different from those observed for annual species.

INTRODUCTION

Domestication is a process of increasing codependence between plants and animals on the one hand, and human societies on the other (Diamond, 1997; Zeder *et al.*, 2006). The key questions relating to the evolutionary processes underlying domestication concern the identity and geographic origin of the wild progenitors of domesticated species (Diamond, 2002), the nature of the genetic changes underlying domestication (Purugganan, Fuller, 2009; Wright, Gaut, 2005), the tempo and mode of domestication (*e.g.*, rapid transition *versus* protracted domestication) (Tenaillon, Manicacci, 2011) and the consequences of domestication for the genetic diversity of the domesticated species (Allaby *et al.*, 2008; Caicedo *et al.*, 2007; Doebley *et al.*, 2006; Gross, Olsen, 2009). An understanding of the domestication process provides insight into the general mechanisms of adaptation and the history of human civilization, but can also guide modern breeding programs aiming to improve crops or livestock species further (Brown *et al.*, 2009; Feuillet *et al.*, 2008).

Plant domestication has mostly been studied in seed-propagated annual crops, in which strong domestication bottlenecks have often been inferred, especially in selfing annuals, such as foxtail millet, wheat and barley (Brown *et al.*, 2009; Glémin, Bataillon, 2009; Kilian *et al.*, 2007; Kovach *et al.*, 2007; Russell *et al.*, 2011; Wang *et al.*, 2010). Genetic data have suggested that domestication or the spread of domesticated traits has been fairly rapid in some annual species (*e.g,* maize or sunflower), with limited numbers of populations or species contributing to current diversity (Blackman *et al.*, 2011; Gross, Olsen, 2009; Harter *et al.*, 2004; Matsuoka *et al.*, 2002; Oumar *et al.*, 2008; Tenaillon *et al.*, 2004). In contrast, a combination of genetics and archaeology suggested a protracted model of domestication for other annual crops, and in particular for the origin of wheat or barley in the Fertile Crescent (Brown *et al.*, 2009; Tanno, Willcox, 2006). However, the genetic consequences of domestication have been little investigated in long-lived perennials, such as fruit trees (Chen *et al.*, 2009a; Miller, Gross, 2011; Miller, Schaal, 2005). Trees have several biological features that make them fascinating and original models for investigating domestication: they are outcrossers with a long lifespan and a long juvenile phase, and tree populations are often large and connected by high levels of gene flow (Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007).

Differences in life-history traits probably result in marked differences in the mode and speed of evolution between trees and seed-propagated selfing annuals (Austerlitz *et al.*, 2000; Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). For example, outcrossing may tend to make domestication more difficult, in part because the probability of fixing selected alleles is lower than in selfing crops (Glémin, Bataillon, 2009; Tenaillon, Manicacci, 2011). The combination of self-incompatibility and a long juvenile phase also results in highly variable progenies, making breeding a slow and expensive process, and rendering crop improvement difficult. The development of vegetative propagation based on cuttings or grafting has been a key element in the domestication of long-lived perennials, allowing the maintenance and spread of superior individuals despite self-incompatibility (Janick, 2005). However, the use of such techniques has further decreased the number of sexual cycles in tree crops since the initial domestication event, adding to the effect of long juvenile phases in limiting the genetic divergence between cultivated trees and their wild progenitors (Janick, 2005; Miller, Schaal, 2006; Zohary, 2004; Zohary, Spiegel-Roy, 1975). Thus, domestication can generally be considered more recent, at least in terms of the number of generations, in fruit tree crops than in seed-propagated selfing annuals.

Given the slow process of selection and the limited number of generations in which humans could exert selection, the protracted nature of the domestication process in trees has probably resulted in limited bottlenecks (Miller, Gross, 2011; Miller, Schaal, 2006) and in a weaker domestication syndrome (Pickersgill, 2007) than in seed-propagated annuals. Nevertheless, many cultivated fruit trees clearly display morphological, phenotypic and physiological features typical of a domestication syndrome, such as large fruits and high sugar or oil content (Juniper, Mabberley, 2006; Zohary, Spiegel-Roy, 1975). Many aspects of fruit tree domestication have been little studied (Miller, Gross, 2011). Consequently, most of the hypotheses concerning the consequences of particular features of trees for their domestication/diversification remain to be tested. Recent studies on grapevines, almond and olive trees have provided illuminating insights, such as the importance of outcrossing and interspecific hybridization (Besnard *et al.*, 2007; Delplancke *et al.*, 2011; Myles *et al.*, 2011), but additional studies of other species are required to draw more general conclusions.

Here, we investigated the origins of the domesticated apple *Malus domestica* Borkh.*,* one of the most emblematic and widespread fruit crops in temperate regions (Juniper, Mabberley, 2006). A form of apple corresponding to extant domestic apples appeared in the Near East around 4,000 years ago (Zohary, Hopf, 2000), at a time corresponding to the first recorded uses of grafting. The domesticated apple was then introduced into Europe and North Africa by the Greeks and Romans and subsequently spread worldwide (Juniper, Mabberley, 2006). While the ancestral progenitor has been clearly identified as being *M. sieversii*, the identity and relative contributions of other wild species present along the Silk route that have contributed to the genetic makeup of apple cultivars remain largely unknown. This is surprising given the potential importance of this knowledge for plant breeding and for our understanding of the process of domestication in fruit trees.

The wild Central Asian species *M. sieversii* (Ldb.) M. Roem has been identified as the main contributor to the *M. domestica* genepool based on similarities in fruit and tree morphology, and genetic data (Coart *et al.*, 2006; Harris *et al.*, 2002; Robinson *et al.*, 2001; Velasco *et al.*, 2010). The Tian Shan forests were identified as the geographic area in which the apple was first domesticated, on the basis of the considerable intraspecific morphological variability of wild apple populations in this region (Dzhangaliev, 2003; Vavilov, 1926). Nucleotide variation for 23 DNA fragments even suggested that *M. sieversii* and *M. domestica* belonged to a single genepool (which would be called *M. pumila* Mill.), with phylogenetic networks showing an intermingling of individuals from the two taxa (Velasco *et al.*, 2010). Some authors have also suggested possible contributions of additional wild species present along the Silk Route: *M. baccata* (L.) Borkh, which is native to Siberia, *M. orientalis* Uglitz., a Caucasian species present along western sections of the ancient trade routes, and *M. sylvestris* Mill*.* (European crabapple), a species native to Europe (Boré, Fleckinger, 1997; Forsline *et al.*, 2002; Luby *et al.*, 2001; Rehder, 1940). These hypotheses were based on the history of human migration and trade, the lack of phylogenetic resolution between *M. domestica* and these four wild species (Harris *et al.*, 2002; Robinson *et al.*, 2001), genetic evidence of hybridization at a local scale between domesticated apple and *M. sylvestris* (Coart *et al.*, 2006), and the recent finding of sequence haplotype sharing between *M. sylvestris* and *M. domestica* (Harrison, Harrison, 2011). However, such secondary contributions remain a matter of debate, mostly due to the difficulty of distinguishing

introgression from incomplete lineage sorting (Harrison, Harrison, 2011; Micheletti *et al.*, 2011; Velasco *et al.*, 2010). The three wild species occurring along the Silk Route all bear small, astringent, tart fruits. None of these species has the fruit quality of *M. sieversii*, but they may have contributed other valuable horticultural traits, such as later flowering, resistance to pests and diseases, capacity for longer storage or climate adaptation. The organoleptic properties of the fruits of these wild species may also have been selected during domestication, for the preparation of apple-based beverages, such as ciders (Forsline *et al.*, 2002; Pereira-Lorenzo *et al.*, 2009). Cider apples are indeed smaller, bitter and more astringent than dessert apples and bear some similarity to *M. sylvestris* apples. There is also evidence to suggest that Neolithic and Bronze Age Europeans were already making use of *M. sylvestris* (Zohary, Hopf, 2000).

In this study, we used a comprehensive set of apple accessions sampled across Eurasia (839 accessions from China to Spain; Figures 1 and S1; Table S1) and 26 microsatellite markers distributed evenly across the genome to investigate the following questions: 1) Is there evidence for population subdivision within and between the five taxa *M. domestica, M. baccata*, *M. orientalis, M. sieversii* and *M. sylvestris*? 2) How large is the contribution of wild species other than the main progenitor, *M. sieversii,* to the genome of *M. domestica*? 3) Does *M. domestica* have a genetic structure associated with its different possible uses (*i.e.*, differences between cider and dessert apples)? 4) What consequences have domestication, subsequent crop improvement and vegetative propagation by grafting had for genetic variation in cultivated apples? Most of our samples of *M. domestica* corresponded to cultivars from Western Europe (Figures 1 and S1), as almost all the cultivars available in modern collections (including American, Australasian cultivars) are of European ancestry and this region is therefore the most relevant area for the detection of possible secondary introgression from the European crabapple.

RESULTS

High diversity and low deviations from random mating expectations within species

Our sampling scheme (Figures 1 and S1), based on the collection of a single tree for each apple variety, was designed to avoid the sampling of clones. However, there may still

be some clonality if some varieties differing by only a few mutations were propagated by grafting. We corrected for this potential clonality, using the clonal assignment procedures implemented in GENODIVE (Meirmans, Van Tienderen, 2004). We found no pair of samples assigned to the same clonal lineage unless using a threshold of 22 pairwise differences between multilocus genotypes, indicating that our samples did not include any clonal genotypes (the threshold corresponds to the maximum genetic distance allowed between genotypes deemed to belong to the same clonal lineage).

Many apple cultivars, including modern cultivars in particular, share recent common ancestors, and siblings or clones of wild species can also be collected unintentionally in the field. Because these features could result in a spurious genetic structure due to the presence of closely related individuals in the dataset, we checked for the presence of groups of related individuals in our dataset between *M. domestica* cultivars and between the individuals of each wild species. The percentage of pairs with a pairwise relatedness (r_{xy}) greater than 0.5 (*i.e.*, full sibs) was: 0.4% in *M. domestica* (*N*=168 pairs), 0.3% in *M. sieversii* (*N*=79), 0.004% in *M. orientalis* (*N*=20), and 0.7% in *M. baccata* (*N*=40). For *M. sylvestris,* no individual pair with r_{xy}>0.5 was identified. However, the distribution of pairwise relatedness r_{xy} among *M. domestica* cultivars did not deviate significantly from a Gaussian distribution centred on 0 and with a low variance (Fisher's exact test, $P\approx1$, standard deviation=0.11, Figure S2). This suggests that closely related cultivars are unlikely to have biased subsequent analyses of population structure. We also checked for the limited effect of relatedness on our conclusions by performing all analyses of population subdivision on both the full dataset and a pruned dataset excluding related individuals (see below).

We tested the null hypothesis of random mating within each species by calculating F_{IS}, which measures inbreeding. All five *Malus* species had relatively low values of F_{IS}, although all were significantly different from zero (Table 1), suggesting that each species corresponded to an almost random mating unit. This is consistent with the self-incompatibility system of these species and indicates a lack of widespread groups of related individuals in *M. domestica*. Low F_{IS} values at species level also indicate a lack of population structure within species. The higher values of F_{IS} observed in *M. baccata* probably resulted from the occurrence of null alleles, as the microsatellite markers were developed in *M. domestica,* to which *M. baccata* is the most distantly related (Table 2). The lowest F_{IS} value

was that obtained for *M. domestica,* reflecting outcrossing between dissimilar parents in breeding programs, or that selection targeted higher levels of heterozygosity (Koopman *et al.*, 2007).

The five *Malus* species form well separated genetic clusters

We used the 'admixture model' implemented in STRUCTURE 2.3 (Pritchard *et al.*, 2000) to infer population structure and introgression. Analyses were run for population structure models assuming K=1 to K=8 distinct clusters (Figure 2). The ΔK statistic, designed to identify the most relevant number of clusters by determining the number of clusters beyond which there is no further increase in likelihood (Evanno *et al.*, 2005), was greatest for K=3 (ΔK=6249, $Pr|ln\ L$=-78590). However, the clusters identified at higher K values may also reveal a genuine and biologically relevant genetic structure, provided that they are well delimited (Vercken *et al.*, 2010). The five *Malus* species were clearly assigned to different clusters for models assuming K≥6 clusters and for a minor clustering solution ("mode") at K=5 (Figure 2). The major mode (*i.e.*, the clustering solution found in more than 60% of the simulation replicates) observed at K=5 grouped together *M. sylvestris* and *M. domestica* genotypes. Increasing the number of clusters above K=6 identified no additional well-delimited clusters corresponding to a subdivision of a previous cluster. Instead, it simply introduced heterogeneity into membership coefficients, indicating that the clustering of the five *Malus* species into separate genepools was the most relevant clustering solution. We checked that the presence of related pairs of cultivars in our dataset did not bias clustering results, by repeating the analysis on a pruned dataset (N=489) excluding all related individuals in wild and cultivated species (*i.e.*, excluding all pairs with r_{xy}≥0.5). Similar results were obtained, with the same five distinct clusters identified as for the full dataset.

We estimated the genetic differentiation between the five *Malus* species by calculating pairwise F_{ST} (Table 2). All F_{ST} values were highly significant (P<0.001) and seemed to indicate a West to East differentiation gradient of *M. domestica* with the wild species. The highest level of differentiation was that between *M. baccata* and the other *Malus* species, and the lowest level of differentiation was that between *M. domestica* and the westernmost species, *M. sylvestris* (Table 2). *Malus domestica* was markedly more differentiated from its

main progenitor *M. sieversii* (F_{ST}=0.0639) than from the European *M. sylvestris* (F_{ST}=0.006) and it was only slightly less differentiated from the Caucasian *M. orientalis* (F_{ST}=0.049).

No bottleneck during apple domestication

We first searched for footprints of a domestication bottleneck by comparing levels of microsatellite variation in *M. domestica* and wild species. There was no significant difference in genetic diversity (as measured by expected heterozygosity, H_E) between *M. domestica* and *M. baccata*, *M. orientalis* or *M. sieversii,* but H_E was significantly higher in *M. sylvestris* than in *M. domestica* (Table 1). Significant differences in allelic richness (A_r) were found between *M. domestica* and *M. orientalis* (Wilcoxon signed rank test, P=0.03) or *M. sylvestris* ($P<10^{-8}$), but not between *M. domestica* and either *M. baccata* (P=0.9) or *M. sieversii* (P=0.9) (Table 1).

We used the method implemented in the BOTTLENECK program (Cornuet, Luikart, 1996), comparing the expected heterozygosity estimated from allele frequencies with that estimated from the number of alleles and the sample size, which should be identical for a neutral locus in a population at mutation-drift equilibrium. Inferences about historical changes in population size are based on the prediction that the expected heterozygosity estimated from allele frequencies decreases faster than that estimated under a given mutation model at mutation-drift equilibrium in populations that have experienced a recent reduction in size. BOTTLENECK analysis showed no significant deviation from mutation-drift equilibrium in any of the five species, under either stepwise or two-phase models of microsatellite evolution (one-tailed Wilcoxon signed rank test, P>0.95). We therefore detected no signal of a demographic bottleneck associated with the domestication of apples.

Variable recent contributions of wild relative species to the *M. domestica* genepool, with the strongest introgression from *M. sylvestris*.

We used the admixture coefficients estimated by STRUCTURE to assess the recent contribution of the various wild species to the *M. domestica* genepool. STRUCTURE analyses of the full dataset showed some admixture among *Malus* species for the minor mode separating the five species at K=5. Admixture coefficients were higher between *M.*

domestica and *M. sylvestris* (α=0.23) than between *M. domestica* and respectively *M. sieversii* (α=0.06), *M. orientalis* (α=0.034) and *M. baccata* (α=0.032).

We further analysed the contribution of each wild species to the genome of *M. domestica* by running STRUCTURE separately on each pair of species including *M. domestica* (Figure 3; Tables 3 and S2). *Malus domestica* genotypes with membership coefficients ≥0.20 in a wild species genepool were considered to display introgression. Using this somehow arbitrary cut-off value, STRUCTURE analyses revealed that 26% of *M. domestica* cultivars displayed introgression from the European crabapple, *M. sylvestris* (Tables 3 and S2). By contrast, only 2%, 3% and 0.02% of the *M. domestica* genotypes displayed introgression from *M. sieversii*, *M. orientalis* and *M. baccata,* respectively (Tables 3 and S2). The *M. domestica* cultivars displaying admixture with the *M. sylvestris* genepool were mostly Russian (*e.g.*, "Antonovka", "Antonovka kamenicka", "Novosibirski Sweet", "Yellow transparent"), French (*e.g.*, "Blanche de St Anne", "St Jean", "Api" and "Michelin") and English (*e.g.*, "Worcester Pearmain" and "Fiesta"). The M9 dwarf apple cultivar ("Paradis jaune de Metz", (Mabberley *et al.*, 2001)) commonly used as a rootstock also appeared to display introgression from the European crabapple (proportion of ancestry in the *M. domestica* genepool: 0.28; Table S2). When French cultivars were removed from the dataset (*N*=89) and pairwise STRUCTURE analyses were repeated for all species pairs including *M. domestica*, 18% of cultivars displayed introgression from *M. sylvestris*, including commercial cultivars such as Granny Smith, Michelin, Antonovka and Ajmi (Figure S3) with a mean membership coefficient of *M. sylvestris* into *M. domestica* genepool of 47%. *Malus sylvestris* thus appears to have made a significant contribution to the *M. domestica* genepool through recent introgression, building on the more ancient contribution (see below) of the Asian wild species *M. sieversii.* We also note that a few *M. domestica* individuals appeared to display introgression from several wild species (Table S2), and that *M. baccata* ornamental cultivars, such as *M. baccata flexilis*, *M. baccata* Hansen's and *M. baccata gracilis*, were partially or even mostly assigned (from 32% to >80%) to the *M. domestica* genepool (Table S3).

Wild Central Asian apple origin of the *M. domestica* genepool

Previous studies (Diego *et al.*, 2011; Harrison, Harrison, 2011; Velasco *et al.*, 2010) identified the Central Asian wild apple *M. sieversii* as the main progenitor of *M. domestica*

on the basis of DNA sequences. Due to the large contribution by *M. sylvestris* detected in our dataset, corresponding mostly to Western European cultivars, *M. domestica* and *M. sylvestris* appeared to be the most closely related pair of species in our analyses of microsatellite markers. We investigated the more ancient contribution of *M. sieversii* to the *M. domestica* genepool, by reassessing the genetic differentiation between species in analyses restricted to "pure" individuals (*i.e.*, assigned at ≥0.9 to their respective genepools) from both wild and cultivated species. All F_{ST} values were highly significant (P<0.001), but the ranking of F_{ST} values between *M. domestica* and the various wild species was affected: the highest differentiation was still observed between *M. domestica* and *M. baccata* (F_{ST}=0.22), but the lowest differentiation was observed between *M. domestica* and *M. sieversii* (F_{ST}=0.11*).* Regarding the differentiation between *M. sylvestris* and *M. domestica*, we observed the opposite of what was found with the full dataset: *M. sylvestris* appeared to be more strongly differentiated (F_{ST}=0.14) from *M. domestica* than *M. sieversii*. Thus, by removing signals of recent introgression between cultivated and wild species we were able to confirm that *M. sieversii* was the initial progenitor of *M. domestica*.

Recent introgression from *M. domestica* into wild species

The finding of a significant level of introgression from wild species into cultivated apple suggested that gene flow might also have occurred in the opposite direction. STRUCTURE analyses of pairs of species confirmed this hypothesis (Figure 3), revealing possible introgression of genetic material into *M. sylvestris*, *M. baccata*, *M. orientalis* and *M. sieversii* from *M. domestica* (mean proportions of ancestry in the *M. domestica* genepool of 0.12, 0.10, 0.03 and 0.23, respectively; Table 3). Considering genotypes with membership coefficients ≥0.9 in the *M. domestica* genepool as misclassified, we found a total of N=31 misclassified wild *Malus* individuals. These results suggest gene flow from the domesticated apple genepool could significantly affect the genetic integrity of wild apple relatives, their future evolution and, possibly, their use as resources for crop improvement.

Inference of demographic history

Model-based Bayesian clustering algorithms, such as that implemented in STRUCTURE, have a high level of power only for the detection of recent introgression events

(Anderson, Thompson, 2002; Excoffier *et al.*, 2005; Pritchard *et al.*, 2000). We therefore investigated the contributions of *M. sylvestris* and *M. orientalis* to the *M. domestica* genepool using approximate Bayesian computation (ABC) methods that offer a more historical perspective on gene flow (Ross-Ibarra *et al.*, 2009). We used a demographic model implementing admixture events (Cornuet *et al.*, 2008).

We compared several admixture models to infer what species pairs underwent introgression events and to estimate introgression rates (Cornuet *et al.*, 2008). *Malus baccata* was not included in these analyses because of its high level of divergence from *M. domestica*. We assumed, as suggested by previous studies, that *M. domestica* derived originally from *M. sieversii*. The most complex model simulated sequential admixtures between *M. domestica* and all wild species. Other models sequentially removed introgression with each wild species, the order being based on F_{ST} values and admixture rates inferred by STRUCTURE. The compared models were the following: (i) the model *a* assumed that *M. domestica* was derived from *M. sieversii* and that the ancestral *M. domestica* population was involved in reciprocal introgression events with *M. orientalis* and *M. sylvestris,* and subsequently introgressed back into *M. sieversii* (Figure 4a), (ii) model *b* was similar to the model *a*, but without introgression events from *M. domestica* into wild species (Figure 4b), (iii) the model *c* included a single introgression event, from *M. sylvestris* into *M. domestica* (Figure 4c), and (iv) the model *d* simulated no admixture (Figure 4d). The number of parameters estimated in the model was limited by fixing the times of admixture with *M. orientalis*, *M. sylvestris* and *M. sieversii* at 600, 200 and 13 generations before the present, respectively. We used the following underlying hypotheses: (i) as the juvenile period of *Malus* lasts five to 10 years, we assumed a generation time of 7.5 years, (ii) admixture between ancestral *M. domestica* and *M. orientalis* in the Caucasus occurred approximately 4,500 years ago, shortly before the appearance of sweet apples in the Middle East (4,000 years ago), (iii) admixture between ancestral *M. domestica* and *M. sylvestris* in Europe occurred approximately 1,500 years ago, soon after the introduction of domesticated apples into Europe by the Greeks and Romans (iv) back-introgression into *M. sieversii* from *M. domestica* occurred approximately 100 years ago, when the cultivation of modern varieties reached Central Asia.

The relative posterior probabilities computed for each model provided strongest statistical support for model *c*, which assumed a single introgression event, from *M. sylvestris* into *M. domestica* (Table 4; posterior probability [*p*]=0.67, 95% confidence interval: 0.63-0.72). Note that the model without admixture (model *d*) had the lowest relative posterior probability (Table 4). In analyses under alternative admixture models (models *a* and *b*), the posterior distributions were flat for introgression between *M. domestica* and *M. orientalis* and highly skewed towards low values for introgression into *M. sylvestris* and *M. sieversii* (not shown), which is consistent with statistical support being highest for model *c*.

Given that the model *c* was clearly favoured, parameter estimates are shown below only for this model (Table 5; prior distributions in Table S4). The contribution of *M. sylvestris* to the *M. domestica* genepool was estimated at about 61% (95% credibility interval [95% CI]: 50-68%). We obtained estimates of effective population sizes of 3,520 (95% CI: 2,090-5,680) for *M. domestica*, 13,200 (95% CI: 6,920-19,300) for *M. sieversii*, 34,600 (95% CI: 15,100-48,000) for *M. sylvestris*, and 28,300 (95% CI: 11,700-64,000) for *M. orientalis*. Using a generation time of 7.5 years, the divergence between *M. domestica* and *M. sieversii* (*T3*) was estimated to have occurred 17,700 years ago (95% CI: 6,225-25,200), which is earlier than previously thought, but we note that the credibility interval is quite large. We estimated that *M. sylvestris* and *M. sieversii* diverged about 83,250 years ago (*T1*, 95% CI: 40,575-334,500), with *M. orientalis* and *M. sieversii* diverging about 20,775 years ago (*T2*, 95% CI: 9,900-47,775).

The results above were obtained using the full dataset. We checked the validity of our inferences by conducting analyses on the dataset without admixed and misclassified individuals and using different times of admixture, by assessing the goodness-of-fit of models to data, and by checking that sufficient power was achieved to discriminate among competing models (Text S1; Tables S5-S7). Overall, ABC analyses all provided clear support for a model with contribution of the European crabapple into the domesticates, although the estimated value of the actual contribution of *M. sylvestris* is probably overestimated here, and should therefore be treated with caution. Indeed, the simulation of a single introgression event hundreds of years ago most likely demanded higher rates of introgression to account for the actual genetic contribution of *M. sylvestris* into *M. domestica* than would be needed under continuous gene flow over a long period.

Weak genetic structure within *M. domestica:* linked to cultivar use or geography?

As cider cultivars produce apples that are smaller, more bitter and astringent than dessert cultivars, we expected to observe genetic differentiation between these two groups of cultivars and a closer genetic proximity of cider cultivars to *M. sylvestris* (Juniper, Mabberley, 2006; Wagner, Weeden, 2000). Neither hypothesis was supported by our data. The classification of apples into "dessert" and "cider" varieties as prior information for STRUCTURE (*Locprior* model) revealed a very weak tendency of cider and dessert cultivars to be assigned to different clusters at K=2 (Figure 5), but increasing K did not further result in clearer differentiation between the two types of cultivars. At K=2, *M. domestica* cider genotypes had a mean membership of 94.7%, and *M. domestica* dessert genotypes had a mean membership of 52.5%. However, STRUCTURE analyses without this prior information gave essentially the same clustering patterns at K=2 (G'=0.95 similarity to analyses using classification to assist clustering). The weak differentiation between cider and dessert cultivars (F_{ST}=0.02) and their high level of admixture in STRUCTURE analyses (Figure 5) indicated a shallow subdivision of the *M. domestica* genepool. Analyses on a pruned dataset from which closely related individuals had been removed (*i.e.*, pairs of genotypes with r_{xy}≥0.5; N=172) revealed the same pattern, confirming that the presence of related cultivars in the dataset did not bias clustering analyses. STRUCTURE was also run on a dataset including all *M. sylvestris* genotypes, to test the hypothesis that cider cultivars would display a higher level of introgression from the European crabapple. However, the opposite pattern was observed: the proportion of genotypes displaying introgression from *M. sylvestris* was actually significantly higher in dessert than in cider cultivars (36.4% and 15.5% respectively, χ^2=16.9, P=4×10^{-5}). Finally, little genetic differentiation was observed between groups of cultivars of different geographic origins (95% CI: -0.8 – 0.6, Table S8).

DISCUSSION

The apple is so deeply rooted in the culture of human populations from temperate regions that it is often not recognized as an exotic plant of unclear origin. We show here that the evolution of the domesticated apple involved more than one geographically restricted wild species. The domesticated apple did not arise from a single event over a short period of

time, but from evolution extending over thousands of years. The genepool of the current domesticated apple varieties has been enriched by the contribution of at least two wild species. *Malus* species have a self-incompatibility system; apple domestication and traditional variety improvement have therefore been based mostly on the selection of the best phenotypes grown from open-pollinated seeds. This breeding strategy has probably favoured the incorporation of genetic material from multiple wild sources and the maintenance of high levels of genetic variation in domesticated apples, despite the extensive use of large-scale vegetative propagation of superior individuals by grafting. Our results are consistent with those reported for the few other woody perennials studied to date, such as grape (Myles *et al.*, 2011), red mombin (Miller, Schaal, 2005) and olive trees (Besnard *et al.*, 2007), and support the view that domestication in long-lived plants differs in many respects from the scenarios described for seed-propagated annuals.

Weak differentiation from wild progenitors and the Central Asian origin of *M. domestica*

Malus sieversii was previously identified as the main contributor to the *M. domestica* genome on the basis of morphological and sequence data (Harris *et al.*, 2002; Velasco *et al.*, 2010). The flanks of the Tian Shan mountains have been identified as a likely initial site of domestication, based on the high morphological variability of the wild apples growing in this region, and their similarity to sweet dessert apples (Dzhangaliev, 2003; Vavilov, 1926). We show here, using a set of rapidly evolving genetic markers distributed throughout the genome and a large sampling, that *M. domestica* now forms a distinct, random mating group, surprisingly well separated from *M. sieversii*, with no difference in levels of genetic variation between the domesticate and its wild progenitor. This contrasts with the pattern previously reported, based on a twenty three-gene phylogenetic network (Velasco *et al.*, 2010), where domesticated varieties of apple appeared nested within *M. sieversii*. After the removal of individuals showing signs of recent admixture, *M. sieversii* and *M. domestica* nevertheless appeared to be the pair of species most closely related genetically, confirming their progenitor-descendant relationship.

Lack of a domestication bottleneck

Apple breeding methods (grafting and “chance seedling” selection), life-history traits

specific to trees and/or the genetic architecture of selected traits have likely played a role in the conservation of levels of genetic diversity in cultivated apples similar to those in wild apples. Some factors, such as “chance seedling” selection (Gardiner *et al.*, 2007), may even have increased genetic diversity, by favouring outcrossing events among domesticates and introgression from wild species (Zohary, Hopf, 2000). The low inbreeding coefficients inferred in domesticated apples and the low level of differentiation between cultivated and wild apple populations (Coart *et al.*, 2006; Coart *et al.*, 2003; Gharghani *et al.*, 2009; Koopman *et al.*, 2007) indicate a high frequency of crosses between individuals of *M. domestica*, *M. sieversii* and other wild relatives hailing from diverse geographic origins. Such a high level of gene flow has likely contributed to maintenance of a high level of genetic diversity in domesticated apples.

The grafting technique, which was probably developed around 3,000 years ago, has made it possible to propagate superior individuals clonally. The spread of grafting, together with the lengthy juvenile phase (5-10 years) and the long lifespan of apples, may have imposed strong limits on the intensity of the domestication bottleneck thereby limiting the loss of genetic diversity (Miller, Schaal, 2006; Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). By decreasing the number of generations since domestication, these factors have probably also helped to restrict the differentiation between domesticates and wild relatives. In theory, grafting may have limited the size of the apple germplasm dispersed early on to a few very popular genotypes, thereby provoking a sudden shrink in effective population size and a loss of diversity. However, we found no evidence that the clonal propagation of apples resulted in a long-lasting decrease in population size or clonal population structure. We can speculate that this may be due to a combination of various factors such as: gene flow with wild species, small-scale propagation (many farmers producing a few grafts each), a large variation in preferences for taste and other quality characteristics between farmers and cultures, large differences in growth conditions leading to the adoption of different sets of genotypes in different regions or the typical behaviour of hobby breeders, who tend to spot particular differences and multiply them. Similarly, for grape, there are huge numbers of old varieties and as much genetic variation in cultivated varieties as in wild-relative progenitors (Myles *et al.*, 2011).

A major secondary contribution from the European crabapple

There has been a long-running debate concerning the possible contribution of other wild species present along the Silk Route to the genetic makeup of *M. domestica* (Coart *et al.*, 2006; Forsline *et al.*, 2002; Ponomarenko, 1991; Rehder, 1940; Wagner, Weeden, 2000). Our results clearly show that interspecific hybridization has been a potent force in the evolution of domesticated apple varieties. Apple thus provides a rare example of the evolution of a domesticated crop over a long period of time and involving at least two wild species (see also the cases of olive tree and avocado (Chen *et al.*, 2009a; Miller, Schaal, 2005; Myles *et al.*, 2011; Olsen, Gross, 2008)). A recent study argued that introgression from *M. sylvestris* into the *M. domestica* genepool was the most parsimonious explanations for shared gene sequence polymorphisms between the two species (Harrison, Harrison, 2011). Using an unprecedentedly large dataset, more numerous and more rapidly evolving markers and a combination of inferential methods, we provide a comprehensive view of the history of domestication in apple. We confirm that *M. sieversii* was the initial progenitor and show that the wild European crabapple *M. sylvestris* has been a major secondary contributor to the diversity of apples, resulting in current varieties of *M. domestica* being more closely related to *M. sylvestris* than to their central Asian progenitor. This situation is reminiscent of that for maize, in which the cultivated crop *Zea mays* is genetically more closely related to current-day highland landraces than to lowland *Z. mays* ssp. *parviglumis* from which the crop was domesticated (van Heerwaarden *et al.*, 2011). This pattern has been attributed to large-scale gene flow from a secondary source, a second subspecies of teosinte, *Z. mays* ssp. *mexicana*, into highland maize populations (van Heerwaarden *et al.*, 2011).

The usefulness of wild relatives for improving elite cultivated crop genepools has long been recognised and the exploitation of wild resources is now considered a strategic priority in breeding and conservation programs for most crops (Brown *et al.*, 2009; Dzhangaliev, 2003; Feuillet *et al.*, 2008). Domesticated apples are unusual in that the contribution of wild relatives probably occurred early and unintentionally in the domestication process, preceding even the use of controlled crosses. The use of genetic markers with lower mutation rates than our set of microsatellites might also make it possible to investigate the

contribution of more phylogenetically distant apple species growing in areas away from the Silk Route to the diversification of modern apple cultivars.

The Romans introduced sweet apples into Europe at a time at which the Europeans were undoubtedly already making cider from the tannin-rich fruits of the native *M. sylvestris* (Juniper, Mabberley, 2006; Orton, 1973). Cider is not typical of Asia (Juniper, Mabberley, 2006), but it was widespread in Europe by the time of Charlemagne (9th century, (Lea, Piggott, 2003)). Large numbers of apple trees were planted for cider production in France and Spain from the 10th century onwards (Boré, Fleckinger, 1997; Pereira-Lorenzo *et al.*, 2009). The very high degree of stringency of cider apples (often to the extent that they are inedible) led to the suggestion that cider cultivars arose from hybridization between *M. sylvestris* and sweet apples (Forsline *et al.*, 2002; Juniper, Mabberley, 2006; Wagner, Weeden, 2000). We show here that the genetic structure within the cultivated apple genepool is very weak, with poor differentiation between cider and dessert apples. Cider cultivars thus appear to be no more closely genetically related to *M. sylvestris* than dessert cultivars. As wild Asian apples are known to cover the full range of tastes (Dzhangaliev, 2003; Forsline *et al.*, 2002), it is possible that fruits with the specific characteristics required for cider production were in fact initially selected in Central Asia and subsequently brought into Europe. There is a long-standing tradition of cider production in some parts of Turkey (Juniper, Mabberley, 2006), for instance, which is potentially consistent with an Eastern origin of cider cultivars. However, the low level of genetic differentiation between dessert and cider apples indicates that, even if different types of apples were domesticated in Asia and brought to Europe, they have not diverged into independent genepools.

Concluding remarks

This study settles a long-running debate by confirming that 1) *M. domestica* was initially domesticated from *M. sieversii*, and 2) *M. domestica* subsequently received a significant genetic contribution from *M. sylvestris*, much larger than previously suspected (Juniper, Mabberley, 2006), at least in Western Europe, where originated most of our samples and most cultivar diversity. The higher level of introgression of the European crabapple into the domesticated apple in this study than in previous studies (Harrison, Harrison, 2011; Micheletti *et al.*, 2011; Velasco *et al.*, 2010) may be attributed to the use of a

larger and more representative set of *M. domestica* genotypes coupled with the genotyping of numerous and rapidly evolving markers known to trace back more recent events.

Our inferences also have important implications for breeding programs and for the conservation of wild species of apple. The major contribution of the various wild species to the *M. domestica* genepool highlights the need to invest efforts into the conservation of these species, which may contain unused genetic resources that could further improve the domesticated apple germplasm (Hajjar, Hodgkin, 2007), such as disease resistance genes or genes encoding specific organoleptic features.

MATERIALS AND METHODS

Sample collection and DNA extraction. Leaf material was retrieved from the collections of various institutes (INRA Angers, France; USDA - ARS, Plant Genetic Resources Unit, Geneva, NY; ILVO Melle, Belgium) and from a private apple germplasm repository in Brittany for *M. domestica* (*N*=368, Figure S1 including only diploid cultivars *N*=299) and from forests for the four wild species (Figure 1; Table S1). *Malus sieversii* (*N*=168) material was collected from 2007 to 2010 in the Chinese Xinjiang province (*N*=26), Kyrgyzstan (*N*=5), Uzbekistan (*N*=1), Tajikistan (*N*=1) and Kazakhstan (*N*=114). *Malus orientalis* (*N*=215) was sampled in 2009 in Armenia (*N*=203), Turkey (*N*=5) and Russia (*N*=5). *Malus sylvestris* (*N*=40) samples were obtained from 15 European countries. *Malus baccata* (*N*=48) was sampled in 2010 in Russia. The origins of *M. domestica* cultivars were: France (*N*=266), Great Britain (*N*=12), USA (*N*=12), Russia (*N*=7), the Netherlands (*N*=6), Australia (*N*=4), Belgium (*N*=4), Germany (*N*=4), Japan (*N*=3), Ukraine (*N*=3), Tunisia (*N*=2), Switzerland (*N*=2), Spain (*N*=2), New Zealand (*N*=2), Israel (*N*=1), Ireland (*N*=1), Canada (*N*=1), Armenia (*N*=2) and unknown/debated (*N*=34). Genomic DNA was extracted with the Nucleo Spin® plant DNA extraction kit II (Macherey & Nagel, Düren, Germany) according to the manufacturer's instructions.

Microsatellite markers and polymerase chain reaction (PCR) amplification. Microsatellites were amplified by multiplex PCR, with the Multiplex PCR Kit (QIAGEN, Inc.). We used 26 microsatellites spread across the 17 chromosomes (one to three microsatellites per chromosome), in 10 different multiplexes previously optimised on a large set of genetically related progenies of *M. domestica* (Patocchi *et al.*, 2009). The four multiplexes (MP01,

MP02, MP03, MP04; Table S9; Lasserre P. unpublished data) were performed in a final reaction volume of 15 µl (7.5 µl of QIAGEN Multiplex Master Mix, 10-20 µM of each primer, with the forward primer labelled with a fluorescent dye and 10 ng of template DNA). We used a touch-down PCR program (initial annealing temperature of 60°C, decreasing by 1°C per cycle down to 55°C). Six other multiplex reactions (Hi6, Hi4a^{b}, Hi5-10, Hi13a, Hi13b, Hi4b) were performed using previously described protocols (Patocchi *et al.*, 2009). Genotyping was performed on an ABI PRISM X3730XL, with 2 µl of GS500LIZ size standard (Applied Biosystems). Alleles were scored with GENEMAPPER® 4.0 software (Applied Biosystems). We retained only multilocus genotypes presenting less than 30% missing data.

Suitability of microsatellites for population genetic analyses. We checked the suitability of the markers for population genetic analyses. None of the 26 microsatellite markers deviated significantly from a neutral equilibrium model, as shown by the non significant *P*-values obtained in Ewen-Watterson tests (Excoffier, Lischer, 2010), and no pair of markers was found to be in significant linkage disequilibrium in any of the species (Raymond, Rousset, 1995; Rousset, 2008). The markers could therefore be considered unlinked and neutral.

Analyses of genetic variation and differentiation between the five species. Apple cultivars may be polyploid (Schuster, Büttner, 1995). We therefore first checked for the presence of polyploidy individuals of *M. domestica* within our dataset. Individuals presenting multiple peaks on electrophoregrams were first re-extracted to eliminate contamination as a possible source of apparent polyploidy. We then checked whether they had been reported to be polyploidy in previous studies (Schuster, Büttner, 1995). After completion of this checking procedure, we removed 69 polyploids (of the 368 samples) from subsequent analyses. We tested for the occurrence of null alleles at each locus with MICROCHECKER 2.2.3 software (Van Oosterhout *et al.*, 2004). Allelic richness and private allele frequencies were calculated with ADZE software (Szpiech *et al.*, 2008), for a sample size of 22. Heterozygosity (expected (H_E) and observed (H_O)), Weir & Cockerham *F*-statistics, deviation from Hardy-Weinberg equilibrium and genotypic linkage disequilibrium were estimated with GENEPOP 4.0 (Raymond, Rousset, 1995; Rousset, 2008). The significance of differences between F_{ST} values was assessed in exact tests carried out with GENEPOP 4.0 (Raymond, Rousset, 1995; Rousset, 2008). Individuals were assigned to clonal lineages with GENODIVE (Meirmans, Van

Tienderen, 2004). We estimated relatedness between pairs of cultivars and between pairs of individuals within each species, by calculating the r_{xy} of Ritland and Lynch (Lynch, Ritland, 1999) with RE-RAT online software (Schwacke *et al.*, 2005). We tested whether the distributions of r_{xy} deviated significantly from a Gaussian distribution with a mean of zero and a standard deviation equal to the observed standard deviation, by comparing observed and simulated distributions in Fisher's exact test (R Development Core Team, URL http://www.R-project.org).

Assessing bottlenecks during apple domestication and diversification. We tested for the occurrence of a bottleneck during apple domestication with the method implemented in BOTTLENECK (Cornuet, Luikart, 1996; Piry *et al.*, 1999). The tests were performed under the stepwise-mutation model (SMM) and under a two-phase model (TPM) allowing for 30% multistep changes. We used Wilcoxon signed rank tests to determine whether a population had a significant number of loci with excess genetic diversity.

Analyses of population subdivision. We used the individual-based Bayesian clustering method implemented in STRUCTURE 2.3.3 (Falush *et al.*, 2003; Hubisz *et al.*, 2009; Pritchard *et al.*, 2000) to investigate species delimitation, intraspecific population structure and admixture. This method is based on Markov Chain Monte Carlo (MCMC) simulations and is used to infer the proportion of ancestry of genotypes in *K* distinct predefined clusters. The algorithm attempts to minimize deviations from Hardy–Weinberg and linkage equilibrium within clusters. Analyses were carried out without the use of prior information, except for analyses of population subdivision within the *M. domestica* genepool for which the "cider"/"dessert" classification of cultivars was used as prior information to assist clustering. *K* ranged from 1 to 8 for analyses of the five-species dataset and the *M. domestica* dataset, and was fixed at *K*=2 for analyses of pairs of species including *M. domestica* and each of the wild species. Ten independent runs were carried out for each *K* and we used 500,000 MCMC iterations after a burn-in of 50,000 steps. We used CLUMPP v1.1.2 (Greedy algorithm) (Jakobsson, Rosenberg, 2007) to look for distinct modes among the 10 replicated runs of each *K*.

STRUCTURE analyses were run for the full dataset (*N*=839) and for two pruned datasets excluding non-pure individuals (*i.e.*, genotypes with <0.9 membership of their

species' genepool) and related individuals (r_{xy}≥0.5).

Inference of demographic history. We used the DIYABC program (Cornuet *et al.*, 2010) to compare different admixture models and infer historical parameters. We simulated microsatellite datasets for 14 loci (Ch01h01, Ch01h10, Ch02c06, Ch02d08, Ch05f06, Ch01f02, Hi02c07, Ch02c09, Ch03d07, Ch04c07, Ch02b03b, MS06g03, Ch04e03, Ch02g01) previously reported to be of the perfect repeat type (Gianfranceschi *et al.*, 1998; Liebhard *et al.*, 2002; Silfverberg-Dilworth *et al.*, 2006). In total, we generated 5 x 10^5 simulated datasets for each model.

A generalized stepwise model (GSM) was used as the mutational model. The model had two parameters: the mean mutation rate (μ) and the mean parameter (P) of the geometric distribution used to model the length of mutation events (in numbers of repeats). As no experimental estimate of microsatellite mutation rate is available for *Malus*, the mean mutation rate was drawn from a uniform distribution by extreme values of 10^{-4} and 10^{-3}, and the mutation rate of each locus was drawn independently from a Gamma distribution (mean=μ; shape=2). The parameter P ranged from 0.1 to 0.3. Each locus L had a possible range of 40 contiguous allelic states (44 for *CH02C06*, 42 for *CH04E03*) and was characterized by individual values for mutation rate (μ_L) and the parameter of the geometric distribution (P_L); μ_L and P_L were drawn from Gamma distributions with the following parameter sets: mean=μ, shape=2, range=5 x 10^{-5}- 5 x 10^{-2} for μ_L, and mean=P, shape=2, range=0.01-0.9 for P_L. As not all allele lengths were multiples of motif length, we also included single-nucleotide insertion-deletion mutations in the model, with a mean mutation rate (μ_{SNI}) and locus-specific rates drawn from a Gamma distribution (mean=μ_{SNI}; shape=2). The summary statistics used were: mean number of alleles per locus, mean genetic diversity (Nei, 1978), genetic differentiation between pairwise groups (F_{ST}; (Weir, Cockerham, 1984)), genetic distances $(\delta\mu)^2$ (Goldstein *et al.*, 1995).

We used a polychotomous logistic regression procedure (Fagundes *et al.*, 2007) to estimate the relative posterior probability of each model, based on the 1% of simulated data sets closest to the observed data. Confidence intervals for the posterior probabilities were computed using the limiting distribution of the maximum likelihood estimators (Cornuet *et al.*, 2008). Once the most likely model was identified, we used a local linear regression to estimate the posterior distributions of parameters under this model (Beaumont *et al.*, 2002).

The 1% simulated datasets most closely resembling the observed data were used for the regression, after the application of a *logit* transformation to parameter values.

Author contributions

Conceived and designed the experiments: TG PG. Performed the experiments: AC. Analyzed the data: AC PG. Contributed reagents/materials/analysis tools: AC PG MJMS IR-R FL BLC AN JC MO LF IG X-GZ TG. Wrote the paper: AC PG MJMS MIT TG. Searched for funding: TG PG AC MIT. Wrote grant proposals: TG PG AC MIT

Acknowledgements

We thank Jacqui A. Shykoff, Rémy Petit, Jérome Enjalbert, Domenica Manicacci, Thierry Robert, Gwendal Restoux and three anonymous reviewers for helpful suggestions and comments. We thank *Plateforme de Génotypage* GENTYANE INRA UMR 1095 *Génétique Diversité et Ecophysiologie des Céréales* for genotyping data, Pauline Lasserre for help with genotyping methods, Aurélien Tellier, Virginie Ravigné, Peter Beerli and Daniel Wegmann for help with, and advice on, data analysis. We thank Eric van de Weg for assistance in the determination of cultivar pedigrees. We thank the following for sampling and providing access to samples: Catherine Peix, Aymar Dzhangaliev and collaborators in Kazakhstan, Evelyne Heyer (*Museum National d'Histoire Naturelle*, France), Marie-Anne Félix (*Institut Jacques Monod*, France) and Emmanuelle Jousselin (*Centre de Biologie et de Gestion des Populations*, France) for *M. sieversii* samples; Dominique Beauvais (*Abbaye de Beauport*, Paimpol, France) and Jean Pierre Roullaud (*Verger Conservatoire d'Arzano*, France) for providing *M. domestica* samples; Ara Hovhannisyan, Karen Manvelyan and Eleonora Gabrielian for *M. orientalis* samples; Alberto Dominici and Emanuela Fabrizi (Monti Simbruini Regional Park, Italy), Jan Kowalczyk and Dzmitry Kahan (Forest Research Institute, Poland), Jörg Kleinschmit and Wilfried Steiner (Northwest German Forest Research Institute, Germany), Thomas Kirisits and Bernhard Kirisits (Institute of Forest Entomology, Forest Pathology and Forest Protection, Vienna, Austria), Heino Konrad (Federal Research and Training Centre for Forests, Vienna, Austria), Dalibor Ballian (Faculty of Forestry, University of Sarajevo, Bosnia-Herzegovina), Petya Gercheva, Argir Zhivondov, Valentina Bojkova and Anna Matova (Fruit-Growing Institute,

Plovdiv, Bulgaria), Anders Larsen (Forest and Landscape, Department for Management of Forest Genetic Resources, Denmark), Stephens Cavers (NERC, Centre for Ecology and Hydrology, UK), Lazlo Nyari (Forest Genetics and Forest Tree Breeding, Göttingen, Germany), Per Avid (Agder Natural History Museum and Botanical Garden, Norway), Lucian Curtus (Transylvania University Brasov, Faculty of Forest Sciences, Romania), Carlos Herrera (Estacion Biologica de Donana, CSIC, Spain), Francisco Donaire (Jardín Botánico La Cortijuela, Sierra Nevada, Granada, Spain), Roman Volansyanchuk (Ukrainian Research Institute of Forestry and Forest Melioration, Ukraine) for providing *M. sylvestris* samples; Ilya Zakarov, Natalia Badmayeva, Irina Kreshchenok for providing *M. baccata* samples. We also thank Thierry Genevet, Frédéric Tournay (*Jardin Botanique de Strasbourg*), Levente Kiss, the East Malling Research Station (UK), Philip Forsline and the Plant Genetic Resources Unit in Geneva (NY).

Figures

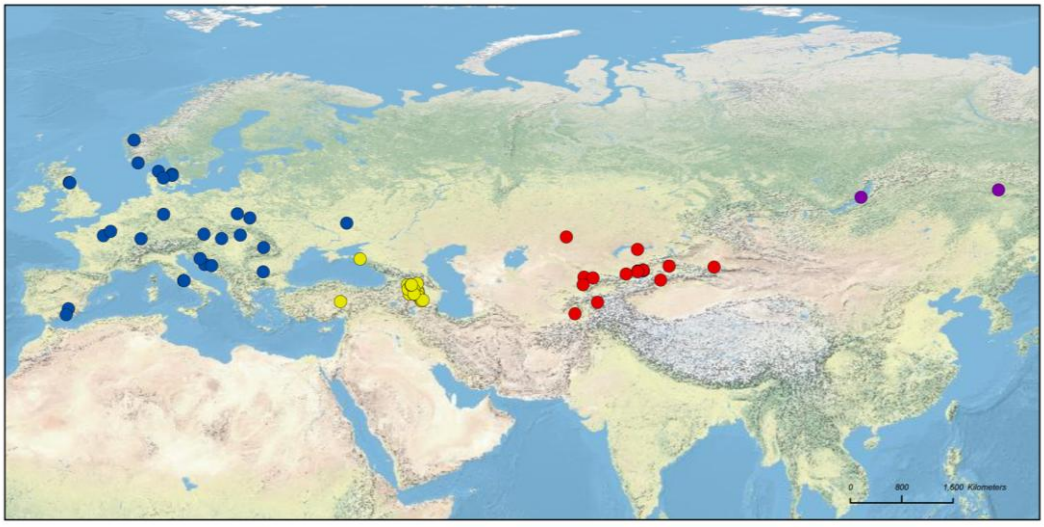


Figure 1. Geographic origins of the samples of the four wild *Malus* species used: *M. sylvestris* (blue), *M. orientalis* (yellow), *M. baccata* (purple) and *M. sieversii* (red). Samples of unknown origin (*N*=28) were not projected onto the map.

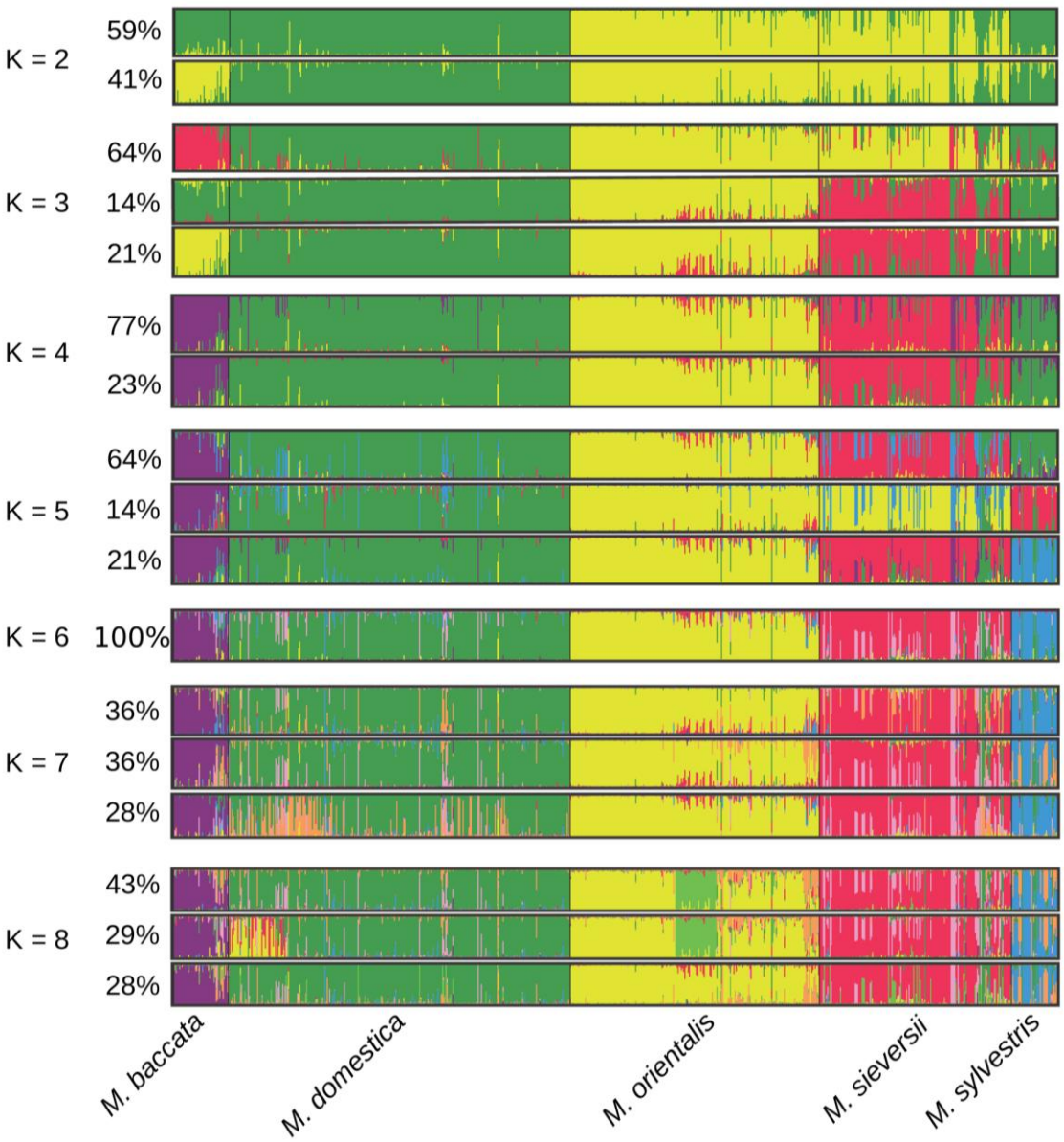


Figure 2. Proportions of ancestry of *Malus* genotypes from five species (*N*=770) from *K*=2 to *K*=8 ancestral genepools ("clusters") inferred with the STRUCTURE program. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome in *K* clusters. When several clustering solutions ("modes") were represented within replicate runs, the proportion of simulations represented by each mode is given.

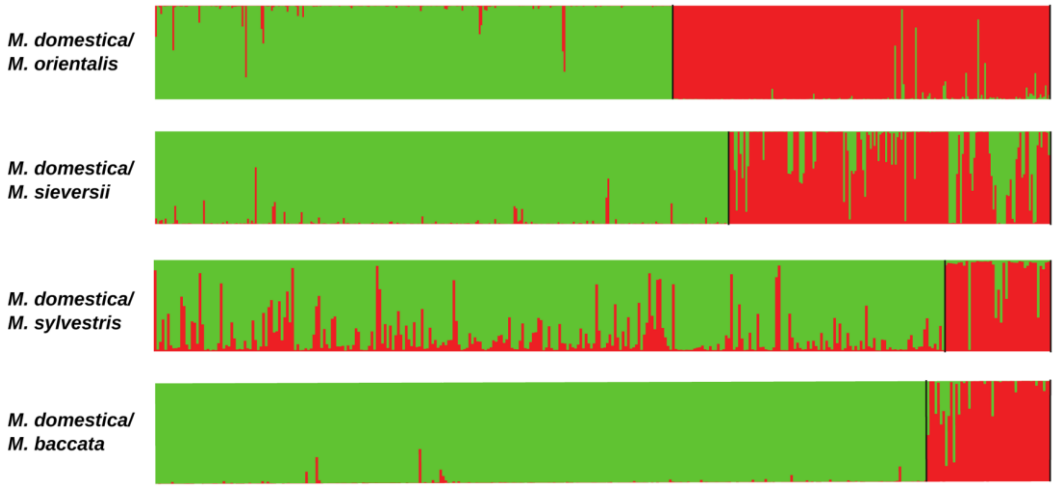


Figure 3. Proportions of ancestry in two ancestral genepools inferred with the STRUCTURE program, based on datasets including *M. domestica* (green, *N*=299) and each of the four wild *Malus* species (red). The x-axis is not to scale (details in Table S2).

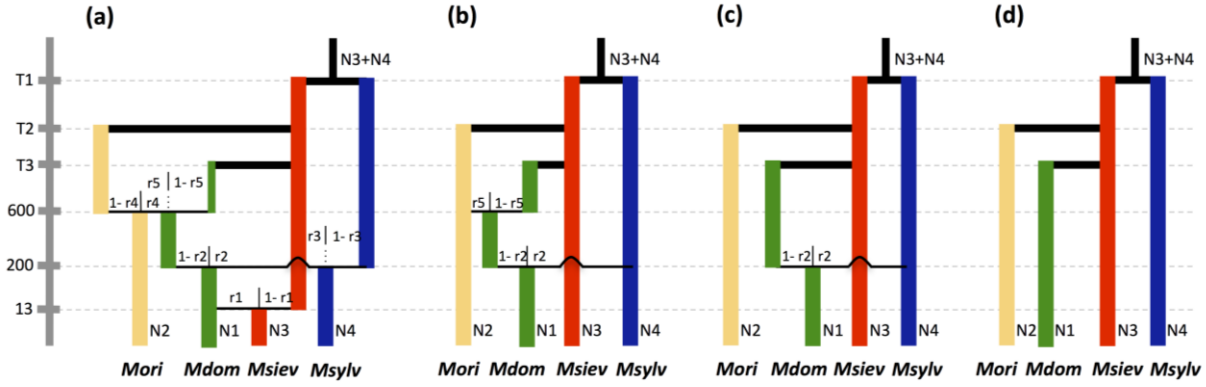


Figure 4. Admixture models compared in approximate Bayesian computations. Model *a* assumes that *M. domestica* is derived from *M. sieversii* and that the ancestral *M. domestica* population was involved in reciprocal introgression events with *M. orientalis* and *M. sylvestris*, and subsequently introgressed back into *M. sieversii*. Model *b* assumes no introgression from *M. domestica* into wild species, model *c* assumes the only admixture event is from *M. sylvestris* into *M. domestica*, and model *d* assumes no admixture. Admixture times between *M. domestica* and the three wild species were fixed (see text). Abbreviations: *Nk*, effective population sizes; *Tk*, divergence times; *r1*, *r3*, *r4* introgression from *M. domestica* into *M. sieversii*, *M. sylvestris*, and *M. orientalis* respectively; *r2*, *r5* introgression from *M. sylvestris* and *M. orientalis,* respectively, into *M. domestica*.

Figure 5. Proportions of ancestry of *M. domestica* genotypes (cider and dessert apples) in two ancestral genepools inferred with the STRUCTURE program.

Tables

Table 1. Summary of genetic variation in the five *Malus* species

	H_O	H_E	F_{IS}	A_r	A_p	A_p*	P_{NA}
M. domestica	0.81	0.83	0.02***	8.0	0.8	1.2	0.02
M. sieversii	0.77	0.82	0.07***	8.0	1.1	1.2	0.03
M. sylvestris	0.75	0.87	0.14***	9.9	1.7	2.5	0.02
M. orientalis	0.79	0.84	0.06***	8.8	2.1	1.9	0.03
M. baccata	0.56	0.75	0.24***	7.8	1.4	2.1	0.12

H_O and H_E: observed and expected heterozygosity, respectively, F_{IS}: inbreeding coefficient, A_r and A_p: allelic richness and private allele richness averaged across loci, respectively, estimated by rarefaction using a standardized sample size of 22, A_p*: private allele richness averaged across loci using the pruned dataset without hybrids in both wild and cultivated species, estimated by rarefaction using a standardized sample size of 12; P_{NA}: proportion of null alleles, ***: *P*-value<0.0001

Table 2. Pairwise differentiation (F_{ST}) between the five *Malus* species

	M. baccata	*M. sylvestris*	*M. domestica*	*M. sieversii*
M. sylvestris	0.1683	-	-	-
M. domestica	0.1505	0.0056	-	-
M. sieversii	0.1457	0.0818	0.0639	-
M. orientalis	0.1337	0.0579	0.0494	0.0393

All F_{ST} values were significant (*P*<0.001)

Table 3. Mean proportions of assignment to each of the two species in species pair comparisons (K=2) including *M. domestica* (Genepool 1) and each of the four wild *Malus* species (Genepool 2).

Species pairs	Genepool 1	Genepool 2
M. domestica	0.841	0.159
M. sylvestris	0.119	0.881
M. domestica	0.993	0.007
M. baccata	0.104	0.896
M. domestica	0.980	0.020
M. orientalis	0.030	0.970
M. domestica	0.981	0.019
M. sieversii	0.231	0.769

Table 4. Relative posterior probabilities (*p*) for the four historical models compared using approximate Bayesian computations. Models are described in Figure 4. CI2.5 and CI97.5 are boundaries of the 95% confidence intervals.

Model	***p***	**CI2.5**	**CI97.5**
a	0.0349	0.0253	0.0445
b	0.2819	0.2509	0.3130
c	0.6832	0.6504	0.7159
d	0.0000	0.0000	0.0000

Table 5. Demographic and mutation parameters estimated using approximate Bayesian computation for model *c*. Posterior distributions are summarized as the mode and boundaries of the 95% credibility intervals (CI2.5 and CI97.5). Demographic parameters are introduced in Figure 4 (note that admixture times are fixed in these analyses). Composite parameters scaled by the mutation rate are also shown. The mutation parameters are μ (mean mutation rate), *p* (mean value of the geometric distribution parameter that governs the number of repeated motifs that increase or decrease the length of the locus during mutation events), *μSNI* (mean single nucleotide indel mutation rate). Species names are abbreviated.

Parameter	**Mode**	**CI2.5**	**CI97.5**
N1 (*M. dom*)	3,520	2,090	5,680
N2 (*M. ori*)	28,300	11,700	64,000
N3 (*M. siev*)	13,200	6,920	19,300
N4 (*M. sylv*)	34,600	15,100	48,000
T1 (*M. siev - M.sylv*)	11,100	5,410	44,600
T2 (*M. siev - M. ori*)	2,770	1,320	6,370
T3 (*M. siev - M. dom*)	2,360	830	3,360
r2 (introgr. by *M. sylv* into *M. dom*)	0.61	0.50	0.68
μ	$2.0.10^{-4}$	$1.1.10^{-4}$	$6.9.10^{-4}$
p	0.3	0.1	0.3
μSNI	$3.0.10^{-8}$	$5.0.10^{-8}$	$5.9.10^{-5}$
$\vartheta 1$ (=4*N1*μ)	0.7	0.5	2.5
$\vartheta 2$ (=4*N2*μ)	6.1	3.1	23.6
$\vartheta 3$ (=4*N3*μ)	2.8	1.9	7.4
$\vartheta 4$ (=4*N4*μ)	6.8	4.4	18.1
$\tau 1$ (=μ*T1*)	2.28	1.30	16.10
$\tau 2$ (=μ*T2*)	0.54	0.31	2.38
$\tau 3$ (=μ*T3*)	0.41	0.19	1.55

Supporting Information

Table S1. Description of the *Malus* species accessions analysed, with their geographic origin and providers.

Figure S1. Geographic origins of diploid *Malus domestica* cultivars (*N*=299).

Figure S2. Distribution of pairwise relatedness coefficients among the *Malus domestica* cultivars.

Table S2. *Malus domestica* cultivars used in the study, with their use, provider, geographic putative origin.

Figure S3. Proportions of ancestry in two ancestral genepools inferred with the STRUCTURE program from datasets including *Malus domestica* and each of the wild *Malus* species except *Malus baccata*.

Table S3. Membership coefficients inferred from the STRUCTURE analysis for *Malus baccata* individuals.

Table S4. Prior distributions used in approximate Bayesian computations.

Text S1. Method used for approximate Bayesian computations on alternative datasets/admixture times.

Table S5. Relative posterior probabilities (*p*) for the four historical models compared using approximate Bayesian computations.

Table S6. Demographic and mutation parameters estimated using approximate Bayesian computation for model *c*.

Table S7. Model checking based on comparisons of test quantities between observed data and 100 pseudo-observed datasets generated using parameter values drawn from posterior distributions.

Table S8. Genetic differentiation (F_{ST}) between cultivars of different geographic origins (*N*=266). Cultivars of unknown origin have been removed.

Table S9. Description of the Multiplex PCRs (MP01, MP02, MP03, MP04) used for microsatellite amplification.

Table S1. Description of the *Malus* species accessions analysed, with their geographic origin and providers.

Species	Nb*	Provider
Malus sylvestris	40	
Austria	2	Thomas Kirisits, Bernhard Kirisits and Heino Konrad
Belgium	4	IRR, CRA-W[1], ILVO[2]
Bosnia-Herzegovina	3	Dalibor Ballian
Bulgaria	1	Petya Gercheva, Argir Zhivondov, Valentina Bojkova and Anna Matova
Denmark	6	Anders Larsen
France	5	INRA[3], USDA-ARS[4]
Germany	5	Jorg Kleinschmit and Wilfried Steiner
UK, Scotland	2	Stephens Cavers
Hungary	2	Lazlo Nyari
Italy	2	Alberto Dominicci and Emanuela Fabrizi
Norway	2	Per Avid
Poland	2	Jan Kowalsky and Dzmitry Kahan
Romania	1	Lucian Curtus
Spain	2	Carlos Ferrera and Francisco Donaire
Ukraine	1	Roman Volansyanchuk
Malus sieversii	168	
Kazakhstan	114	BLC, FL, PG, Emmanuelle Jousselin, Marie-Anne Félix, Catherine Peix and Aymar Dzhangaliev
	28	USDA-ARS[4]
China	26	BLC, PG, XGZ
Kirghizstan	5	Evelyne Heyer
Tajikistan	1	USDA-ARS[4]
Uzbekistan	1	USDA-ARS[4]
Malus orientalis	215	
Armenia	203	PG, JC, AN, IG, Ara Hovhannisyan, Karen Manvelyan and Eleonora Gabrielian

Russia	5	USDA-ARS[4]
Turkey	5	USDA-ARS[4]
Unknown	2	USDA-ARS[4]
M. baccata	48	
Unknown	10	USDA-ARS[4], EMR[6] , ILVO[2]
Russia (Transbaikal Region)	36	MO, Ilya Zakarov, Natalia Badmayeva and Irina Kreshchenok
Romania	1	ILVO[2]
Hungary	1	ILVO[2]
Malus domestica	368	INRA[3], CRA-W[1], USDA-ARS[4], Dominique Beauvais [5] and Jean Pierre Roullaud[7]
Diploid	299	
Unknown	28	INRA[3], CRA-W[1], USDA-ARS[4]
Australia	3	INRA[3]
Belgium	4	INRA[3], CRA - W[1]
Canada	1	INRA[3]
France	209	INRA[3], CRA-W[1], Dominique Beauvais [5] and Jean Pierre Roullaud[7]
Germany	4	INRA[3], CRA - W[1]
Great Britain	9	INRA[3], CRA - W[1]
Ireland	1	USDA-ARS[4]
Israel	1	INRA[3]
Japan	2	INRA[3]
Netherlands	6	INRA[3], USDA-ARS[4]
New Zealand	2	INRA[3]
Russia	7	INRA[3], USDA-ARS[4]
Spain	2	INRA[3]
Switzerland	2	INRA[3]

Tunisia	2	INRA[3]
Ukraine	3	INRA[3], CRA-W[1]
USA	11	INRA[3]
Armenia	2	PG, JC, AN, IG, Ara Hovhannisyan, Karen Manvelyan and Eleonora Gabrielian
Triploid	**69**	
Unknown	6	INRA[3]
Australia	1	INRA[3]
France	58	INRA[3], Dominique Beauvais [5] and Jean Pierre Roullaud[7]
Great Britain	2	INRA[3]
Japan	1	INRA[3]
USA	1	INRA[3]
Malus domestica		
Cider	119	
Desserts	180	

*Number of trees sampled

[1] CRA - W	Centre Wallons de Recherches Agronomiques, Belgium
[2] ILVO - PLANT	Plant -Growth and Development, Melle, Belgium
[3] INRA	Institut de Recherche en Horticulture et Semences, Angers, France
[4] USDA - ARS	Plant Genetic Resources Unit, Geneva (NY)
[5] Abbaye de Beauport	Conservatory Orchards of ancient apple varieties, Paimpol, France.
[6] EMR	East Malling Research, Kent, UK
[7]Verger Conservatoire d'Arzano	Conservatory Orchards of ancient apple varieties, Brittany, France.

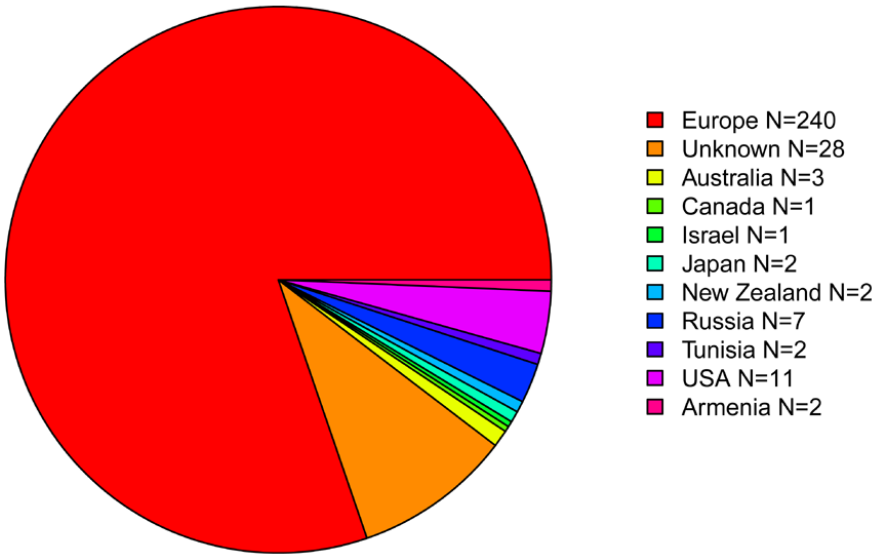


Figure S1. Geographic origins of diploid *Malus domestica* cultivars (*N*=299).

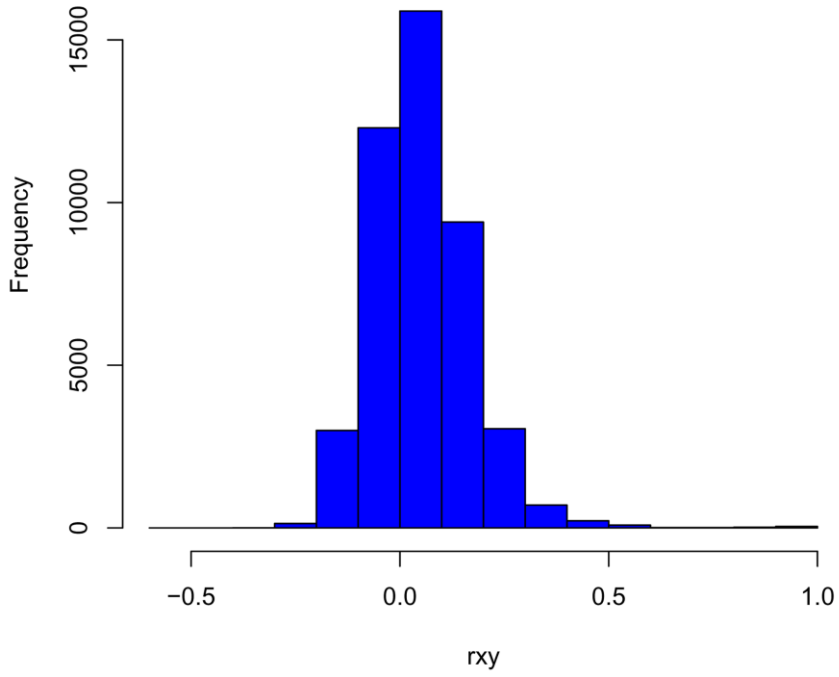


Figure S2. Distribution of pairwise relatedness coefficients (Lynch, Ritland, 1999) among the *Malus domestica* cultivars. r_{xy} values among cultivars are normally distributed around a mean of zero, with a low variance between pairs of cultivars (Fisher's exact test, $P \approx 1$).

Table S2. *Malus domestica* cultivars used in the study, with their use, provider, geographic putative origin. Details of the STRUCTURE analysis summarized in Table 3 are also provided.

Sample	Use	Geographical Origin	Provider	*M. domestica*	*M. sylvestris*	*M. domestica*	*M. baccata*	*M. domestica*	*M. sieversii*	*M. domestica*	*M. orientalis*
borovitsky	dessert	Russia	INRA	0.067	0.933	0.001	0.999	0.94	0.06	0.669	0.331
reinette ananas	dessert	?	INRA	0.98	0.02	0.004	0.996	0.997	0.003	0.998	0.002
udarre zagarra	dessert	France	INRA	0.887	0.113	0.102	0.898	0.964	0.036	0.984	0.016
usta gorria	dessert	France	INRA	0.628	0.352	0.001	0.999	0.956	0.044	0.912	0.088
orleans	dessert	Netherlands	INRA	0.983	0.017	0.001	0.999	0.996	0.004	0.998	0.002
camuesa verde	dessert	Spain	INRA	0.585	0.415	0.049	0.951	0.942	0.058	0.993	0.007
saint-bernard d29	dessert	France	CRA - W	0.934	0.066	0.001	0.999	0.997	0.003	0.996	0.004
purpurroter cousinot c21	dessert	Germany	CRA - W	0.966	0.034	0.003	0.997	0.997	0.003	0.997	0.003
pigeonnet a12	dessert	France	CRA - W	0.964	0.036	0.001	0.999	0.997	0.003	0.998	0.002
mc intosh	dessert	Canada	INRA	0.955	0.045	0.001	0.999	0.992	0.008	0.997	0.003
douce de sfax	dessert	?	INRA	0.397	0.603	0.001	0.999	0.804	0.196	0.523	0.477
calville duquesne b14	dessert	Belgium	CRA - W	0.504	0.496	0.001	0.999	0.958	0.042	0.984	0.016
pun d'boche b56	dessert	?	CRA - W	0.926	0.074	0.001	0.999	0.995	0.005	0.996	0.004
keuleman jaune d15	dessert	?	CRA - W	0.978	0.022	0.114	0.886	0.997	0.003	0.999	0.001
reinette rouge etoilée b11	dessert	Belgium	CRA - W	0.669	0.331	0.002	0.998	0.994	0.006	0.993	0.007
drap d'or a40	dessert	France	CRA - W	0.917	0.083	0.002	0.998	0.996	0.004	0.998	0.002
grand alexandre a34	dessert	Ukraine	CRA - W	0.147	0.853	0.001	0.999	0.992	0.008	0.99	0.01
président damseaux b17	dessert	?	CRA - W	0.702	0.298	0.001	0.999	0.993	0.007	0.99	0.01
keiing.rode d13	dessert	?	CRA - W	0.984	0.016	0.012	0.988	0.997	0.003	0.998	0.002
pladei d40	dessert	?	CRA - W	0.972	0.028	0.003	0.997	0.997	0.003	0.998	0.002
reinette de geer d22	dessert	Belgium	CRA - W	0.989	0.011	0.001	0.999	0.997	0.003	0.998	0.002
maréchal d62	dessert	?	CRA - W	0.981	0.019	0.352	0.648	0.997	0.003	0.998	0.002
belle du bois c44	dessert	?	CRA - W	0.987	0.013	0.001	0.999	0.997	0.003	0.999	0.001

blanc braibant s02	dessert	?	CRA - W	0.909	0.091	0.001	0.999	0.997	0.003	0.997	0.003
api k01	**dessert**	**France**	**CRA - W**	**0.255**	**0.745**	**0.029**	**0.971**	**0.744**	**0.256**	**0.819**	**0.181**
streeping q16	dessert	?	CRA - W	0.97	0.03	0.001	0.999	0.994	0.006	0.997	0.003
belle fontaine s01	dessert	?	CRA - W	0.932	0.068	0.001	0.999	0.995	0.005	0.998	0.002
apez sagarra	dessert	France	INRA	0.945	0.055	0.002	0.998	0.995	0.005	0.998	0.002
rambour de himbsel k31	dessert	?	CRA - W	0.68	0.32	0.001	0.999	0.991	0.009	0.997	0.003
reinette struel k39	dessert	?	CRA - W	0.857	0.143	0.003	0.997	0.997	0.003	0.998	0.002
patte de loup	dessert	France	INRA	0.979	0.021	0.275	0.725	0.997	0.003	0.998	0.002
saint-omer k48	dessert	France	CRA - W	0.972	0.028	0.001	0.999	0.995	0.005	0.995	0.005
type eisdener meulemans o32	dessert	?	CRA - W	0.978	0.022	0.022	0.978	0.997	0.003	0.998	0.002
franc bon pommier	dessert	France	INRA	0.883	0.117	0.005	0.995	0.994	0.006	0.996	0.004
idared	dessert	USA	INRA	0.965	0.035	0.333	0.667	0.984	0.016	0.991	0.009
golden delicious	dessert	USA	INRA	0.976	0.024	0.027	0.973	0.997	0.003	0.998	0.002
reinette etoilee	dessert	Netherlands	INRA	0.659	0.341	0.81	0.19	0.97	0.03	0.966	0.034
rome beauty	dessert	USA	INRA	0.876	0.124	0.011	0.989	0.994	0.006	0.993	0.007
calville d'angleterre	dessert	Great Britain	INRA	0.985	0.015	0.347	0.653	0.996	0.004	0.998	0.002
calville blanc	dessert	Switzerland/Germany	INRA	0.946	0.054	0.845	0.155	0.998	0.002	0.998	0.002
grand alexandre	dessert	Ukraine	INRA	0.582	0.418	0.298	0.702	0.996	0.004	0.997	0.003
james grieve	dessert	Scotland	INRA	0.943	0.057	0.431	0.569	0.99	0.01	0.994	0.006
reinette de landsberg	dessert	Germany	INRA	0.52	0.48	0.015	0.985	0.997	0.003	0.998	0.002
bismark	dessert	Australia	INRA	0.44	0.56	0.326	0.674	0.992	0.008	0.975	0.025
romarin blanc	dessert	?	INRA	0.793	0.207	0.002	0.998	0.996	0.004	0.99	0.01
winesap	dessert	USA	INRA	0.778	0.222	0.002	0.998	0.992	0.008	0.991	0.009
cox's orange pippin	dessert	Great Britain	INRA	0.454	0.546	0.536	0.464	0.994	0.006	0.995	0.005
benoni	dessert	USA	INRA	0.875	0.125	0.988	0.012	0.943	0.057	0.996	0.004
democrat	dessert	?	INRA	0.628	0.372	0.999	0.001	0.993	0.007	0.994	0.006

reinette bergamotte	dessert	Russia	INRA	0.343	0.657	0.998	0.002	0.963	0.037	0.778	0.222
worcester pearmain	**dessert**	**Great Britain**	**INRA**	**0.683**	**0.317**	**0.993**	**0.007**	**0.99**	**0.01**	**0.996**	**0.004**
ajmi	dessert	?	INRA	0.089	0.911	0.999	0.001	0.386	0.614	0.235	0.765
newton pippin	dessert	USA	INRA	0.985	0.015	0.997	0.003	0.996	0.004	0.998	0.002
reine des reinettes	dessert	Germany	INRA	0.922	0.078	0.999	0.001	0.998	0.002	0.999	0.001
melrose	dessert	USA	INRA	0.987	0.013	0.998	0.002	0.998	0.002	0.999	0.001
jonathan	dessert	USA	INRA	0.984	0.016	0.999	0.001	0.996	0.004	0.998	0.002
millers seedling	dessert	?	INRA	0.965	0.035	0.999	0.001	0.997	0.003	0.998	0.002
reinette du mans	dessert	France	INRA	0.984	0.016	0.849	0.151	0.997	0.003	0.998	0.002
chantecler	dessert	France	INRA	0.967	0.033	0.991	0.009	0.986	0.014	0.997	0.003
akane	dessert	Japan	INRA	0.818	0.182	0.997	0.003	0.994	0.006	0.998	0.002
chahla	dessert	Tunisia	INRA	0.505	0.495	0.999	0.001	0.81	0.19	0.752	0.248
aziza	dessert	Tunisia	INRA	0.391	0.609	0.998	0.002	0.763	0.237	0.597	0.403
fuji	dessert	Japan	INRA	0.962	0.038	0.998	0.002	0.998	0.002	0.999	0.001
granny smith	dessert	Australia	INRA	0.753	0.247	0.999	0.001	0.997	0.003	0.995	0.005
grifer	dessert	?	INRA	0.92	0.08	0.995	0.005	0.995	0.005	0.993	0.007
elstar	dessert	Netherlands	INRA	0.968	0.032	0.997	0.003	0.997	0.003	0.998	0.002
shiemer	dessert	?	INRA	0.643	0.357	0.999	0.001	0.869	0.131	0.946	0.054
anna	dessert	Israel	INRA	0.617	0.383	0.999	0.001	0.997	0.003	0.985	0.015
reinette marbree	dessert	?	INRA	0.943	0.057	0.999	0.001	0.99	0.01	0.998	0.002
gala	dessert	New-Zealand	INRA	0.947	0.053	0.999	0.001	0.997	0.003	0.998	0.002
falstaff	dessert	Great Britain	INRA	0.938	0.062	0.999	0.001	0.996	0.004	0.995	0.005
braeburn	dessert	New-Zealand	INRA	0.99	0.01	0.998	0.002	0.997	0.003	0.998	0.002
grenadier	dessert	Great Britain	INRA	0.934	0.066	0.997	0.003	0.996	0.004	0.997	0.003
pink lady	dessert	Australia	INRA	0.979	0.021	0.999	0.001	0.996	0.004	0.998	0.002
ariane	dessert	France	INRA	0.93	0.07	0.994	0.006	0.983	0.017	0.991	0.009

beauty of bath	dessert	Great Britain	INRA	0.746	0.254	0.999	0.001	0.872	0.128	0.936	0.064
pinova	dessert	Germany	INRA	0.783	0.217	0.997	0.003	0.996	0.004	0.996	0.004
belle de boskoop	dessert	Netherlands	INRA	0.982	0.018	0.986	0.014	0.997	0.003	0.998	0.002
reinette baumann	dessert	Belgium	INRA	0.985	0.015	0.999	0.001	0.996	0.004	0.998	0.002
winter banana	dessert	USA	INRA	0.986	0.014	0.999	0.001	0.998	0.002	0.999	0.001
anisha hosta	dessert	France	INRA	0.937	0.063	0.999	0.001	0.973	0.027	0.991	0.009
cachao sagarra	dessert	France	INRA	0.78	0.22	0.998	0.002	0.986	0.014	0.89	0.11
margil	dessert	?	INRA	0.981	0.019	0.999	0.001	0.995	0.005	0.999	0.001
michelin	**dessert**	**France**	**INRA**	**0.319**	**0.681**	**0.978**	**0.022**	**0.93**	**0.07**	**0.879**	**0.121**
worcester	dessert	Great Britain	INRA	0.912	0.088	0.998	0.002	0.997	0.003	0.999	0.001
fiesta	**dessert**	**Great Britain**	**INRA**	**0.701**	**0.299**	**0.999**	**0.001**	**0.993**	**0.007**	**0.997**	**0.003**
prima	dessert	USA	INRA	0.963	0.037	0.999	0.001	0.992	0.008	0.998	0.002
M9	**dessert**	**France**	**INRA**	**0.717**	**0.283**	**0.998**	**0.002**	**0.996**	**0.004**	**0.998**	**0.002**
delicious	dessert	?	INRA	0.983	0.017	0.998	0.002	0.998	0.002	0.999	0.001
unamed	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.685	0.315	0.001	0.999	0.995	0.005	0.993	0.007
trojen hir	cider	France	Verger Conservatoire d'Arzano, Brittany	0.822	0.178	0.999	0.001	0.996	0.004	0.991	0.009
unamed	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.559	0.441	0.996	0.004	0.994	0.006	0.994	0.006
drap d'or à tort	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.813	0.187	0.999	0.001	0.995	0.005	0.992	0.008
guillevic	cider	France	Verger Conservatoire d'Arzano, Brittany	0.982	0.018	0.985	0.015	0.992	0.008	0.998	0.002
mickellic	cider	France	Verger Conservatoire d'Arzano, Brittany	0.715	0.285	0.998	0.002	0.976	0.024	0.993	0.007
bacon melen cotro melen	cider	France	Verger Conservatoire d'Arzano, Brittany	0.864	0.136	0.998	0.002	0.997	0.003	0.998	0.002
judin	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.943	0.057	0.997	0.003	0.997	0.003	0.998	0.002
tockec	dessert	France	Verger Conservatoire	0.681	0.319	0.998	0.002	0.935	0.065	0.997	0.003

			d'Arzano, Brittany								
skouarn gat	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.952	0.048	0.997	0.003	0.995	0.005	0.998	0.002
s2 moulin du roch	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.945	0.055	0.931	0.069	0.988	0.012	0.997	0.003
s3 moulin du roch	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.536	0.464	0.999	0.001	0.995	0.005	0.996	0.004
bacon ru	cider	France	Verger Conservatoire d'Arzano, Brittany	0.972	0.028	0.999	0.001	0.997	0.003	0.998	0.002
c'huello mverger moulin du roch	cider	France	Verger Conservatoire d'Arzano, Brittany	0.966	0.034	0.999	0.001	0.996	0.004	0.998	0.002
inconnue guy padan	cider	France	Verger Conservatoire d'Arzano, Brittany	0.742	0.258	0.999	0.001	0.994	0.006	0.998	0.002
dous coumoulen	cider	France	Verger Conservatoire d'Arzano, Brittany	0.919	0.081	0.999	0.001	0.997	0.003	0.999	0.001
kermerrien	cider	France	Verger Conservatoire d'Arzano, Brittany	0.874	0.126	0.999	0.001	0.986	0.014	0.993	0.007
dous bohars	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.96	0.04	0.998	0.002	0.997	0.003	0.998	0.002
pomme orange	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.934	0.066	0.999	0.001	0.998	0.002	0.999	0.001
poul poche	cider	France	Verger Conservatoire d'Arzano, Brittany	0.902	0.098	0.999	0.001	0.997	0.003	0.999	0.001
pomme cloc'h	cider	France	Verger Conservatoire d'Arzano, Brittany	0.931	0.069	0.971	0.029	0.997	0.003	0.999	0.001
commère	cider	France	Verger Conservatoire d'Arzano, Brittany	0.876	0.124	0.998	0.002	0.997	0.003	0.999	0.001
chuero bris	cider	France	Verger Conservatoire d'Arzano, Brittany	0.873	0.127	0.997	0.003	0.995	0.005	0.994	0.006
Blanche de ste anne	**dessert**	**France**	**Verger Conservatoire d'Arzano, Brittany**	**0.221**	**0.779**	**0.998**	**0.002**	**0.985**	**0.015**	**0.974**	**0.026**
tardive de la sarthe	cider	France	Verger Conservatoire d'Arzano, Brittany	0.666	0.334	0.999	0.001	0.997	0.003	0.997	0.003
bouteille	cider	France	Verger Conservatoire d'Arzano, Brittany	0.927	0.073	0.977	0.023	0.993	0.007	0.997	0.003
lost cam	cider	France	Verger Conservatoire d'Arzano, Brittany	0.984	0.016	0.997	0.003	0.997	0.003	0.998	0.002

fil rouge	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.975	0.025	0.998	0.002	0.995	0.005	0.997	0.003
caot plom	cider	France	Verger Conservatoire d'Arzano, Brittany	0.959	0.041	0.999	0.001	0.993	0.007	0.996	0.004
dessus du paillé	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.801	0.199	0.999	0.001	0.996	0.004	0.996	0.004
coat hir	cider	France	Verger Conservatoire d'Arzano, Brittany	0.914	0.086	0.999	0.001	0.996	0.004	0.999	0.001
fil jaune	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.845	0.155	0.998	0.002	0.997	0.003	0.998	0.002
dous ribote	cider	France	Verger Conservatoire d'Arzano, Brittany	0.89	0.11	0.999	0.001	0.976	0.024	0.994	0.006
ste anne rouge	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.892	0.108	0.999	0.001	0.973	0.027	0.991	0.009
dous kernao	cider	France	Verger Conservatoire d'Arzano, Brittany	0.974	0.026	0.997	0.003	0.996	0.004	0.998	0.002
justine	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.986	0.014	0.998	0.002	0.997	0.003	0.998	0.002
dous glas	cider	France	Verger Conservatoire d'Arzano, Brittany	0.97	0.03	0.999	0.001	0.985	0.015	0.997	0.003
grise dieppoise	cider	France	Verger Conservatoire d'Arzano, Brittany	0.954	0.046	0.999	0.001	0.995	0.005	0.996	0.004
bienvenue	cider	France	Verger Conservatoire d'Arzano, Brittany	0.88	0.12	0.999	0.001	0.997	0.003	0.999	0.001
chailleux	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.895	0.105	0.998	0.002	0.995	0.005	0.997	0.003
matheline	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.827	0.173	0.999	0.001	0.996	0.004	0.998	0.002
doffage	cider	France	Verger Conservatoire d'Arzano, Brittany	0.82	0.18	0.999	0.001	0.998	0.002	0.999	0.001
tillet	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.907	0.093	0.998	0.002	0.998	0.002	0.998	0.002
dessus du paillé	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.867	0.133	0.996	0.004	0.994	0.006	0.992	0.008
carabille	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.625	0.375	0.999	0.001	0.994	0.006	0.993	0.007
dous grise	cider	France	Verger Conservatoire d'Arzano, Brittany	0.961	0.039	0.977	0.023	0.997	0.003	0.998	0.002

reinette d'armorique	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.975	0.025	0.999	0.001	0.997	0.003	0.998	0.002
50 margueilt coz	cider	France	Verger Conservatoire d'Arzano, Brittany	0.978	0.022	0.999	0.001	0.994	0.006	0.997	0.003
dous gwen kerijan	cider	France	Verger Conservatoire d'Arzano, Brittany	0.883	0.117	0.999	0.001	0.998	0.002	0.997	0.003
bacon ru hatif	cider	France	Verger Conservatoire d'Arzano, Brittany	0.693	0.307	0.999	0.001	0.917	0.083	0.979	0.021
dous kervidan	cider	France	Verger Conservatoire d'Arzano, Brittany	0.965	0.035	0.994	0.006	0.996	0.004	0.998	0.002
kignet fri	cider	France	Verger Conservatoire d'Arzano, Brittany	0.972	0.028	0.999	0.001	0.995	0.005	0.998	0.002
reinette grise type de saintonge	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.898	0.102	0.999	0.001	0.996	0.004	0.997	0.003
carabine	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.794	0.206	0.997	0.003	0.997	0.003	0.998	0.002
bris canic	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.79	0.21	0.998	0.002	0.994	0.006	0.997	0.003
dous moën	cider	France	Verger Conservatoire d'Arzano, Brittany	0.636	0.364	0.998	0.002	0.995	0.005	0.996	0.004
dous minotte	cider	France	Verger Conservatoire d'Arzano, Brittany	0.659	0.341	0.994	0.006	0.972	0.028	0.991	0.009
rené vert	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.987	0.013	0.973	0.027	0.996	0.004	0.999	0.001
bacon bihan bris	cider	France	Verger Conservatoire d'Arzano, Brittany	0.958	0.042	0.998	0.002	0.996	0.004	0.998	0.002
fouesnen gwen	cider	France	Verger Conservatoire d'Arzano, Brittany	0.724	0.276	0.998	0.002	0.995	0.005	0.997	0.003
ru galand	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.964	0.036	0.999	0.001	0.995	0.005	0.998	0.002
pen ognon	cider	France	Verger Conservatoire d'Arzano, Brittany	0.984	0.016	0.999	0.001	0.997	0.003	0.998	0.002
dous miliner	cider	France	Verger Conservatoire d'Arzano, Brittany	0.983	0.017	0.997	0.003	0.996	0.004	0.997	0.003
st quidic	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.695	0.305	0.999	0.001	0.995	0.005	0.943	0.057
dous mad	cider	France	Verger Conservatoire d'Arzano, Brittany	0.906	0.094	0.998	0.002	0.985	0.015	0.995	0.005

couhen	cider	France	Verger Conservatoire d'Arzano, Brittany	0.913	0.087	0.999	0.001	0.995	0.005	0.999	0.001
dous bleud	cider	France	Verger Conservatoire d'Arzano, Brittany	0.984	0.016	0.999	0.001	0.997	0.003	0.998	0.002
eistek trink	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.914	0.086	0.998	0.002	0.992	0.008	0.997	0.003
reinette de pont farcy	Pont	France	Verger Conservatoire d'Arzano, Brittany	0.954	0.046	0.999	0.001	0.997	0.003	0.998	0.002
mirebloaz	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.968	0.032	0.999	0.001	0.994	0.006	0.996	0.004
dous mann	cider	France	Verger Conservatoire d'Arzano, Brittany	0.924	0.076	0.998	0.002	0.997	0.003	0.998	0.002
san adrian	cider	France	Verger Conservatoire d'Arzano, Brittany	0.749	0.251	0.987	0.013	0.995	0.005	0.998	0.002
pomme de vin moëlan	cider	France	Verger Conservatoire d'Arzano, Brittany	0.966	0.034	0.998	0.002	0.993	0.007	0.996	0.004
chuero jégo	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.804	0.196	0.998	0.002	0.99	0.01	0.997	0.003
blanc duret	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.91	0.09	0.999	0.001	0.996	0.004	0.997	0.003
médaille d'or	cider	France	Verger Conservatoire d'Arzano, Brittany	0.969	0.031	0.998	0.002	0.996	0.004	0.998	0.002
eistek lanester	cider	France	Verger Conservatoire d'Arzano, Brittany	0.911	0.089	0.998	0.002	0.998	0.002	0.999	0.001
st jean	**dessert**	**France**	**Verger Conservatoire d'Arzano, Brittany**	**0.266**	**0.734**	**0.999**	**0.001**	**0.977**	**0.023**	**0.968**	**0.032**
pommier auberge du cleuziou	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.771	0.229	0.999	0.001	0.96	0.04	0.962	0.038
locard vert	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.963	0.037	0.999	0.001	0.993	0.007	0.994	0.006
coat plom	cider	France	Verger Conservatoire d'Arzano, Brittany	0.958	0.042	0.996	0.004	0.995	0.005	0.997	0.003
sac'h biniou	cider	France	Verger Conservatoire d'Arzano, Brittany	0.906	0.094	0.998	0.002	0.997	0.003	0.998	0.002
dous veg bris	cider	France	Verger Conservatoire d'Arzano, Brittany	0.815	0.185	0.999	0.001	0.997	0.003	0.998	0.002
gwen penker diffon	cider	France	Verger Conservatoire	0.937	0.063	0.999	0.001	0.997	0.003	0.998	0.002

			d'Arzano, Brittany								
avallou belein	cider	France	Verger Conservatoire d'Arzano, Brittany	0.481	0.519	0.999	0.001	0.992	0.008	0.994	0.006
chuero ru mentec	cider	France	Verger Conservatoire d'Arzano, Brittany	0.847	0.153	0.999	0.001	0.998	0.002	0.997	0.003
coat skorn	cider	France	Verger Conservatoire d'Arzano, Brittany	0.983	0.017	0.998	0.002	0.997	0.003	0.998	0.002
carter	cider	France	Verger Conservatoire d'Arzano, Brittany	0.779	0.221	0.998	0.002	0.947	0.053	0.949	0.051
joachim	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.943	0.057	0.999	0.001	0.997	0.003	0.998	0.002
pont kor	cider	France	Verger Conservatoire d'Arzano, Brittany	0.624	0.376	0.999	0.001	0.991	0.009	0.996	0.004
greffen	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.97	0.03	0.999	0.001	0.996	0.004	0.998	0.002
la galeuse	dessert	France	Verger Conservatoire d'Arzano, Brittany	0.955	0.045	0.999	0.001	0.996	0.004	0.999	0.001
bolomic	cider	France	Verger Conservatoire d'Arzano, Brittany	0.925	0.075	0.998	0.002	0.997	0.003	0.998	0.002
pen du	cider	France	Verger Conservatoire d'Arzano, Brittany	0.471	0.529	0.997	0.003	0.997	0.003	0.995	0.005
bienvenue	cider	France	Verger Conservatoire d'Arzano, Brittany	0.984	0.016	0.999	0.001	0.998	0.002	0.999	0.001
dous bloc'hic	cider	France	Verger Conservatoire d'Arzano, Brittany	0.818	0.182	0.998	0.002	0.997	0.003	0.997	0.003
ein shemer	dessert	Russia	USDA - PI 280401	0.681	0.319	0.998	0.002	0.995	0.005	0.987	0.013
antonovka 172670-b	**cider**	**Russia**	**USDA - PI 589956**	**0.145**	**0.855**	**0.995**	**0.005**	**0.808**	**0.192**	**0.695**	**0.305**
korichnoe polosatoje	dessert	?	USDA - PI 589491	0.61	0.39	0.999	0.001	0.831	0.169	0.792	0.208
gravenstein washington red	dessert	?	USDA - PI 588837	0.757	0.243	0.999	0.001	0.974	0.026	0.997	0.003
yellow transparent	**dessert**	**Russia**	**USDA - PI 588859**	**0.217**	**0.783**	**0.999**	**0.001**	**0.961**	**0.039**	**0.957**	**0.043**
antonovka kamenichka	**dessert**	**Russia**	**USDA - PI 588995**	**0.211**	**0.789**	**0.999**	**0.001**	**0.837**	**0.163**	**0.949**	**0.051**
irish peach	dessert	Ireland	USDA - PI 104727	0.742	0.258	0.996	0.004	0.994	0.006	0.993	0.007
koningszuur	dessert	Netherlands	USDA - PI 188517	0.846	0.154	0.999	0.001	0.975	0.025	0.991	0.009
poeltsamaa winter apple	dessert	?	USDA - PI 383515	0.941	0.059	0.997	0.003	0.995	0.005	0.996	0.004

ingol	dessert	?	USDA - PI 589441	0.98	0.02	0.999	0.001	0.998	0.002	0.999	0.001
Novosibirski sweet	**dessert**	**Russia**	**USDA - PI 589478**	**0.263**	**0.737**	**0.999**	**0.001**	**0.976**	**0.024**	**0.977**	**0.023**
x a17 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.973	0.027	0.999	0.001	0.997	0.003	0.998	0.002
x a18 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.983	0.017	0.999	0.001	0.995	0.005	0.998	0.002
peau de chien a19	cider	France	Abbaye de Beauport (France, Brittany)	0.982	0.018	0.999	0.001	0.996	0.004	0.998	0.002
x a22 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.986	0.014	0.998	0.002	0.998	0.002	0.999	0.001
x a23 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.982	0.018	0.997	0.003	0.996	0.004	0.998	0.002
x a28 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.983	0.017	0.998	0.002	0.997	0.003	0.999	0.001
gros pigeonnet a29	cider	France	Abbaye de Beauport (France, Brittany)	0.983	0.017	0.999	0.001	0.998	0.002	0.999	0.001
x a3 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.968	0.032	0.999	0.001	0.996	0.004	0.998	0.002
x a4 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.958	0.042	0.998	0.002	0.995	0.005	0.998	0.002
kemerrien a7	cider	France	Abbaye de Beauport (France, Brittany)	0.954	0.046	0.999	0.001	0.997	0.003	0.998	0.002
kemerrien a8	cider	France	Abbaye de Beauport (France, Brittany)	0.937	0.063	0.998	0.002	0.996	0.004	0.998	0.002
gros pigeonnet b1	cider	France	Abbaye de Beauport (France, Brittany)	0.975	0.025	0.997	0.003	0.998	0.002	0.999	0.001
x b12 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.986	0.014	0.999	0.001	0.997	0.003	0.999	0.001
poire b7 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.943	0.057	0.998	0.002	0.982	0.018	0.996	0.004
reinette de pontrieux b8	cider	France	Abbaye de Beauport (France, Brittany)	0.985	0.015	0.995	0.005	0.997	0.003	0.998	0.002
x b9r ap	cider	France	Abbaye de Beauport (France, Brittany)	0.985	0.015	0.996	0.004	0.995	0.005	0.997	0.003
x c123 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.917	0.083	0.995	0.005	0.997	0.003	0.998	0.002
x c125 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.988	0.012	0.999	0.001	0.997	0.003	0.999	0.001

x c134 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.977	0.023	0.997	0.003	0.998	0.002	0.999	0.001
x c146 ap1	cider	France	Abbaye de Beauport (France, Brittany)	0.824	0.176	0.996	0.004	0.997	0.003	0.998	0.002
x c146 ap2	cider	France	Abbaye de Beauport (France, Brittany)	0.856	0.144	0.998	0.002	0.996	0.004	0.998	0.002
x c147 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.156	0.844	0.998	0.002	0.975	0.025	0.966	0.034
x c148 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.982	0.018	0.998	0.002	0.998	0.002	0.999	0.001
x c156 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.953	0.047	0.998	0.002	0.996	0.004	0.999	0.001
tete de vache c160	cider	France	Abbaye de Beauport (France, Brittany)	0.485	0.515	0.998	0.002	0.991	0.009	0.968	0.032
x d1 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.861	0.139	0.997	0.003	0.992	0.008	0.994	0.006
gros pigeonnet d11	cider	France	Abbaye de Beauport (France, Brittany)	0.987	0.013	0.999	0.001	0.998	0.002	0.999	0.001
gros pigeonnet d11 ap2	cider	France	Abbaye de Beauport (France, Brittany)	0.981	0.019	0.999	0.001	0.997	0.003	0.999	0.001
x d3 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.794	0.206	0.994	0.006	0.965	0.035	0.998	0.002
x d4 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.977	0.023	0.998	0.002	0.998	0.002	0.998	0.002
x e1 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.986	0.014	0.999	0.001	0.997	0.003	0.999	0.001
x e4 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.59	0.41	0.999	0.001	0.966	0.034	0.981	0.019
x f2 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.953	0.047	0.985	0.015	0.996	0.004	0.997	0.003
x f3 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.949	0.051	0.998	0.002	0.996	0.004	0.997	0.003
x f5 ap	cider	France	Abbaye de Beauport (France, Brittany)	0.975	0.025	0.986	0.014	0.994	0.006	0.997	0.003
x rear ap	cider	France	Abbaye de Beauport (France, Brittany)	0.984	0.016	0.999	0.001	0.997	0.003	0.998	0.002
x repi ap	cider	France	Abbaye de Beauport (France, Brittany)	0.988	0.012	0.998	0.002	0.998	0.002	0.999	0.001

x ropl ap	cider	France	Abbaye de Beauport (France, Brittany)	0.956	0.044	0.974	0.026	0.996	0.004	0.998	0.002
M8	dessert	Armenia	Field sampling	0.958	0.042	0.931	0.069	0.913	0.087	0.963	0.037
Shagarkeni	dessert	Armenia	Field sampling	0.838	0.162	0.938	0,062	0.909	0.091	0.768	0.232
fenouillet gris	dessert	France	INRA	0.985	0.015	0.995	0.005	0.997	0.003	0.998	0.002
belle fleur jaune	dessert	USA	INRA	0.985	0.015	0.997	0.003	0.998	0.002	0.998	0.002
calville blanc d'hiver	dessert	Switzerland	INRA	0.987	0.013	0.972	0.028	0.998	0.002	0.999	0.001
gros api	dessert	France	INRA	0.587	0.413	0.999	0.001	0.948	0.052	0.992	0.008
reinette de cuzy	dessert	France	INRA	0.841	0.159	0.999	0.001	0.997	0.003	0.998	0.002
bec d'oie	dessert	France	INRA	0.973	0.027	0.993	0.007	0.997	0.003	0.998	0.002
amere de berthecourt	cider	France	INRA	0.904	0.096	0.999	0.001	0.996	0.004	0.995	0.005
armagnac	cider	France	INRA	0.963	0.037	0.662	0.338	0.996	0.004	0.999	0.001
bassard	cider	France	INRA	0.979	0.021	0.998	0.002	0.998	0.002	0.999	0.001
bedange rouge	cider	France	INRA	0.977	0.023	0.998	0.002	0.997	0.003	0.998	0.002
binet blanc	cider	France	INRA	0.887	0.113	0.998	0.002	0.99	0.01	0.993	0.007
binet gris	cider	France	INRA	0.971	0.029	0.999	0.001	0.998	0.002	0.999	0.001
blanc mollet	cider	France	INRA	0.882	0.118	0.998	0.002	0.996	0.004	0.999	0.001
cahoua	cider	France	INRA	0.941	0.059	0.999	0.001	0.994	0.006	0.998	0.002
michelin	cider	France	INRA	0.983	0.017	0.999	0.001	0.996	0.004	0.998	0.002
moulin a vent de l'eure	cider	France	INRA	0.797	0.203	0.998	0.002	0.996	0.004	0.998	0.002
moulin a vent du calvados	cider	France	INRA	0.983	0.017	0.999	0.001	0.995	0.005	0.998	0.002
petit gilet rouge de janze	cider	France	INRA	0.883	0.117	0.998	0.002	0.983	0.017	0.987	0.013
petite sorte du parc dufour	cider	France	INRA	0.954	0.046	0.998	0.002	0.997	0.003	0.997	0.003
reinette d'armorique	cider	France	INRA	0.979	0.021	0.999	0.001	0.997	0.003	0.998	0.002
stang ru	cider	France	INRA	0.634	0.366	0.999	0.001	0.995	0.005	0.992	0.008
cartigny	cider	France	INRA	0.91	0.09	0.999	0.001	0.998	0.002	0.998	0.002
mettais	cider	France	INRA	0.692	0.308	0.999	0.001	0.994	0.006	0.998	0.002
sebin blanc	cider	France	INRA	0.983	0.017	0.999	0.001	0.997	0.003	0.998	0.002

clos renaux	cider	France	INRA	0.965	0.035	0.997	0.003	0.973	0.027	0.997	0.003
binet violet	cider	France	INRA	0.796	0.204	0.999	0.001	0.992	0.008	0.997	0.003
clara	cider	Spain	INRA	0.939	0.061	0.998	0.002	0.995	0.005	0.998	0.002
bisquet	cider	France	INRA	0.985	0.015	0.997	0.003	0.995	0.005	0.998	0.002
egyptia	cider	France	INRA	0.941	0.059	0.997	0.003	0.997	0.003	0.999	0.001
saint martin	cider	France	INRA	0.986	0.014	0.995	0.005	0.997	0.003	0.998	0.002
jeanne renard	cider	France	INRA	0.884	0.116	0.999	0.001	0.997	0.003	0.999	0.001
judin	cider	France	INRA	0.976	0.024	0.999	0.001	0.997	0.003	0.999	0.001
api etoile	dessert	France	INRA	0.497	0.503	0.997	0.003	0.777	0.223	0.979	0.021
chuero ru	cider	France	INRA	0.945	0.055	0.997	0.003	0.993	0.007	0.996	0.004
avrolles	cider	France	INRA	0.959	0.041	0.999	0.001	0.997	0.003	0.998	0.002
doux veret de carrouges	cider	France	INRA	0.852	0.148	0.999	0.001	0.998	0.002	0.999	0.001
blanc sur	cider	France	INRA	0.988	0.012	0.997	0.003	0.997	0.003	0.999	0.001
petit rouget de dol	cider	France	INRA	0.989	0.011	0.997	0.003	0.997	0.003	0.999	0.001
chevalier jaune	cider	France	INRA	0.869	0.131	0.999	0.001	0.996	0.004	0.999	0.001
petit jaune	cider	France	INRA	0.874	0.126	0.997	0.003	0.996	0.004	0.999	0.001
guillevic	cider	France	INRA	0.98	0.02	0.999	0.001	0.996	0.004	0.998	0.002
rene martin	cider	France	INRA	0.955	0.045	0.998	0.002	0.997	0.003	0.999	0.001
jaune de vitre	cider	France	INRA	0.98	0.02	0.999	0.001	0.997	0.003	0.999	0.001
binet rouge	cider	France	INRA	0.977	0.023	0.999	0.001	0.994	0.006	0.998	0.002
doux joseph	cider	France	INRA	0.984	0.016	0.999	0.001	0.998	0.002	0.999	0.001
crollon	cider	France	INRA	0.976	0.024	0.998	0.002	0.997	0.003	0.999	0.001
mariennet	cider	France	INRA	0.981	0.019	0.97	0.03	0.997	0.003	0.999	0.001
petit amer	cider	France	INRA	0.98	0.02	0.739	0.261	0.995	0.005	0.998	0.002
amere saint jacques	cider	France	INRA	0.87	0.13	0.999	0.001	0.996	0.004	0.998	0.002
doux eveque briz	cider	France	INRA	0.823	0.177	0.999	0.001	0.997	0.003	0.998	0.002
chuero ru bihan	cider	France	INRA	0.956	0.044	0.999	0.001	0.93	0.07	0.987	0.013
chuero ru mod koz	cider	France	INRA	0.978	0.022	0.884	0.116	0.998	0.002	0.999	0.001
treujenn hir	cider	France	INRA	0.949	0.051	0.997	0.003	0.995	0.005	0.996	0.004

prat yeot	cider	France	INRA	0.947	0.053	0.999	0.001	0.998	0.002	0.998	0.002
avalou belein	cider	France	INRA	0.759	0.241	0.999	0.001	0.996	0.004	0.997	0.003
doux au gobet	cider	France	INRA	0.637	0.363	0.999	0.001	0.993	0.007	0.997	0.003
cossa	cider	France	INRA	0.947	0.053	0.999	0.001	0.979	0.021	0.996	0.004
doux corier	cider	France	INRA	0.964	0.036	0.999	0.001	0.997	0.003	0.999	0.001
marseigna	cider	France	INRA	0.915	0.085	0.996	0.004	0.994	0.006	0.997	0.003
patte de loup	dessert	France	INRA	0.981	0.019	0.999	0.001	0.997	0.003	0.999	0.001
colapuis	dessert	Ukraine	INRA	0.718	0.282	0.998	0.002	0.989	0.011	0.996	0.004
non pareil	dessert	France	INRA	0.987	0.013	0.999	0.001	0.997	0.003	0.998	0.002

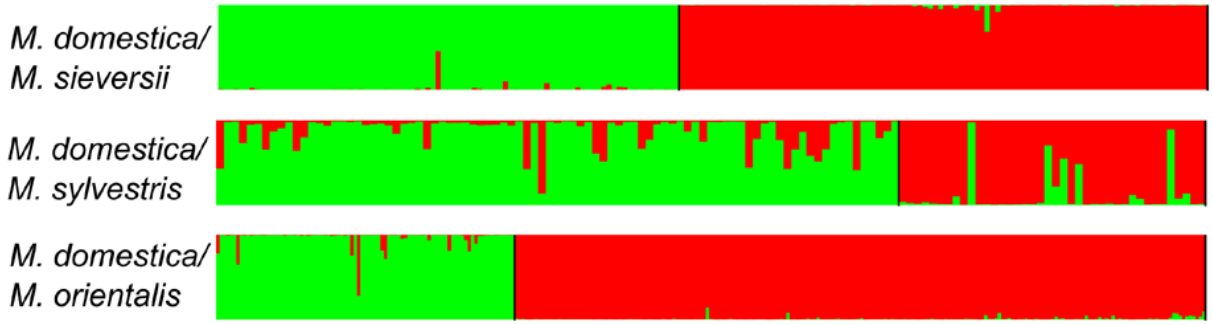


Figure S3. Proportions of ancestry in two ancestral genepools inferred with the STRUCTURE program from datasets including *Malus domestica* (green, *N*=89) and each of the wild *Malus* species (red) except *Malus baccata*. The x-axis is not at scale.

Table S3. Membership coefficients inferred from the STRUCTURE analysis for *Malus baccata* individuals.

Origin/cultivar name*	*Malus baccata*	*Malus domestica*
Russia	0.999	0.001
Russia	0.996	0.004
Russia	0.898	0.102
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.951	0.049
Russia	0.999	0.001
Russia	0.997	0.003
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.886	0.114
Russia	0.998	0.002
Russia	0.999	0.001
Russia	0.998	0.002
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.988	0.012
Russia	0.997	0.003
Russia	0.999	0.001
Russia	0.648	0.352
Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.971	0.029

Russia	0.999	0.001
Russia	0.999	0.001
Russia	0.998	0.002
Russia	0.999	0.001
Russia	0.997	0.003
Russia	0.725	0.275
Russia	0.999	0.001
Russia	0.978	0.022
Russia	0.995	0.005
Russia	**0.667**	**0.333**
Romania	0.973	0.027
Hungary	**0.19**	**0.81**
unknown (EMR[1])	0.989	0.011
flexilis (*EMR[1]*)	**0.653**	**0.347**
*gracilis (*EMR[1]*)*	0.155	0.845
*jackii (*EMR[1]*)*	0.702	0.298
mandshurica (USDA-ARS[2]. gmal35)	**0.569**	**0.431**
rockii (USDA-ARS[2]. gmal423)	0.985	0.015
flexilis (USDA-ARS[2]. gmal1605)	**0.674**	**0.326**
unknown (*USDA-ARS[2]*. gmal1617)	0.998	0.002
jackii (USDA-ARS[2]. gmal2460)	0.998	0.002
Hansen's (USDA-ARS[2]. gmal2477)	**0.464**	**0.536**

* Samples whose origin is indicated as “Russia” were effectively collected in Russia during field trip. Samples from apple collections from Romania or Hungary are actually of unknown geographic origin. For all other samples, variety names are given, with germplasm repository names in brackets.

EMR[1] East Malling Research, Kent, UK

USDA-ARS[2] Plant Genetic Resources Unit, Geneva (NY)

Table S4. Prior distributions used in approximate Bayesian computations. Prior distributions are uniform between lower and upper bound. Parameters are introduced in Figure 4 and Table 5. Species names are abbreviated.

Parameter	Lower bound	Upper bound
N1 (*M. dom*)	1	6,000
N2 (*M. ori*)	1,000	70,000
N3 (*M. siev*)	1,000	20,000
N4 (*M. sylv*)	1,000	50,000
T1 (*M. siev - M.sylv*)	615	50,000
T2 (*M. siev - M. ori*)	615	7,000
T3 (*M. siev - M. dom*)	614	3,500
r1 (introgr. by *M. dom* into *M. siev*)	0.001	0.700
r2 (introgr. by *M. sylv* into *M. dom*)	0.001	0.400
r3 (introgr. by *M. dom* into *M. sylv*)	0.001	0.700
r4 (introgr. by *M. dom* into *M. ori*)	0.001	0.400
r5 (introgr. by *M. ori* into *M. dom*)	0.001	0.700
μ	10^{-4}	10^{-3}
p	0.10	0.30
μSNI	10^{-8}	10^{-4}

Text S1. Method used for approximate Bayesian computations on alternative datasets/admixture times.

Approximate Bayesian computations on alternative datasets/admixture times

We conducted two additional sets of approximate Bayesian computations: (i) on a pruned dataset with misclassified wild individuals and individuals with a recent admixed ancestry removed, (ii) on the full dataset, but assuming that admixture between ancestral *M. domestica* and *M. sylvestris* was more recent (67 generations, 500 ybp) than in original analyses (200 generations, 1500 ybp). Analyses were conducted using the same prior sets than the main dataset (Table S4). In analyses on the pruned dataset, relative posterior probabilities of models *b* and *c* were not significantly different from each other (Table S5; model *b*: 0.5135, 95% confidence interval: 0.4778-0.5492; model *c*: 0.4857, 95% confidence interval: 0.4500-0.5214). However, since introgression between *M. domestica* and *M. orientalis* could not be estimated under model *b* (not shown), only parameter estimates for model *c* are reported in Table S6. In analyses assuming an alternative admixture time

between *M. domestica* and *M. sylvestris*, the posterior probability of model *b* was only slightly lower than that of model *c* (Table S5), but, again, introgression between *M. domestica* and *M. orientalis* could not be estimated accurately under model *b* (model *b*: 0.5486, 95% confidence interval: 0.5089-0.5882; model *c*: 0.4154, 95% confidence interval: 0.3765-0.4543). Both analyses resulted in very minor changes to the point estimates for all parameters (Table S6).

Model checking

We assessed the goodness-of-fit of all model parameter posterior combinations. For each combination, 100 datasets were simulated using parameter values drawn posterior distributions. Summary statistics of the observed data were then ranked against the distributions obtained from simulated datasets (Cornuet et al. 2010). To avoid overestimating the quality of the fit by using the same statistics twice, model checking was based on test quantities summary statistics that have not been used in parameter inferences. Results are shown in Table S7. We found that none of the test quantities had significant tail-area probabilities under model *a*, *b*, *c*. Two test quantities (proportion of shared alleles between *M. domestica* and *M. sylvestris*, and between *M. orientalis* and *M. sieversii*) showed significant, or marginally significant, tail-area probabilities under model *d*. These results suggest that observed data are more plausible under the posterior predictive distributions generated under admixture models *a*, *b* and *c*, than under the posterior predictive distributions generated under model *d*.

Confidence in model choice

The performance of the method to discriminate among competing historical models was assessed by analyzing test datasets simulated with the same number of loci and individuals than in observed datasets (*i.e.*, pseudo-observed datasets). One hundred of such test data sets were simulated under each competing model, using parameter values drawn in the same prior distributions than those for original analyses. Relative posterior probabilities of competing models were evaluated for each pseudo-observed dataset, using the same methodology as described for the observed dataset. Confidence in model choice was then estimated using the proportion of cases a given scenario has not the highest

posterior probability among competing scenarios when it is actually the true scenario (type I error) and the proportion of cases a given scenario has the highest posterior probability when it is actually not the true scenario (type II error).

For main analyses, results indicated a good power of our methodology to discriminate among the four competing models. For model *c*, the type I error rate amounted to 0.54, and the mean type II error rate was 0.05 (range: 0 – 14). Analyses on alternative datasets/parameter sets also indicated a good power to discriminate among competing models. For model *c* with the pruned dataset, the type I error rate amounted to 0.57 and the mean type II error rate was 0.08 (range: 2-19). For model *c* using an alternative admixture time, type I error rate was 0.47, and the mean type II error was 0.067 (range: 0 – 18).

Table S5. Relative posterior probabilities (*p*) for the four historical models compared using approximate Bayesian computations. Models are described in Figure 4. CI2.5 and CI97.5 are boundaries of the 95% confidence intervals. (A) Analyses on a pruned dataset with misclassified wild individuals and individuals with a recent admixed ancestry removed. (B) Analyses on the full dataset, assuming that admixture between ancestral *M. domestica* and *M. sylvestris* was more recent (67 generations – 500 ybp) than in original analyses (200 generations – 1,500 ybp).

Treatment/Model	***p***	**CI2.5**	**CI97.5**
A			
a	0.0008	0.0005	0.0011
b	0.5135	0.4778	0.5492
c	0.4857	0.4500	0.5214
d	0.0000	0.0000	0.0000
B			
a	0.0360	0.0240	0.0480
b	0.4154	0.3765	0.4543
c	0.5486	0.5089	0.5882
d	0.0000	0.0000	0.0000

Table S6. Demographic and mutation parameters estimated using approximate Bayesian computation for model *c*. Posterior distributions are summarized as the mode and boundaries of the 95% credibility intervals (CI2.5 and CI97.5). Demographic parameters are introduced in Figure 4 (note that admixture times are fixed in these analyses). Composite parameters scaled by the mutation rate are also shown. The mutation parameters are μ (mean mutation rate), *p* (mean value of the geometric distribution parameter that governs the number of repeated motifs that increase or decrease the length of the locus during mutation events), μSNI (mean single nucleotide indel mutation rate). Species names are abbreviated. (A) Analyses on a pruned dataset with misclassified wild individuals and individuals with a recent admixed ancestry removed. (B) Analyses on the full dataset, assuming that admixture between ancestral *M. domestica* and *M. sylvestris* was more recent (67 generations – 500 ybp) than in original analyses (200 generations – 1,500 ybp).

	Treatment A			**Treatment B**		
Parameter	**Mode**	**CI2.5**	**CI97.5**	**Mode**	**CI2.5**	**CI97.5**
N1 (*M. dom*)	1,470	844	4,250	1,410	860	5,090
N2 (*M. ori*)	40,200	17,000	67,700	27,000	10,900	63,500
N3 (*M. siev*)	4,800	2,480	12,300	14,200	6,710	19,400
N4 (*M. sylv*)	33,700	16,300	48,700	32,900	16,000	48,400
T1 (*M. siev - M.sylv*)	9,270	4,900	45,000	11,600	5,060	44,500
T2 (*M. siev - M. ori*)	2,340	1,100	6,260	2,480	1,220	6,070
T3 (*M. siev - M. dom*)	1,690	746	3,400	1,560	758	3,300
r2 (introgr. by *M. sylv* into *M. dom*)	0.65	0.50	0.69	0.63	0.51	0.68
μ	$2.2.10^{-4}$	$1.1.10^{-4}$	$7.8.10^{-4}$	$1.7.10^{-4}$	$1.1.10^{-4}$	$6.4.10^{-4}$
p	0.3	0.1	0.3	0.3	0.1	0.3
μSNI	$3.0.10^{-8}$	$4.0.10^{-8}$	$5.2.10^{-5}$	$4.0.10^{-8}$	$7.0.10^{-8}$	$6.1.10^{-5}$
$\vartheta 1$ (=$4N1\mu$)	0.43	0.21	1.54	0.36	0.19	0.17
$\vartheta 2$ (=$4N2\mu$)	10.8	4.8	34.2	4.6	2.8	22.0
$\vartheta 3$ (=$4N3\mu$)	1.5	0.9	3.8	2.7	1.8	7.0
$\vartheta 4$ (=$4N4\mu$)	8.3	4.8	22.0	6.7	4.2	17.9
$\tau 1$ (=$\mu T1$)	3.2	1.4	18.8	2.6	1.2	15.4
$\tau 2$ (=$\mu T2$)	0.6	0.3	2.6	0.5	0.3	2.2
$\tau 3$ (=$\mu T3$)	0.5	0.2	1.7	0.3	0.2	1.3

Table S7. Model checking based on comparisons of test quantities between observed data and 100 pseudo-observed datasets generated using parameter values drawn from posterior distributions. (A) Analyses on the full dataset, (B) Analyses on a pruned dataset with misclassified wild individuals and individuals with a recent admixed ancestry removed. (C) Analyses on the full dataset, assuming that admixture between ancestral *Malus domestica* and *M. sylvestris* was more recent (67 generations – 500 ybp) than in (A).

	Tail-area probability														
	Treatment A					Treatment B					Treatment C				
Test quantity	Observed value	*Model a*	*Model b*	*Model c*	*Model d*	Observed value	*Model a*	*Model b*	*Model c*	*Model d*	Observed value	*Model a*	*Model b*	*Model c*	*Model d*
VAR1	32.05	0.34	0.36	0.33	0.62	29.74	0.25	0.31	0.24	0.86	32.05	0.32	0.39	0.38	0.77
VAR2	19.18	0.51	0.25	0.16	0.2	17.58	0.68	0.45	0.49	0.45	19.18	0.42	0.23	0.18	0.23
VAR3	21.17	0.6	0.35	0.26	0.37	15.15	0.69	0.52	0.57	0.62	21.17	0.59	0.39	0.29	0.45
VAR4	32.02	0.38	0.39	0.38	0.35	30.63	0.25	0.33	0.29	0.29	32.02	0.345	0.43	0.35	0.49
N2P12	23.57	0.25	0.36	0.275	0.605	21.00	0.27	0.355	0.41	0.68	23.57	0.27	0.265	0.245	0.52
N2P13	20.64	0.245	0.19	0.17	0.61	16.64	0.23	0.265	0.32	0.725	20.64	0.24	0.21	0.185	0.59
N2P14	21.79	0.33	0.45	0.415	0.36	19.64	0.3	0.27	0.33	0.26	21.79	0.415	0.4	0.46	0.34
N2P23	21.00	0.295	0.46	0.32	0.45	17.36	0.38	0.375	0.41	0.38	21.00	0.33	0.46	0.39	0.29
N2P24	24.14	0.29	0.34	0.285	0.45	23.86	0.28	0.33	0.38	0.395	24.14	0.33	0.3	0.33	0.395
N2P34	21.86	0.355	0.29	0.28	0.405	20.50	0.315	0.28	0.365	0.39	21.86	0.39	0.27	0.275	0.41
H2P12	0.871	0.17	0.115	0.135	0.445	0.860	0.285	0.26	0.305	0.5	0.871	0.2	0.175	0.175	0.345
H2P13	0.866	0.25	0.09	0.165	0.58	0.852	0.315	0.35	0.395	0.755	0.866	0.2	0.17	0.23	0.57
H2P14	0.862	0.135	0.1	0.145	0.585	0.854	0.175	0.2	0.285	0.67	0.862	0.1	0.15	0.215	0.53
H2P23	0.832	0.245	0.115	0.155	0.12	0.807	0.3	0.245	0.275	0.125	0.832	0.27	0.17	0.135	0.09
H2P24	0.852	0.175	0.09	0.13	0.11	0.843	0.235	0.25	0.24	0.135	0.852	0.19	0.155	0.13	0.09
H2P34	0.847	0.38	0.175	0.23	0.23	0.826	0.39	0.345	0.325	0.35	0.847	0.435	0.195	0.155	0.255

DAS12	0.091	0.69	0.795	0.77	0.45	0.089	0.475	0.645	0.54	0.325	0.091	0.585	0.72	0.795	0.535
DAS13	0.098	0.67	0.85	0.795	0.415	0.089	0.535	0.53	0.445	0.18	0.098	0.62	0.775	0.75	0.435
DAS14	0.095	0.7	0.81	0.74	0.965*	0.084	0.73	0.725	0.72	0.975*	0.095	0.575	0.765	0.77	0.925
DAS23	0.148	0.765	0.85	0.84	0.89	0.174	0.68	0.78	0.75	0.915	0.148	0.76	0.835	0.84	0.94
DAS24	0.057	0.52	0.67	0.655	0.545	0.050	0.475	0.69	0.64	0.635	0.057	0.515	0.675	0.655	0.61
DAS34	0.063	0.535	0.83	0.75	0.685	0.051	0.515	0.705	0.62	0.665	0.063	0.555	0.755	0.745	0.7

1: *M. domestica*; 2: *M. orientalis*; 3: *M. sieversii*; 4: *M. sylvestris*. Tail-area probability was computed for each test quantities (*tq*) as p and $1 - p$ for $p \leq 0.5$ and > 0.5, respectively, with p = *Prob[tq(simulated)<tq(observed)]* (Cornuet et al. 2010). *VARi* = mean allelic size variance in population *i*, *N2Pij* = mean number of alleles in populations *i* and *j*, *H2Pij* = mean gene diversity in populations *i* and *j*, *DASij* = proportion of shared alleles between populations *i* and *j*. * $p>0.95$.

Table S8. Genetic differentiation (F_{ST}) between cultivars of different geographic origins (*N*=266). Cultivars of unknown origin have been removed.

	Germany	UK	Australia	Belgium	Canada	Spain	France	Japan	New Zealand	Netherlands	Russia	United-States	Switzerland	Tunisia	Ukraine
UK	0.0089														
Australia	0.0025	-0.0142													
Belgium	0.0379	0.0306*	0.0095												
Canada	0.0523	0.0269	0.0610	0.0865											
Spain	0.0484	0.0001	0.0552	0.0230	0.1160										
France	0.0260*	0.0155*	-0.0007*	0.0123	0.0153*	-0.0093									
Japan	-0.0008	-0.0295	0.0137	0.0313	0.0156	0.0247	0.0020*								
New Zealand	0.0280	0.0168	0.0846	0.0618	0.1595	0.0893	0.0278*	0.0463							
Netherlands	0.0116	0.0064	0.0258	-0.0199	0.0430	0.0233	0.0071	0.0237	-0.0080						
Russia	0.0377*	0.0369*	0.0155	0.0318	-0.0409	-0.0011	0.0332*	-0.0359	0.0230	0.0354					
United States	0.0168	0.0248*	0.0198	0.0314*	0.0744	0.0290*	0.0299*	0.0045	-0.0062	-0.0043	0.0369*				
Switzerland	0.0259	-0.0089	0.0699	0.0750	0.5647	0.2087	0.0283	0.0235	0.2575	0.0489	0.0219	-0.0177			
Tunisia	0.0838	0.0575*	0.0796	0.0901	0.2072	0.0969	0.0517*	0.0695	0.1158	0.0782*	0.0229	0.0294	0.1877*		
Ukraine	0.0726	0.0365*	0.0502	0.1295*	0.1183	0.1345	0.0629*	0.0280	0.1334	0.1029*	0.0262	0.0757*	0.1323	0.1415	
Israel	0.0267	-0.0766	0.0631	0.0767	0.5122	0.1384	-0.0221	-0.0793	0.1539	0.0184	-0.0019	-0.0343	0.4605	0.1828	0.1293

* : P<0.05

Table S9. Description of the Multiplex PCRs (MP01, MP02, MP03, MP04) used for microsatellite amplification.

MP01		
		Volume (µL)
Mix qiagen		7.50
CH01h10 (10µM)	VIC	0.15
		0.15
CH01h01 (10µM)	PET	0.15
		0.15
CH01f03b (10µM)	6-FAM	0.15
		0.15
CH02c06 (20µM)	NED	0.30
		0.30
dWater		4
DNA		2
Total		15

MP02		
		Volume (µL)
Mix qiagen		7.5
NZ05g08 (20µM)	HEX	0.30
		0.30
CH05f06 (10µM)	PET	0.30
		0.30
CH02d08 (10µM)	NED	0.15
		0.15
CH04e05 (20µM)	6-FAM	0.30
		0.30
dWater		3.40
DNA		2
Total		15

MP03		
		Volume (µL)
Mix qiagen		7.5
Hi02c07 (10µM)	VIC	0.15
		0.15
CH01f02 (20µM)	6FAM	0.30
		0.30
CH02c11 (10µM)	NED	0.15
		0.15
dWater		4.30
DNA		2
Total		15

MP04		
		Volume (µL)
Mix qiagen		7.5
CH04c07 (10µM)	VIC	0.15
		0.15
GD12 (10µM)	PET	0.15
		0.15
CH03d07 (20µM)	6FAM	0.30
		0.30
CH02c09 (10µM)	NED	0.15
		0.15
dWater		4
DNA		2
Total		15

References

Allaby RG, Fuller DQ, Brown TA (2008) The genetic expectations of a protracted model for the origins of domesticated crops. *PNAS* **105**, 13982-13986.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217-1229.

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000) Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics* **154**, 1309-1321.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in population genetics. *Genetics* **162**, 2025-2035.

Besnard G, Rubio de Casas R, Vargas P (2007) Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). *J Biogeogr* **34**, 736-752.

Blackman BK, Scascitelli M, Kane NC*, et al.* (2011) Sunflower domestication alleles support single domestication center in eastern North America. *PNAS* **108**, 14360-14365.

Boré JM, Fleckinger J (1997) *Pommiers à cidre: variétés de France* INRA éditions, Paris, FRANCE (1997) (Monographie).

Brown TA, Jones MK, Powell W, Allaby RG (2009) The complex origins of domesticated crops in the Fertile Crescent. *Trends Ecol. Evol.* **24**, 103-109.

Caicedo AL, Williamson SH, Hernandez RD*, et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3**, 1745-1756.

Chen H, Morrell PL, Ashworth VETM, de la Cruz M, Clegg MT (2009) Tracing the geographic origins of major avocado cultivars. *J. Hered.* **100**, 56-65.

Coart E, Van Glabeke S, De Loose M, Larsen AS, Roldán-Ruiz I (2006) Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Molecular Ecology* **15**, 2171-2182.

Coart E, Vekemans X, Smulders MJM*, et al.* (2003) Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Mol Ecol* **12**, 845-857.

Cornuet J-M, Ravigne V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics* **11**, 401.

Cornuet J-M, Santos F, Beaumont MA*, et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**, 2713-2719.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001-2014.

Delplancke M, Alvarez N, Espíndola A*, et al.* (2011) Gene flow among wild and domesticated almond species: insights from chloroplast and nuclear markers. *Evolutionary Applications* **5**, 317-329.

Diamond J (1997) *Guns, Germs, and Steel: The Fates of Human Societies* Norton, W. W. & Company, Inc. Sales.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700-707.

Diego M, Michela T, Francesco S*, et al.* (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1044-1045.

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* **127**, 1309-1321.

Dzhangaliev AD (2003) The wild apple tree of Kazakhstan. In: *Hortic Rev* pp. 63-303. John Wiley & Sons, Inc.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Fagundes NJR, Ray N, Beaumont M*, et al.* (2007) Statistical evaluation of alternative models of human evolution. *PNAS* **104**, 17614-17619.

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.

Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. *Trends Genet* **24**, 24-32.

Forsline PL, Aldwinckle HS, Dickson EE, Luby JJ, Hokanson SC (2002) Collection, maintenance, characterization and utilization of wild apples of Central Asia. In: *Hortic Rev* pp. 1-61. John Wiley & Sons, Inc.

Gardiner SE, Bus VGM, Rusholme RL*, et al.* (2007) Fruits and Nuts: Apple (ed. Kole C), pp. 1-62. Springer Berlin Heidelberg.

Gharghani A, Zamani Z, Talaie A*, et al.* (2009) Genetic identity and relationships of Iranian apple (*Malus domestica* Borkh.) cultivars and landraces, wild *Malus* species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genet Resour Crop Ev* **56**, 829-842.

Gianfranceschi L, Seglias N, Tarchini R, Komjanc M, Gessler C (1998) Simple sequence repeats for the genetic analysis of apple. *Theoritical Applied Genetics* **96**, 1069-1076.

Glémin S, Bataillon T (2009) A comparative view of the evolution of grasses under domestication. *New Phytol.* **183**, 273-290.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences* **92**, 6723-6727.

Gross BL, Olsen KM (2009) Genetic perspectives on crop domestication. *Trends Plant Sci.* **15**, 529-537.

Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* **156**, 1-13.

Harris SA, Robinson JP, Juniper BE (2002) Genetic clues to the origin of the apple. *Trends Genet* **18**, 426-430.

Harrison N, Harrison RJ (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1043-1044.

Harter AV, Gardner KA, Falush D*, et al.* (2004) Origin of extant domesticated sunflowers in eastern North America. *Nature* **430**, 201-205.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* **9**, 1322-1332.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Janick J (2005) The origins of fruits, fruit growing, and fruit breeding. In: *Plant Breeding Reviews*, pp. 255-321. John Wiley & Sons, Inc.

Juniper BE, Mabberley DJ (2006) *The story of the apple* Imber Press, Inc.

Kilian B, Özkan H, Walther A*, et al.* (2007) Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during *Triticum monococcum* (Einkorn) domestication: implications for the origin of agriculture. *Mol. Biol. Evol.* **24**, 2657-2668.

Koopman WJM, Li Y, Coart E*, et al.* (2007) Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. *Mol Ecol* **16**, 243-256.

Kovach MJ, Sweeney MT, McCouch SR (2007) New insights into the history of rice domestication. *Trends Genet.* **23**, 578-587.

Lea AGH, Piggott JR (2003) *Fermented Beverage Production* Kluwer Academic/Plenum.

Liebhard R, Gianfranceschi L, Koller B*, et al.* (2002) Development and characterisation of 140 new microsatellites in apple (*Malus* x *domestica* Borkh.). *Molecular Breeding* **10**, 217-241.

Luby JJ, Alspach PA, Bus VGM, Oraguzie NC (2001) Field resistance to fire blight in a diverse apple (*Malus* sp.) germplasm collection. *J Am Soc Hortic Sci* **127**, 245–253.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753-1766.

Mabberley DJ, Jarvis CE, Juniper BE (2001) The name of the apple. *Telopea* **9**, 2001.

Matsuoka Y, Vigouroux Y, Goodman MM*, et al.* (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *PNAS* **99**, 6080-6084.

Meirmans PG, Van Tienderen PH (2004) Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes* **4**, 792-794.

Micheletti D, Troggio M, Salamini F*, et al.* (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1044-1045.

Miller A, Gross BL (2011) From forest to field: perennial fruit crops domestication. *Am. J. Bot.* **98**, 1389-1414.

Miller A, Schaal B (2005) Domestication of a Mesoamerican cultivated fruit tree, *Spondias purpurea*. *PNAS* **102**, 12801-12806.

Miller AJ, Schaal BA (2006) Domestication and the distribution of genetic variation in wild and cultivated populations of the Mesoamerican fruit tree *Spondias purpurea* L. (Anacardiaceae). *Mol. Ecol.* **15**, 1467-1480.

Myles S, Boyko AR, Owens CL*, et al.* (2011) Genetic structure and domestication history of the grape. *PNAS* **108**, 3530-3535.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583-590.

Olsen KM, Gross BL (2008) Detecting multiple origins of domesticated crops. *PNAS* **105**, 13701-13702.

Orton V (1973) *The American cider Book: The story of America's natural beverage* Farrar, Straus and Giroux edn. North Point Press.

Oumar I, Mariac C, Pham J-L, Vigouroux Y (2008) Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theor. Appl. Genet.* **117**, 489-497.

Patocchi A, Fernàndez-Fernàndez F, Evans K*, et al.* (2009) Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. *Tree Genet Genomes* **5**, 211-223.

Pereira-Lorenzo S, Ramos-Cabrer AM, Fischer M (2009) Breeding Apple (*Malus x Domestica* Borkh). Breeding Plantation Tree Crops: Temperate Species, pp. 33-81. Springer New York.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Evol. Syst.* **37**, 187-214.

Pickersgill B (2007) Domestication of plants in the Americas: insights from mendelian and molecular genetics. *Ann Bot* **100**, 925-940.

Piry S, Luikart G, Cornuet JM (1999) Computer note. BOTTLENECK: a computer program for detecting recent reductions in the effective size using allele frequency data. *Journal of Heredity* **90**, 502-503.

Ponomarenko V (1991) On a little known species *Malus* x *asiatica* (Rosaceae). *Bot Zhurn* **76**, 715-720.

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

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. *Nature* **457**, 843-848.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* **86**, 248-249.

Rehder A (1940) *Manual of cultivated trees and shrubs*, 2 edn. Macmillan, New York.

Robinson JP, Harris SA, Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. *Plant Syst. Evol.* **226**, 35-58.

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the Genus *Zea*. *Genetics* **181**, 1399-1413.

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* **8**, 103-106.

Russell J, Dawson IK, Flavell AJ*, et al.* (2011) Analysis of >1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol.* **191**, 564-578.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Curr Opin Plant Biol* **10**, 162-167.

Schuster M, Büttner R (1995) Chromosome numbers in the *Malus* wild species collection of the genebank Dresden-Pillnitz. *Genet Resour Crop Ev* **42**, 353-361.

Schwacke L, Schwacke J, Rosel P (2005) RE-RAT: relatedness estimation and rarefaction analysis tool. http://people.musc.edu/~schwaclh/.

Silfverberg-Dilworth E, Matasci C, Van de Weg W*, et al.* (2006) Microsatellite markers spanning the apple (*Malus* x *domestica* Borkh.) genome. *Tree Genet Genomes* **2**, 202-224.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-2504.

Tanno K-i, Willcox G (2006) How Fast Was Wild Wheat Domesticated? *Science* **311**, 1886.

Tenaillon MI, Manicacci D (2011) Maize origins: an old question under the spotlights. In: *Advances in Maize (Essential Reviews in Experimental Biology)* (eds. Prioul J-L, Thévenot C, Molnar T), pp. 89-110. The Society for Experimental Biology.

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214-1225.

van Heerwaarden J, Doebley J, Briggs WH*, et al.* (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *PNAS* **108**, 1088-1092.

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.

Vavilov NI (1926) Studies on the origin of cultivated plants. *Trudy Byuro. Prikl. Bot.* **16**, 139–245.

Velasco R, Zharkikh A, Affourtit J*, et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics* **42**, 833-839.

Vercken E, Fontaine MC, Gladieux P*, et al.* (2010) Glacial refugia in pathogens: European genetic structure of anther smut pathogens on *Silene latifolia* and *Silene dioica*. *PLoS Pathogens* **6**, e1001229.

Wagner I, Weeden NF (2000) Isozyme in *Malus sylvestris*, *Malus* x *domestica* and in related *Malus* species. *Acta Horticulturae* **538**, 51-56.

Wang C, Chen J, Zhi H*, et al.* (2010) Population genetics of foxtail millet and its wild ancestor. *BMC Genet.* **11**, 90.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* **22**, 506-519.

Zeder MA, Emshwiller E, Smith BD, Bradley DG (2006) Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* **22**, 139-155.

Zohary D (2004) Unconscious selection and the evolution of domesticated plants. *Econ Bot* **58**, 5-10.

Zohary D, Hopf M (2000) *Domestication of plants in the Old World*, 3 edn. New York: Oxford University Press.

Zohary D, Spiegel-Roy P (1975) Beginnings of Fruit Growing in the Old World. *Science*, 319-327.

Manuscrit B: Phylogéographie de pommier sauvage européen (*Malus sylvestris*)

Manuscript B : Post-glacial recolonization history of the European crabapple (*Malus sylvestris* Mill.), a wild contributor to the domesticated apple. Resubmitted to Molecular Ecology

Cornille A.[1], Giraud T.[1], Bellard C.[1], Tellier A.[2], Le Cam B.[3], Smulders M.J.M.[4], Kleinschmit J.[5], Roldan-Ruiz I.[6], Gladieux P.[1,7]

Corresponding author: amandine.cornille@gmail.com

1. CNRS, Laboratoire Ecologie Systématique et Evolution - UMR8079, Bâtiment 360, 91405 Orsay, France; Univ. Paris Sud, 91405 Orsay, France; AgroParisTech, 91405 Orsay, France; 2. Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Deutschlan; 3. INRA, IRHS, PRES UNAM, SFR QUASAV, Rue G. Morel F-49071 Beaucouzé, France ; 4. Plant Research International, Wageningen UR Plant Breeding, PO Box 16, 6700 AA Wageningen, The Netherlands; 5. Northwest German Forest Research Institute, Department of Forest Genetic Resources, Germany; 6. ILVO, Plant – Growth and Development, Caritasstraat 21, 9090 Melle, Belgium; 7. Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA.

Abstract

An understanding of the way in which the climatic oscillations of the Quaternary Period have shaped the distribution and genetic structure of extant tree species can provide insight into the processes driving species diversification, distribution and survival. Deciphering the genetic consequences of past climatic change is also critical for the conservation and sustainable management of forest and tree genetic resources, a timely endeavor as the Earth heads into a new period of climate change. We used a combination of genetic data and ecological niche models to investigate the historical patterns of biogeographic range expansion for a wild fruit tree, the European crabapple (*Malus sylvestris*), a wild contributor to the domesticated apple (*Malus domestica*). While climatic predictions suggested that this species maintained a large and continuous distribution during the last glacial maximum (LGM), analyses of microsatellite variation indicated that *M. sylvestris* experienced range contraction and fragmentation. Bayesian clustering analyses indeed revealed a clear pattern of genetic structure, with a first genetic cluster spanning a large area in Western Europe and two genetic clusters having more limited distributions in Eastern Europe, one located in an area around the Carpathian Mountains and the other restricted to the Balkan Peninsula. Approximate Bayesian computation appeared to be a powerful technique for inferring the history of these clusters, supporting a scenario of simultaneous differentiation of three separate glacial refugia. Admixture between these three populations was found in their suture zones. A weak isolation by distance pattern was detected within each population, indicating a high dispersal capacity for the European crabapple.

Key words: fruit trees, intra-species diversification, microsatellite, approximate Bayesian computation, ecological niche modeling

Introduction

The climatic oscillations of the Quaternary Period clearly affected the distribution and genetic structure of many species (Hewitt, 2004; Lascoux *et al.*, 2004). In Europe, many temperate species faced with the expansion of ice sheets over the northern part of the continent survived periods of glaciation in small ice-free spots in southern and south-eastern Europe (called glacial refugia), in which conditions were more favorable (Hewitt, 2004; Schmitt, 2007). These range contractions were followed by range expansions during the interglacial periods, with populations at the northern boundaries of refugia expanding into the large territories further north, which became increasingly hospitable as the climate rapidly warmed. Such shifts in distribution following rapid climate change are excellent models for understanding the mechanisms by which intra- and interspecific genetic differentiation arises and is maintained (Excoffier *et al.*, 2009; Hewitt, 1990; Hewitt, 1996; Hewitt, 2004; Petit *et al.*, 2004a; Petit, Excoffier, 2009; Schmitt, 2007). Identification of the sites of the refugia that existed during the last glacial maximum (LGM) can also help set priorities for the conservation and management of genetic resources (Hampe, Petit, 2005), and an understanding of past population dynamics is important for predictions of the effects of climate change on species distribution and survival (Nogués-Bravo *et al.*, 2008; Petit *et al.*, 2004a; Petit *et al.*, 2008; Provan, Bennett, 2008).

Trees are fascinating models for investigating the impact of climatic fluctuations on species colonization, adaptation, speciation and extinction (Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). Retrospective analyses based on paleoecological and genetic data have greatly increased our understanding of tree phylogeography (Davis, Shaw, 2001; Hewitt, 2000; Lascoux *et al.*, 2004; Petit *et al.*, 2004a; Stewart *et al.*, 2010). A combination of population genetic studies and explicit species distributions determined from fossil pollen and macrofossil evidence has revealed that trees recolonized higher latitudes within the Holocene, following expansion from southern glacial refugia (Brewer *et al.*, 2002; Cheddadi *et al.*, 2006; Magri *et al.*, 2006), but also from small populations that persisted in more northern “cryptic” refugia (Anderson *et al.*, 2006; Parducci *et al.*, 2012). However, additional studies of other species are required for more general conclusions to be drawn regarding the impact of historical contingencies and differences in climatic niches (from southern

temperate to southern boreal), and mechanisms of gene dispersal on interspecific differences in postglacial histories. In particular, the post-glacial evolutionary histories of wind-dispersed tree species growing in high-density populations has been extensively documented (*e.g.*, Heuertz *et al.*, 2006; Petit *et al.*, 2002), but fewer studies have focused on low-density, animal-dispersed fruit trees (Jolivet, Degen, 2011; Oddou-Muratorio *et al.*, 2004; Palmé, Vendramin, 2002). Animal-dispersed species would be expected to have stronger genetic structure and weaker dispersal capacities than wind-dispersed species, whereas low-density species would be expected to have higher dispersal capacities than high-density species (Hardy *et al.*, 2006; Vekemans, Hardy, 2004). It remains unclear how the balance between these two life-history traits characteristic of fruit trees — animal dispersal and a low density — as well as the importance of outcrossing, shapes the genetic diversity and genetic structure of populations during post-LGM recolonization.

Most previous studies investigating the phylogeography of European tree species were based on chloroplast DNA sequences. An advantage is that maternally inherited markers often display stronger differentiation between populations than biparental markers, due to their smaller effective size and lack of recombination. However, it has repeatedly been argued that this lack of recombination, together with selective pressures and interspecific organelle introgressions can bias biogeographic inferences (Ballard, Whitlock, 2004). Conversely, microsatellites have proved powerful for the elucidation of species history and recent hybridization (Lu *et al.*, 2001; Randi, Lucchini, 2002; Uwimana *et al.*, 2012). Although sequences can be more useful for dating ancient demographic events, studies using microsatellites should help to unravel recent demographic processes, which have received little attention to date, such as gene flow between recolonization wave fronts.

Post-glacial recolonization routes have been inferred from patterns of population structure (Lascoux *et al.*, 2004; Schmitt, 2007; Taberlet *et al.*, 1998). However, until recently, the lack of suitable and tractable statistical methods for distinguishing between different demographic scenarios and estimating demographic and population parameters (effective population size, divergence time between populations and the amount of gene flow between populations) (Richards *et al.*, 2007) hindered the accurate and precise inference of complex scenarios. Approximate Bayesian computation (ABC) provides a robust framework, allowing more powerful inferences of a species' past demography (Beaumont *et al.*, 2002;

Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010). In particular, it makes possible probabilistic comparisons of alternative demographic scenarios (Bertorelle *et al.*, 2010). A new framework has recently been developed, in which ABC methods are used to test ecological niche models (ENMs), to identify habitats that were suitable for particular species in the past (Brown, Knowles, 2012; Carstens, Richards, 2007). This approach, which is complementary to genetic inference, may further improve inferences concerning the post-glacial history of species (Richards *et al.*, 2007; Waltari *et al.*, 2007).

The European crabapple (*Malus sylvestris* Mill), a woody fruit species occurring across Western and Central Europe (Larsen *et al.*, 2006), is a tree of the Rosaceae family, mainly pollinated by bees and flies (Syrphidae), that grows in low-density populations in natural habitats. A great diversity of wild animals feeds on the fruit but their respective efficiencies as seed dispersal vectors are unknown (Larsen *et al.*, 2006). *Malus sylvestris* has been identified as a major contributor to the genome of the cultivated apple *M. domestica* (Cornille *et al.*, 2012b) and is considered to be endangered (Jacques *et al.*, 2009). Efforts to decipher the genetic structure and demographic history of this tree species is thus timely in programs aiming to conserve the genetic resources of *M. sylvestris:* 1) to mitigate the effects of climate change and prevent negative effects of habitat fragmentation on this endangered species, and 2) to increase the genetic resources available for apple breeding programs, through the identification of genes conferring resistance to pathogens or tolerance of diverse abiotic stress conditions. The genetic structure of the European crabapple has been studied only in limited areas (Coart *et al.*, 2003; Larsen *et al.*, 2006) and appears to be weak at this scale, suggesting a high dispersal capacity. No study has yet investigated the phylogeography of this emblematic wild European species across its entire distribution range. We therefore used population genetic analyses and ENM to investigate the glacial refugia and post-glacial recolonization history of *M. sylvestris*. For genetic analyses, we used a comprehensive set of *M. sylvestris* individuals sampled throughout Europe and 26 microsatellite markers. We addressed the following questions: 1) Did wild apples survive the last glacial period in a single refuge area in Europe or at multiple sites? Did all relict populations contribute equally to the post-glacial recolonization of Europe by wild apples? 2) Can we detect the genetic consequences of successive founder events during postglacial colonization, *i.e.,* does genetic diversity decline with increasing distance from refugia? 3) Can

we detect genetic patterns of isolation-by-distance (IBD) and obtain information about dispersal capacities? 4) Did admixture occur between recolonizing populations during expansion of the post-glacial range of this species? 5) Can we reconstruct habitats that were suitable for European crabapple in the past by ENM methods?

Materials and methods

DNA extraction and microsatellite genotyping

DNA was extracted with the Nucleo Spin® plant DNA extraction kit II (Macherey & Nagel, Düren, Germany). PCR amplification was performed with the Multiplex PCR kit (QIAGEN, Inc.). We used 26 microsatellites, dispersed over the 17 chromosomes, typed in 10 different multiplex reactions as previously described (Cornille *et al.*, 2012b; Patocchi *et al.*, 2009). We retained only multilocus genotypes with fewer than 25% missing data. We checked the suitability of the markers for population genetic analysis with ARLEQUIN (Excoffier, Lischer, 2010). None of the 26 microsatellite markers deviated significantly from the neutral equilibrium model, as shown by the non significant *P*-values obtained in Ewen-Watterson tests, and no pair of markers was in significant linkage disequilibrium (Raymond, Rousset, 1995; Rousset, 2008). The markers used may therefore be considered to be unlinked and to be evolving in a neutral manner.

Sampling

We have previously demonstrated the existence of gene flow from the cultivated apple *Malus domestica* to the European wild apple (Cornille *et al.*, 2012b). We therefore initially carried out a STRUCTURE analysis, including 40 reference *M. domestica* cultivars previously identified as displaying no introgression from European crabapple (*i.e*, with membership coefficients >0.9 to the *M. domestica* genepool) and the entire *M. sylvestris* dataset. We retained only individuals assigned with a value of >0.9 to the *M. sylvestris* genepool for further analyses. Leaf material of the retained trees was collected at 37 sites across Europe (*N*=381, Table S1 and Figure S1), covering most part of the geographical distribution of the European crabapple except in Spain and Sweden (see Euforgen map of the European crabapple http://www.euforgen.org/distribution_maps.html).

Descriptive statistics

We tested for the occurrence of null alleles at each locus with MICROCHECKER 2.2.3 (Van Oosterhout *et al.*, 2004). Allelic richness and private allele richness were calculated with ADZE (Szpiech *et al.*, 2008), at the site (*i.e*, geographical locations) and cluster (*i.e*, population inferred by TESS analyses including hybrids up to 0.55 membership coefficient in the given cluster) levels, using sample sizes of *N*=12 (6 individuals x two chromosomes) and *N*=200 (100 individuals x two chromosomes), corresponding to the smallest number of observations for sites and clusters, respectively. Heterozygosity, Weir and Cockerham *F*-statistics and Hardy-Weinberg genotypic linkage equilibrium were assessed with GENEPOP 4.0 (Raymond, Rousset, 1995; Rousset, 2008). Only sampling sites with at least four successfully genotyped specimens were included in site-specific computations (25 sites in total).

Population subdivision, genetic variability and isolation by distance (IBD)

We used the individual-based Bayesian clustering methods implemented in STRUCTURE 2.3.3 (Pritchard *et al.*, 2000) and TESS 2.1 (Chen *et al.*, 2007) to investigate population subdivision. These methods are based on the use of Markov Chain Monte Carlo (MCMC) simulations to infer the assignment of genotypes into *K* distinct clusters. The underlying algorithms attempt to minimize deviations from Hardy–Weinberg and linkage disequilibria within each cluster. The clustering procedure of TESS also includes a spatial component, such that genotypes from geographically closer locations are considered more likely to belong to the same cluster. In TESS analyses, we used the conditional autoregressive (CAR) Gaussian model of admixture with a linear trend surface, setting the spatial interaction parameter (ρ) at 0.6. These parameters (ρ and trend) affect the weighting assigned to spatial proximity when clustering genotypes.

For both methods, ten independent analyses were carried out for each number of clusters *K* ($2 \leq K \leq 6$ for TESS and $1 \leq K \leq 6$ for STRUCTURE), with 500,000 MCMC iterations after a burn-in of 50,000 steps. Outputs were processed with CLUMPP v1.1.2 (Jakobsson, Rosenberg, 2007), for the identification of potentially distinct modes (*i.e.*, clustering solutions) in replicated runs for each *K*. A *G'*-statistic greater than 80% was used to assign replicates to a common TESS mode. We determined the amount of additional information

explained by increasing *K*, by using the *ΔK* statistic (Evanno *et al.*, 2005) for STRUCTURE analyses and the rate of change of the deviation index criterion (DIC) when increasing *K* for TESS analyses.

We checked for IBD patterns as previously described (Loiselle *et al.*, 1995). A Mantel test with 10,000 random permutations was performed between the individual coefficient of relatedness F_{ij} and the matrix of the natural logarithm of geographic distance. These analyses were performed with SPAGeDI 1.3 (Hardy, Vekemans, 2002). Spatial patterns of genetic variability were visualized by mapping variation in allelic richness across space with the interpolation kriging function in ARCINFO (ESRI, Redlands, CA), using a spherical semi-variogram model.

Approximate Bayesian inference

We found weak genetic differentiation in the European crabapple between the north-eastern and south-eastern populations identified with TESS (see results). We therefore investigated whether the observed pattern of genetic diversity resulted from: (*1*) the simultaneous expansion of three independent populations (*i.e*, western, south-eastern and north-eastern; referred to as W, SE and NE, respectively) or (*2*) simultaneous expansions of the W and SE populations, followed by the divergence and expansion of a NE population derived from the SE population (*i.e*, resulting from a more recent colonization wave front). We used ABCtoolBox (Wegmann *et al.*, 2010) to compare these two different scenarios, with and without the occurrence of gene flow in each case (Figure 1). The juvenile period of *M. domestica* lasts five to 10 years and no data for this parameter are available for *M. sylvestris*. We therefore assumed a generation time of 7.5 years. We estimated the effective size of each population (N_W, N_{NE}, N_{SE}), the rate of migration between populations for each generation ($m_{x\text{-}y}$: migration rate from population x to y per generation), the exponential growth rate of each population (G_W, G_{NE}, G_{SE}), and the divergence times (T_{EXP} and $T_{NE\text{-}SE}$). The boundaries of the uniform (or log-uniform) prior distributions (Table S2) were chosen based on preliminary analyses run with very large priors (not shown). The "populations" used in these analyses are the main clusters identified by Bayesian clustering methods.

For all models, identical microsatellite datasets were simulated for 14 of our loci (Ch01h01, Ch01h10, Ch02c06, Ch02d08, Ch05f06, Ch01f02, Hi02c07, Ch02c09, Ch03d07,

Ch04c07, Ch02b03b, MS06g03, Ch04e03, Ch02g01) that had been reported to carry perfect repeats, increasing confidence in the simulated mutation model (Gianfranceschi *et al.*, 1998; Liebhard *et al.*, 2002; Silfverberg-Dilworth *et al.*, 2006). We checked that TESS yielded the same pattern of population structure with these 14 markers as observed with the full set of markers. Using CLUMPP (Jakobsson, Rosenberg, 2007), we showed that both sets of markers (14 *versus* 26 SSR) gave similar clustering patterns (data not shown, similarity index *G'*=96% for *K*=3). We generated 2 x 10^6 genetic datasets from coalescent simulations, using population parameters drawn from a prior distribution (Table S2) under the four previously specified scenarios. For each simulation, we calculated two summary statistics per population: *H*, the mean heterozygosity across loci; *K*, the mean number of alleles across loci. We also calculated pairwise F_{ST} (Weir, Cockerham, 1984) and genetic distances $(\delta\mu)^2$ (Goldstein *et al.*, 1995) between pairs of populations. We conducted a preliminary principal component analysis (PCA) with R software (MASS package, function prcomp), based on 3,000 simulated datasets for *M. sylvestris,* for the four models, to establish and check correlations between the main parameters of the model and the chosen summary statistics (Tellier *et al.*, 2011).

We assumed a generalized stepwise model of microsatellite evolution (Estoup *et al.*, 2002). The mutation rate was allowed to vary across loci, with locus-specific mutation rates being drawn from a gamma distribution (α, α/μ) in which μ is the mutation rate per generation and α is a shape parameter. We assumed a log-uniform prior distribution for μ [0.00001, 0.02] and a uniform distribution for α [1, 30].

We compared the four models by calculating their Bayes factors (Wegmann *et al.*, 2010) and estimating their relative posterior probabilities, based on the 1% of simulated datasets most closely matching the observed data (*i.e.*, 2,000 simulated datasets). Once the best model had been chosen, we estimated demographic parameters under this scenario, using a general linear model (ABC-GLM) post-sampling regression adjustment for the 2,000 retained simulations (Leuenberger, Wegmann, 2010; Wegmann *et al.*, 2010). We report the mode and 95% highest posterior density (HPD) interval for each model parameter estimate.

The performance of the method for discriminating between competing historical models was assessed by analyzing test datasets simulated with the same number of loci and individuals as for the observed datasets (*i.e.*, pseudo-observed datasets). We simulated

2,000 such datasets for each competing model, using parameter values drawn from the same prior distributions as for the original analyses. We determined the relative posterior probabilities of competing models for each pseudo-observed dataset, using the model choice procedure, as described above (Wegmann *et al.*, 2010). Confidence in model choice was then estimated from the likelihood that a given scenario did not have the highest posterior probability of the competing scenarios when it was actually the true scenario (type I error), and the likelihood of a given scenario having the highest posterior probability when it was not the true scenario (type II error).

Projections of M. sylvestris *distribution during LGM*

We modeled the climatic niche of *M. sylvestris* from the current species distribution with BIOMOD package (v 1.1-7.00, 2011-08.03, (Thuiller *et al.*, 2009)) using 19 and eight bioclimatic variables download from WorldClim dataset v. 1.4 (http://www.worldclim.org/; (Hijmans *et al.*, 2005)), and records of the presence of *M. sylvestris* obtained for sampled individuals (*N*=381). We removed duplicated coordinate data points, resulting in 79 presences in total (Dataset S1) to evaluate the distribution at the LGM. These models make the assumptions that climate is one of the main factors driving species distribution and that the climatic niche of this species has remained largely unchanged in recent centuries. Details on ecological niche models (ENM) for each step are provided in Text S1.

Results

Population structure

Summary statistics for genetic variability are shown in Tables S3 and S4. For the 25 sites with at least four samples, the mean number of genotypes was 13.8±8.4, allelic richness was 3.7±0.4 (range: 2.6–4.4) and genetic diversity was 0.84±0.14 (range: 0.55–0.89), on average, across markers. F_{IS} values were low, with a mean of 0.03±0.07 per site and per marker, and 20 of the 25 populations having values below 0.10. Heterozygote deficit, estimated over the whole dataset, was low (F_{IS}=0.03, P<0.001). Between-site differences in allelic frequencies, estimated by calculating the mean F_{ST} across loci, were small (F_{ST}=0.10; range: 0.008-0.280) but significant, for all pairs of populations (P<0.02, Table S5).

The results of TESS analyses are shown in Figure 2. For *K*=2, the analyses revealed a clear west/east partitioning. The simulations for *K*=3 split the eastern cluster into north-eastern (NE) and south-eastern (SE) clusters. A similar pattern was obtained for *K*=4, whereas for *K*=5 a fourth cluster was identified in Bosnia-Herzegovina (Figures 2 and 3). STRUCTURE analyses generated congruent clustering patterns (Figures S2 and S3).

In TESS analyses, DIC values decreased monotonically from *K*=2 to *K*=6. Thus, increasing the number *K* of clusters continually improved the fit of the model to the data. However, DIC seemed to decrease more slowly after *K*=3 (Figure S4), suggesting that further increases in *K* provided little relevant information. In STRUCTURE analyses, the mode of the *Δ*K statistic was observed at *K*=2 (*ΔK*=377.1, *Pr*|*ln L*=-37040, Figure S5), but *ΔK* was still high at *K*=3 *(ΔK*=480, *Pr*|*ln L* = -37084.4), suggesting further improvement in the fit of the model. Based on the narrow geographic distribution of the clusters inferred at *K*>3, and the *ΔK* and DIC values obtained, *K*=3 was considered the most biologically relevant clustering solution for subsequent historical inference.

We used the TESS membership coefficient inferred at *K*=3 to define the three populations used in subsequent analyses. Genotypes were assigned to a given population if their membership coefficient for that population exceeded 0.55. Five genotypes could not be assigned to any population and were not included in subsequent analyses. The three populations are referred to hereafter as the “western” (W, red, *N*=213), the “north-eastern” (NE, blue, *N*=90) and “south-eastern” (SE, green, *N*=73) populations. The western population was relatively homogeneous, with 91% of genotypes having membership coefficients >0.9 for that population (Figures 2 and 3). The NE and the SE populations presented higher levels of admixture, with 31% (*N*=28) and 26% (*N*=19) of genotypes, respectively, having membership coefficients <0.9 (Figures 2 and 3). Admixed genotypes in the NE population could be assigned to both the W (*N*=7) and SE (*N*=14) populations, whereas admixture in the W and SE populations occurred only with the NE population (Figures 2 and 3).

Population-specific polymorphism summary statistics are shown in Table 1. Genetic differentiation was weakest between the two eastern populations (F_{ST}=0.04, *P*<0.001), and higher between the W and NE (F_{ST}=0.09, *P*<0.001) or W and SE (F_{ST}=0.12, *P*<0.001) populations. The W population had a level of genetic variability significantly lower than that of the NE and SE populations, in terms of both genetic diversity (H_E) (Wilcoxon signed rank

(WSR) tests: V=271, P=0.007 and V=229, P=0.008 respectively) and allelic richness (A_r) (WSR: V=351, P<0.0001 and V=345, P<0.0001, respectively) (Table 1). The W population contained a similar number of private alleles to the NE population (WSR: P=0.08), but significantly fewer than the SE population. The NE and SE populations had similar levels of genetic diversity (H_E) (WSR: P=0.7), but the SE population had a higher allelic richness (WSR: V=320, P<0.0001) and a larger number of private alleles (WSR: V=338, P<0.0001) (Table 2).

The map of interpolated allelic richness (Figure 4) showed that genetic diversity decreased with increasing latitude, being highest in the Balkans, Carpathians, Italy, and the Iberian Peninsula and lowest in Northern Europe. Within each of the three populations, genetic differentiation and geographic distance were significantly correlated (Table 1), consistent with an IBD model. The *Sp* statistic can be used to quantify spatial structure and can be compared between populations and/or species. Lower *Sp* values are associated with greater dispersal capacities and effective population sizes. Spatial *Sp* values were very low both overall and within each population (Table 1), lying close to 0 (Sp=0.04, P<0.0001). This suggests that each population had a high dispersal capacity and a large effective population size.

Models of population expansion and estimation of demographic parameters

We used ABC to determine statistically which of the following scenarios provided the most likely explanation for the existence of the three genetically differentiated populations of *M. sylvestris* in Europe: (*1*) the simultaneous expansion of three independent populations (W, SE and NE) that had differentiated during the last period of glaciations, with (model *a*) or without (model *b*) gene flow, (*2*) divergence and expansions of the W and SE populations, followed by the divergence and expansion of a NE population derived from the SE population (*i.e*, resulting from a more recent colonization wave front), with (model *c*) or without (model *d*) gene flow (Figure 1). The correlations between demographic parameters and summary statistics are presented in Figure S6. The relative posterior probabilities calculated for each model provided the strongest statistical support for model *a*, suggesting that the three populations (W, SE, NE) diverged simultaneously, with all populations growing exponentially and bidirectional gene flow occurring between each pair of populations during expansion (Table 2; Bayes factor for model *a*=4.37). The models without gene flow (models *c*

and *d*) had the lowest relative posterior probabilities (Table 2). The migration rates per generation were estimated at $m_{W\text{-}SE}$=0.01 [95% HPD]: [0-0.21], $m_{W\text{-}NE}$=0.03 [0–0.21], $m_{SE\text{-}NE}$=0.01 [0–0.22], $m_{SE\text{-}W}$=1.01 x10^{-15} [0–0.17], $m_{NE\text{-}SE}$=1.01 x 10^{-15} [0–0.21], and $m_{NE\text{-}W}$=0.05 [0–0.27]. We obtained estimates of effective population sizes of 40,883 [524-828,327] for N_W, 20,691 [505–565,916] for N_{NE}, and 40,882 [524-828,327] for N_{SE}. Using a generation time of 7.5 years, we estimated the population split to have occurred 303,016 years ago [120,963–545,649].

We also checked that the power of the analysis was sufficiently high to discriminate between the competing models: For model *a*, the type I error rate was 0, and the mean type II error rate was 0. Overall, ABC analyses provided clear strong support for gene flow between recolonizing refugia and for the simultaneous divergence of the three populations.

Ecological niche modeling

Model performance, as assessed from the AUC (Area Under the receiver operating characteristic Curve), was high for all six algorithms (Table S6), with an AUC value of 0.98±0.01, indicating that all six models fit the data well (Allouche *et al.*, 2006; Fieldings, Bell, 1997; Monserud, Leemans, 1992). With thresholds maximizing the TSS (True Skill Statistic), *M. sylvestris* had a good TSS value of 0.79±0.04. We ran ecological niche models with both sets of past climate data, CCSM2 and MIROC, but only MIROC gave consistent results across Europe. We therefore present projections based on MIROC data (Figures 5b and S7b). The projection onto current climate layers identified a putative suitable climate area essentially located in Western Europe for *M. sylvestris* (Figures 5a and S7a). The MIROC model predicted that the areas suitable for this species during the LGM were limited to lower latitudes than those considered suitable today (Figures 5 and S7). The climatic model suggested that populations of the European crabapple may have been maintained in areas further north than the typical glacial refugia (Hewitt, 2004), with possible continuity between the populations from Iberia, Italy and the Balkans. The predicted distribution however does not show a refugial distribution in isolated places as expected if the species survived in nunatacks in northern latitudes.

DISCUSSION

Paleodistribution modeling and genetic data allowed inferences on the phylogeography of the European crabapple, an endangered species and valued genetic resource for apple breeding. The distribution predicted based on climatic data was overall consistent with population genetic analyses, altogether suggesting population and range contractions of *M. sylvestris* during the last glaciation. However, while the climatic models predicted a continuous distribution, Bayesian analyses indicated the existence of three differentiated populations in Europe, likely resulting from differentiation in different glacial refugia. Iberia and the Balkans were easily identified as refugia, whereas the *M. sylvestris* populations from the Carpathian Mountains may have resulted from two alternative demographic scenarios of recolonization (*i.e*, later wave of recolonization from Balkans or from a Carpathian Mountains refugium). The ABC analyses made it possible to distinguish between competing scenarios, suggesting that the Carpathian Mountains were also a glacial refugium, as its split from the other populations was simultaneous. Spatially explicit analyses combining ENM and genetic data (Brown, Knowles, 2012; Carstens, Richards, 2007) would have been more powerful for inferences of precise locations of refugia, but the continuous distribution predicted by ENM impaired such analyses.

Glacial refugia for *Malus sylvestris*

Malus sylvestris displays a clear geographic pattern of population structure, with three strongly differentiated populations in Europe: (*i*) a western population (W) spanning a huge area from Iberia to Sweden, and an eastern group subdivided into (*ii*) a north-eastern (NE) population surrounding the Carpathian Mountains and (*iii*) a south-eastern (SE) population located at the north-east edge of the Balkan Peninsula. The strong differentiation between the three populations and the decreasing allele richness with increasing distance north from Southern Europe suggest that the European crabapple contracted its range to southern glacial refugia, at least one of which probably was in the Iberian Peninsula and another in the Balkans. Evidence for admixture was found at intermediate latitudes, but not in the southern-most populations, indicating that the greater variability observed in the south was due to relict populations from glacial refugia in these regions rather than admixture following secondary contact between recolonizing fronts.

The population growing around the Carpathian Mountains (NE population) displayed a low level of genetic differentiation from the Balkans population (SE population), but a high level of allelic richness. There were two possible origins for this population: a refugium from north-eastern Europe that came into secondary contact with the Balkans population, or a wave front from the Balkan refugium during the recolonization of Europe. ABC analyses evaluating the fit of various demographic models to microsatellite data showed that the most strongly supported scenario was the simultaneous divergence of the three populations, with gene flow between each pair of populations. The NE population, therefore, probably originated from isolation in a glacial refugium rather than during the postglacial recolonization of Europe. The inferred divergence dates are consistent with such a scenario, although they should be taken with caution, as not much data was available on generation time in *M. sylvestris* and overlapping generations were not modeled.

The geographic pattern of population structure uncovered in *M. sylvestris* is consistent with those found in other animal, plant and fungal taxa (Hewitt, 2004; Lascoux *et al.*, 2004; Schmitt, 2007; Vercken *et al.*, 2010) and, in particular, with the patterns commonly found in temperate forest trees (Heuertz *et al.*, 2006; Heuertz *et al.*, 2004; Magri *et al.*, 2006; Petit *et al.*, 2002). Indeed, the existence of a large Western European population and differentiated eastern populations in the Balkans Peninsula has been reported in *Quercus* sp*., Abies alba* and *Fraxinus excelsior*. In some other tree species, such as *Alnus glutinosa* (King, Feris, 1998) and the common beech *Fagus sylvatica* (Magri *et al.*, 2006)*,* the existence of additional refugia in the eastern part of Europe, particularly in the Carpathian Mountains, has been suggested on the basis of the high level of genetic diversity in cpDNA and pollen fossil records (Magri *et al.*, 2006). We clearly identified a distinct genetic cluster around the Carpathian Mountains in the European crabapple, on the basis of microsatellite markers. A similar situation has been suggested for the common ash *Fraxinus excelsior* (Heuertz *et al.*, 2004), but no clear origin of the Carpathian population could be established due to a lack of samples in this area. Our ABC analyses indicated that the NE populations originated at the same time as the other recolonizing populations, thus suggesting that the Carpathian Mountains may have acted as a glacial refugium for temperate forest tree species, as for other temperate species, including mammals, reptiles and amphibians (Provan, Bennett, 2008; Stewart *et al.*, 2009). The existence of such a “northern” glacial refugium would be

consistent with the results of ENM, suggesting that *M. sylvestris* may have survived at high latitudes. In some tree species, such as *Fagus sylvatica*, it has been suggested that such a northern refugium may have served as the main source population for the recolonization of Eastern Europe after the LGM, whereas the Balkan population located farther south spread over a much more limited area during post-glacial recolonization (Hu *et al.*, 2008).

For ash, silver fir, oaks and beech, Italy has been identified as a possible additional glacial refugium, isolated from other lineages by the Alpine barrier (Taberlet *et al.*, 1998). We detected no footprint of an Italian refugium in *M. sylvestris*. This may be due to the lack of pure *M. sylvestris* (*i.e*, membership coefficient >0.9) from this region. Our samples from Italy indeed were all introgressed by the cultivated apple *M. domestica*, which may be linked to the introduction of the cultivated apple in Europe by the Romans in Italy 3000 years ago (Juniper, Mabberley, 2006). For the western population, which probably expanded from an Iberian refugium, the level of allelic richness was high, but significantly lower than that in the eastern populations. These differences in genetic variability may be due to much more severe climatic episodes (*i.e*., arid and cold) during the Quaternary Period in this region than in other parts of Europe (Petit *et al.*, 2003). Tree populations that survived successive ice ages in the Iberian Peninsula were restricted to a few small suitable areas and were thus smaller than those in other parts of Europe.

Suture zones and recolonization fronts for the European crabapple

There were two main waves of recolonization from the glacial refugia in Europe: the whole of Western Europe, right into the north, was probably recolonized by a population from the Iberian Peninsula, whereas the population from the Carpathian Mountains spread, albeit to a lesser extent, northwards in Eastern Europe. The population from the Balkans refugium has not recolonized large areas.

Our microsatellite markers provided evidence of admixture in all three of the populations identified. The Iberian (Western) population contained the smallest number of admixed individuals and displayed the highest level of genetic differentiation from the other populations. This pattern suggests higher levels of recent genetic exchange between the Balkan (SE) and Carpathian (NE) populations. The lack of samples from Central Europe may have resulted in an underestimation of the number of individuals with admixed ancestry in

the western and eastern populations, but ABC analyses provided strong support and demonstrated a high level of confidence for the choice of the model assuming gene flow between populations, although migration rates could not be estimated precisely.

The biogeographic scenario uncovered in our study, with two main recolonization fronts in Western and Eastern Europe, has been demonstrated for many other temperate tree species (Heuertz *et al.*, 2004; Lascoux *et al.*, 2004). Suture zones in Central Europe, as detected for the European crabapple, are also typical of other temperate tree species (Heuertz *et al.*, 2006; Heuertz *et al.*, 2004; Liepelt *et al.*, 2009; Magri *et al.*, 2006; Petit *et al.*, 2002). However, the clear suture between the SE and NE populations and the evidence for admixture have not been reported before. These findings demonstrate the utility of nuclear microsatellite markers for retracing the ancient demographic history of populations and the extent of admixture in phylogeographic studies. Indeed, these markers appear to provide additional and more precise insight into the demographic history of species than the more frequently used organelle markers (Heuertz *et al.*, 2004; Magri *et al.*, 2006). In apple trees, chloroplast markers are not polymorphic enough for a high degree of resolution between closely related species (Robinson *et al.*, 2001), whereas microsatellite markers clearly discriminate between the species (Cornille *et al.*, 2012b).

Historical gene flow and dispersal capacity in the European crabapple

This species is dispersed by animals, but the weak spatial genetic structure within each population at the European scale, weak IBD patterns and low *Sp* values suggest that the European crabapple may dispersed over large distances. These results are consistent with those previously reported for scattered temperate tree species (Oddou-Muratorio *et al.*, 2004; Oddou-Muratorio, Klein, 2008). The high dispersal capacity of the European crabapple and its current wide distributional range extending into the northern-most parts of Europe identify this species as a rapid colonizer (Svenning, Skov, 2007), consistent with its high pioneering capacity (Larsen *et al.*, 2006).

Ecological niche models and reconstruction of the recolonization history of the European crabapple on the basis of its paleodistribution

ENM projections onto current climate layers resulted in a current suitable habitat for the European crabapple covering most of Europe, with the exception of southern Spain and Italy, consistent with the Euforgen distribution map (http://www.euforgen.org/distribution_maps.html), which extends further eastern than the region studied in this paper. The projection onto paleoclimate layers showed that the distribution of the European crabapple, like those of many other European tree species (Svenning *et al.*, 2008), was affected by climate changes during the Quaternary Period, with a contraction of the northern European crabapple populations. The modeling results suggested possible refugia in the European crabapple in the northeast, southeast and west, concordant with the distribution of the three genetic clusters. ABC results were also in agreement with these results showing that the three clusters expanded from three independent sources (Hewitt, 1996). Climatic conditions may have allowed *M. sylvestris* to grow at intermediate latitudes in Europe, extending up to northern parts of France and connecting glacial refugia during the LGM. Such northern distributions were demonstrated on the basis of pollen fossil records and genetic data, mostly in cold-tolerant tree species of the genera *Betula* and *Alnus* (Lascoux *et al*., 2004). The genetic data showed a possible high-latitude refugium in the Carpathian Mountains as well as besides and in southern refugia for *M. sylvestris*. Despite the overall concordance between ENM and genetic data, the predicted past distribution extends northwards up within the extent of the glaciers. This raises questions about the accuracy of ENM paleoreconstructions and their interpretation, including limitations concerning the available predictor variables (Araújo, Guisan, 2006). Indeed, in these models, the species niche is defined in terms of temperature and precipitation data only. Although climate is one of the main factors underlying species distribution, other factors will need to be taken into account in the future (Pearson, Dawson, 2003), including dispersal processes (using hybrid approaches), biotic interactions (*e.g*, competitors and pollinators) and stages of succession in forests. One of the main issues of niche modeling is the assumption of niche conservatism. If the species niche shifts over time, the species may not respond to climate change in a predictable way (Jezkova *et al.*, 2011). In addition, calibrating the climatic niche of species under current conditions and projecting them to non-analogous conditions in the past would lead to spurious response curves and therefore to naïve projection (Thuiller *et al.*, 2004). However, while limits currently exist in

paleoreconstruction distribution, ENM is becoming a valuable complementary approach to genetic demographic inference in many cases (Brown, Knowles, 2012),

Author contributions

Conceived and designed the experiments: GT, GP. Found the funding: GT, CA, GP. Performed the experiments: CA. Analyzed the data: CA. Contributed analysis tools: BC, GP, TA. Provide samples: LCB, KJ, RRI. Wrote the paper: CA, GT, LCB, TA, RMJM, RRI, GP.

Acknowledgments

We thank the *Région Ile de France* (PICRI), IDEEV and SBF (*Société Botanique de France*) for funding, *Plateforme de Génotypage* GENTYANE INRA UMR 1095 and Pauline Lasserre for assistance with genotyping. We thank all those who helped with sampling (Table S1). We thank Alex Edelman & Associates for editing English and three anonymous referees for very useful suggestions.

Figures

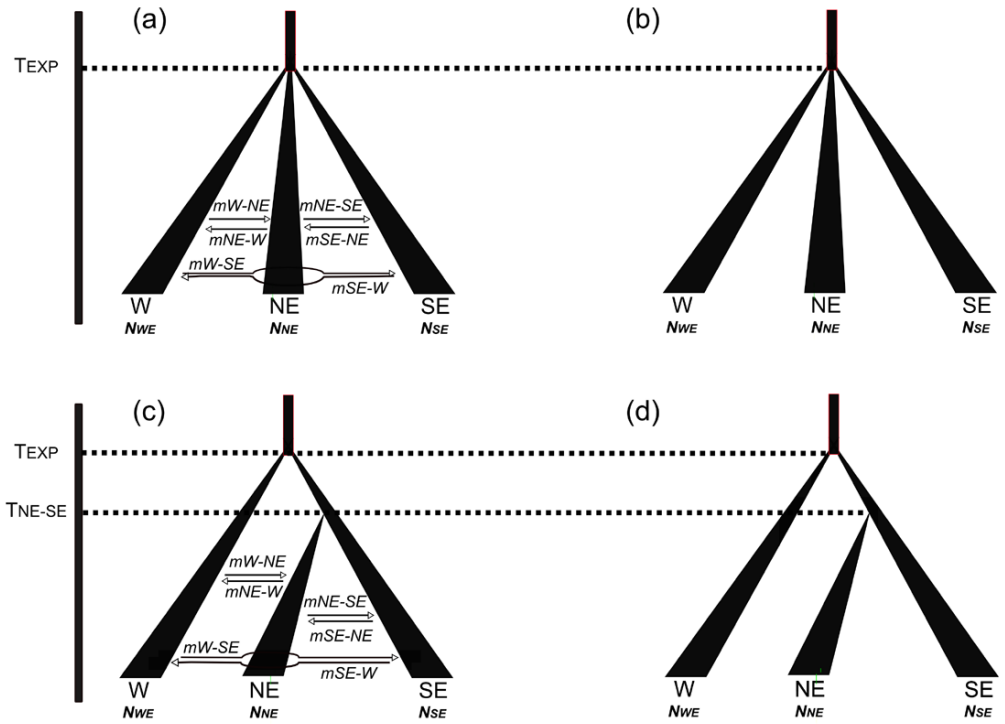


Figure 1. Demographic models compared in approximate Bayesian computations. Model a assumes that three populations (western (W), south-eastern (SE), north-eastern (NE)) diverged and underwent simultaneous exponential growth; bidirectional gene flow was assumed between all pairs of populations. Model *b* is identical to model *a* but with no gene flow. Model *c* assumes that the W and SE populations expanded simultaneously and that the NE population subsequently diverged from the SE population, with all populations undergoing exponential growth; bidirectional gene flow was assumed between all population pairs. Model *d* is identical to model *c* but without gene flow. *Nx:* Effective population size of population x*,* $m_{x\text{-}y}$: gene flow per generation from population x to population y; T_{EXP}: divergence time at the beginning of the Holocene; $T_{NE\text{-}SE}$: time at which the north-eastern population split from the south-eastern population.

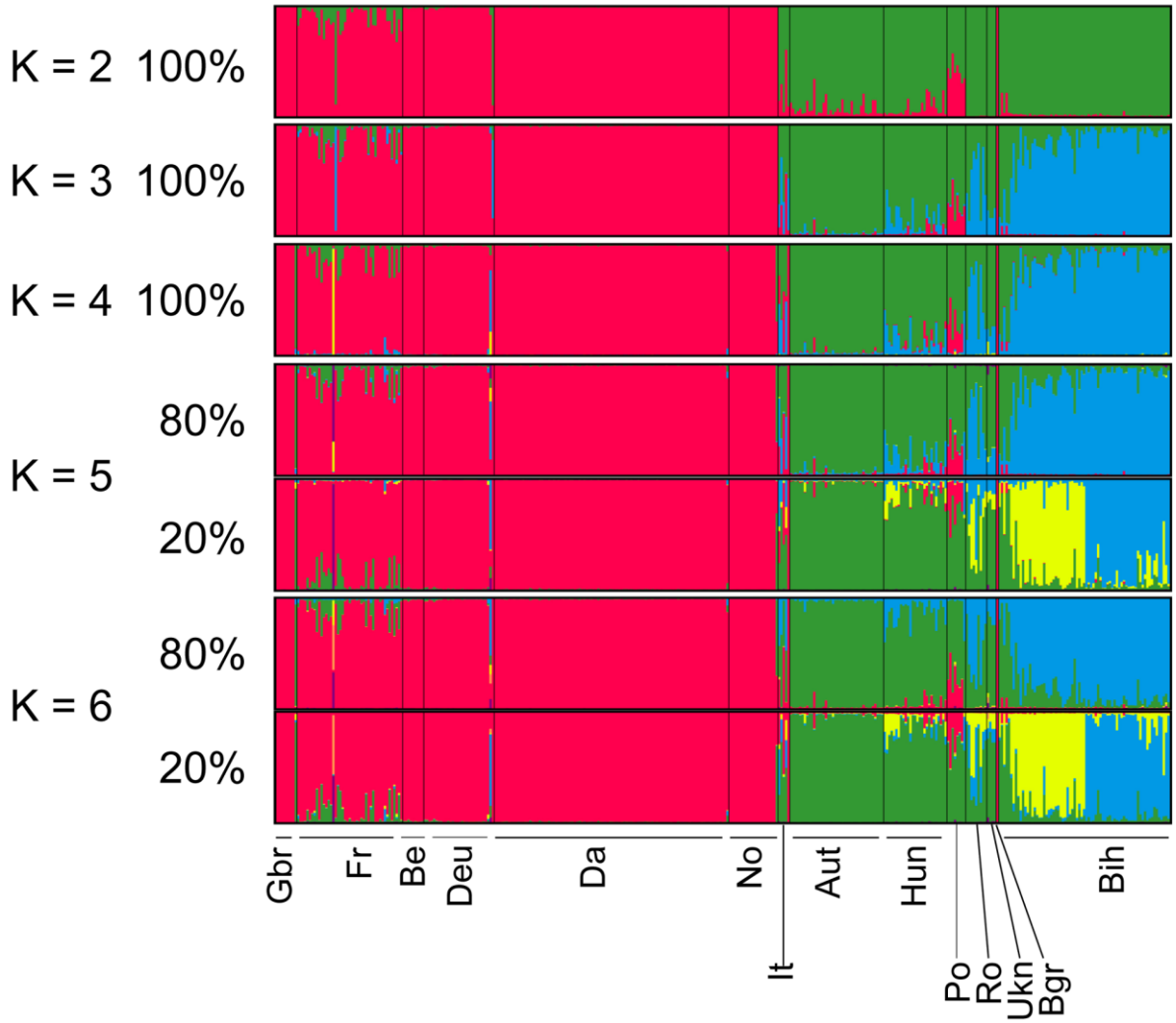


Figure 2. Population structure in European crabapple *Malus sylvestris* (*N*=381), inferred with the Bayesian clustering algorithm implemented in TESS. Each individual is represented by a vertical bar, partitioned into *K* segments representing the proportions of ancestry of its genome in *K* clusters. When several clustering solutions ("modes") were represented within replicate runs, the proportion of simulations represented by each mode is shown. Aut: Austria, Be: Belgium, Bgr: Bulgaria, Bih: Bosnia Herzegovina, Da: Denmark, Fr: France, Deu: Germany, Gbr: Great Britain, Hun, Hungary, It: Italy, No: Norway, Po: Poland, Ro: Romania, Ukn: Ukraine.

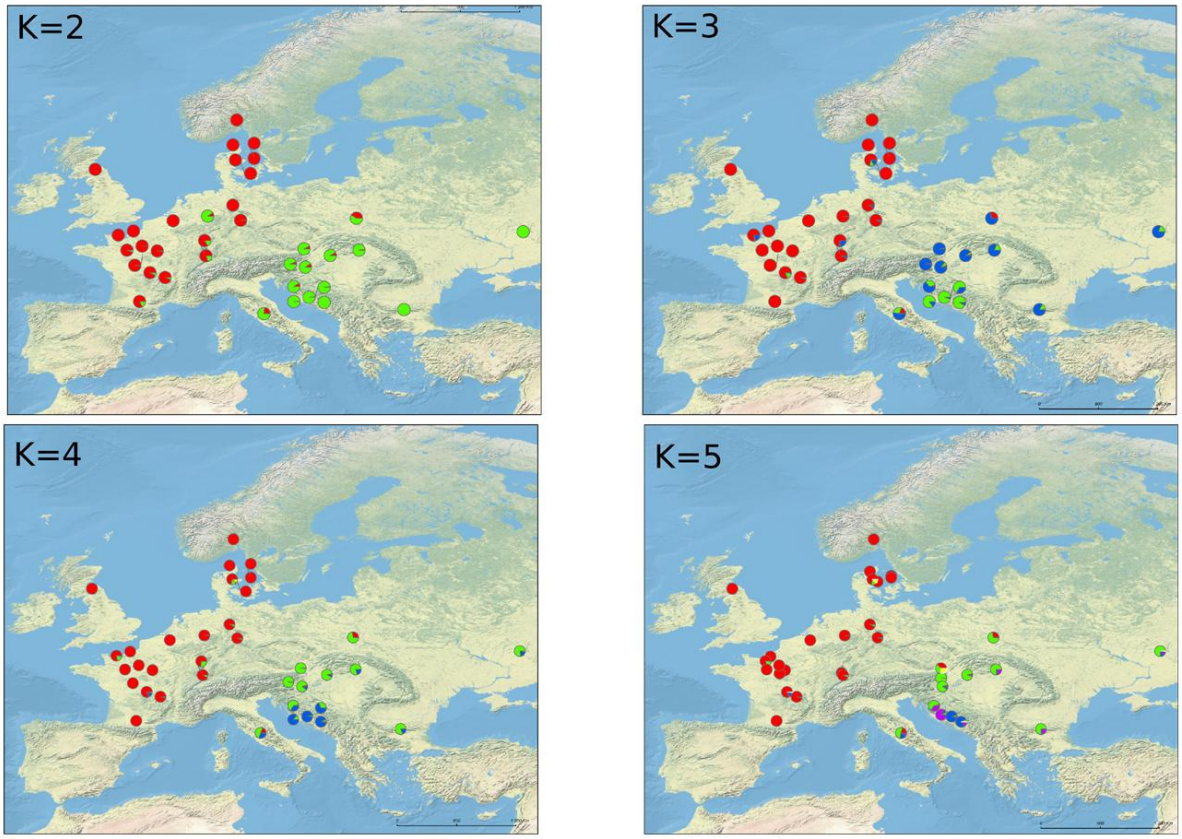


Figure 3. Maps representing the mean membership proportions for *K* clusters, for samples of *Malus sylvestris* collected from the same site. Membership proportions were inferred with the Bayesian clustering algorithm implemented in TESS. At *K*=5 the results presented are those of the minor clustering solution ("mode"), showing the geographic location of the fourth previously identified cluster (Figure 2).

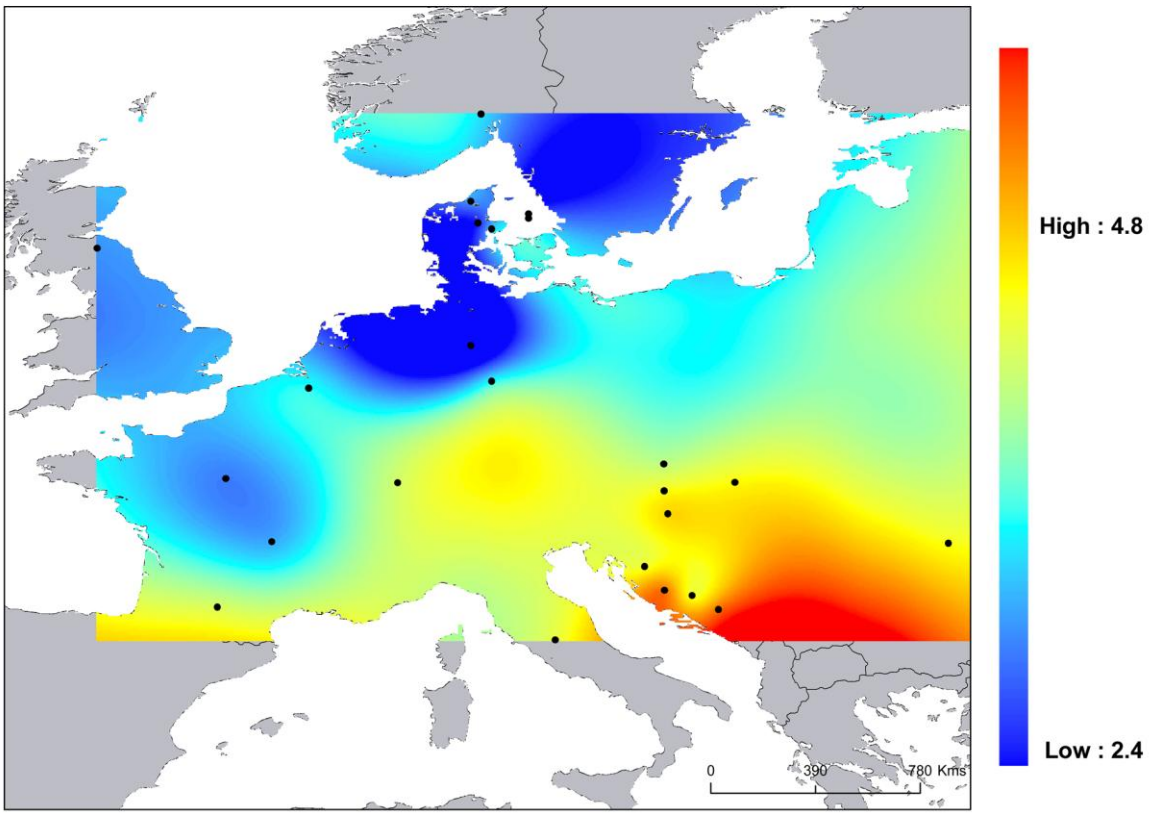


Figure. 4. Map of overall allelic richness at 22 sites.

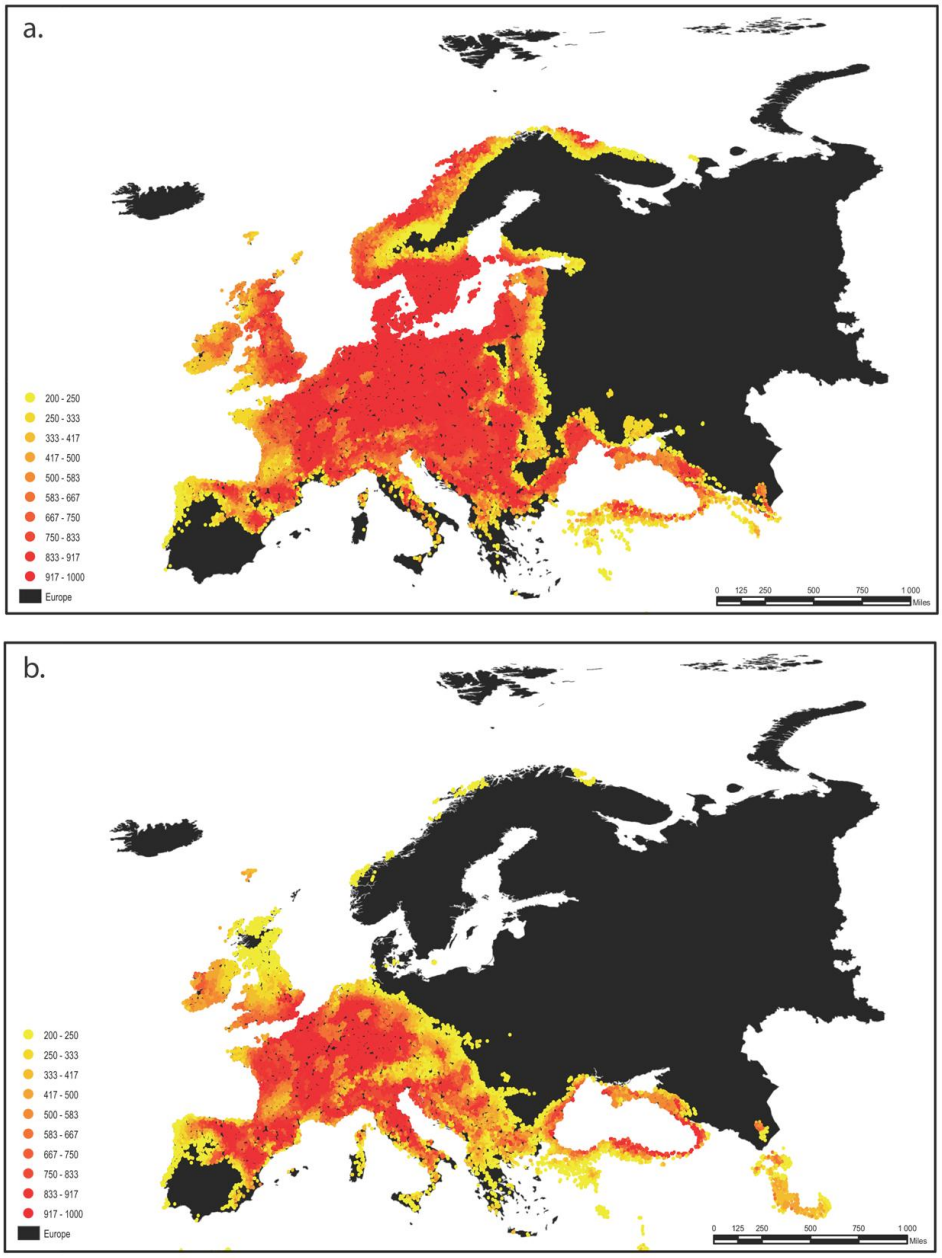


Figure 5. Ensemble forecasting of the six different algorithms predicting the current (a) and LGM (b) suitable climate area distribution for *Malus sylvestris*. The probabilities (x 1000) of a suitable habitat are given in the legend.

Tables

Table 1. Genetic polymorphism and spatial pattern of differentiation within each cluster in *Malus sylvestris*

	Cluster		
	W (red)	NE (blue)	SE (green)
Microsatellite polymorphism			
N	213	90	73
Ho	0.69	0.77	0.80
H_E	0.74	0.85	0.88
F_{IS}	0.07***	0.10***	0.09***
A_r	11.5	15.3	17.3
Ap	1.3	2.1	4.4
Spatial pattern (logarithm)			
Sp	0.005	0.004	-0.003
mean(Ln(*dist*))	1.3	1.3	0. 8
b	-0.006***	-0.004***	-0.001***
F(1)	0.03	0.01	0.02
r^2	0.008	0.008	0.009

N: sample size of each cluster, H_O and H_E: observed and expected heterozygosity, F_{IS}: inbreeding coefficient, A_r: mean allelic richness for loci, corrected by the rarefaction method, estimated for a sample size of 100, A_p: number of private alleles, corrected by the rarefaction method, estimated for a sample size of 100, *Sp*: spatial *Sp* parameter, mean(Ln(*dist*)): mean of the logarithm of the geographic distance between genotypes, *b* regression slope between F_{ij} and the logarithm of geographic distance, $F_{(1)}$: mean F_{ij} between individuals from the first distance class, *: $0.05<P\leq0.01$, **: $0.01<P\leq0.001$, ***: $P<0.001$, r^2: squared correlation coefficient.

Table 2. Relative posterior probabilities (*p*) and Bayes factor (*BF*) for the four historical models compared by approximate Bayesian computations. The models are described in Figure 1.

Model	*p*	*BF*
a	0.81	4.37
b	0.19	0.23
c	$2.3e^{-12}$	$2.3e^{-12}$
d	$1.0297e^{-56}$	$1.03e^{-56}$

Supporting information

Table S1. Description of the *Malus sylvestris* accessions analysed with their geographical origin and providers and acknowledgement.

Figure S1. Sampling of the different *Malus sylvestris* locations through Europe

Table S2. Prior distributions used in approximate Bayesian computations.

Dataset S1. X/Y coordinates of presences of *M. sylvestris* in Europe used for ENM

Text S1. Ecological Niche Models methodology used to project paleodistribution of *Malus sylvestris*

Table S3. Summary statistics for each *Malus sylvestris* sampling site.

Table S4. Summary statistics for the 26 microsatellite loci in *Malus sylvestris***.**

Table S5. Pairwise genetic differentiation (F_{ST}) among the 25 localities**.**

Figure S2. Bayesian clustering results of *Malus sylvestris* in Europe (N=381) using the program STRUCTURE from *K*=2 to *K*=6.

Figure S3. Maps of mean membership probabilities per locality from the STRUCTURE analysis for *Malus sylvestris* assuming 2 to 5 clusters.

Figure S4. Estimated number of populations in *Malus sylvestris* from TESS analyses using the DIC.

Figure S5. Estimated number of populations in *Malus sylvestris* from STRUCTURE analyses using the *ΔK*.

Figure S6. PCA on 3,000 simulations for *Malus sylvestris*.

Table S6. AUC Index for ENM ran with eight and 19 bioclimatic variables for each of the six models and each repetition

Figure S7. Ensemble forecast using *M. sylvestris* presence records and pseudo-absences projected onto the map of Europe and Western Russia using 19 bioclimatic variables.

Table S1. Description of the *Malus sylvestris* accessions analysed with their geographical origin and providers.

Country	Nb	Provider
Austria	40	Heino Konrad (Federal Research and Training Centre for Forests, Vienna, Austria), Thomas and Bernhard Kirisits (Institute of Forest Entomology, Forest Pathology and Forest Protection, Vienna, Austria)
Belgium	9	ILVO*, IRR
Bosnia y Herzegovina	73	Dalibor Ballian (Faculty of Forestry, University of Sarajevo, Bosnia-Herzegovina)
Bulgaria	1	Petya Gercheva, Argir Zhivondov, Valentina Bojkova and Anna Matova (Fruit-Growing Institute, Plovdiv, Bulgaria)
Denmark	100	Anders Larsens (Forest and Landscape, Department for Management of Forest Genetic Resources, Denmark)
France	45	BLC, François Laurens, TG, PG, Pascal Heitlzer, Dominique Beauvais, Nicolas Feau, Aurélien Cabaret, ILVO*, IRR, Jean-Pierre Rioult (Université de Caen Basse-Normandie, EREM - Equipe de Recherche et d'Etudes en Mycologie, France), Nicolas Feau (Forest Sciences Centre, UBC, Vancouver, Canada), François Laurens (INRA, Equipe FruitQual et CaDiPom, France)
Germany	30	JK and Wilfried Steiner (Northwest German Forest Research Institute, Germany)
Great Britain	9	Stephens Cavers (NERC, Centre for Ecology and Hydrology, UK)
Hungary	27	Lazlo Nyari (Forest Genetics and Forest Tree Breeding, Göttingen, Germany)
Italy	5	Stefano Porta, Alfredo Martignoni, Alberto Dominicci and Emanuela Fabrizi (Monti Simbruini Regional Park, Italy), François Salomone,
Norway	21	Per Avid (Agder Natural History Museum and Botanical Garden, Norway)
Poland	8	Jan Kowalsky and Dzmitry Kahan (Forest Research Institute, Poland)
Romania	9	Luian A Curtus (Transylvania University Brasov, Faculty of Forest Sciences, Romania)
Ukraine	4	Roman Volansyanchuk

Nb : Number of sampled individuals

* ILVO - PLANT Plant -Growth and Development, Melle, Belgium

We also want to thank Dominique Beauvais (*Abbaye de Beauport*, Paimpol, France); Alberto Dominici and Emanuela Fabrizi (Monti Simbruini Regional Park, Italy), Carlos Herrera (Estacion Biologica de Donana, CSIC, Spain), Francisco Donaire (Jardín Botánico La Cortijuela, Sierra Nevada, Granada, Spain), Roman Volansyanchuk (Ukrainian Research Institute of Forestry and Forest Melioration, Ukraine) for providing *M. sylvestris* samples. We also thank Thierry Genevet, Frédéric Tournay (*Jardin Botanique de Strasbourg*), Levente Kiss, the East Malling Research Station (UK), Philip Forsline and the Plant Genetic Resources Unit in Geneva (NY).

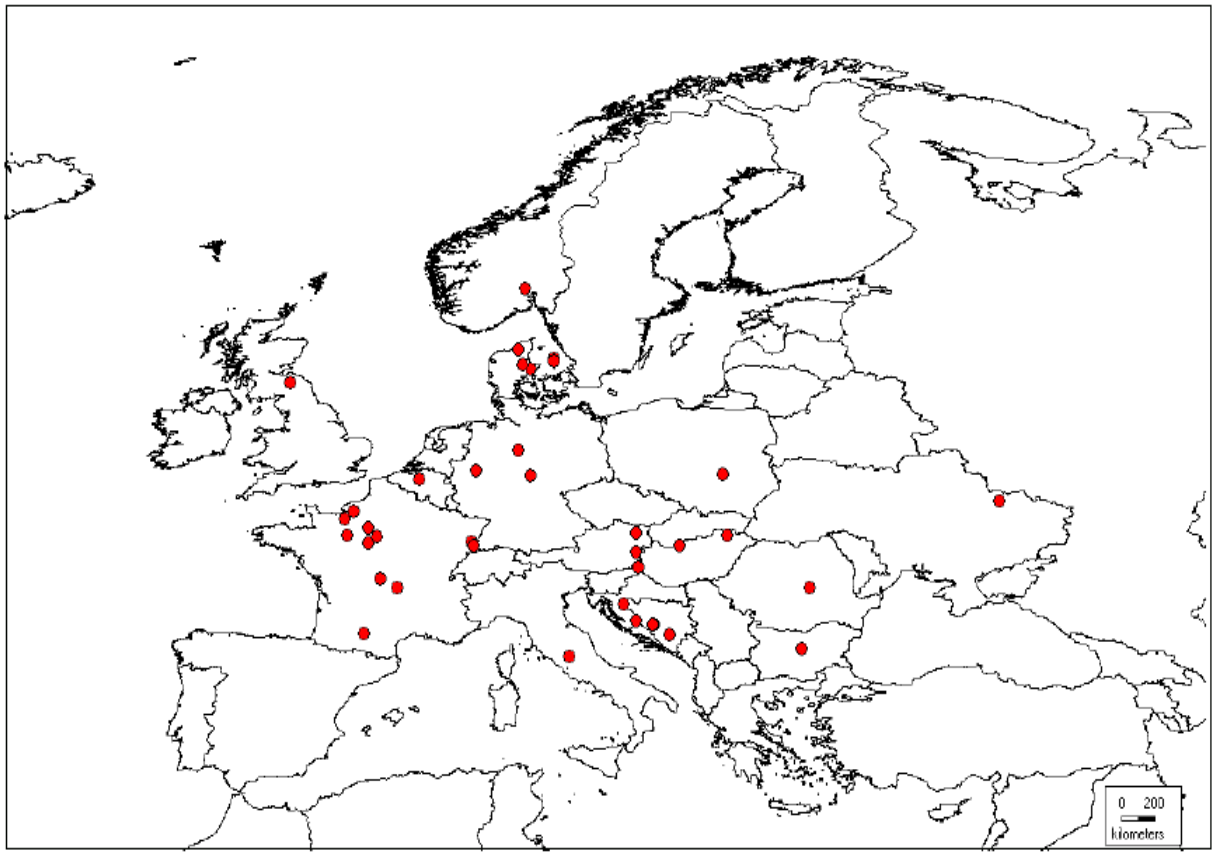


Figure S1. Sampling of the different *Malus sylvestris* locations through Europe

Table S2. Prior distributions used in approximate Bayesian computations. Prior distributions are uniform between lower and upper bound. Parameters are introduced in Fig.1.

Parameter	distribution	Lower Bound	Upper Bound
N_{SE}	log unif	500	1,000,000
N_{NE}	log unif	500	1,000,000
N_{W}	log unif	500	1,000,000
T_{EXP}	log unif	0	50,000
$T_{NE\text{-}SE}$	log unif	0	50,000
$m_{W\text{-}SE}$	log unif	0	0.05
$m_{W\text{-}NE}$	log unif	0	0.05
$m_{SE\text{-}NE}$	log unif	0	0.05
$m_{SE\text{-}W}$	log unif	0	0.05
$m_{NE\text{-}SE}$	log unif	0	0.05
$m_{NE\text{-}W}$	log unif	0	0.05
G_{W}	unif	-1	0
G_{NE}	unif	-1	0
G_{SE}	unif	-1	0
μ	logunif	0.00001	0.02
α	unif	1	30

log unif : log-uniform distribution; unif: uniform distribution; μ : mutation rate*;* α: shape parameter of the gamma distribution

Dataset S1. X/Y coordinates of presences of *Malus sylvestris* in Europe used for ENM

X	Y
-2.756	55.611
-2.756	55.611
-2.755	55.612
-2.755	55.612
-2.754	55.612
-2.754	55.612
-2.754	55.612
-2.754	55.612
0.261	49.042
0.373	48.267
0.373	48.268
0.373	48.268
0.717	49.450
1.270	43.604
1.499	48.643
1.544	47.901
2.034	48.199
2.214	46.228
3.087	45.790
4.327	50.939
6.316	61.550
6.900	58.367
7.186	47.957
7.305	47.760
7.460	51.364
8.160	62.960
9.750	57.183
9.751	52.358
9.904	44.808
9.983	56.465
10.048	59.987
10.298	59.128
10.330	63.000
10.351	59.898
10.368	59.783
10.416	60.211
10.439	59.744
10.450	56.260
10.452	51.166
10.454	59.293
10.536	59.664
10.612	60.074
10.654	59.777

10.661	59.319
10.743	60.056
10.770	60.114
10.864	60.115
10.885	59.963
10.922	59.951
11.317	59.187
11.683	56.766
13.080	42.000
13.270	41.900
13.350	41.900
15.573	44.971
16.114	47.885
16.115	47.883
16.117	47.881
16.118	47.881
16.119	47.881
16.217	48.386
16.232	44.159
16.346	46.729
17.162	43.995
17.195	44.004
18.049	43.520
18.596	47.780
20.702	51.415
20.731	51.253
20.734	51.257
20.750	51.270
20.760	51.264
20.765	51.026
20.811	51.240
21.216	48.268
22.564	50.656
25.299	42.867
25.746	45.746
36.132	49.948

Text S1. Ecological Niche Models methodology used to project paleodistribution of *Malus sylvestris*

Current bioclimatic and past LGM climate data were downloaded from WorldClim dataset v. 1.4 (http://www.worldclim.org/; (Hijmans *et al.*, 2005)), at a 2.5 arc-minute resolution. LGM climate data were selected from general circulation model (GCM) simulations from two models: the Community Climate System Model (CCSM) (Collins *et al.*, 2006) and the Model for Interdisciplinary Research On Climate (MIROC, version 3.2) (Hasumi, Emori, 2004).

For Ecological Niche Models (ENM), we first examined 19 bioclimatic variables. We assessed the pairwise correlation of values for the 19 bioclimatic variables by calculating Pearson's correlation coefficients, to prevent autocorrelation and the overfitting of our data. We retained only the variables that were not strongly correlated with each other (*i.e*, Pearson's correlation coefficient<0.75). The eight climate variables retained were: mean diurnal temperature range, isothermality (*i.e.,* mean diurnal temperature range/annual range of temperature), minimum temperature of the coldest month, annual range of temperature, mean temperature of wettest quarter, mean temperature of the driest quarter, seasonality of precipitation, and precipitation during the coldest quarter. These variables summarize the means and variation of temperature and precipitation and therefore probably summarize the dimensions of climate determining the distribution of *M. sylvestris*. Estimated distributions based on the model including eight bioclimatic variables were not markedly different from those obtained with the model based on all 19 of the initially considered bioclimatic variables (see results).

Species records. Records of the presence of *M. sylvestris* were obtained for sampled individuals (*N*=381). We removed duplicated coordinate data points, resulting in 79 presences in total (Dataset S1). As most of the models required presence/absence data, we generated 79 pseudo-absences by the Species Range Envelope (SRE) strategy in BIOMOD and, because different selections can provide different results, the models were run with 10 different sets of pseudo absences. SRE strategy consists to generate random selection of points from all points outside the suitable area estimated by a rectilinear surface envelope from the presence sample (*i.e.* SRE) (Thuiller *et al.*, 2009). Bioclimatic variables were

downscaled to a resolution of 2.5 minutes, by bilinear interpolation for all records from the global climate model (GCM) output listed above.

Ecological niche models (ENM). Climatic niche models were constructed with the BIOMOD package (package BIOMOD v 1.1-7.00, 2011-08.03, (Thuiller *et al.*, 2009)). Details are provided in supplementary material. Many different ENM algorithms exist which may provide substantially deviating predictions (Pearson *et al.*, 2006). There is currently no consensus concerning the most suitable ENM algorithm for species distribution modeling (Araújo, Guisan, 2006; Segurado, Araújo, 2004; Svenning *et al.*, 2011; Tsoar *et al.*, 2007), but combining ENM projections in an ensemble forecast framework has been proposed as a strategy to reduce model uncertainty (Araújo, New, 2007). We therefore performed simulations with multiple ENM algorithms, to predict the current and past (CCSM2 and MIROC models) distribution of *M. sylvestris*. We created ensembles consisting of projections derived from six statistical algorithms available in the BIOMOD package: GAM (General Additive Model), a flexible and automated approach to identifying and describing non-linear relationships between predictors and response, GLM (General Linear Model), a regression method that is a generalization of the multiple regression model that used the so-called link function to accommodate response variables that are distributed other than normally, GBM (Generalized Boosting Model), a machine learning method which combines a boosting algorithm and a regression tree algorithm to construct an 'ensemble' of trees, RF (Random Forest), a machine learning method which is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest, CTA (Classification Tree Analysis), a classification method running a 50-fold cross-validation to select the best trade-off between the number of leaves of the tree and the explained deviance, and MARS (Multivariate Adaptive Regression Splines), a generalization of stepwise linear regression for large numbers of predictor variables, in the *R* v.2.14.0 environment (R, 2011). More details about these modeling techniques can be found in Thuiller *et al.* (2009). We first performed the calibration procedure with 80% of the data. We then evaluated the predictive model performance of a species distribution mode, we used a random subset of 80% of the data to calibrate the model, then we used the remaining 20% for evaluation, using a threshold independent method, AUC. The data splitting approach was then replicated 10 times, from

which we calculated the mean AUC as well as the mean TSS (True Skill Statistic) value. The TSS is the sum of the sensitivity (proportion of actual positives which are correctly identified as such) and the specificity (proportion of negatives which are correctly identified). The final calibration of every model for making predictions uses 100% of available data.

Validation of model robustness*.* We evaluated the robustness of the predictive performance of each algorithm, by calculating the Area Under the receiver operating characteristic Curve (AUC) (Fieldings, Bell, 1997; Zweig, Campbell, 1993), which quantifies the performance of the model. The value of AUC indicates the proportion of the time a random selection from the positive group will score higher than a random selection from the negative group (Fieldings, Bell, 1997). Using the eight selected bioclimatic variables, we produced, for each climatic dataset, a total of 60 models for the current time (6 models x 10 repetitions), from which we calculated an ensemble forecast scenario by weighted average consensus based on AUC values (Araújo, New, 2007; Marmion *et al.*, 2009). The potential problems raised on the use of AUC (Lobo *et al.*, 2008) as a measure of model performance were potentially minor here because AUC was used to select the best models for a given species within a fixed geographical area and using the same pseudo-absences. For each LGM model (CCSM2 and MIROC), we calculated an ensemble forecast from the current prediction and then projected probabilities of presence onto a map, using ArcGis (ESRI, Redlands, CA).

Table S3. Summary statistics for each *Malus sylvestris* sampling site with more than four individuals

Site ID	Country	*N*	*Ho*	*He*	F_{IS}	*Ar*
Aut_hern	Austria	10	0.78	0.81	0.04 NS	4.1
Aut_stock	Austria	30	0.77	0.82	0.05***	3.9
Bel	Belgium	9	0.73	0.75	0.03 NS	3.6
Bih_jab	Bosnia	6	0.80	0.80	0.002NS	4.0
Bih_liv	Bosnia	27	0.83	0.87	0.04***	4.4
Bih_olo	Bosnia	18	0.74	0.81	0.09***	4.0
Bih_pre	Bosnia	18	0.80	0.85	0.06***	4.4
Da_he	Denmark	17	0.66	0.69	0.05***	3.5
Da_k	Denmark	27	0.63	0.69	0.08***	3.2
Da_ka	Denmark	28	0.67	0.69	0.04***	3.4
Da_no	Denmark	28	0.66	0.69	0.04***	3.4
Fr_clermt	France	6	0.63	0.71	0.11*	3.4
Fr_Hom	France	14	0.78	0.79	0.015*	3.9
Fr_orl	France	4	0.81	0.70	-0.17NS	3.3
Fr_tou	France	6	0.83	0.77	-0.08NS	3.9
Deu_pop0	Germany	9	0.71	0.75	0.06*	3.6
Deu_pop1	Germany	8	0.66	0.67	0.01NS	3.1
Deu_pop2	Germany	12	0.62	0.55	-0.13NS	2.6
Gbr	Great Britain	9	0.70	0.73	0.04NS	3.4
Hun_p2	Hungary	13	0.75	0.84	0.11***	4.1
Hun_p3	Hungary	6	0.74	0.84	0.11***	4.2
It	Italy	5	0.80	0.82	0.03NS	4.0
No	Norway	21	0.68	0.73	0.07***	3.5
Ro	Romania	9	0.75	0.83	0.10***	4.1
Ukn	Ukraine	4	0.82	0.86	0.04NS	4.3
Full dataset		344	0.73	0.76	0.03***	

N: Sample size at the site, H_O and H_E: observed and expected heterozygosities, F_{IS}: inbreeding coefficient, A_r: mean allelic richness for loci, corrected by the rarefaction method, estimated for a sample size of 6, *: $0.05<P\leq0.01$, **: $0.01<P\leq0.001$, ***: $P<0.001$, NS: non-significant

Table S4. Summary statistics for the 26 microsatellite loci in *Malus sylvestris*.

	Range	H_O	H_E	F_{IS}	A_r (N=6)	NA
Ch01f03	139-247	0.72	0.86	0.15***	4.33	0.08
Ch01h01	104-160	0.87	0.91	0.04***	4.81	0.02
Ch01h10	84-144	0.84	0.89	0.06***	4.68	0.05
Ch02c06	204-290	0.74	0.95	0.22***	5.29	0.11
Ch02d08	199-258	0.78	0.83	0.06***	4.11	0.04
Ch05f06	160-198	0.73	0.88	0.17***	4.57	0.09
Ch01f02	159-225	0.86	0.93	0.07***	5.07	0.04
Ch02c11	202-266	0.80	0.87	0.08***	4.47	0.04
Hi02c07	97-148	0.79	0.87	0.10***	4.45	0.04
Ch02c09	202-266	0.73	0.83	0.12***	4.17	0.05
Ch03d07	97-148	0.67	0.79	0.15*	3.92	0.06
Ch04c07	234-262	0.61	0.76	0.19***	3.78	0.06
GD12	167-228	0.66	0.78	0.15***	3.73	0.07
CH02b03b	98-159	0.85	0.87	0.03***	4.49	0.02
CH02b12	149-193	0.73	0.85	0.15***	4.30	0.06
Hi03a10	75-199	0.52	0.70	0.26***	3.35	0.11
MS06g03	114-166	0.64	0.87	0.26***	4.42	0.12
CH Vf1	203-302	0.75	0.85	0.11***	4.28	0.06
CH03d12	132-222	0.69	0.85	0.18***	4.26	0.09
CH05a05	128-177	0.70	0.81	0.14***	4.08	0.05
CH05f04	98-163	0.36	0.58	0.39***	2.80	0.14
CH05g08	191-253	0.80	0.87	0.08***	4.54	0.04
NH009b	148-174	0.76	0.86	0.11***	4.28	0.06
CH04e03	153-216	0.78	0.90	0.13***	4.79	0.04
CH02g01	138-171	0.69	0.83	0.16***	4.07	0.08
NZ02b01	182-237	0.76	0.91	0.17***	4.89	0.09
Total		0.73	0.84	0.14***		

Range: size range of SSR alleles in bp, H_O and H_E: observed and expected heterozygosity, F_{IS}: inbreeding coefficient, A_r: mean allelic richness for loci, corrected by rarefaction method estimated for sample a size of 6, NA: Null alleles.

Table S5. Pairwise genetic differentiation (F_{ST}) among the 25 localities.

Site	AutHern	Aut.stock	Bel	Bih.jab	Bih.liv	Bih.olo	Bih.pre	Da.he	Da.k	Da.ka	Da.no	Fr.clmt	Fr.Hom	Fr.or	Fr.tou	Deu.pop0	Deu.pop1	Deu.pop2	Gbr	Hun.p2	Hun.p3	It	No	Ro
Aut.stock	0.01																							
Bel	0.11	0.11																						
Bih.jab	0.08	0.08	0.12																					
Bih.liv	0.08	0.08	0.11	0.05																				
Bih.olo	0.11	0.11	0.13	0.10	0.06																			
Bih.pre	0.09	0.08	0.13	0.08	0.03	0.04																		
Da.he	0.14	0.14	0.05	0.15	0.14	0.17	0.16																	
Da.k	0.15	0.15	0.06	0.16	0.15	0.18	0.18	0.03																
Da.ka	0.14	0.13	0.05	0.15	0.14	0.18	0.17	0.03	0.04															
Da.no	0.15	0.14	0.05	0.15	0.15	0.18	0.18	0.01	0.02	0.02														
Fr.clmt	0.12	0.12	0.06	0.14	0.13	0.15	0.15	0.06	0.06	0.07	0.07													
Fr.Hom	0.09	0.08	0.05	0.10	0.09	0.11	0.11	0.08	0.07	0.09	0.09	0.05												
Fr.orl	0.13	0.11	0.03	0.12	0.11	0.15	0.14	0.05	0.05	0.04	0.04	0.05	0.05											
Fr.tou	0.09	0.08	0.04	0.08	0.08	0.12	0.11	0.06	0.08	0.05	0.06	0.05	0.07	0.07										
Deu.pop0	0.12	0.12	0.03	0.12	0.10	0.13	0.13	0.02	0.03	0.04	0.04	0.05	0.03	0.05	0.04									
Deu.pop1	0.16	0.14	0.08	0.15	0.13	0.18	0.17	0.04	0.04	0.07	0.05	0.10	0.07	0.08	0.09	0.04								
Deu.pop2	0.24	0.21	0.11	0.24	0.19	0.23	0.23	0.08	0.09	0.10	0.10	0.17	0.14	0.16	0.17	0.06	0.11							
Gbr	0.12	0.12	0.02	0.11	0.11	0.15	0.14	0.05	0.06	0.05	0.05	0.05	0.07	0.03	0.02	0.05	0.08	0.14						
Hun.p2	0.03	0.04	0.10	0.04	0.05	0.08	0.06	0.14	0.15	0.14	0.15	0.13	0.08	0.11	0.09	0.11	0.14	0.20	0.11					
Hun.p3	0.05	0.05	0.11	0.03	0.05	0.07	0.07	0.14	0.15	0.15	0.15	0.11	0.08	0.09	0.09	0.12	0.15	0.25	0.11	0.02				
It	0.09	0.08	0.09	0.09	0.08	0.11	0.09	0.14	0.15	0.15	0.15	0.13	0.09	0.11	0.09	0.11	0.17	0.25	0.11	0.08	0.06			
No	0.12	0.12	0.04	0.13	0.12	0.15	0.15	0.02	0.04	0.01	0.01	0.04	0.06	0.03	0.03	0.03	0.07	0.10	0.02	0.12	0.12	0.13		
Ro	0.09	0.09	0.13	0.07	0.05	0.07	0.07	0.16	0.17	0.15	0.16	0.16	0.12	0.13	0.11	0.13	0.17	0.24	0.14	0.06	0.06	0.10	0.14	
Ukn	0.06	0.05	0.11	0.05	0.05	0.09	0.06	0.15	0.16	0.15	0.15	0.12	0.09	0.12	0.09	0.12	0.16	0.26	0.13	0.03	0.05	0.07	0.12	0.05

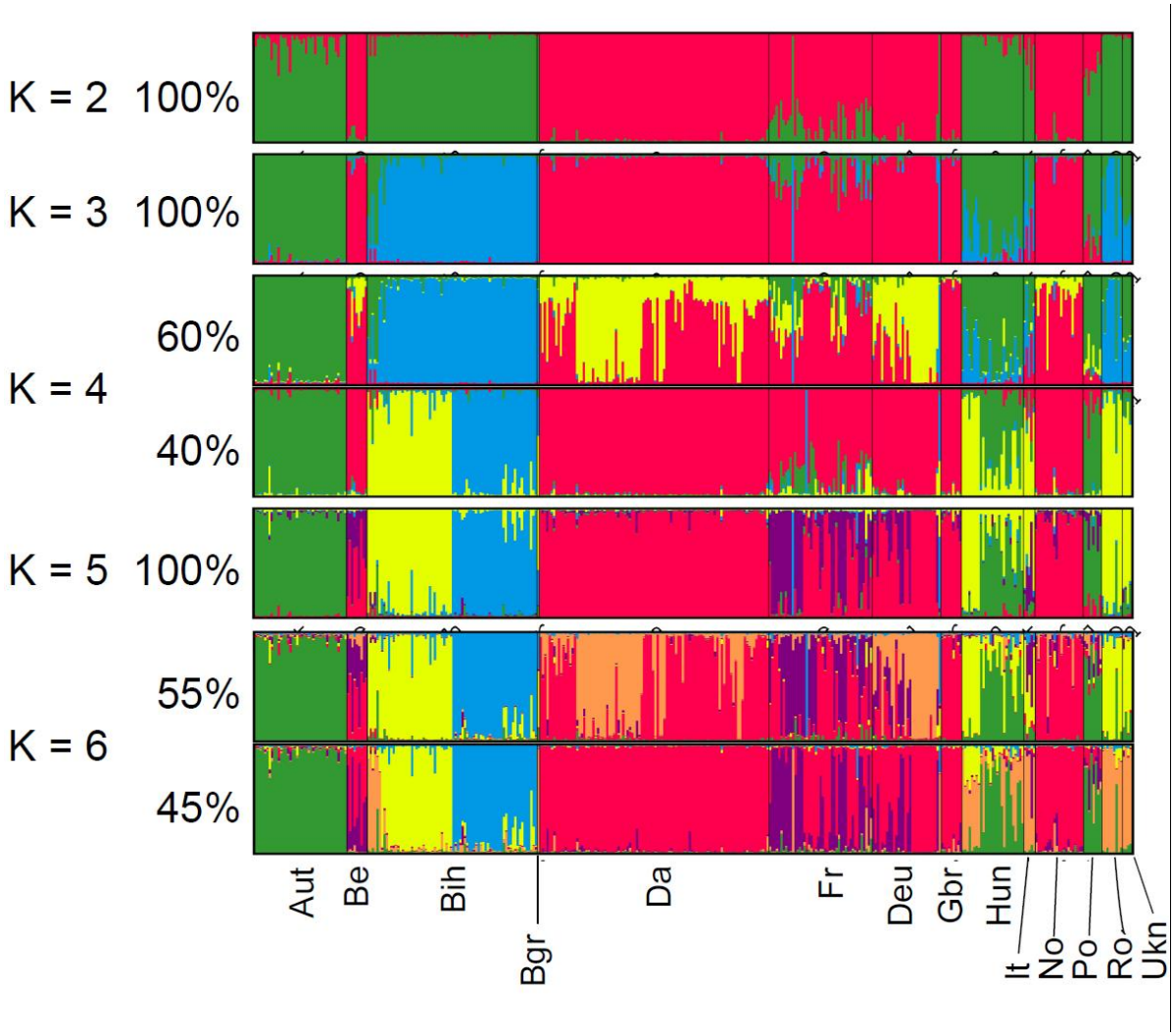


Figure S2. Bayesian clustering results of *Malus sylvestris* in Europe (*N*=381) using the program STRUCTURE from *K*=2 to *K*=6. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome in *K* clusters. When several clustering solutions ("modes") were represented within replicate runs, we give the proportion of simulations represented by each mode. For better visualisation we grouped localities from the same country together. Aut: Austria, Be: Belgium, Bih: Bosnia Herzegovina, Da: Denmark, Fr: France, Deu: Germany, Gbr: Great Britain, Hun, Hungary, It: Italy, No: Norway, Po: Poland, Ro: Romania, Ukn: Ukraine, Bgr: Bulgaria.

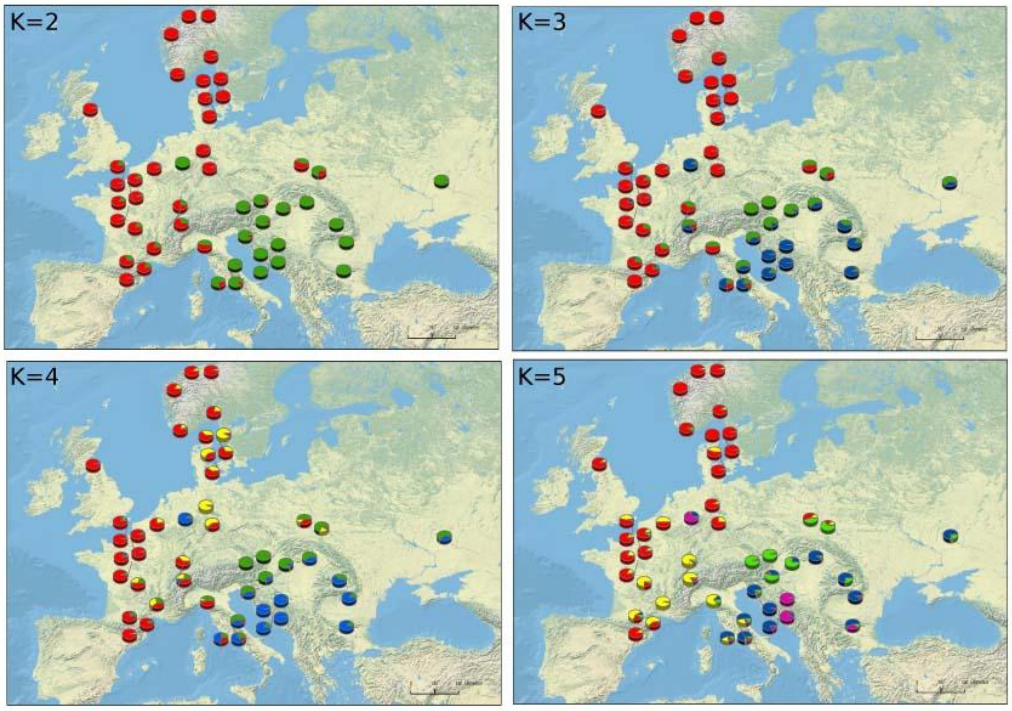


Figure S3. Maps of mean membership probabilities per locality from the STRUCTURE analysis for *Malus sylvestris* assuming 2 to 5 clusters

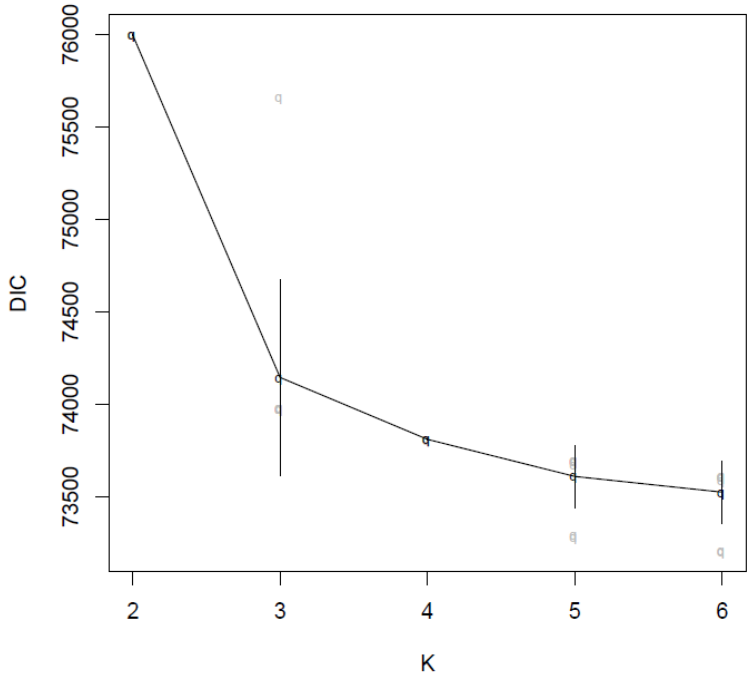


Figure S4. Estimated number of populations in *Malus sylvestris* from TESS analyses using the DIC.

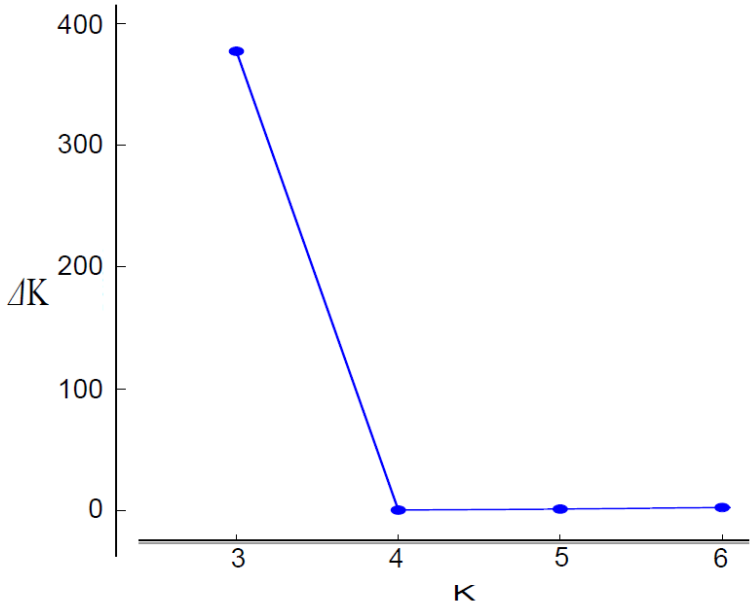


Figure S5. Estimated number of populations in *Malus sylvestris* from STRUCTURE analyses using the *ΔK*.

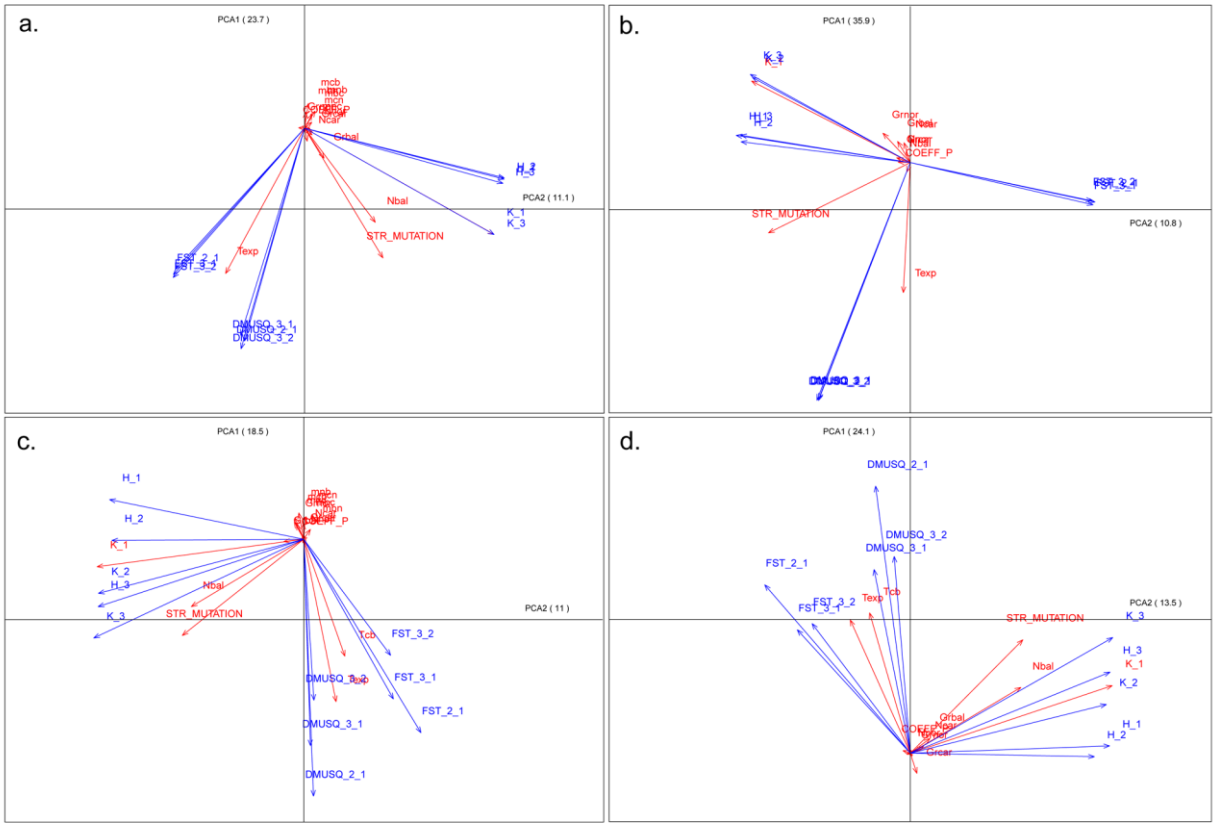


Figure S6. PCA on 3,000 simulations for *Malus sylvestris*. PCA axes 1 and 2 are shown in a) model a b) model b c) model c d) model d. The percentage of the total variance explained is indicated for each PCA axe. For clarity, the simulated datasets are not indicated, and we show in blue the representation of the summary statistics and in red the parameters of the models. In blue, the summary statistics : $F_{ST}\_x\_y$: pairwise F_{ST} between the populations x and y, K_x : mean number of alleles over loci and population x, H_x : mean heterozygosity over loci and population x, $DMUSQ\_x$: mean $\delta\mu^2$ (Goldstein *et al.*, 1995) over loci in the population x. In red, the model parameters: $m_{x\text{-}y}$: migration rate per generation from the population x to the population y, *STR_MUTATION* : mutation rate of microsatellites, N_x : effective population size of the population x, Gr_x : Growth rate of the population x, $T_{x\text{-}y}$: divergence time between the population x and the population y, *COEFF_P* : Value of the geometric parameter for the Generalized Stepwise Mutational model

Table S6. AUC Index for ENM ran with eight and 19 bioclimatic variables for each of the six models and each repetition.

	rep1	rep2	rep3	rep4	rep5	rep6	rep7	rep8	rep9	rep10
Roc Curve (AUC) 8 variables										
CTA	0.961	0.963	0.963	0.961	0.971	0.968	0.969	0.969	0.955	0.96
GAM	0.983	0.984	0.985	0.984	0.987	0.986	0.987	0.984	0.98	0.984
GBM	0.988	0.989	0.99	0.988	0.989	0.991	0.991	0.989	0.985	0.989
GLM	0.983	0.984	0.986	0.984	0.987	0.987	0.987	0.984	0.98	0.984
MARS	0.982	0.98	0.982	0.981	0.983	0.982	0.984	0.981	0.975	0.98
RF	0.993	0.994	0.994	0.993	0.995	0.995	0.995	0.994	0.993	0.994
Roc Curve (AUC) 19 variables										
CTA	0.856	0.765	0.794	0.858	0.906	0.846	0.829	0.875	0.938	0.893
GAM	0.93	0.93	0.936	0.923	0.923	0.952	0.851	0.978	0.996	0.901
GBM	0.957	0.96	0.971	0.96	0.969	0.96	0.949	0.919	0.978	0.96
GLM	0.91	0.849	0.923	0.938	0.875	0.945	0.875	0.993	0.939	0.91
MARS	0.89	0.824	0.787	0.904	0.879	0.976	0.941	0.873	0.919	0.914
RF	0.967	0.949	0.952	0.96	0.993	0.949	0.96	0.978	0.996	0.971

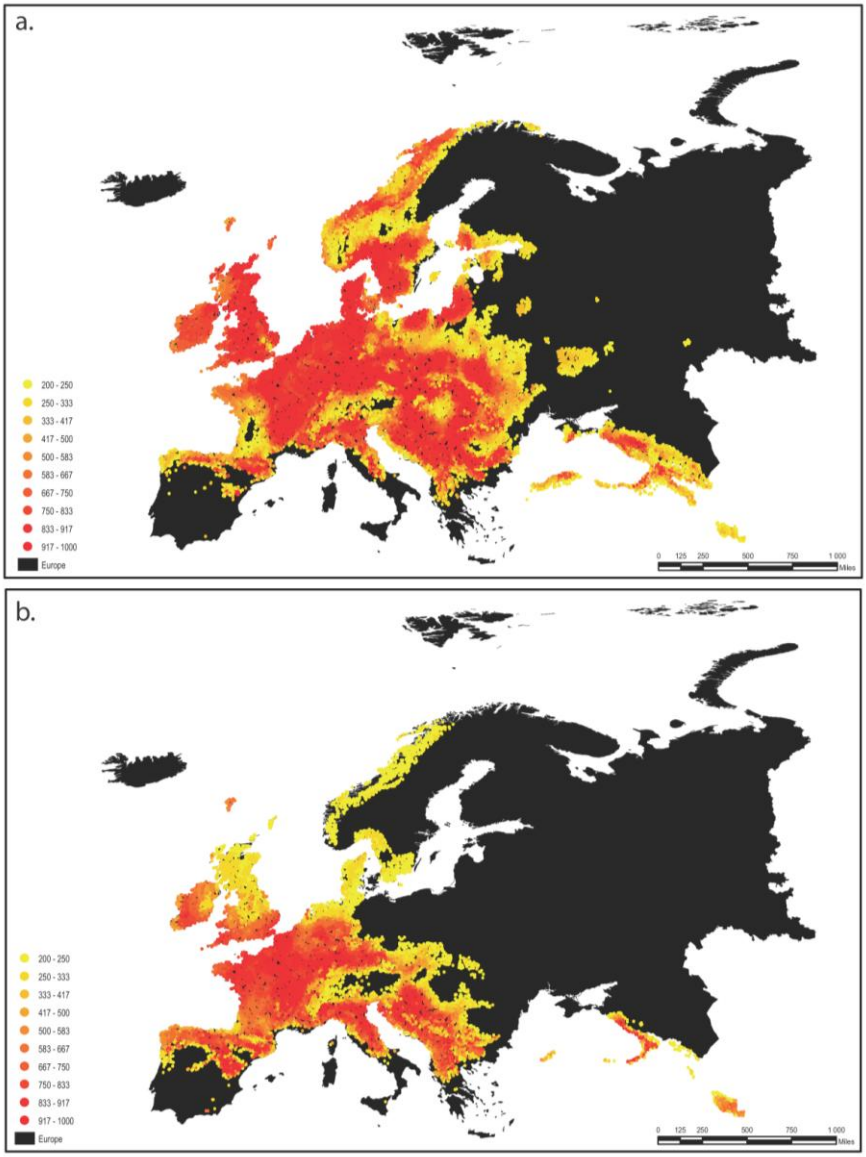


Figure S7. Ensemble forecast using *Malus sylvestris* presence records and pseudo-absences projected onto the map of Europe and Western Russia using 19 bioclimatic variables.

References

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43**, 1223-1232.

Anderson LL, Hu FS, Nelson DM, Petit RJ, Paige KN (2006) Ice-age endurance: DNA evidence of a white spruce refugium in Alaska. *Proceedings of the National Academy of Sciences* **103**, 12447-12450.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33**, 1677-1688.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729-744.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* **19**, 2609-2625.

Brewer S, Cheddadi R, de Beaulieu JL, Reille M (2002) The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management* **156**, 27-48.

Brown JL, Knowles LL (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of Pleistocene glaciation and recent climate change on the American Pika. *Molecular Ecology* **21**, 3757-3775.

Carstens BC, Richards CL (2007) Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* **61**, 1439-1454.

Cheddadi R, Vendramin GG, Litt T*, et al.* (2006) Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. *Global Ecology and Biogeography* **15**, 271-282.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**, 747-756.

Coart E, Vekemans X, Smulders MJM*, et al.* (2003) Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length

polymorphism and microsatellite markers. *Molecular Ecology* **12**, 845-857.

Cornille AC, Gladieux P, Smulders MJM*, et al.* (2012) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* **8**, e1002703.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410-418.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. *Science* **292**, 673-679.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11**, 1591-1604.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40**, 481-501.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Fieldings AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**, 38-49

Gianfranceschi L, Seglias N, Tarchini R, Komjanc M, Gessler C (1998) Simple sequence repeats for the genetic analysis of apple. *Theoritical Applied Genetics* **96**, 1069-1076.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences* **92**, 6723-6727.

Hampe A, Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters* **8**, 461–467.

Hardy OJ, Maggia L, Bandou E*, et al.* (2006) Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. *Molecular Ecology* **15**, 559-571.

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**,

618-620.

Heuertz M, Carnevale S, Fineschi S*, et al.* (2006) Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Molecular Ecology* **15**, 2131-2140.

Heuertz M, Hausman J-F, Hardy OJ*, et al.* (2004) Nuclear microsatellites reveal contrasting patterns of genetic structure between Western and Southeastern European populations of the common ash (*Fraxinus excelsior* L.) *Evolution* **58**, 976-988.

Hewitt GM (1990) Divergence and speciation as viewed from an insect hybrid zone. *Canadian Journal of Zoology* **68**, 1701-1715.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58**, 247-276.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-917.

Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 183-195.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-1978.

Hu FS, Hampe A, Petit RJ (2008) Paleoecology meets genetics: deciphering past vegetational dynamics. *Frontiers in Ecology and the Environment* **7**, 371-379.

Jacques D, Vandermijnsbrugge K, Lemaire S, Antofie A, Lateur M (2009) Natural distribution and variability of the wild apple (*Malus sylvestris*) in Belgium *Belgian Journal of Botany* **142**, 39-49

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jezkova T, Olah-Hemmings V, Riddle BR (2011) Niche shifting in response to warming climate after the last glacial maximum: inference from genetic data and niche assessments in the chisel-toothed kangaroo rat (*Dipodomys microps*). *Global Change Biology* **17**, 3486-3502.

Jolivet C, Degen B (2011) Spatial genetic structure in wild cherry (*Prunus avium* L.): II. Effect

of density and clonal propagation on spatial genetic structure based on simulation studies. *Tree Genetics & Genomes* **7**, 541-552.

Juniper BE, Mabberley DJ (2006) *The Story of the Apple* Imber Press, Inc.

King AR, Feris C (1998) Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. *Molecular Ecology* **7**, 1151-1161.

Larsen A, Asmussen C, Coart E, Olrik D, Kjær E (2006) Hybridization and genetic variation in Danish populations of European crab apple (*Malus sylvestris*). *Tree Genetics & Genomes* **2**, 86-97.

Lascoux M, Palmé AE, Cheddadi R, Latta RG (2004) Impact of Ice Ages on the genetic structure of trees and shrubs. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 197-207.

Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. *Genetics* **184**, 243-252.

Liebhard R, Gianfranceschi L, Koller B*, et al.* (2002) Development and characterisation of 140 new microsatellites in apple (*Malus* x *domestica* Borkh.). *Molecular Breeding* **10**, 217-241.

Liepelt S, Cheddadi R, de Beaulieu J-L*, et al.* (2009) Postglacial range expansion and its genetic imprints in *Abies alba* (Mill.) A synthesis from palaeobotanic and genetic data. *Review of Palaeobotany and Palynology* **153**, 139-149.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **Vol. 82**, 1420-1425

Lu G, Basley DJ, Bernatchez L (2001) Contrasting patterns of mitochondrial DNA and microsatellite introgressive hybridization between lineages of lake whitefish (*Coregonus clupeaformis*); relevance for speciation. *Molecular Ecology* **10**, 965-985.

Magri D, Vendramin GG, Comps B*, et al.* (2006) A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist* **171**, 199-221.

Monserud RA, Leemans R (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling* **62**, 275-293.

Nogués-Bravo D, Rodríguez J, Hortal J, Batra P, Araújo MB (2008) Climate Change, Humans,

and the Extinction of the Woolly Mammoth. *PLoS Biology* **6**, e79.

Oddou-Muratorio S, Demesure-Musch B, Pelissier R, Gouyon PH (2004) Impacts of gene flow and logging history on the local genetic structure of a scattered tree species, *Sorbus torminalis* L. Crantz. *Molecular Ecology* **13**, 3689-3702.

Oddou-Muratorio S, Klein EK (2008) Comparing direct vs. indirect estimates of gene flow within a population of a scattered tree species. *Molecular Ecology* **17**, 2743-2754.

Palmé AE, Vendramin GG (2002) Chloroplast DNA variation, postglacial recolonization and hybridization in hazel, *Corylus avellana*. *Molecular Ecology* **11**, 1769-1779.

Parducci L, Jørgensen T, Tollefsrud MM*, et al.* (2012) Glacial survival of boreal trees in Northern Scandinavia. *Science* **335**, 1083-1086.

Patocchi A, Fernàndez-Fernàndez F, Evans K*, et al.* (2009) Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. *Tree Genetics Genomes* **5**, 211-223.

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* **12**, 361-371.

Petit RJ, Aguinagalde I, de Beaulieu J-L*, et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**, 1563-1565.

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

Petit RJ, Csaikl UM, Bordás Sn*, et al.* (2002) Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management* **156**, 5-26.

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends in ecology & evolution (Personal edition)* **24**, 386-393.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37**, 187-214.

Petit RJ, Hu FS, Dick CW (2008) Forests of the past: a window to future changes. *Science* **320**, 1450-1452.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using

multilocus genotype data. *Genetics* **155**, 945-959.

Provan J, Bennett KD (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in ecology & evolution* **23**, 564-571.

Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics* **3**, 29-43.

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

Richards CL, Carstens BC, Lacey Knowles L (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34**, 1833-1845.

Robinson JP, Harris SA, Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. *Plant Syst. Evol.* **226**, 35-58.

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinion Plant Biology* **10**, 162-167.

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology* **4**, 11.

Silfverberg-Dilworth E, Matasci C, Van de Weg W*, et al.* (2006) Microsatellite markers spanning the apple (*Malus* x *domestica* Borkh.) genome. *Tree Genetics Genomes* **2**, 202-224.

Stewart JR, Lister AM, Barnes I, Dalén L (2009) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**, 661-671.

Stewart JR, Lister AM, Barnes I, Dalén L (2010) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**, 661-671.

Svenning J-C, Normand S, Kageyama M (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology* **96**, 1117-1127.

Svenning J-C, Skov F (2007) Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology Letters* **10**, 453-460.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-2504.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464.

Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* **108**, 17052-17057.

Thuiller W, Brotons L, Araújo MB, Lavorel S (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165-172.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369-373.

Uwimana B, D'Andrea L, Felber F*, et al.* (2012) A Bayesian analysis of gene flow from crops to their wild relatives: cultivated (*Lactuca sativa* L.) and prickly lettuce (*L. serriola* L.) and the recent expansion of *L. serriola* in Europe. *Molecular Ecology* **21**, 2640-2654.

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.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921-935.

Vercken E, Fontaine MC, Gladieux P*, et al.* (2010) Glacial refugia in pathogens: European genetic structure of anther smut pathogens on *Silene latifolia* and *Silene dioica*. *PLoS Pathogens* **6**, e1001229.

Waltari E, Hijmans RJ, Peterson AT*, et al.* (2007) Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**, e563.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Manuscrit C: Histoires de spéciations dans le genre *Malus*.

Manuscrit C : Speciation histories of four wild apple species and the cultivated apple in Eurasia. In prep.

Synthèse

Dans ce chapitre nous nous intéressons aux mécanismes de diversification interspécifique dans le genre *Malus*. Plus particulièrement, nous nous sommes concentrés sur les relations de parenté, le degré de différenciation génétique et l'existence de flux de gènes durant la divergence entre cinq espèces apparentées de pommiers dont une est cultivée, respectivement: *Malus sieversii* (Asie Centrale), *M. orientalis* (Caucase), *M. sylvestris* (Europe), *M. baccata* (Sibérie) et *M. domestica*. Les études précédentes basées sur quelques marqueurs chloroplastiques et nucléaires n'avaient pas résolu les relations phylogénétiques entre les pommiers sauvages caucasiens, européens et asiatiques suggérant une divergence récente entre ces espèces ou des flux de gènes post-divergence. *Malus baccata* est une espèce génétiquement plus différenciée dans ce groupe de quatre espèces, avec un phénotype aussi plus divergent, présentant des petites pommes type « cerise ». Dans le manuscrit A (Histoire de la domestication du pommier cultivé), nous avons montré que ces cinq espèces se différenciaient sur la base de marqueurs microsatellites, et ces derniers nous ont aussi permis de détecter des introgressions importantes du pommier cultivé par le pommier sauvage européen. Dans cette étude en cours, l'utilisation de 13 fragments nucléaires amplifiés sur 45 individus nous permettra sans doute d'apporter un nouvel éclairage sur l'histoire de la domestication du pommier cultivé et sur la diversification interspécifique des cinq espèces étudiées. Nous pourrons également comparer les résultats obtenus sur la base des marqueurs microsatellites et ceux obtenus sur la base de ces séquences nucléaires. Ce manuscrit est en préparation. Les résultats préliminaires montrent que les espèces sont peu différenciées génétiquement, en particulier *M. sieversii* et *M. orientalis*, et *M. baccata* est plus différenciée des autres espèces. Nous sommes en train d'utiliser l'approche ABC (*approximate Bayesian computation*) afin de comparer des scénarios de spéciation avec et sans flux de gènes. La construction des scénarios a été basée sur les études phylogénétiques précédentes (Harris *et al.*, 2002; Robinson *et al.*, 2001). Les analyses ABC sont en cours ; les résultats ne sont donc pas encore présentés dans le manuscrit.

Introduction

Speciation with gene flow has long been dismissed, but still remains one of the hottest topics in evolutionary biology. Accumulating evidence suggest that ecological and phenotypic divergence can occur in the presence of gene flow (Feder *et al.*, 2012; Nosil, 2008). The emergence of a new methodological framework, known as "divergence population genetics" (Kliman *et al.*, 2000), using DNA variation at multiple loci combined with coalescent population genetic models, allowed inferences on past demography histories of species, in particular on the degree and timing of gene flow among them. Such approaches helped unravelling whether divergence has occurred in the face of gene flow or whether species were isolated and exchanged genes after secondary contact. The approximate Bayesian computation (ABC) framework (Beaumont *et al.*, 2002; Cornuet *et al.*, 2008; Csilléry *et al.*, 2010; Wegmann, Excoffier, 2010) has become a valuable tool in this context as it provides both estimates of historical demographic parameters (*e.g.*, population mutation parameters, divergence times and migration rates) and allows probabilistic comparison of alternative models (Bertorelle *et al.*, 2010). This approach has been successfully applied to various model organisms in testing speciation scenarios (Bertorelle *et al.*, 2010; François *et al.*, 2008; Li *et al.*, 2011; Row *et al.*, 2011; St. Onge *et al.*, 2011; Thornton, Andolfatto, 2006)

With their high levels of inter-population gene flow associated with large effective population sizes, long generation times and frequent hybridizations, trees represent fascinating and original models to study speciation with gene flow (Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). These features are thought to play a major role in this existence of species complexes characterized by extensive shared polymorphisms and low levels of interspecific divergence (Li *et al.*, 2010; Savolainen, Pyhäjärvi, 2007). Studies on several tree species showed complex histories of speciation with variable levels of gene flow among related species (Li *et al.*, 2010; Li *et al.*, 2011; Wachowiak *et al.*, 2011).

Here, we aimed at elucidating the recent diversification history of the apple genus (*Malus* spp.), and in particular at deciphering the relationships among the closest wild relatives of the domesticated apple and at testing whether the divergence between these species occurred with or without gene flow. A previous study based on microsatellite data pinpointed three wild apple species as being contributor to the genome of the cultivated

apple (*Malus domestica*): *Malus sieversii* (Central Asia), *M. orientalis* (Caucasus), and *M. sylvestris* (Europe) (Cornille *et al.*, 2012b). *Malus sieversii* is the ancestral progenitor of the cultivated apple (Cornille *et al.*, 2012a; Velasco *et al.*, 2010), while *M. sylvestris* and *M. orientalis* (to a lesser extent) have been involved in post-domestication introgressions (Cornille *et al.*, 2012b). We investigated the demographic and divergence histories between five *Malus* species: the cultivated apple (*Malus domestica),* the three wild contributor species, and a more divergent species, *M. baccata*, found in Siberia. Previous phylogenetic studies based on chloroplast and nuclear markers did not resolve the genetic relationships among these five species, except in showing that *M. baccata* was the most divergent species (Harris *et al.*, 2002; Robinson *et al.*, 2001). We used a set of 45 apple accessions sampled across Eurasia (Table S1) and 13 nuclear DNA fragments to investigate the following specific questions: 1) What is the level of genetic differentiation and the genetic relationships among the five *Malus* species? 2) Did divergence among the five species occur in the face of gene flow? 3) Can these new data and new analyses bring new insights into the history of domestication of the domesticated apple?

Materials and methods

Plant material and DNA extraction

Leaf material was retrieved from the collections of various institutes (INRA Angers, France; USDA – ARS, Geneva, USA; ILVO Melle, Belgium) and *in situ* (Table S1). *Malus sieversii* (*N*=9) material was collected from 2007 to 2010 in Kazakhstan. *Malus orientalis* (*N*=9) was sampled in 2009 in Armenia. *Malus sylvestris* (*N*=10) samples were obtained from seven European countries (Bosnia Herzegovina, Denmark, Spain, Great Britain, Ukraine, Austria and Belgium). *Malus baccata* (*N*=9) was sampled in 2010 in Russia. *Malus domestica* cultivars (*N*=8) were Worcester, Belle de Boskoop, Cox's Orange Pippin, Jonathan, Fuji, Granny Smith, Macintosh and Golden Delicious varieties. Genomic DNA was extracted with the Nucleo Spin® plant DNA extraction kit II (Macherey & Nagel, Düren, Germany).

Sequencing

Thirteen DNA fragments were sequenced (205, 30, 166, 262, 437, 444, 477, 559, 583, 586,

983, 1027, 1322 (Velasco *et al.*, 2010)) (Table S2). Sequence reactions were outsourced at Cogenics (Grenoble, France) using standard techniques. Forward sequence data were base-called, assembled and edited using Codon Code Aligner v. 3.0.1 (CodonCode Corporation). All putative polymorphic sites were validated by visual inspection of electrophoregrams. All indels were excluded from analyses. All sequence data have been deposited with the EMBL/GenBankData (Table S1). The resulting contigs were processed with automated shell and Perl programs and aligned using MAFFT (Katoh *et al.*, 2005) as implemented in Jalview 9.5 (Clamp *et al.*, 2004).

Polymorphism and divergence

The gametic phase of diploid genotypes was inferred using PHASE (Stephens *et al.*, 2001), as implemented in DnaSP v. 4.0 (Rozas *et al.*, 2003). Standard population genetics parameters and neutrality test statistics were computed for each locus with DnaSP v. 4.0 (Rozas *et al.*, 2003). For each species, intraspecific variation was estimated using number of haplotypes (H), haplotypic diversity (H_d) (Nei 1987), and two estimators of the population mutation parameter $4Ne\mu$ (where Ne is the effective population size and μ the neutral mutation rate): π, that uses the average number of pairwise differences (Tajima, 1983) and ϑ_W, that uses the number of polymorphic sites (Watterson, 1975). The standard neutral model was tested using Tajima's D (Tajima, 1989), Fu and Li's D^* and F^* (Fu, Li, 1993).

Population structure

We used an analog of Wright's fixation index (Weir, Cockerham, 1984) that also takes into account the distance among haplotypes, φ_{ST} (Excoffier et al., 1992), to assess differentiation among species. φ_{ST} values were estimated with an analysis of molecular variance (AMOVA) as implemented in Arlequin v. 3.5 (Excoffier, Lischer, 2010). Results are reported as averages over loci weighted by the total variance in allele frequency as well as locus-by-locus. Significance of φ_{ST} values was tested by permuting haplotypes among populations.

We used the individual-based Bayesian clustering method implemented in STRUCTURE 2.3.3 (Pritchard et al., 2000) to investigate species delimitation, intraspecific population structure and admixture. This method is based on Markov Chain Monte Carlo

(MCMC) simulations and is used to infer the proportion of ancestry of genotypes in K distinct predefined clusters. The algorithm attempts to minimize deviations from Hardy–Weinberg and linkage equilibrium within clusters. Analyses were carried out with the use of prior information (i.e., species identification) to assist clustering. K ranged from 1 to 8 for analyses including the five-species dataset. Ten independent runs were carried out for each K and we used 500,000 MCMC iterations after a burn-in of 50,000 steps. We used CLUMPP v1.1.2 (Greedy algorithm) (Jakobsson, Rosenberg, 2007) to look for distinct modes among the 10 replicated runs of each K.

Demographic scenario of speciation

We used ABCtoolBox (Wegmann *et al.*, 2010) to compare scenarios with and without gene flow (Figure 1). Models *a* and *b* assumed that *M. domestica* and *M. sieversii* diverged after *M. sieversii* and *M. orientalis*, while models *c* and *d* assumed that *M. sieversii* and *M. orientalis* diverged after *M. domestica* and *M. sieversii*. Bidirectionnal gene flow was assumed between each species in models *a* and *c*. The juvenile period of *M. domestica* lasts five to 10 years and no data for this parameter are available for *M. sylvestris*. We therefore assumed a generation time of 7.5 years. We estimated the effective size of each species (N_{siev}, N_{syl}, N_{ori}, N_{dom}, N_{bacc}), the rate of migration between species for each generation ($m_{x\text{-}y}$: migration rate from species x to y per generation), divergence times ($T_{x\text{-}y}$: divergence time between species x and y).

We will simulate 2 x 10^6 datasets, using population parameters drawn from a prior distribution under the four previously specified scenarios. For each simulation, we will compute two summary statistics per population: *S*, the number of sites with segregating substitutions output for each species,π: the mean number of pairwise differences for each species. We will also compute pairwise F_{ST} (Weir, Cockerham, 1984) between pairs of species. We will conduct a preliminary principal component analysis (PCA) in R environment (MASS package, function prcomp), based on 3,000 simulated datasets including the five *Malus* species, for the four models, to establish and check correlations between the main parameters of the model and the chosen summary statistics (Tellier *et al.*, 2011).

The mutation rate will be allowed to vary across loci, with locus-specific mutation rates being drawn from a gamma distribution (α, α/μ) in which μ is the mutation rate per

generation and α is a shape parameter. We will assume a log-uniform prior distribution for μ [0.000001, 0.0001] and a uniform distribution for α [1, 30].

We will compare the four models by calculating their Bayes factors (Wegmann *et al.*, 2010) and estimating their relative posterior probabilities, based on the 1% of simulated datasets most closely matching the observed data (*i.e.*, 2,000 simulated datasets). Once the best model are chosen, we will estimate demographic parameters under this scenario, using a general linear model (ABC-GLM) post-sampling regression adjustment for the 2,000 retained simulations (Leuenberger, Wegmann, 2010; Wegmann *et al.*, 2010). We will report the mode and 95% highest posterior density (HPD) interval for each model parameter estimate.

The performance of the method for discriminating between competing historical models will be assessed by analyzing test datasets simulated with the same number of loci and individuals as for the observed datasets (*i.e.*, pseudo-observed datasets). We will simulate 2,000 such datasets for each competing model, using parameter values drawn from the same prior distributions as for the original analyses. We will determine the relative posterior probabilities of competing models for each pseudo-observed dataset, using the model choice procedure, as described above (Wegmann *et al.*, 2010). Confidence in model choice will then be estimated from the likelihood that a given scenario does not have the highest posterior probability of the competing scenarios when it is actually the true scenario (type I error), and the likelihood of a given scenario having the highest posterior probability when it is not the true scenario (type II error).

Results

Polymorphism summary

All 13 loci were polymorphic in all species. A mean number of 16 alleles per locus were sequenced per species. Of the 5409 bp sequenced, ca. 40% was in coding regions. Summaries of variation are given in Tables 1 and S3. Genetic variation was not significatively different among species (Wilcoxon's Sign Rank test, P>0.1 for all comparison between each species pairs for π, ϑ_W *and* H_d). For all species, the allele frequency spectrum, as measured by

Tajima's *D* (Tajima, 1983) and Fu and Li's *D** and *F** (Fu, Li, 1993), generally conformed to expectations under a standard neutral model of molecular evolution (Table S3)

Population differentiation

Population differentiation values (φ_{ST}) are presented in Table 1 and locus-by-locus AMOVA are presented in Table S4. *Malus baccata* appears the most differentiated species, the φ_{ST} values with each of the four other species being >0.4. φ_{ST} values between *M. orientalis* and *M. sieversi* were very low (0.03) indicating low differentiation between the two species. The differentiation between *M. domestica* and each of its three wild contributors lies within the same order of magnitude, $\varphi_{ST} \approx 0.2$, with *M. sylvestris* being the closest one.

STRUCTURE analyses with *prior* information on species status revealed structure among the five species (but analyses without using *prior* information did not – data not shown). A split between *M. orientalis*/*M. sieversii* from the one hand and *M. domestica*/*M. baccata*/*M. sylvestris* on the other hand was observed at *K*=2 (Figure 2). At *K*=3, *M. baccata* individualized in the main mode, while a minor mode instead splitted *M. domestica* in a new cluster. At *K*=4, the species *M. baccata*, *M. sylvestris*, *M. domestica* formed well separated clusters whereas *M. sieversii* and *M. orientalis* were clustered together. Increasing *K* yielded no obvious further structuring, and instead introduced some heterogeneity in individual membership proportions (Figure 2). The *ΔK* statistic, designed to identify the most relevant number of clusters by determining the number of clusters beyond which there is no further increase in likelihood (Evanno *et al.*, 2005), was greatest for *K*=7 (*ΔK*=1.31, Pr|lnL=-1318.9). However, membership coefficients do not allow identifying more than four populations and *ΔK* at *K*=4 was close to its value at *K*=7 (*ΔK*=0.850128, Pr|lnL=-1313.1) (Figure S1). Altogether, these results indicate weak genetic differentiation among the five wild species, particularly between *M. orientalis* and *M. sieversii*.

Acknowledgments

We thank the *Région Ile de France* (PICRI), IDEEV, *Fondation Dufrenoy*, SBF (*Société Botanique de France*), CNRS and *Université Paris Sud* for funding, *Plateforme de Génotypage* GENTYANE

INRA UMR 1095 *Génétique Diversité et Ecophysiologie des Céréales* and Pauline Lasserre. We thank René Smulders for helpful suggestions and comments. We thank the following for sampling and providing access to samples: Dalibor Ballian (Faculty of Forestry, University of Sarajevo, Bosnia-Herzegovina), Anders Larsen (Forest and Landscape, Department for Management of Forest Genetic Resources, Denmark), Joanne Clavel, Francisco Donaire (ILVO, Plant – Growth and Development, Belgium), Stephen Cavers (NERC, Centre for Ecology and Hydrology, UK), Roman Volansyanchuck (Ukrainian Research Institute of Forestry and Forest Melioration, Ukraine), François Laurens and Bruno Le Cam (INRA, IRHS, PRES UNAM, SFR QUASAV, Beaucouzé, France), Thomas Kirisits and Bernhard Kirisits (Institute of Forest Entomology, Forest Pathology and Forest Protection, Vienna, Austria). Isabel Roldán-Ruiz (ILVO, Plant – Growth and Development, Belgium), Marie-Anne Félix (*Institut Jacques Monod*, France). We also thank Thierry Genevet, Frédéric Tournay (*Jardin Botanique de Strasbourg*), Levente Kiss, the East Malling Research Station (UK), Philip Forsline and the Plant Genetic Resources Unit in Geneva (NY).

Figures

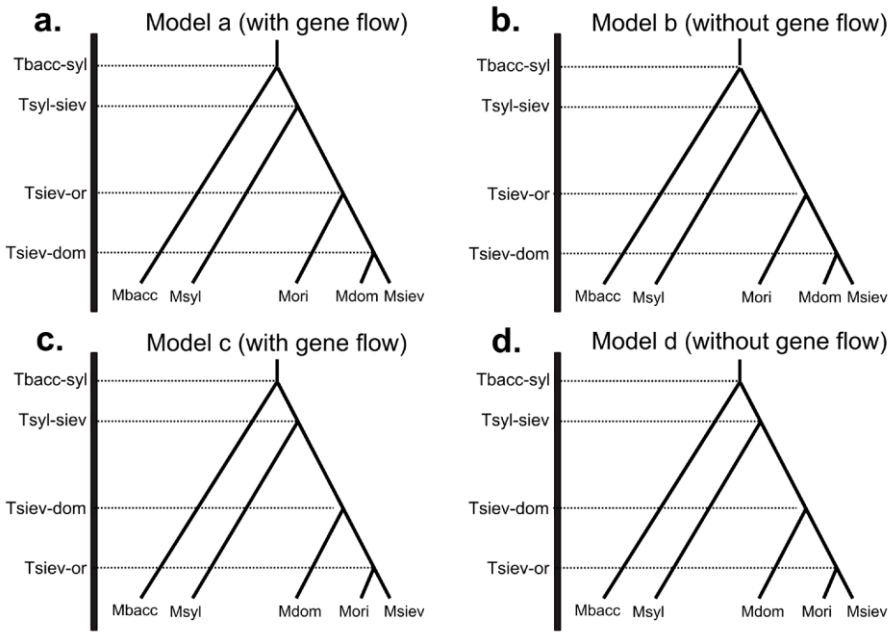


Figure 2. Demographic models compared using approximate Bayesian computations. Models *a* and *b* assume divergence of *Malus orientalis* and *Malus sieversii* followed by a divergence of *Malus sieversii* and *Malus domestica*. Models *c* and *d* assumes divergence of *Malus orientalis* and *Malus sieversii* followed by a divergence of *Malus sieversii* and *Malus domestica*. Bidirectional gene flow was assumed between all pairs of species in the Models *a* and *c*. Model *b* is identical to model *a* but without gene flow. Model *d* is identical to model *c* but without gene flow. $T_{x\text{-}y}$: divergence between species x and y. *Mori*: *Malus orientalis*, *Msiev*: *Malus sieversii*, *Mdom: Malus domestica*, *Msyl*: *Malus sylvestris*, *Mbacc*: *Malus baccata*.

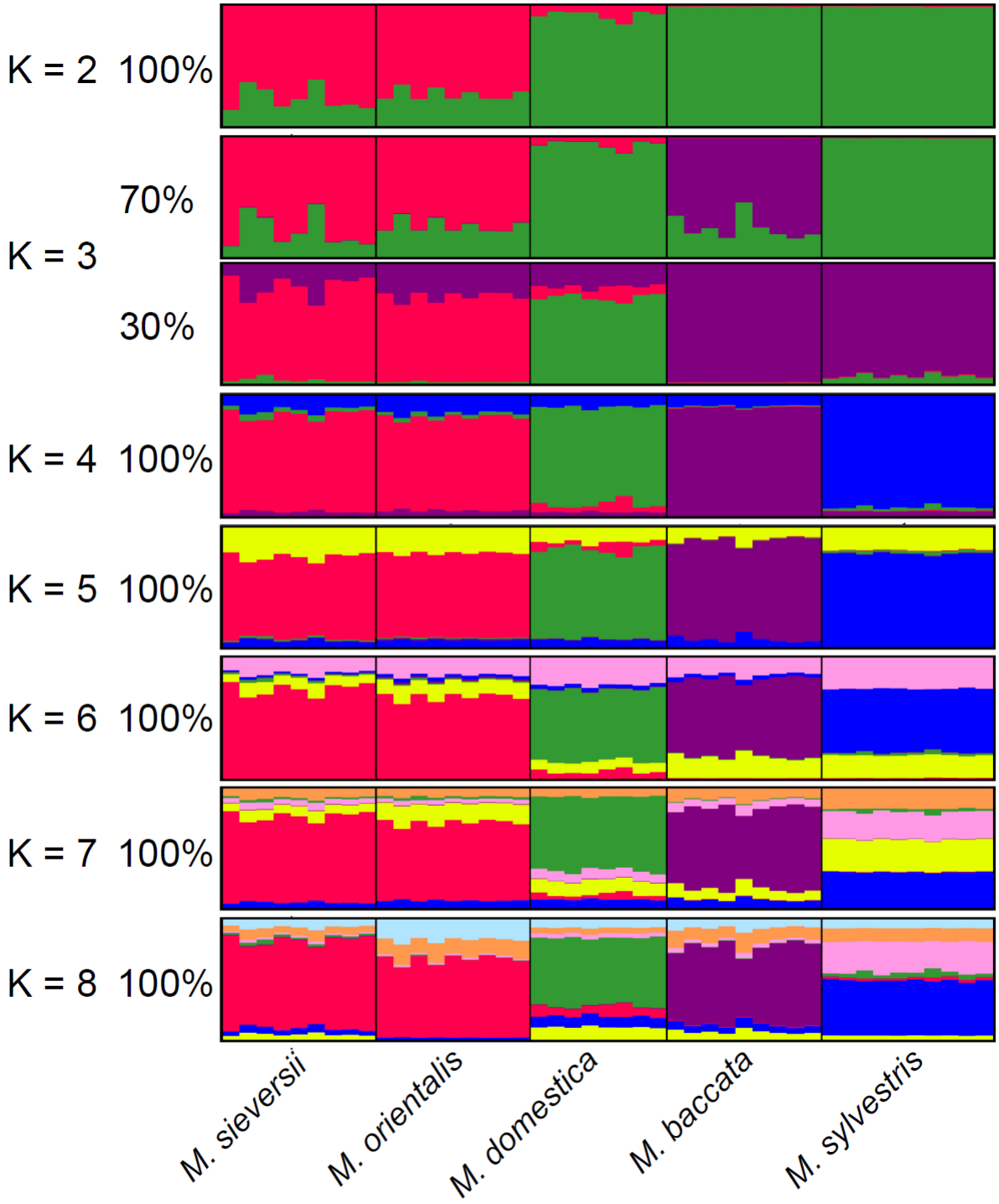


Figure 3. STRUCTURE analysis of the five *Malus* species when *K*=2-8 clusters are assumed. Each individual is represented by a vertical bar, partitioned into *K* segments representing the proportions of ancestry of its genome in *K* clusters. When several clustering solutions ("modes") were represented within replicate runs, the proportion of simulations represented by each mode is shown.

Tables

Table 1. Summary statistics of nucleotide variation within each wild *Malus* species

Species	N^a	S^b	H^c	H_d^d	$\vartheta_W{}^e$(%)	π^e(%)
Malus baccata	14.5	5.6	4	0.59	0.45	0.57
Malus domestica	14.3	5.3	4.1	0.59	0.43	0.46
Malus orientalis	17.4	3.3	3.7	0.47	0.24	0.29
Malus sieversii	16.5	4.5	4.1	0.45	0.32	0.31
Malus sylvestris	17.1	7.4	5.6	0.64	0.54	0.50

[a] Mean sample size over loci, [b]: Number of segregating sites, [c]: Number of haplotypes, [d]: Haplotypic diversity, [e]: Estimates of the population mutation parameter based on the number of polymorphic sites (θ_W, (Watterson, 1975)) and the average number of pairwise differences (π; (Tajima, 1983)), respectively.

Table 2. Φ_{ST} values over all loci among the five *Malus* species. All values were significant, P<0.001.

	M. dom	*M. bacc*	*M. or*	*M. siev*
M. bacc	0.45			
M. or	0.23	0.53		
M. siev	0.22	0.58	0.03	
M. syl	0.18	0.44	0.38	0.37

M. bacc: Malus baccata, M. syl: Malus sylvestris, M. or: Malus orientalis, M. siev: Malus sieversii, M. dom: Malus domestica.

Supporting information

Table S1. List of the sampled individuals for the five *Malus* species accessions.

Table S2. List of primers used to amplify the loci analysed.

Table S3. Summary of nucleotide variation for each locus within each wild *Malus* species

Figure S1. Estimated number of populations using the five species dataset from STRUCTURE analyses using the *ΔK*

Table S4. Locus-by-locus AMOVA including the five *Malus* species.

Table S1. List of the sampled individuals for the five *Malus* species accessions with their geographical origin and providers.

Species	Origin	location/varieties
Malus baccata	*N*=9	
BACC10RUS19	Ilya Zahharov	Russia
BACC10RUS21	Ilya Zahharov	Russia
BACC10RUS29	Marina Olonova	Russia
BACC10RUS35	Marina Olonova	Russia
BACC10RUSW12	Ilya Zahharov	Russia
BACC10RUSW5	Ilya Zahharov	Russia
BACC10RUSW6	Ilya Zahharov	Russia
BACCRUSS28	Marina Olonova	Russia
BACCRUSW7	Ilya Zahharov	Russia
Malus orientalis	*N*=9	
OR092639	PG, Joanne Clavel,	Armenia
OR092646	PG, Joanne Clavel,	Armenia
OR092720	PG, Joanne Clavel,	Armenia
OR092826	PG, Joanne Clavel,	Armenia
OR09POI13	PG, Joanne Clavel,	Armenia
OR09POI27	PG, Joanne Clavel,	Armenia
OR09POI3	PG, Joanne Clavel,	Armenia
OR09POI60	PG, Joanne Clavel,	Armenia
USMORT2	USDA*	?
Malus sieversii	*N*=9	
SI06KZL1199	Bruno Le Cam, PG, François Laurens	Kazakhstan
SI06KZL325	Bruno Le Cam, PG, François Laurens	Kazakhstan
SI07KZL15AT6	Bruno Le Cam, PG, François Laurens	Kazakhstan
SI07KZL3T1	Bruno Le Cam, PG, François Laurens	Kazakhstan
SI08KZL1T2	Bruno Le Cam, PG, François Laurens	Kazakhstan
SI08KZL2T2	Bruno Le Cam, PG, François Laurens	Kazakhstan
SIEV10KZLW40	Marie-Anne Félix	Kazakhstan
SIEV10KZLW49	Marie-Anne Félix	Kazakhstan
USMSIEVT19	USDA*	Kazakhstan
Malus sylvestris	*N*=10	
SY10BIHJAB28	Dalibor Ballian	Bosnia Herzegovina
SY10BIHLIV2	Dalibor Ballian	Bosnia Herzegovina
SY10DAK15	Anders Larsen	Denmark
SY10ESSY3421	Francisco Donaire	Spain
SY10GBR10	Stephen Cavers	Great Britain
SY10GBR14	Stephen Cavers	Great Britain
SY10UKN10	Roman Volansyanchuck	Ukraine
SYAUT19	Thomas Kirisits and Bernhard Kirisits	Austria

SYDANO26	Anders Larsen	Denmark
SYBE453	ILVO**, Isabel Roldán-Ruiz	Belgium
Malus domestica	*N*=8	
X1954	INRA***	Cox's Orange Pippin
X2656	INRA***	Jonathan
09MXDT3	INRA***	Worcester
7190	INRA***	Belle de Boskoop
X3069	INRA***	Granny Smith
X557	INRA***	MacIntosh
X972	INRA***	Golden Delicious
X3049	INRA***	Fuji

* USDA: Plant Genetic Resources Unit, Geneva (NY)

**ILVO - PLANT: Plant -Growth and Development, Melle, Belgium

***INRA: Institut de Recherche en Horticulture et Semences, Angers, France

N: sample size

Table S2. List of primers used to amplify the loci analysed (Velasco *et al.*, 2010).

Amplicon	Associated SNP	Primer Forward (5'-3')	Primer Reverse (5'-3')	LG	gene ID
MDE01108.1	GDSNP00477	GGCATGATAGCTTTTGGGAG	CACTTGTACATGCCATTCGG	3	MDP0000291654
MDE02156.2	GDSNP00583	TCCTTTGCAAGTTGGTGACA	CAAATTTGGGACCATTCCAC	1	MDP0000258088
MDE02757.1	GDSNP00030	CCACAAACCATAGGTGGGTC	CTTGCAACAATCCCAGTGTG	16	MDP0000212368
MDE04897.1	GDSNP00983	GCAGGTGAAAGGCAAGACTC	GGACTGAGAAACCAACGGAA	15	MDP0000171994
MDE14436.1	GDSNP00262	ACTTGAAGTTGGGCGAATTG	CCAAAAGCACGAGCATAACA	17	MDP0000191398
MDE00758.1	GDSNP00437	TAGCTTTTTCGGGTCATGCT	AAACGGTGAGTTTTGATGGC	3	MDP0000291654
MDE00787.2	GDSNP00444	TGTTGGAGCTGCAAAAACTG	GTGGTGGGGTCTGTGATTCT	16	MDP0000170922
MDE01932.1	GDSNP00559	TTTGCCAGAAAGTGTCGATG	TAAACGCAGCTTTTGGAGGT	5	MDP0000230999
MDE02181.1	GDSNP00586	TGGGAAGCACATGAATGAAA	CTCAGAGGACACCCCAATGT	14	MDP0000898700
MDE04610.2	GDSNP00205	AGATTTGGTCATGTCCGAGG	TCGATTGGATTGGTTTCACA	9	MDP0000181482
MDE05301.1	GDSNP01027	TGAACAAACAACAGACCGGA	GGTAATGTGTTCGTCGCCTT	1	MDP0000124654
MDE12995.1	GDSNP01322	TTCTTCGCTCCCCTCTTACA	AAGCTTCAATTCCTTCCGGT	2	MDP0000022758
MDE01422.2	GDSNP00166	AGGGGGAAAAGCAAAAAGAA	TCTCCTTAAGTCCCGGTGTG	6	MDP0000835914

LG: linkage group, Gene ID: Identification for the target sequenced gene

Table S3. Summary of nucleotide variation for each locus within each wild *Malus* species

Species	Locus	*N*	*S*	*H*	H_d	ϑ_W (%)	π (%)	*D*	*D**	*F**
Malus baccata	205	18	6	7	0.84	0.46	0.49	0.186	-0.104	-0.028
	30	14	4	6	0.28	0.30	0.19	-1.164	-0.555	-0.813
	166	18	14	7	0.85	1.01	1.48	1.748	1.150	1.527
	262	16	4	4	0.71	0.26	0.36	1.132	1.141	1.304
	437	18	11	4	0.40	0.96	0.96	0.013	**1.438***	1.196
	444	14	2	2	0.54	0.15	0.26	1.933	0.935	1.353
	477	2	0	0	0.00	0.00	0.00	0.000	0.000	0.000
	559	16	5	5	0.60	0.42	0.45	0.240	1.351	1.578
	583	14	2	3	0.39	0.32	0.13	-0.011	0.935	0.787
	586	18	6	3	0.45	0.34	0.18	-1.553	-2.149	-2.285
	983	18	9	5	0.71	0.52	0.52	0.015	-0.115	-0.090
	1027	10	5	4	0.78	0.53	0.67	1.082	1.300	1.397
	1322	12	5	2	0.55	0.64	1.06	2.386	1.261	**1.752****
	Average	14.5	5.6	4	0.59	0.45	0.56			
Malus domestica	205	16	5	4	0.35	0.40	0.22	-1.491	-1.808	-1.974
	30	14	5	3	0.69	0.37	0.59	**2.055***	1.234	**1.653***
	166	8	4	4	0.75	0.38	0.37	-0.121	-0.176	-0.180
	262	16	2	3	0.58	0.12	0.13	0.201	-0.504	-0.364
	437	16	9	5	0.58	0.82	0.92	0.457	0.429	0.502
	444	16	7	3	0.59	0.51	0.31	-1.380	-2.196	-2.266
	477	16	2	3	0.49	0.14	0.19	1.085	0.907	1.088
	559	12	7	8	0.91	0.64	0.89	1.510	1.351	1.578
	583	12	3	2	0.30	0.28	0.25	0.003	1.105	0.855
	586	16	8	9	0.89	0.47	0.50	0.283	0.829	0.781
	983	16	10	3	0.43	0.60	0.64	0.262	0.535	0.529
	1027	16	6	5	0.68	0.70	0.74	0.158	0.612	0.561

	1322	12	1	2	0.41	0.12	0.15	0.541	0.752	0.787
	Average	14.3	5.3	4.2	0.59	0.43	0.45			
Malus orientalis	205	18	5	6	0.56	0.39	0.32	-0.504	-0.359	-0.459
	30	18	1	2	0.52	0.06	0.11	1.505	0.667	1.011
	166	18	3	4	0.65	0.22	0.35	1.653	1.024	1.369
	262	18	4	4	0.77	0.24	0.29	0.642	1.123	1.140
	437	18	0	0	0.00	0.00	0.00	0.000	0.000	0.000
	444	18	2	3	0.22	0.14	0.05	-1.508	-1.989	-2.130
	477	18	3	4	0.61	0.20	0.16	-0.508	-0.881	1.024
	559	16	6	6	0.81	0.50	0.69	1.310	0.612	0.919
	583	18	1	2	0.11	0.15	0.06	-1.165	-1.499	-1.612
	586	16	2	3	0.43	0.12	0.09	-0.578	0.907	0.592
	983	18	4	2	0.11	0.23	0.09	**-1.853***	**-2.525***	**2.691***
	1027	18	11	10	0.88	0.86	1.04	0.774	0.164	0.388
	1322	14	1	2	0.50	0.11	0.17	1.212	0.716	0.953
	Average	17.4	3.3	3.7	0.47	0.25	0.28			
Malus sieversii	205	16	4	4	0.62	0.32	0.22	-0.966	1.522	-1.572
	30	18	1	2	0.01	0.06	0.10	1.166	0.667	0.910
	166	18	5	5	0.55	0.36	0.38	0.165	-0.359	-0.247
	262	18	4	4	0.53	0.24	0.16	-0.743	-1.613	-1.675
	437	16	1	2	0.33	0.09	0.10	0.156	0.688	0.627
	444	16	7	4	0.35	0.51	0.29	-1.798	-1.808	-2.074
	477	16	2	3	0.24	0.14	0.06	-1.498	-1.915	-2.060
	559	12	6	5	0.76	0.55	0.66	0.752	0.101	0.304
	583	14	2	2	0.14	0.18	0.08	-1.481	-1.827	-1.974
	586	16	3	3	0.59	0.18	0.15	-0.387	-1.122	-1.060
	983	18	9	3	0.31	0.52	0.36	-1.065	-0.614	-0.854
	1027	18	11	13	0.97	0.86	1.22	1.542	**1.438***	**1.695***
	1322	18	3	4	0.50	0.22	0.22	-0.872	-1.309	-1.369

	Average	16.5	4.5	4.2	0.45	0.32	0.31			
Malus sylvestris	205	20	8	9	0.65	0.60	0.47	-0.743	-0.354	-0.538
	30	20	7	4	0.62	0.46	0.52	0.422	0.674	0.697
	166	16	15	6	0.68	1.12	0.61	-1.787	-2.246	-2.440
	262	20	3	4	0.28	0.17	0.06	-1.723	-2.386	-2.535
	437	20	11	8	0.82	0.93	0.73	0.156	0.688	0.627
	444	16	8	7	0.87	0.58	0.87	1.753	**1.356***	**1.683***
	477	20	4	4	0.66	0.26	0.24	-0.138	1.108	0.879
	559	16	9	8	0.88	0.75	0.63	-0.570	-1.011	-1.023
	583	20	4	3	0.49	0.32	0.40	0.727	1.108	1.155
	586	18	7	7	0.86	0.39	0.44	0.443	0.701	0.725
	983	20	9	4	0.49	0.50	0.64	0.920	0.862	1.018
	1027	14	10	7	0.82	0.84	0.85	0.025	0.582	0.495
	1322	16	1	2	0.23	0.11	0.09	-0.448	0.688	0.450
	Average	18.2	7.4	5.6	0.64	0.54	0.50			

[a]:Sample size, [b]: Number of segregating sites, [c]: Number of haplotypes, [d]: Haplotypic diversity, [e]: Estimators of the population mutation parameter based on the number of polymorphic sites (ϑ_W; (Watterson, 1975)) and the average number of pairwise differences (π; (Tajima, 1989)), respectively, [f]: Tajima's *D* and Fu and Li's *D** and *F** statistics to test for the standard neutral model ((Fu, Li, 1993; Tajima, 1989))

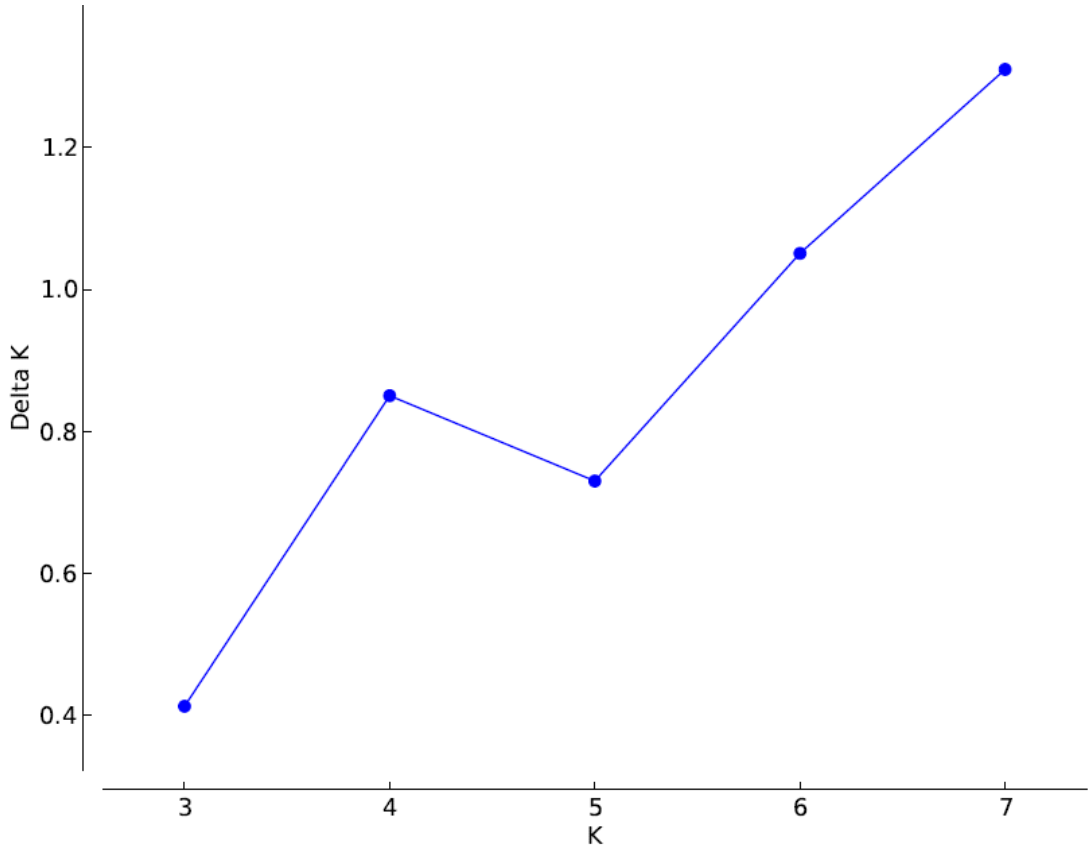


Figure S1. Estimated number of populations using the five species dataset from STRUCTURE analyses using the *ΔK*

Table S4. Locus-by-locus AMOVA including the five *Malus* species.

Locus	Variation	SS	Var.	*F*	*P-value*
30	Among	55.7	0.87	0.49	0
	Within	72.0	0.90		
166	Among	33.7	0.46	0.25	0
	Within	99.2	1.36		
205	Among	26.6	0.34	0.34	0
	Within	55.2	0.66		
262	Among	46.8	0.64	0.60	0
	Within	35.0	0.42		
437	Among	34.4	0.44	0.33	0
	Within	75.5	0.91		
444	Among	39.2	39.17	0.44	0
	Within	39.2	0.73		
477	Among	10.5	0.17	0.31	0
	Within	10.5	0.37		
559	Among	43.1	78.81	0.36	0
	Within	78.8	1.18		
583	Among	29.1	16.75	0.66	0
	Within	16.7	0.23		
586	Among	29.1	0.46	0.66	0
	Within	16.7	0.23		
1027	Among	25.6	0.38	0.34	0
	Within	52.8	0.74		
1322	Among	12.2	0.18	0.34	0
	Within	24.9	0.35		

SS: the sum squares. *Var*: the variance component. *F*: F-statistics.

References

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* **19**, 2609-2625.

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427.

Cornille A, Giraud T, Bellard C*, et al.* (2012a) Post-glacial recolonization history of the European crabapple (Malus sylvestris Mill.), a wild contributor to the domesticated apple. *Molecular Ecology* **in revision**.

Cornille A, Gladieux P, Smulders MJM*, et al.* (2012b) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* **8**, e1002703.

Cornuet J-M, Santos F, Beaumont MA*, et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**, 2713-2719.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410-418.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics* **28**, 342-350.

François O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet* **4**, e1000075.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Harris SA, Robinson JP, Juniper BE (2002) Genetic clues to the origin of the apple. *Trends Genet* **18**, 426-430.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Katoh K, Kuma K-i, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511-518.

Kliman RM, Andolfatto P, Coyne JA*, et al.* (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**, 1913-1931.

Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. *Genetics* **184**, 243-252.

Li Y, Stocks M, Hemmilä S*, et al.* (2010) Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution* **27**, 1001-1014.

Li Z, Zou J, Mao K*, et al.* (2011) Population genetic evidence for complex evolutionary histories of four high altitude *Juniper* species in the Qinghai-Tibetan plateau. *Evolution*, no-no.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology* **17**, 2103-2106.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37**, 187-214.

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

Robinson JP, Harris SA, Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. *Plant Syst. Evol.* **226**, 35-58.

Row JR, Brooks RJ, MacKinnon CA*, et al.* (2011) Approximate Bayesian computation reveals the factors that influence genetic diversity and population structure of foxsnakes. *Journal of Evolutionary Biology* **24**, 2364-2377.

Rozas J, SÃ¡nchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinion Plant Biology* **10**, 162-167.

St. Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE (2011) Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Molecular Ecology* **20**, 3306-3320.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* **68**, 978-989.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations *Genetics* **105**, 437-460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* **108**, 17052-17057.

Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**, 1607-1619.

Velasco R, Zharkikh A, Affourtit J*, et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics* **42**, 833-839.

Wachowiak W, PalmÉ AE, Savolainen O (2011) Speciation history of three closely related pines *Pinus mugo* (T.), *P. uliginosa* (N.) and *P. sylvestris* (L.). *Molecular Ecology* **20**, 1729-1743.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256-276.

Wegmann D, Excoffier L (2010) Bayesian Inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* **27**, 1425-1435.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Manuscrit D : Hybridations interspécifiques du pommier cultivé vers les trois espèces contributrices sauvages et estimation de la structuration génétique spatiale chez l'espèce asiatique et l'espèce caucasienne.

Manuscrit D : Crop-to-wild gene flow and spatial genetic structure in the closest wild relatives of the cultivated apple. Submitted in Evolutionnary Applications.

Amandine Cornille [1,2†], Pierre Gladieux[1,2,3*], Tatiana Giraud[1,2*]

† Corresponding author: amandine.cornille@gmail.com

*jointly directed the work

1. CNRS, Laboratoire Ecologie Systématique et Evolution - UMR8079, Bâtiment 360, 91405 Orsay, France; 2. Univ. Paris Sud, 91405 Orsay, France; 3. Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA.

Abstract

Gene flow and introgression from cultivated to wild plant populations have important evolutionary and ecological consequences and require careful consideration in conservation programs for wild genetic resources of potential use in breeding programs and in assessments of the risk of transgene escape into natural ecosystems. Using 26 microsatellites and a comprehensive set of 1181 trees, we investigated the extent of introgression from the cultivated apple, *Malus domestica*, to its three closest wild relatives, *M. sylvestris* in Europe, *M. orientalis* in the Caucasus and *M. sieversii* in Central Asia. We found footprints of introgression from *M. domestica* to *M. orientalis* (3.2% of hybrids), *M. sieversii* (14.8%) and *M. sylvestris* (36.7%). *Malus sieversii* and *M. orientalis* both presented weak but significant spatial genetic structures across their geographic range. *Malus orientalis* displayed genetic differentiation across a north-south gradient, with three differentiated populations, in Turkey, Armenia and Russia. *Malus sieversii* consisted of a main population spread over Central Asia and a smaller population in the Tian Shan Mountains. The low values of *Sp* obtained indicate high dispersal capacities. Our findings are important for the *in situ* conservation of wild apple species and for the establishment of seed orchards based on wild genetic resources.

Key words: hybridization, *Venturia inaequalis*, apple scab, conservation, isolation by distance, tree

Introduction

The anthropogenization of landscapes and ecosystems, with landscape fragmentation and the introduction of crops over extended areas, has greatly increased the likelihood of contact between domesticated and related wild taxa that were previously isolated either geographically or ecologically (Kareiva *et al.*, 2007). In the last 20 years, an increasing number of studies have documented introgression from crops into their wild or weedy relatives (Arnaud *et al.*, 2003; Ellstrand, 2003; Ellstrand *et al.*, 1999). Crop-to-wild gene flow thus appears to be more frequent than previously thought on the basis of the assumption that domesticated traits probably reduce fitness in natural conditions (Ellstrand, 2003). Thirteen major food crops have been shown to hybridize with their wild relatives (Ellstrand *et al.*, 1999). Hybridization events are facilitated by the frequent lack of a strong intrinsic reproductive barrier between domesticated crops and their wild relatives (Ellstrand *et al.*, 1999; Gepts, Papa, 2003).

Crop-to-wild introgression may greatly affect the evolution and ecology of wild relatives of domesticated plants (Ellstrand, 1992). The most direct negative consequences of crop-to-wild gene flow include a loss of wild population integrity, resulting in a loss of gene pools important for ecosystem function and potentially useful in controlled introgression strategies for crop improvement (Ellstrand, 2005). Modern landscapes are characterized by a mosaic of wild and cultivated populations, promoting hybridization and potentially threatening wild related species already jeopardized by a loss of their natural habitats (Sagnard *et al.*, 2011). Detailed investigations of gene flow between crops and wild relatives are required for the development of conservation plans for wild crop relatives as genetic resources for breeding purposes and for the assessment of risks associated with transgene escape into natural ecosystems (Ellstrand, 2003; Papa, 2005). Hybridization and introgression rates vary considerably between populations and species (Ellstrand, 2003), hindering the development of general conservation programs. There is therefore an urgent need to quantify crop-to-wild relative gene flow for the various domesticated species and their wild relatives. It is also important to evaluate the dispersal capacities of wild crop relatives exposed to seed and pollen flow from domesticated taxa, for the evaluation of potential crop-to-wild gene flow (Krutovsky *et al.*, 2012). Dispersal capacity and the extent of

gene flow between wild populations strongly affect the spread and ultimate distribution of domesticated alleles or transgenes in the landscape. Dispersal capacities can be estimated indirectly by analyzing intraspecies spatial genetic structure (SGS) (Vekemans, Hardy, 2004). The investigation of SGS may also reveal cryptic genetic discontinuities, making it possible to identify priority areas for conservation, to maintain genetic resources for future breeding programs (Manel *et al.*, 2003).

Trees generally have exceptionally high long-distance dispersal capacities (Kremer *et al.*, 2012) and are therefore important models for studies of crop-to-wild gene flow. Gene flow through long-distance dispersal may increase the likelihood of hybridization between wild and crop populations, thereby increasing the risk of introgression of domesticated alleles or potential transgenes over a large scale. Interest in gene flow from cultivated trees to their wild relatives has increased in the last decade, with trees and shrubs accounting for 16% of the 48 plant species for substantial evidence of crop to wild gene flow was obtained over this period (Ellstrand, 2003).

The cultivated apple (*Malus domestica)* is one of the most widely grown fruit crops of temperate regions, with an annual worldwide production of about 40 million tons. Apple trees are self-incompatible, favoring intra- and interspecific gene flow. The cultivated apple was domesticated in Central Asia from *Malus sieversii* (Cornille *et al.*, 2012b; Velasco *et al.*, 2010) and was brought to Europe through human migrations about 3000 years ago (Juniper, Mabberley, 2006). The migration of domesticated apples westwards from Asia may have involved contact and hybridization with local wild *Malus* relatives growing along the Silk Route. Previous investigations of the evolutionary history of the cultivated apple have provided evidence for wild-to-crop gene flow, with a major contribution of the European crabapple, *M. sylvestris,* to the genetic makeup of modern domesticated apples, and a possible contribution of the Caucasian *M. orientalis* to the genome of some Mediterranean cultivars (Cornille *et al.*, 2012b). These three wild species are small insect-pollinated trees of the Rosaceae family. *M. sylvestris* and *M. orientalis* grow in low-density populations in natural habitats, whereas *M. sieversii* forms high-density populations in the Tian Shan Mountains (Jackson, Weng, 1999). The three wild relatives are mostly pollinated by bees and flies (Syrphidae). Diverse wild animals, including mammals and large birds, feeds on the fruit, but their respective efficiencies as seed dispersal vectors are unknown (Juniper, Mabberley,

2006; Larsen *et al.*, 2006). Hybrids cannot be differentiated from cultivated apples or pure wild individuals on the basis of morphological characteristics alone, molecular tools are therefore required to investigate hybridization. Previous molecular studies have identified a few wild *M. sylvestris* trees displaying introgression from *M. domestica* in populations from Denmark and Belgium (Coart *et al.*, 2006; Larsen *et al.*, 2006; Larsen, Kjær, 2009). However, we still know little about crop-to-wild gene flow, SGS and dispersal capacities across the full geographic ranges of the wild relatives that have contributed to the cultivated apple genome.

We used 26 microsatellite markers and a comprehensive set of 1181 samples of wild apples collected in Europe, the Caucasus and Central Asia to investigate the extent of introgressive hybridization from the cultivated apple into the genomes of *M. sieversii*, *M. orientalis* and *M. sylvestris*, together with a previously characterized set of pure *M. domestica* reference genotypes (Cornille *et al.*, 2012b). We also investigated the SGS of the three wild relatives of the cultivated apple. We addressed the following specific questions: 1) How much crop-to-wild gene flow has occurred from the cultivated apple to its wild relatives? 2) What are the SGS and dispersal capacities of the three wild species? The SGS and dispersal capacity of the European crabapple have recently been investigated (Cornille *et al.*, 2012a) and are therefore not presented here.

Materials and methods

Sampling, DNA extraction and microsatellite genotyping

Leaf material was collected from 1) *M. sylvestris* (*N*=796) at 56 sites across Europe, including 38 samples from the apple collection of the Plant Genetic Resources Unit (ILVO); 2) *M. orientalis* (*N*=217) at 27 sites across the Caucasus (Armenia, Russia and Turkey); 3) *M. sieversii* at 28 sites across Asia (Kazakhstan, China, Kyrgyzstan, Tajikistan and Uzbekistan) (*N*=168). Details of the sampling locations are provided in Table S1. We chose 40 *M. domestica* individuals previously identified as displaying no introgression from *M. sylvestris*, *M. sieversii* or *M. orientalis*, *i.e.*, with membership coefficients exceeding 0.9 for the *M. domestica* genepool in a STRUCTURE analysis (Cornille *et al.*, 2012b). These 40 individuals

were used as the reference "pure" *M. domestica* genepool, for the identification of hybrids between wild species and the domesticated apple.

DNA was extracted with the Nucleo Spin® plant DNA extraction kit II (Macherey & Nagel, Düren, Germany). Multiplex microsatellite PCR amplifications were performed with a Multiplex PCR Kit (QIAGEN, Inc.). We used 26 microsatellites spread across the 17 chromosomes (one to three microsatellites per chromosome) using 10 different multiplex reactions, as previously described (Cornille *et al.*, 2012b; Patocchi *et al.*, 2009). We retained only multilocus genotypes with less than 25% missing data.

Bayesian inference of population structure and hybridization

We used the individual-based Bayesian clustering methods implemented in STRUCTURE 2.3.3 (Pritchard *et al.*, 2000). Preliminary Bayesian analyses with STRUCTURE showed that the most relevant numbers of populations (K_w) for the wild species were K_w=3 for *M. sylvestris* (Cornille *et al.*, 2012a); K_w=2 for *M. sieversii* and K_w=3 for *M. orientalis* (data not shown). We then estimated introgression from *M. domestica* into the wild species. We used STRUCTURE 2.3.3 (Pritchard *et al.*, 2000) on three datasets, each encompassing the 40 reference *M. domestica* cultivars and a wild species, without using the origin of samples as prior information to assist clustering. For each dataset, we considered models with K=2 to K= K_w+1 clusters, K_w corresponding to the number of genetically differentiated populations identified in the focal wild species in preliminary analyses. We considered individuals assigned to a wild genepool with a membership coefficient <0.9 to display introgression from *M. domestica*. The membership coefficients of wild genotypes into the *M. domestica* and individual wild species gene pools are denoted $P_{domestica}$ and $P_{wild,}$ respectively.

For the analysis of within-species population structure, we excluded hybrids and used the individual-based spatially explicit Bayesian clustering method implemented in TESS 2.1 (Chen *et al.*, 2007). These analyses were performed for *M. sieversii* and *M. orientalis* only, as the population structure of *M. sylvestris* has already been investigated (Cornille *et al.*, 2012a). Both STRUCTURE and TESS make use of Markov Chain Monte Carlo (MCMC) simulations to infer the proportion of ancestry of genotypes from K distinct clusters. The underlying algorithms attempt to minimize deviations from Hardy–Weinberg and linkage disequilibria. TESS also incorporates a spatial component into the clustering procedure, such

that genotypes from areas located closer together geographically are considered more likely to belong to the same cluster. In TESS analyses, we used the conditional autoregressive (CAR) Gaussian model of admixture with linear trend surface, setting the spatial interaction parameter (ρ) to 0.6. These parameters (ρ and trend) affect the weight given to spatial distance when clustering genotypes. Ten independent analyses were carried out for each number of clusters K ($2 \leq K \leq 6$ for TESS), using 500,000 MCMC iterations after a burn-in of 50,000 steps. Outputs were processed with CLUMPP v1.1.2 (Jakobsson, Rosenberg, 2007) to identify distinct modes (*i.e.*, clustering solutions) in the replicated runs of each K.

Genetic variation within M. sieversii *and* M. orientalis

We used ARLEQUIN (Excoffier, Lischer, 2010) to check the suitability of the markers for population genetic analyses in each species. None of the 26 microsatellite markers significantly deviated from a neutral equilibrium model, as shown by the non significant P-values obtained in Ewen-Watterson tests, and no pair of markers was in significant linkage disequilibrium (Raymond, Rousset, 1995; Rousset, 2008). The markers were therefore considered to be unlinked and to be subject to neutral evolution, in each species.

We tested for the occurrence of null alleles at each locus with MICROCHECKER 2.2.3 (Van Oosterhout *et al.*, 2004). Allelic richness was calculated with ADZE software (Szpiech *et al.*, 2008), at the site and cluster levels, using sample sizes of N=14 (seven individuals x two chromosomes) for *M. orientalis* and N=6 (three individuals x two chromosomes) for *M. sieversii*, corresponding to the smallest number of observations for sites and clusters, respectively. Heterozygosity, Weir and Cockerham F-statistics and Hardy-Weinberg genotypic linkage equilibrium were assessed in GENEPOP 4.0 (Raymond, Rousset, 1995; Rousset, 2008). Only sampling sites with at least six successfully genotyped specimens were included in site-specific calculations (seven sites for *M. sieversii* and 14 for *M. orientalis*).

We checked for isolation by distance (IBD) patterns, as previously described (Loiselle *et al.*, 1995). A Mantel test with 10,000 random permutations was performed between the individual coefficient of relatedness F_{ij} and the matrix of the natural logarithm of geographic distance. These analyses were performed with SPAGeDI 1.3 (Hardy, Vekemans, 2002), separately for the main clusters identified by Bayesian clustering algorithms in both *M. sieversii* (Main.siev) and *M. orientalis* (C.or and S.Or). Spatial patterns of genetic variability

were visualized by mapping variation in allelic richness over space with the interpolation kriging function in ARCINFO (ESRI, Redlands, CA), using a spherical semi-variogram model.

Results

Estimation of introgression from the cultivated apple to the wild species

We ran STRUCTURE analyses on three datasets, each including the 40 reference *M. domestica* genotypes and one wild species, setting the number of clusters (*K*) to values corresponding to the number of clusters (K_w) determined in preliminary analyses on the focal wild species, plus one for *M. domestica*, *i.e.* $K = K_w+1$, with *K*=3 for *M. sieversii* and *K*=4 for both *M. orientalis* and *M. sylvestris.* STRUCTURE analyses consistently detected hybrid genotypes in wild species (Figure 1). The assignments obtained at lower *K* values confirmed that running STRUCTURE at *K* values below the most relevant value can create spurious assignments (Kalinowski, 2011) (Figure 1). For instance, some of the *M. sieversii* individuals partly assigned to the *M. domestica* gene pool at *K*=2 actually formed a distinct pure *M. sieversii* population at *K*=3, without footprints of admixture with the *M. domestica* genepool. Not taking into account the population structure of the wild species can thus lead to the detection of spurious footprints of introgression (Kalinowski, 2011). Setting *K* at higher values revealed no further substructure within species, indicating that the admixture detected at K_w+1 was not an artifact.

Each wild species included genotypes showing signs of admixture with *M. domestica*, (*i.e*, individuals with coefficients of <0.9 for membership of the wild species gene pool) (Figure 2). However, the proportion of hybrids differed considerably between the wild species: only seven hybrid genotypes were identified in *M. orientalis* (3.2% of the sample), 25 hybrids were identified in *M. sieversii* (14.8% of the sample), and *M. sylvestris* included many hybrids, with 292 admixed individuals (36.7% of the sample). Misidentified individuals (*i.e*., pure *M. domestica* genotypes) were detected within *M. sylvestris* (*N*=37) and *M. sieversii* (*N*=4). The distributions of the proportion of admixture between wild species and *M. domestica* were obtained by summing, across the populations identified within each species, the number of genotypes with a membership coefficient corresponding to a given membership class. The distribution of admixture proportions varied among wild species

(χ^2=74.6, P<0.01). All *M. orientalis* trees displaying introgression from *M. domestica* had low proportions of admixture, *M. sieversii* included some individuals with higher admixture proportions, whereas *M. sylvestris* presented the whole range of admixture values (Figure 2). *Malus sylvestris* presented a large number of intermediate hybrid genotypes, with trees displaying introgression from *M. domestica* ($0.1<P_{domestica}<0.45$), and even included trees that appeared to be hybrids backcrossed with *M. domestica* or feral *M. domestica* displaying introgression from *M. sylvestris* ($0.55<P_{domestica}<0.9$) across its distribution range.

Population subdivision in Malus sieversii *and* Malus orientalis

Hybrids were removed from the datasets for within-species analyses. Summary statistics for population structure and diversity are shown for *M. sieversii* and *M. orientalis* in Table 1. Within *M. sieversii*, for the eight sites for which at least six samples were available, the mean sample size was 10.9±6.1 genotypes. On average, across sites and markers, allelic richness was 3.5±0.3 [min: 3.1 – max: 3.9] and gene diversity was 0.75±0.03 [min: 0.69 – max: 0.76]. F_{IS} values were low, with a mean of 0.02±0.04 per site and marker. The species-wide heterozygote deficit was highly significant (P<0.001) but low (F_{IS}=0.05). Between-site differences in allelic frequencies, estimated from the mean F_{ST} across loci, were low (F_{ST}=0.02; min: -0.008 – max: 0.06), but significant for 15 pairs of populations (P<0.05, Table S2).

Within *M. orientalis*, for the 14 sites for which at least six samples were available, the mean sample size was 12.9±3.5 genotypes. On average, across sites and markers, allelic richness was 6.2±0.4 [min: 5.4 – max: 7.1] and gene diversity was 0.81±0.02 [min: 0.69 – max: 0.76]. F_{IS} values were low, with a mean value of 0.04±0.06 per site and marker. The species-wide heterozygote deficit was highly significant (P<0.001) but low (F_{IS}=0.06). Between-site differences in allelic frequencies, estimated from the mean F_{ST} across loci, were low (F_{ST}=0.02; min: -0.02- max: 0.07), but significant for 58 pairs of populations (P<0.05, Table S3).

Spatially explicit clustering analyses with TESS revealed two and three well defined clusters for *M. sieversii* and *M. orientalis,* respectively, even when setting K>2 (Figure S1) and K>3 (Figure S2), respectively. We thus concluded that K=2 and K=3 were the biologically most relevant numbers of populations, for *M. sieversii* and *M. orientalis* respectively, as

observed in preliminary analyses with full datasets. For *M. sieversii*, clustering patterns at *K*=2 revealed a large cluster spreading across Central Asia (green) and a small cluster (red) consisting mostly of individuals from the *Kaz Kuz* population in the Tian Shan Mountains (Figure 3). For *M. orientalis*, clustering patterns at *K*=2 indicated a north/south genetic differentiation between a cluster encompassing the Turkish and Armenian *Shikahogh* populations (green) and another cluster comprising most of the Armenian populations (red) (Figure S2). At *K*=3, a third cluster (blue) appeared, including the Russian population and the Armenian *Jermouck* population (Figure 4).

For subsequent analyses, genotypes were assigned to the geographic population for which their membership coefficient was the highest, provided that this coefficient exceeded 0.55. For *M. orientalis*, six genotypes could not be assigned to any population and were therefore not included in subsequent analyses. The three populations identified in *M. orientalis* are referred to hereafter as the "Northern" ("*N.or*", blue, *N*=4), the "Central" ("*C.or*", red, *N*=159) and the "Southern" ("*S.or*", green, *N*=38) populations. For *M. sieversii*, the two populations are referred to as the "Mountains" ("*Mount.siev*", red at *K*=2, *N*=9) and "Main" ("*Main.siev*", green at *K*=2, *N*=92) populations.

Within the three populations for which *N*>10 (*i.e*, *C.or*, *S.or* and *Main.siev*), genetic differentiation and geographic distance were significantly correlated, consistent with an IBD model. The spatial *Sp* statistic can be used to quantify spatial structure and is useful for comparisons between populations and/or species. Lower *Sp* values are associated with greater dispersal capacities and effective population sizes. *Sp* values were very low, close to 0, but were significant in the *M. sieversii Main.siev* population (*Sp*=-0.002, *P*<0.05) and in *M. orientalis* (*C.or*: *Sp*=0.003, *P*<0.001; *S.Or*: *Sp*=0.002*, P*<0.001). These results suggest that the wild *Malus* species have high dispersal capacities and/or a large effective population size.

Discussion

Wild relatives of the cultivated apple represent a valuable genetic resource for enriching and improving the cultivated apple gene pool, by incorporating alleles or allelic combinations providing greater tolerance to abiotic factors or resistance to pest and diseases. Knowledge of the population structure and dispersal capacities of wild apples and

of the degree of hybridization of these wild species with the domesticated apple is essential for the development of sound conservation plans and innovative breeding strategies. Our findings reveal substantial gene flow from *M. domestica* to wild apple relatives from Europe and the Caucasus (*M. sylvestris* and *M. orientalis*) and to the main Asian progenitor of cultivated apple (*M. sieversii*). The results obtained in this study, together with our previous results for the European crabapple (*M. sylvestris*) (Cornille *et al.*, 2012a), indicate that the wild species have a weak spatial genetic structure, with large-scale isolation by distance, entirely consistent with their high dispersal capacities.

Evidence for crop-to-wild gene flow: introgression from *M. domestica* to its three main wild contributors

Introgression from the cultivated apple to wild relatives has been investigated in *M. sylvestris,* but at a local geographic scale (Coart *et al.*, 2006; Coart *et al.*, 2003; Larsen *et al.*, 2006). Here, we analyzed crop-to-wild gene flow across the entire geographic range of three wild *Malus* species, demonstrating much higher introgression levels. Indeed, our results suggest that substantial crop-to-wild gene flow occurs: 37% of *M. sylvestris*, 15% of *M. sieversii* and 3%, of *M. orientalis* samples were found to be hybrids. Thus, large-scale introgression from *M. domestica* to its wild relatives may occur spontaneously. The lower rates of hybridization observed in previous studies in Belgium and Germany for *M. sylvestris* (4% to 14%) (Coart *et al.*, 2006; Coart *et al.*, 2003) and in Denmark (2%) may be due to the lower power for detecting the *M. domestica* gene pool in these studies, due to the use of fewer markers and/or smaller reference samples, or potential differences in landscape features (*e.g.,* distance to *M. domestica* orchards or forest fragmentation). Unfortunately, the available information concerning landscape features for our sample was insufficiently detailed for assessments of the influence of specific factors.

Previous studies reporting low rates of hybridization between the cultivated apple and the European crabapple had suggested that the low reproductive fitness of interspecific hybrids might account for this finding (Coart *et al.*, 2003). By contrast, our findings of high rates of introgression and backcrossing suggest that hybrids may actually be reasonably fit in natural habitats. Gene flow from domestic to wild species has been investigated in various plant and animal models and recurrent gene flow from crop to wild species has often been

demonstrated (Arrigo *et al.*, 2011; De Andrés *et al.*, 2012; Delplancke *et al.*, 2011; Ellstrand, 2003; Goedbloed *et al.*, 2012; Hübner *et al.*, 2012; Sagnard *et al.*, 2011). The proportion of introgressive hybridization found in the European wild apple was of a similar order of magnitude to that reported in wild relatives of other crops, such as goatgrasses (*Aegilops*, 25% of these wild relatives of wheat are hybrids) (Arrigo *et al.*, 2011), wild grapevine (*Vitis vinifera ssp. sylvestris*, 19%) (De Andrés *et al.*, 2012) and prickly lettuce (*Lactuca serriola*, 7%). In wheat, crop-to-wild gene flow levels may differ between species due to differences in mating systems, with *Aegilops* species, which are the most allogamous, showing the highest levels of hybridization (Arrigo *et al.*, 2011). In lettuce (*Lactuca sativa*), an allele from the crop, delaying flowering, might confer a selective advantage on hybrids in natural conditions (Hartman *et al.*, 2012).

The previous studies reporting low rates of hybridization in *M. sylvestris* also put forward isolation by distance as a barrier to hybridization between *M. domestica* and *M. sylvestris* (Larsen *et al.*, 2006). Other factors, such as overlapping flowering times, have also been proposed as potential barriers to hybridization. We show here that the rates of hybridization between *M. domestica* and wild apple trees are much higher than previously thought, suggesting that hybrids are often viable and that all these potential barriers to interspecific gene flow are actually quite weak.

Weak spatial genetic structure (SGS) in wild contributors and high dispersal capacities

The wild apple species *M. sieversii*, *M. orientalis* and *M. sylvestris* had a weak SGS across a wide geographic range, suggestive of high dispersal capacities. The SGS of the European crabapple has been described elsewhere (Cornille *et al.*, 2012b). The SGS of *M. sylvestris* suggests an ancient contraction followed by expansion since the last glacial maximum in Europe. Three principal populations have been identified in *M. sylvestris*: a large population spreading through Western Europe and two populations with narrower distributions located in the Carpathian Mountains and the Balkans. We investigated the SGS of *M. sieversii* and *M. orientalis* across their geographic ranges, in Central Asia and the Caucasus, respectively. Our findings contrast with previous estimates based on fewer samples and markers, which had suggested a much more pronounced regional structure (Richards *et al.*, 2009; Volk *et al.*, 2008). *Malus sieversii* displayed weak SGS over a large

geographic scale, with two distinct populations identified: one with a broad distribution throughout Central Asia and the other restricted to the Tian Shan Mountains in Kazakhstan. The differentiation of the *Kaz Kuz* population is particularly interesting when compared with that of the fungal pathogen *Venturia inaequalis*, which causes apple scab disease on the domesticated apple and its wild relatives. *Venturia inaequalis* displays a similar pattern of genetic differentiation in the *M. sieversii* forests of the eastern mountains of Kazakhstan (Gladieux *et al.*, 2010b). This correspondence suggests possible co-structuring of the populations of the main apple progenitor *M. sieversii* and its pathogen. The *V. inaequalis* population in the Tian Shan Mountains is thought to be a relic of the ancestral populations infecting *M. sieversii* before the domestication of *M. domestica* (Gladieux *et al.*, 2010b). *Malus orientalis* displayed a weak north-south pattern of SGS, with three distinct populations: a large population corresponding to most of the Armenian samples and two more narrowly distributed populations: one in the Southern Caucasus (Turkey and southern Armenia) and the other in more northerly latitudes (in Russia).

Isolation by distance patterns were detected for both *M. orientalis* and *M. sieversii*, as previously reported for *M. sylvestris* (Cornille *et al.*, 2012b). The three wild species, with their different geographic distributions and densities, displayed similarly low values of the *Sp* statistic, consistent with high dispersal capacities and large effective population sizes. *Sp* estimates can be used to compare the levels of gene flow between wild apples and other tree species. The *Sp* values of the three wild apple species (all animal-dispersed) were of a similar order of magnitude to those estimated for wind-dispersed trees, such as *Larix laricina*, *Quercus robur* and *Fraxinus excelsior* (Vekemans, Hardy, 2004), but were lower than those found for animal-dispersed shrub species (*e.g*, *Sorbus torminalis, Sp*=0.02).

Several features may account for these high *Sp* values and weak SGS. Apples are allogamous, with self-incompatibility systems favoring gene flow and low levels of population genetic differentiation, as typically observed in trees (Kremer *et al.*, 2012). However, animal seed-dispersal syndrome might have been expected to limit gene flow (Vekemans, Hardy, 2004). Animal dispersers of apples include honeybees, bumblebees, leaf-cutter bees and mason-bees for pollen transport and large mammals, such as wild cattle, brown bear and humans for seed transport (Juniper, Mabberley, 2006). These dispersers can travel over long distances, much greater than those covered by birds (while carrying fruits),

for example, potentially accounting for the large dispersal distances, similar to those for wind-dispersed species, obtained for this animal-dispersed species.

The combined dispersal capacities of the seed and pollen dispersers may account for the low *Sp* values obtained for apple trees, and for the large effective population sizes, typical of trees, reported elsewhere (Cornille *et al.*, 2012b). Low density has also been put forward as an explanation for long-distance dispersal and weak SGS (Vekemans, Hardy, 2004). However, whereas *M. orientalis* and *M. sylvestris* are present in low-density populations throughout their geographic ranges, mostly being scattered in forests, *M. sieversii* forms large populations with high densities in the Tian Shan Mountains. As the three wild species studied had similar *Sp* values, density does not appear to be an important factor underlying the weak SGS. It should be noted that *Sp* provides an unbiased estimate when SGS is measured at the appropriate spatial scale, is due exclusively to isolation by distance and has reached equilibrium (Hardy *et al.*, 2006; Vekemans, Hardy, 2004). This might not be the case here, and the SGS may also be explained by historical factors, such as ancient colonization events, bottlenecks or population extinctions.

Conclusion and perspectives

A good knowledge of the existence and location of differentiated populations and of admixture zones is essential to guide breeding and conservation programs (Pautasso, 2009). In breeding programs, differentiated populations are often used to improve the local adaptation of cultivars or to enhance resistance to pathogens of cultivars, by selecting for sources of disease resistance from wild populations (Lenne, Wood, 1991). In programs aiming to conserve wild genetic resources, the identification of interesting populations for potential conservation *in situ* (*e.g.*, conservation of several genetically differentiated populations) or *ex situ* (*e.g.*, establishment of orchards or seed-based core collections from pure wild individuals) is essential. Regions corresponding to “melting pots” of genetic diversity (Jay *et al.*, 2012; Liepelt *et al.*, 2009; Petit *et al.*, 2003) may also include interesting genotypes, *e.g.,* recombinants adapted to new environmental conditions and capable of meeting the challenge of global warming. The genetic uniqueness of southern “rear-edge” populations is also of key importance for long-term conservation purposes (Petit, Hampe, 2006). As wild relatives and contributors to the cultivated apple (Cornille *et al.*, 2012b;

Velasco *et al.*, 2010), *M. orientalis*, *M. sieversii* and *M. sylvestris* are targets for conservation and sustainable management programs for genetic resources (Jacques *et al.*, 2009; Zhang *et al.*, 2007). In these three wild species, differentiated populations have been detected across the corresponding distribution ranges, and these populations may harbor a number of valuable horticultural traits, including resistance to fire blight (*Erwinia amylovora*), apple scab resistance (*V. inaequalis*), fruit quality and drought tolerance. The discovery of signs of a co-structure between *M. sieversii* and *V. inaequalis*, with a small and specific population in the Tian Shan Mountains for both the host and the pathogen species, is of particular interest in terms of breeding programs aiming to increase resistance to apple scab. Our results suggest high levels of historical gene flow throughout the geographic range, but contemporary gene flow may have been drastically reduced by population fragmentation and reduction. *Malus sieversii* is, indeed, suffering from forest destruction, with the restriction of its populations to areas that have been rapidly decreasing in size over the last decade (Zhang *et al.*, 2007).

Crop-to-wild gene flow leads to the spread of crop genes into the wild gene pool and may have various consequences for the evolution of the wild species. Crop-to-wild gene flow can lead to genetic swamping (Le Normand, 2002) and, thus, to the loss of genetic diversity in wild taxa. Such a decline in diversity has been demonstrated in wild *Fragaria virginiana* (Westman et al., 2004), and it has even been suggested that wild cotton, *Gossypium darwinii* and *Gossypium tomentosum,* disappeared through hybridization with the crop *Gossypium hirsutum* (Ellstrand *et al.*, 1999). The problem of crop-to-wild gene flow is of particular importance in forest trees, given their high dispersal capacity, particularly since the development of genetic modification techniques, introducing a risk of transgene flow (Strauss, 2011). The spread of transgenes into the wild gene pool can confer selective advantages on hybrids in the wild (Ellstrand, 2003), even if these genes prove to be deleterious in the long term. Transgene spread is thought to have occurred from transgenic rice (*Oryza sativa*) to red rice (*Oryza sativa f. spontanea*), a weed species (Oard *et al.*, 2000). Some crop alleles may not affect the fitness of hybrids, facilitating their introgression, and resulting in a loss of wild species integrity. This has been demonstrated for the transfer of the BAR transgene into *Oryza sativa f. spontanea*. The first genetically modified apples are probably still a long way from approval for market release, but scientists are already trying

to model the risk and impact of transgene flow in apples (Tyson *et al.*, 2011), and our data will help to provide credible parameter values. There is also too little information available to predict the possible evolutionary consequences of hybridization for wild relatives of the cultivated apple. The current challenge is to evaluate hybrid fitness and to determine whether certain genomic regions are preferentially sites of introgression for crop alleles, that is, whether introgression occurs mostly in non-coding regions or whether introgressed alleles show footprints of recent selective sweeps, indicating positive selection.

Another question warranting further investigation is whether introgressive hybridization in wild apple populations is affected by distance to orchards or sources of *M. domestica* populations or by other ecological or geographic factors. Our results also highlight the need to use molecular markers to elucidate the taxonomic status of *Malus* individuals sampled in the field. Indeed, it is not possible to distinguish between *M. domestica* and *M. sylvestris* reliably on the basis of morphological characters alone, and this may lead to frequent misidentification in the field.

Author contributions

Conceived and designed the experiments: TG PG. Performed the experiments: AC. Analyzed the data: AC. Wrote the paper: AC PG TG. Searched for funding: TG PG AC Wrote grant proposals: TG PG AC

Acknowledgments

We thank the *Région Ile de France* (PICRI), IDEEV, *Fondation Dufrenoy*, SBF (*Société Botanique de France*), CNRS and *Université Paris Sud* for funding, *Plateforme de Génotypage* GENTYANE INRA UMR 1095 *Génétique Diversité et Ecophysiologie des Céréales* and Pauline Lasserre. We thank the following for sampling and providing access to samples: Bruno Le Cam and François Laurens (INRA, IRHS, PRES UNAM, SFR QUASAV, Beaucouzé, France), Xiu-Guo Zhang (Department of Plant Pathology, Shandong Agricultural University, Taiwan, China), Catherine Peix, Aymar Dzhangaliev and collaborators in Kazakhstan, Evelyne Heyer (*Museum National d'Histoire Naturelle*, France), Marie-Anne Félix (*Institut Jacques Monod*, France) and Emmanuelle Jousselin (*Centre de Biologie et de Gestion des Populations*, France) for *M.*

sieversii samples; Joanne Clavel, Ara Hovhannisyan, Karen Manvelyan and Eleonora Gabrielian for *M. orientalis* samples; Dominique Beauvais (*Abbaye de Beauport*, Paimpol, France); Alberto Dominici and Emanuela Fabrizi (Monti Simbruini Regional Park, Italy), Aurélien Cabaret, Jean-Pierre Rioult (Université de Caen Basse-Normandie, EREM - Equipe de Recherche et d'Etudes en Mycologie, France), Isabel Roldán-Ruiz (ILVO, Plant – Growth and Development, Belgium), Nicolas Feau (Forest Sciences Centre, UBC, Vancouver, Canada), Pascal Heitzler (Institut de génétique et de biologie moléculaire et cellulaire, France), François Salomone, Stephano Porta, Jan Kowalczyk and Dzmitry Kahan (Forest Research Institute, Poland), Wilfried Steiner (Northwest German Forest Research Institute, Germany), Thomas Kirisits and Bernhard Kirisits (Institute of Forest Entomology, Forest Pathology and Forest Protection, Vienna, Austria), Heino Konrad (Federal Research and Training Center for Forests, Vienna, Austria), Dalibor Ballian (Faculty of Forestry, University of Sarajevo, Bosnia-Herzegovina), Petya Gercheva, Argir Zhivondov, Valentina Bojkova and Anna Matova (Fruit-Growing Institute, Plovdiv, Bulgaria), Anders Larsen (Forest and Landscape, Department for Management of Forest Genetic Resources, Denmark), Stephens Cavers (NERC, Centre for Ecology and Hydrology, UK), Lazlo Nyari (Forest Genetics and Forest Tree Breeding, Göttingen, Germany), Per Avid (Agder Natural History Museum and Botanical Garden, Norway), Lucian Curtus (Transylvania University Brasov, Faculty of Forest Sciences, Romania), Carlos Herrera (*Estacion Biologica de Donana*, CSIC, Spain), Francisco Donaire (*Jardín Botánico La Cortijuela*, Sierra Nevada, Granada, Spain), Roman Volansyanchuk (Ukrainian Research Institute of Forestry and Forest Amelioration, Ukraine) for providing *M. sylvestris* samples. We also thank Thierry Genevet, Frédéric Tournay (*Jardin Botanique de Strasbourg*), Levente Kiss, the East Malling Research Station (UK), Philip Forsline and the Plant Genetic Resources Unit in Geneva (NY).

Figures

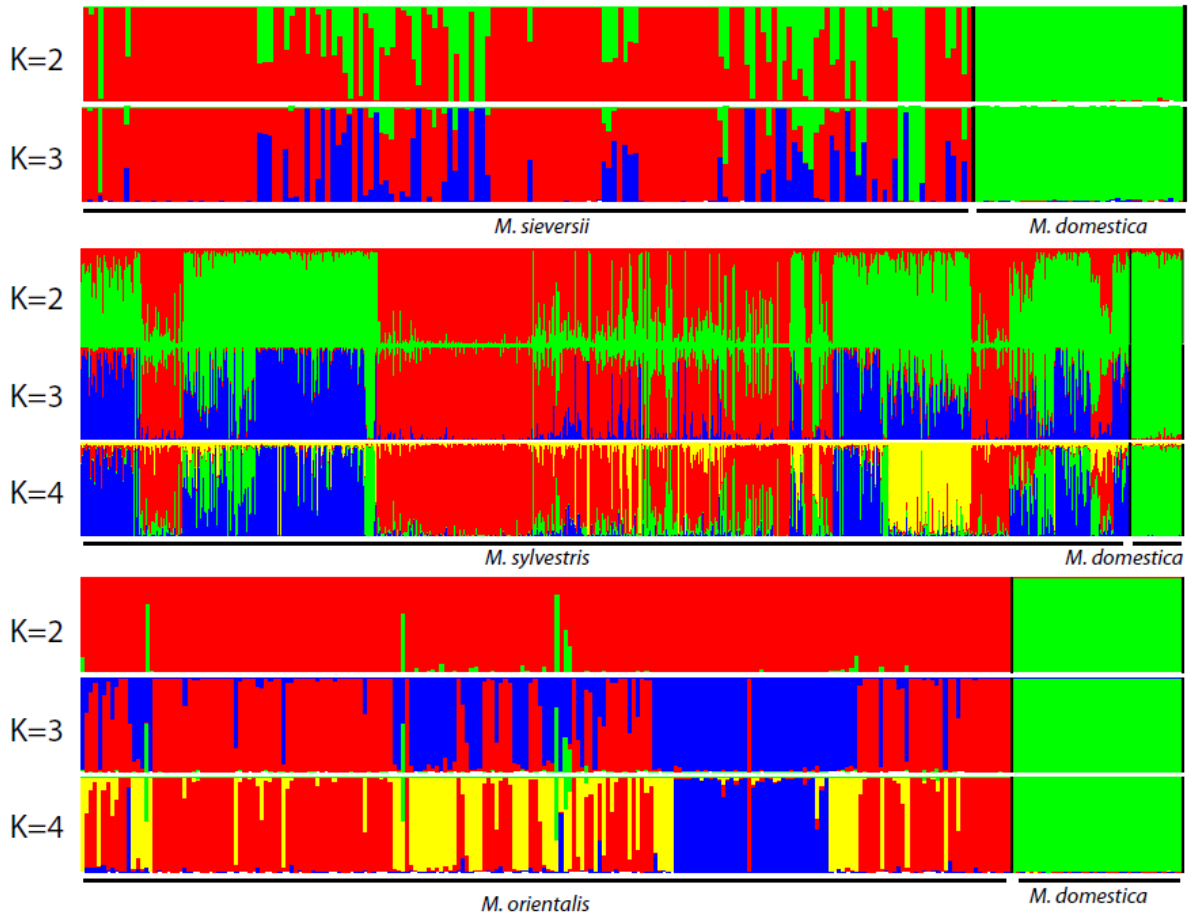


Figure 1. Coefficients of membership in various gene pools inferred with the STRUCTURE program, based on datasets including 40 *Malus domestica* reference samples (green, *N*=40) and one of the three wild *Malus* species in each case. The *x*-axis is not shown to scale. Hybrids were detected by running STRUCTURE from *K*=2 to *K*=3 for *Malus sieversii* and up to *K*=4 for *Malus sylvestris* and *Malus orientalis*.

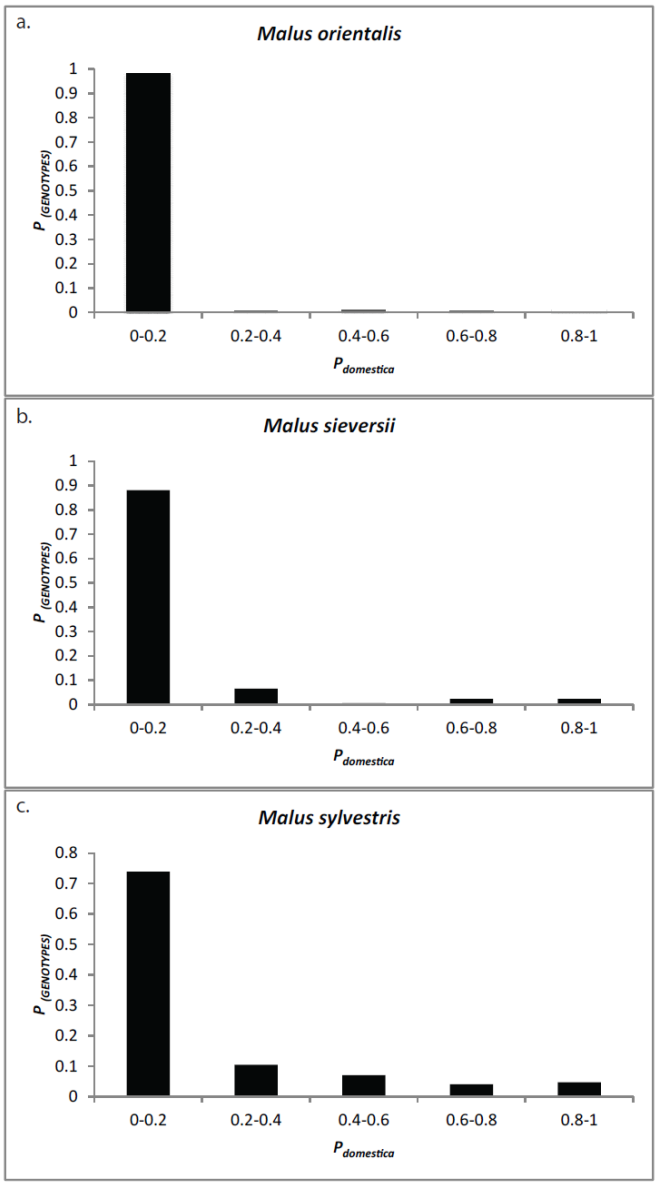


Figure 2. Distribution of membership coefficients in the *Malus domestica* gene pool ($P_{domestica}$) inferred by STRUCTURE for (a) *Malus orientalis* at *K*=4, (b) for *Malus sieversii* at *K*=3 (c) for *Malus sylvestris* at *K*=4. Membership coefficients of each population within the species were summed to obtain $P_{domestica}$. $P_{(GENOTYPES)}$: proportion of genotypes.

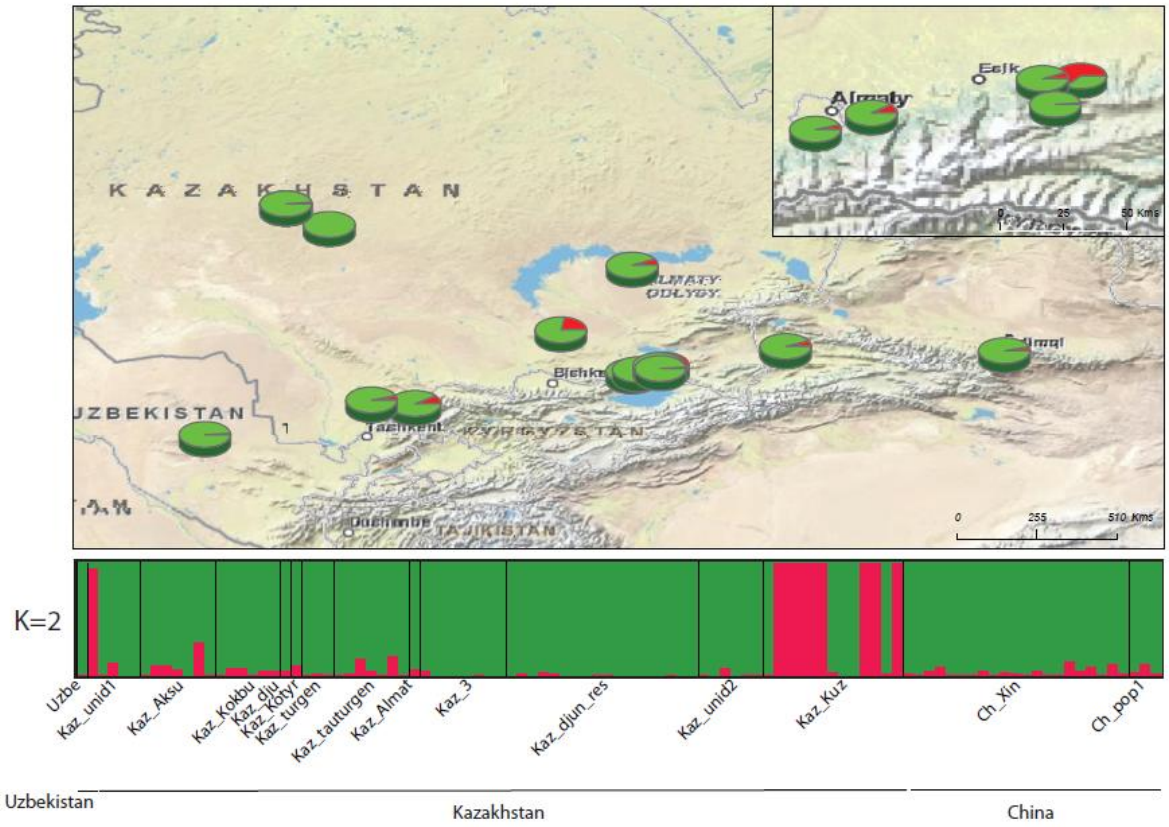


Figure 3. Bayesian clustering results for *Malus sieversii* (*N*=101) in Central Asia, obtained with TESS at *K*=2, and associated map of mean membership probabilities per site. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome corresponding to *K* clusters. Visualization was improved by sorting genotypes by site.

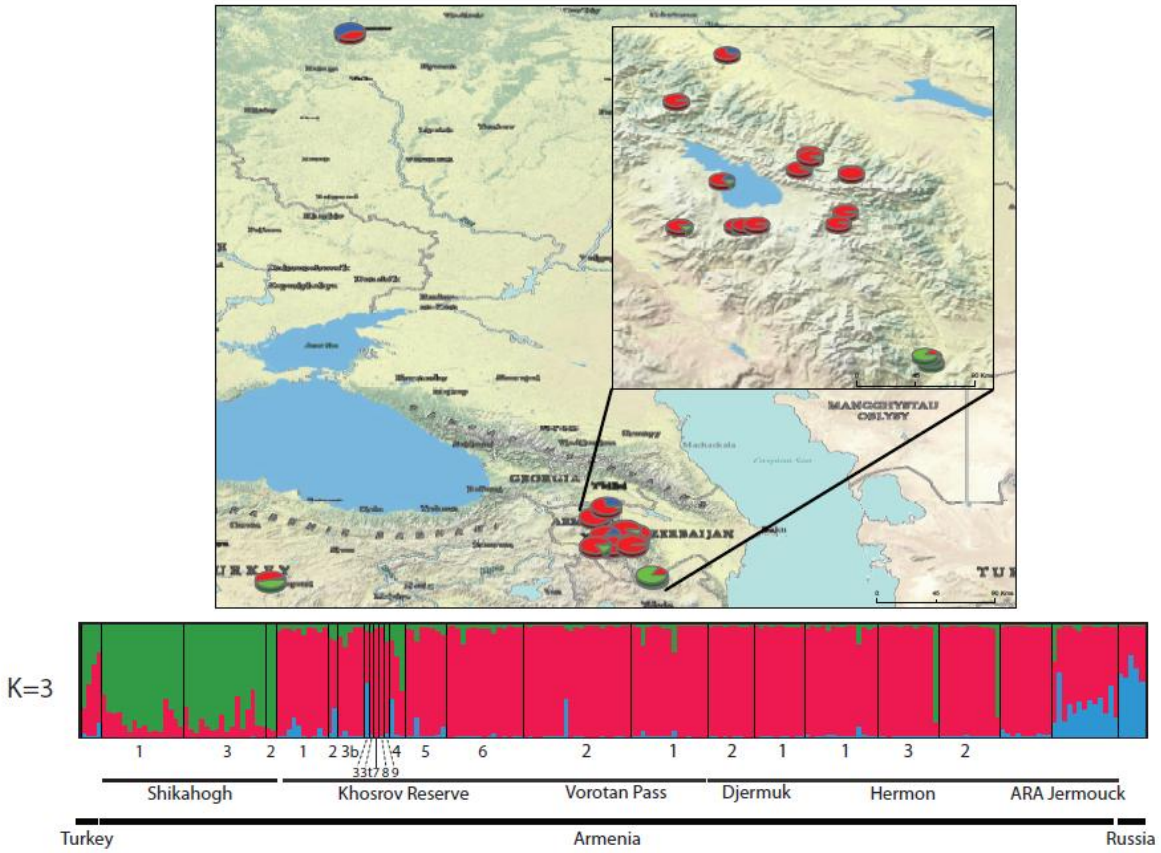


Figure 4: Bayesian clustering results for *Malus orientalis* in the Caucasus (*N*=217) obtained with TESS at *K*=3 and an associated map of mean membership probabilities per site. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome corresponding to *K* clusters. Visualization was improved by sorting the genotypes by site.

Tables

Table 1. Genetic variation within *Malus sieversii* and *Malus orientalis*

Species	Site	*N*	H_O	H_E	F_{IS}	A_R
*M. sieversii**	Ch Xinj	21	0.71	0.75	0.05***	3.5
	Kaz3	8	0.76	0.74	-0.03	3.5
	Kaz Aksu	7	0.72	0.72	0.007	3.9
	Kaz Kuz	13	0.69	0.72	0.05***	3.1
	Kaz djun res	19	0.7	0.80	0.05	3.5
	Kaz tauturgen	7	0.75	0.78	0.04*	3.2
	Kaz unid2	6	0.71	0.77	0.07	3.9
	Overall	101	0.73	0.77	0.05***	
*M. orientalis***	ARA	10	0.82	0.81	-0.02	6.1
	Djermuk1	10	0.71	0.82	0.14***	5.7
	Djermuk2	9	0.74	0.77	0.03	5.4
	Hermon1	14	0.81	0.80	-0.02	6.5
	Hermon2	12	0.78	0.81	0.04***	6.2
	Hermon3	12	0.84	0.80	-0.06	6.1
	Jermouck	13	0.77	0.84	0.08***	7.1
	Khosrov Reserve 1	10	0.74	0.85	0.13***	6.5
	Khosrov Reserve 5	8	0.81	0.84	0.03	6.5
	Khosrov Reserve 6	15	0.80	0.82	0.03	6.3
	Shikahogh 1	16	0.76	0.83	0.08***	6
	Shikahogh 3	16	0.79	0.81	0.03	6
	Vorotanpass 1	15	0.80	0.83	0.04***	6.3
	Vorotanpass 2	21	0.77	0.78	0.02	5.9
	Overall	211	0.79	0.83	0.06***	

N: sample size of each cluster, H_O and H_E: observed and expected heterozygosities, F_{IS}: inbreeding coefficient, A_R: mean allelic richness across loci, corrected by the rarefaction method. *For. *Malus sieversii*, 5 sites (*Ch pop1*; *Kaz Kokbu*; *Kaz Kotyr*; *Uzbe*; *Kaz Almat*) with $N<6$ were excluded from these analyses. **For *Malus orientalis,* 14 sites (*KhosrovReserve 2*, *3*, *3bis*, *3ter*, *4*, *7*, *8, 9*, *unidentified*, *Arm unknown*, *Kaz Kokbu*, *Turkey*, *Russia*, *Shikahogh2*) with $N<6$ were excluded from the analyses. *$0.05<P\leq0.01$; ** $0.01<P\leq0.001$; *** $P<0.001$

Supporting information

Table S1. Description of the *Malus* accessions analysed with their geographical origins and providers.

Table S2. Pairwise genetic differentiation (F_{ST}) among *Malus sieversii* sites (*N*>6)

Table S3. Pairwise genetic differentiation (F_{ST}) among *Malus orientalis* sites (*N*>6)

Figure S1. Bayesian clustering results for *Malus sieversii* in Central Asia (*N*=101) using the program TESS from *K*=2 to *K*=6.

Figure S2. Bayesian clustering results for *Malus orientalis* in the Caucasus (N=217) using the program TESS from K=2 to K=6.

Figure S3. Maps representing the mean membership proportions for *K* clusters, for samples of *Malus sylvestris* collected from the same site

Table S1. Description of the *Malus* accessions analysed with their geographical origins and providers.

Species and locations	Nb	Providers
Malus sylvestris	381	
Austria	40	Thomas Kirisits, Bernhard Kirisits and Heino Konrad
Belgium	9	Isabel Roldán-Ruiz, CRA-W[1], ILVO[2]
Bosnia-Herzegovina	73	Dalibor Ballian
Bulgaria	1	Petya Gercheva, Argir Zhivondov, Valentina Bojkova and Anna Matova
Denmark	100	Anders Larsen
France	45	INRA[3], USDA-ARS[4?] Aurélien Cabaret, Nicolas Feau, Jean-Pierre Rioult, Pascal Heitzler
Germany	30	Jorg Kleinschmit and Wilfried Steiner
UK, Scotland	9	Stephens Cavers
Hungary	27	Lazlo Nyari
Italy	5	Alberto Dominicci and Emanuela Fabrizi, François Salomone, Stephano Porta
Norway	21	Per Avid
Poland	8	Jan Kowalsky and Dzmitry Kahan
Romania	9	Lucian Curtus
Spain	0	Carlos Ferrera and Francisco Donaire
Ukraine	4	Roman Volansyanchuk
Malus sieversii	168	
Kazakhstan	114	Bruno Le Cam, François Laurens, PG, Emmanuelle Jousselin, Marie-Anne Félix, Catherine Peix and Aymar Dzhangaliev
	28	USDA-ARS[4]
China	26	Bruno Le Cam, PG, Xiu-Guo Zhang
Kirghizstan	5	Evelyne Heyer
Tajikistan	1	USDA-ARS[4]
Uzbekistan	1	USDA-ARS[4]

Malus orientalis	217	
		PG, Joanne Clavel, Anush Nersesyan, Ivan Gabrielyan, Ara Hovhannisyan, Karen Manvelyan and Eleonora Gabrielian
Armenia	205	
Russia	5	USDA-ARS[4]
Turkey	5	USDA-ARS[4]
Unknown	2	USDA-ARS[4]
Malus domestica	40	INRA[3], CRA-W[1], USDA-ARS[4], Dominique Beauvais [5] and Jean Pierre Roullaud[7]

Number of trees sampled

[1] CRA - W	Centre Wallons de Recherches Agronomiques, Belgium
[2] ILVO - PLANT	Plant -Growth and Development, Melle, Belgium
[3] INRA	Institut de Recherche en Horticulture et Semences, Angers, France
[4] USDA - ARS	Plant Genetic Resources Unit, Geneva (NY)
[5] Abbaye de Beauport	Conservatory Orchards of ancient apple varieties, Paimpol, France.
[6] EMR	East Malling Research, Kent, UK
[7]Verger Conservatoire d'Arzano	Conservatory Orchards of ancient apple varieties, Brittany, France.

Table S2. Pairwise genetic differentiation (F_{ST}) among *Malus sieversii* sites (*N*>6)

	Kaz_Kuz	Kaz_3	Kaz_djun	Kaz_Aksu	Ch_Xinj	Kaz_taut
Kaz_3	-0.005					
Kaz_djun	0.0336***	0.0152***				
Kaz_Aksu	0.0574***	0.0317***	0.0536***			
Ch_Xinj	0.0343***	0.0142***	0.0262	0.0172		
Kaz_taut	0.023***	0.0151*	0.0063	0.0391***	0.0076**	
Kaz_unid	0.0183	0.0203	0.0017	0.06***	0.0157**	-0.008
Kaz_Kokb	0.0157***	-0.0037	0.0102	0.0515***	-0.0021	-0.0016

*0.05<P≤0.01 ; ** 0.01<P≤0.001 ; *** P<0.001

Table S3. Pairwise genetic differentiation (F_{ST}) among *Malus orientalis* sites (*N*>6)

Site	ARA	Djermuk1	Djermuk2	Hermon1	Hermon2	Hermon3	Jermouck	KhosrovR1	KhosrovR5	KhosrovR6	Shikahog1	Shikahog3
Djermuk1	0.0258***											
Djermuk2	0.0293***	0.0010										
Hermon1	0.0282***	0.0123**	-0.0135									
Hermon2	0.0152	-0.0051	-0.0063	0.0044								
Hermon3	0.0324***	0.0127***	0.0068	0.0101	0.0167							
Jermouck	0.0131*	0.0120**	0.0109	0.0129**	0.0043	0.0255***						
KhosrovR1	0.0081	0.0222***	0.0256***	0.0242***	0.0179**	0.0335***	-0.0024					
KhosrovR5	0.0148***	0.0114	0.0273	0.0184*	0.0200	0.0389*	0.0135	0.0103				
KhosrovR6	0.0192	0.0123**	0.0229***	0.0176***	0.0103***	0.0231***	0.0124***	0.0122	-0.0031			
Shikahog1	0.0403***	0.0434***	0.0559***	0.0463***	0.0460***	0.0646***	0.0241***	0.0210***	0.0269*	0.0371**		
Shikahog2	0.0658***	0.0248**	0.0563***	0.0504***	0.0530**	0.0692***	0.0172*	0.0273***	0.0449*	0.0471***	0.0078	
Vorotanp1	0.0126	-0.0007	-0.0005	0.0026	-0.0155	0.0145*	0.0045	0.0093	0.0030	0.0044	0.0363***	0.0371***
Vorotanp2	0.0312***	0.0196***	0.0048*	0.0083**	-0.0017	0.0225***	0.0260***	0.0311***	0.0251**	0.0215**	0.0594***	0.0574***

*0.05<P≤0.01 ; ** 0.01<P≤0.001 ; *** P<0.001

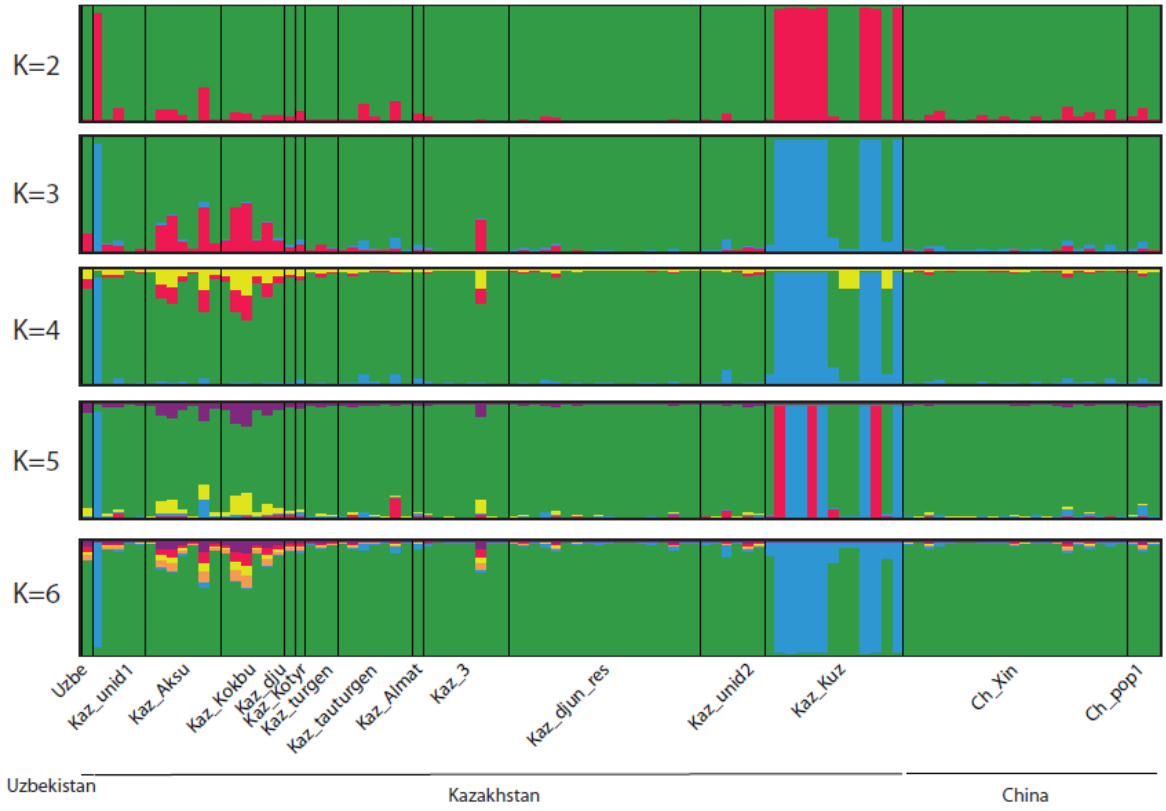


Figure S1. Bayesian clustering results for *Malus sieversii* in Central Asia (*N*=101) using the program TESS from *K*=2 to *K*=6. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome in *K* clusters. Visualization was improved by sorting genotypes by site.

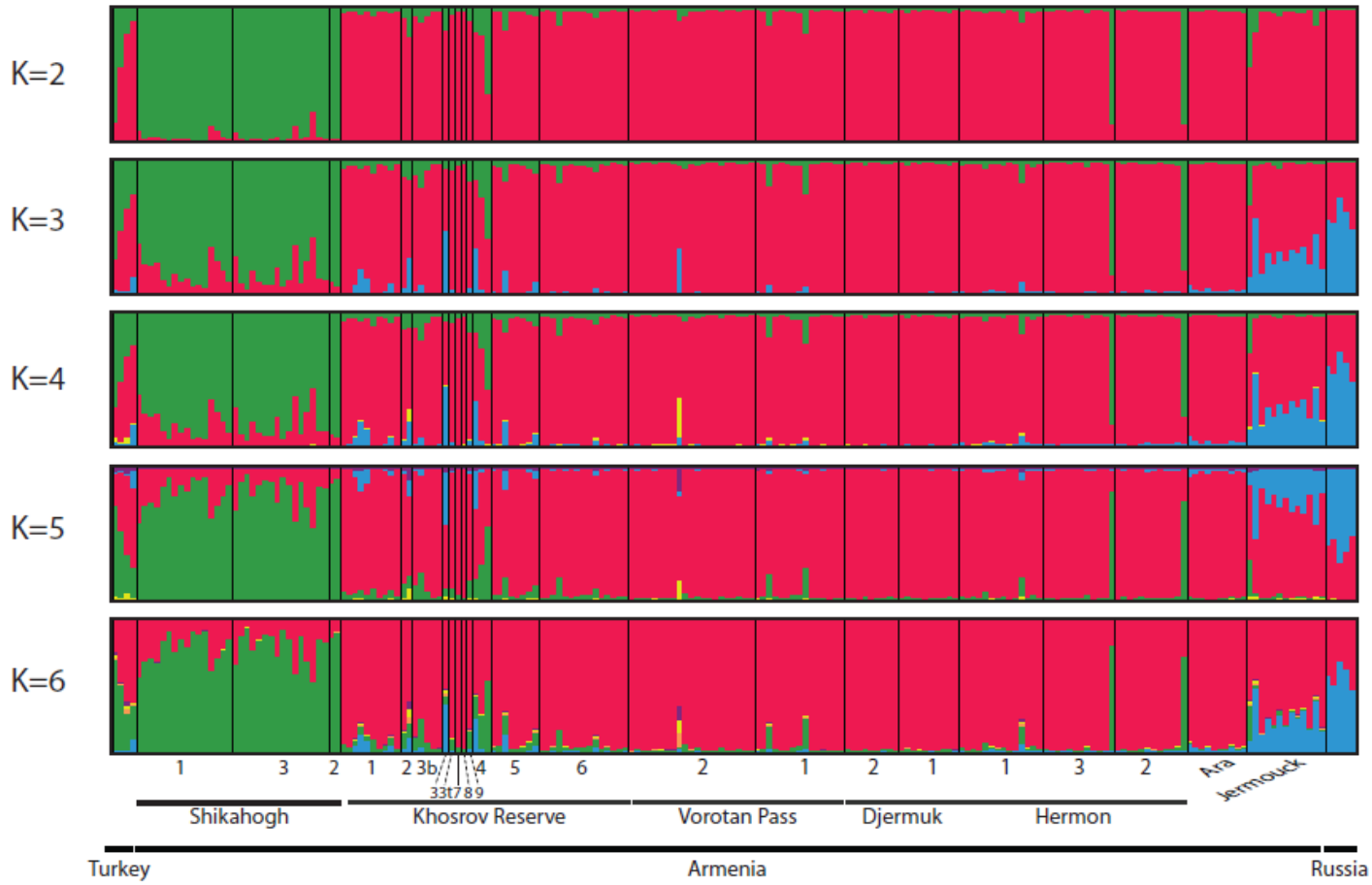


Figure S2. Bayesian clustering results for *Malus orientalis* in the Caucasus (*N*=217) using the program TESS from *K*=2 to *K*=6. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of ancestry of its genome in *K* clusters. Visualization was improved by sorting genotypes by site.

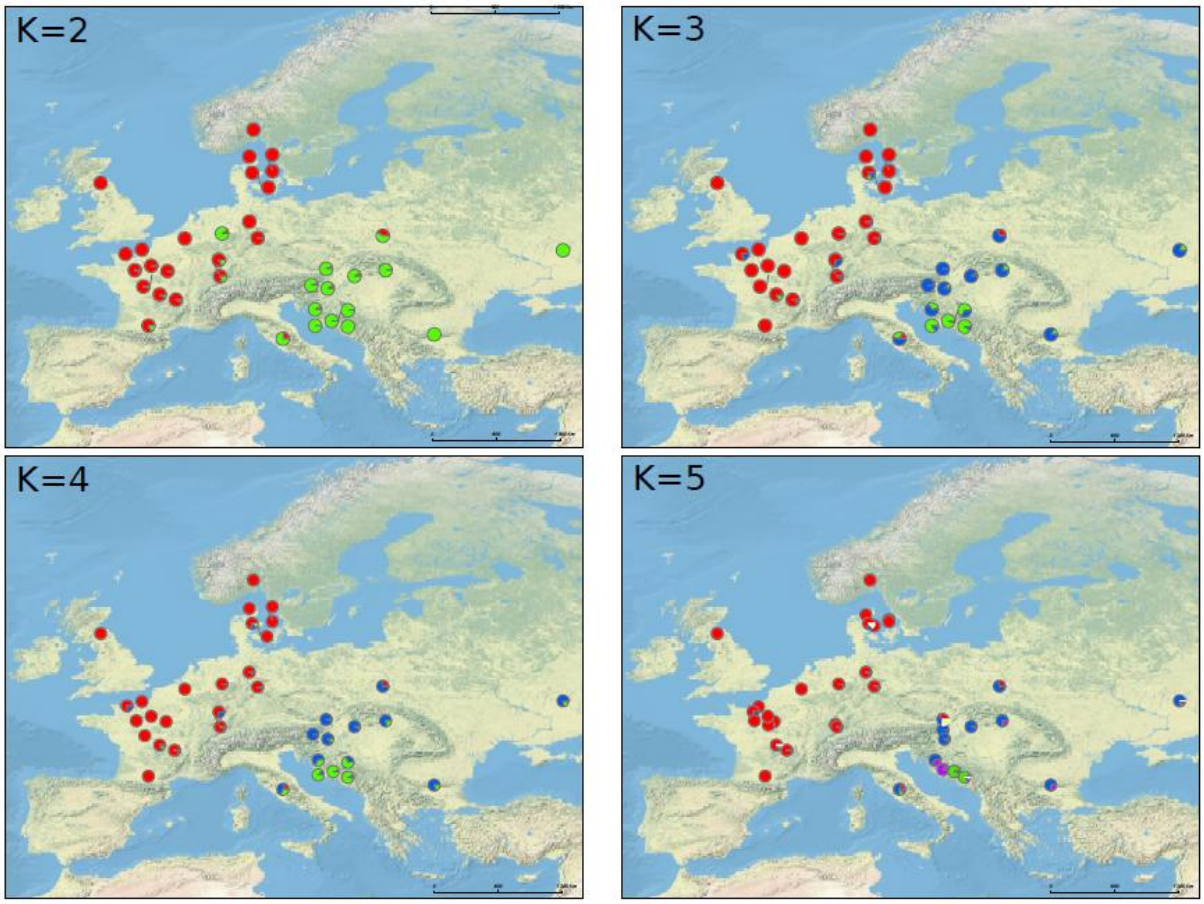


Figure S3. Maps representing the mean membership proportions for *K* clusters, for samples of *Malus sylvestris* collected from the same site. Membership proportions were inferred with the Bayesian clustering algorithm implemented in TESS. At *K*=5 the results presented are those of the minor clustering solution ("mode"), showing the geographic location of the fourth previously identified cluster (Cornille *et al.*, 2012a).

References

Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 1565-1571.

Arrigo N, Guadagnuolo R, Lappe S*, et al.* (2011) Gene flow between wheat and wild relatives: empirical evidence from *Aegilops geniculata*, *Ae. neglecta* and *Ae. triuncialis*. *Evolutionary Applications* **4**, 685-695.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**, 747-756.

Coart E, Van Glabeke S, De Loose M, Larsen AS, Roldán-Ruiz I (2006) Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Molecular Ecology* **15**, 2171-2182.

Coart E, Vekemans X, Smulders MJM*, et al.* (2003) Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Molecular Ecology* **12**, 845-857.

Cornille A, Giraud T, Bellard C*, et al.* (2012a) Post-glacial recolonization history of the European crabapple (Malus sylvestris Mill.), a wild contributor to the domesticated apple. *Molecular Ecology* **in revision**.

Cornille A, Gladieux P, Smulders MJM*, et al.* (2012b) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* **8**, e1002703.

De Andrés MT, Benito A, Pérez-Rivera G*, et al.* (2012) Genetic diversity of wild grapevine populations in Spain and their genetic relationships with cultivated grapevines. *Molecular Ecology* **21**, 800-816.

Delplancke M, Alvarez N, Espíndola A*, et al.* (2011) Gene flow among wild and domesticated almond species: insights from chloroplast and nuclear markers. *Evolutionary Applications* **5**, 317-329.

Ellstrand NC (1992) Gene flow by pollen: implications for plant conservation genetics. *Oikos* **63**, 77-86.

Ellstrand NC (2003) Current knowledge of gene flow in plants: implications for transgene flow. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **358**, 1163-1170.

Ellstrand NC (2005) *Dangerous Liaisons?: When Cultivated Plants Mate with Their Wild Relatives* Johns Hopkins University Press.

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **Vol. 30**, 539-563.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Gepts P, Papa R (2003) Possible effects of (trans)gene flow from crops on the genetic diversity from landraces and wild relatives. *Environmental Biosafety Research* **2**, 89-103.

Gladieux P, Zhang XG, Roldàn-Ruiz I*, et al.* (2010) Evolution of the population structure of *Venturia inaequalis*, the apple scab fungus, associated with the domestication of its host. *Molecular Ecology* **19**, 658-674.

Goedbloed DJ, Megens HJ, Van Hooft P*, et al.* (2012) Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. *Molecular Ecology*, in press.

Hardy OJ, Maggia L, Bandou E*, et al.* (2006) Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. *Molecular Ecology* **15**, 559-571.

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620.

Hartman Y, Hooftman DAP, Uwimana B*, et al.* (2012) Genomic regions in crop–wild hybrids of lettuce are affected differently in different environments: implications for crop breeding. *Evolutionary Applications*, in press.

Hübner S, Günther T, Flavell A*, et al.* (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Molecular Ecology* **21**, 1115-1129.

Jackson ST, Weng C (1999) Late Quaternary extinction of a tree species in eastern North America. *Proceedings of the National Academy of Sciences* **96**, 13847-13852.

Jacques D, Vandermijnsbrugge K, Lemaire S, Antofie A, Lateur M (2009) Natural distribution and variability of the wild apple (*Malus sylvestris*) in Belgium *Belgian Journal of Botany* **142**, 39-49

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jay F, Manel S, Alvarez N*, et al.* (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology* **21**, 2354-2368.

Juniper BE, Mabberley DJ (2006) *The Story of the Apple* Imber Press, Inc.

Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* **106**, 625-632.

Kareiva P, Watts S, McDonald R, Boucher T (2007) Domesticated nature: shaping landscapes and ecosystems for human welfare. *Science* **316**, 1866-1869.

Kremer A, Ronce O, Robledo-Arnuncio JJ*, et al.* (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters* **15**, 378-392.

Krutovsky KV, Burczyk J, Chybicki I*, et al.* (2012) Gene flow, spatial structure, local adaptation, and assisted migration in trees

Genomics of Tree Crops, pp. 71-116. Springer New York.

Larsen A, Asmussen C, Coart E, Olrik D, Kjær E (2006) Hybridization and genetic variation in Danish populations of European crab apple (*Malus sylvestris*). *Tree Genetics & Genomes* **2**, 86-97.

Larsen A, Kjær E (2009) Pollen mediated gene flow in a native population of *Malus sylvestris* and its implications for contemporary gene conservation management. *Conservation Genetics* **10**, 1637-1646.

Le Normand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution* **17**, 183-189.

Lenne JM, Wood D (1991) Plant diseases and the use of wild germplasm. *Annual Review of Phytopathology* **29**, 35-63.

Liepelt S, Cheddadi R, de Beaulieu J-L*, et al.* (2009) Postglacial range expansion and its genetic imprints in *Abies alba* (Mill.) A synthesis from palaeobotanic and genetic data. *Review of Palaeobotany and Palynology* **153**, 139-149.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **Vol. 82**, 1420-1425

Manel Sp, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**, 189-197.

Oard J, Cohn MA, Linscombe S, Gealy D, Gravois K (2000) Field evaluation of seed production, shattering, and dormancy in hybrid populations of transgenic rice (*Oryza sativa*) and the weed, red rice (*Oryza sativa*). *Plant Science* **157**, 13-22.

Papa R (2005) Gene flow and introgression between domesticated crops and their wild relatives. In: *Proceedings of the International Workshop on the Role of Biotechnology for the Characterisation and Conservation of Crop, Forestry, Animal and Fishery Genetic Resources*, Turin, Italy.

Patocchi A, Fernàndez-Fernàndez F, Evans K*, et al.* (2009) Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. *Tree Genetics Genomes* **5**, 211-223.

Pautasso M (2009) Geographical genetics and the conservation of forest trees. *Perspectives in Plant Ecology, Evolution and Systematics* **11**, 157-189.

Petit RJ, Aguinagalde I, de Beaulieu J-L*, et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**, 1563-1565.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37**, 187-214.

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

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

Richards C, Volk G, Reilley A*, et al.* (2009) Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. *Tree Genetics & Genomes* **5**, 339-347.

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Sagnard F, Deu M, Dembélé D*, et al.* (2011) Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region. *TAG Theoretical and Applied Genetics* **123**, 1231-1246.

Strauss (2011) Transgenic biotechnology in forestry: what a long strange trip it's been. *BMC Proceedings* **5**.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-2504.

Tyson RC, Wilson JB, Lane WD (2011) A mechanistic model to predict transgenic seed contamination in bee-pollinated crops validated in an apple orchard. *Ecological Modelling* **222**, 2084-2092.

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.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921-935.

Velasco R, Zharkikh A, Affourtit J*, et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics* **42**, 833-839.

Volk GM, Richards CM, Reilley AA*, et al.* (2008) Genetic diversity and disease resistance of wild *Malus orientalis* from Turkey and Southern Russia. *Journal of the American Society for Horticultural Science* **133**, 383-389.

Westman AL, Medel S, Spira TP*, et al.* (2004) Molecular genetic assessment of the potential for gene escape in strawberry, a model perennial study crop. In: *Introgression from genetically modified plants into wild relatives*, pp. 75-88. H. C. M.

Zhang C, Chen X, He T*, et al.* (2007) Genetic structure of *Malus sieversii* population from Xinjiang, China, revealed by SSR markers. *Journal of Genetics and Genomics* **34**, 947-955.

Discussion et Perspectives

Discussion et Perspectives

Le genre *Malus* s'est révélé être un bon modèle pour l'étude des processus de diversification à différentes échelles évolutives (domestication, diversifications intraspécifiques et interspécifiques) : il inclut une espèce domestiquée, *Malus domestica,* et présente une histoire de diversification récente avec des espèces phylogénétiquement très proches et possibles contributrices au génome du pommier domestiqué : *M. orientalis*, *M. sieversii*, *M. sylvestris* et *M. baccata*. D'autre part, l'étude de ce complexe d'espèces a permis d'apporter des éléments sur l'importance des flux de gènes dans les processus de diversification interspécifiques et intraspécifiques chez un genre d'arbres fruitiers.

1. Diversification récente dans le genre *Malus*

La grande variabilité morphologique observée dans le genre *Malus*, non discriminante entre les espèces, et les probables nombreuses hybridations interspécifiques, rendent le nombre d'espèces très controversé. En particulier, les relations phylogénétiques entre *M. domestica, M. orientalis*, *M. baccata*, *M. sieversii* et *M. sylvestris* restaient encore peu résolues (Harris *et al.*, 2002; Robinson *et al.*, 2001). Les résultats issus du manuscrit A montrent que, sur la base de marqueurs microsatellites, les cinq espèces de pommiers forment des groupes génétiques distincts les uns des autres, malgré leurs faibles différenciations génétiques interspécifiques (Manuscrit A). Les séquences nucléaires (13 loci) montrent une différenciation interspécifique encore plus faible qu'avec les microsatellites, les espèces apparaissant différenciées cette fois uniquement avec l'utilisation d'*a priori* sur l'appartenance des individus à une espèce (Manuscrit C, analyses *STRUCTURE* avec utilisation de distributions *priors* informatives). *Malus baccata* présente la plus forte différenciation génétique par rapport aux autres espèces étudiées, à la fois au niveau des marqueurs microsatellites et des séquences nucléaires. Ce résultat est en accord avec la différenciation phénotypique observée entre cette espèce et les autres espèces sauvages, asiatique (*M. sieversii*), caucasienne (*M. orientalis*) et européenne (*M. sylvestris*), notamment en ce qui concerne la morphologie des fruits. En effet, *M. baccata* présente des pommes ressemblant à des « cerises » (diamètre <1cm), petites et rouges avec pédicelles flexibles sensibles au vent, typiques d'une dispersion par les oiseaux (Juniper, Mabberley,

2006) (Figure 1). En comparaison, les pommes de *M. sylvestris* et *M. orientalis* présentent un diamètre plus élevé (diamètre≈2-5cm), sont vertes/rougeâtres et amères, et dispersées par des herbivores domestiqués ou sauvages (Khoshbakht, Hammer, 2006; Larsen *et al.*, 2006). *Malus sieversii* est l'espèce qui présente la plus grande diversité morphologique de pommes, allant de phénotypes type « *M. sylvestris* » au phénotype de la fameuse variété « Golden Delicious », qui sont elles aussi dispersées par de gros mammifères (Figure 2) (Juniper, Mabberley, 2006). De plus, *M. baccata* a une distribution plus marginale que les autres espèces, dans l'Est de l'Eurasie. Cette plus forte différenciation de *M. baccata,* à la fois sur les plans génétiques et morphologiques, est en accord avec les études précédentes, utilisant des séquences ITS et chloroplastiques, incluant *M. baccata* dans une série différente (série *Baccatae*) de celles de *M. domestica*, *M. sieversii*, *M. orientalis* et *M. sylvestris* (série *Malus*) (Juniper, Mabberley, 2006; Robinson *et al.*, 2001). Bien que les modèles testant les scénarios de spéciation entre les espèces soient encore en cours d'analyse, les données préliminaires suggèrent que ces dernières espèces ont divergé récemment.

Chez les arbres, les tailles efficaces de population (*Ne*) extrêmement élevées laissent penser que la dérive génétique intervient peu dans l'évolution de ces populations/espèces (Petit, Hampe, 2006; Savolainen, Pyhäjärvi, 2007). Mes estimations des tailles efficaces (*Ne*) sur la base des marqueurs microsatellites (2090<*Ne*<64000, Manuscrit A) révèlent que les cinq espèces étudiées ont effectivement des tailles efficaces élevées en comparaison à d'autres espèces de plantes (Petit, Hampe, 2006). Concernant les arbres, mes estimations sont du même ordre de grandeur que les tailles efficaces de certaines espèces de genévriers (Li *et al.*, 2011), et moitié moins élevées que celles de certaines espèces de pins (Wachowiak *et al.*, 2011) et d'épicéas (Chen *et al.*, 2009b). Ces fortes tailles efficaces ont dû jouer un rôle dans la diverfication du genre *Malus*, en particulier dans l'histoire évolutive des cinq espèces étudiées dans cette thèse. En effet, le genre *Malus* serait originaire de Chine et se serait dispersé à partir de là le long du corridor de forêts tempérées distribuées jusqu'en Europe de l'Ouest (Juniper, Mabberley, 2006), qui sont maintenant extrêmement fragmentées. Les tailles efficaces élevées, ont pu contribuer à limiter le niveau de différenciation entre les espèces isolées par la fragmentation du corridor forestier. Le test de scénarios alternatifs par *approximate Bayesian computation*, visant à estimer la part de polymorphisme ancestral et

de flux de gènes dans le niveau de différenciation interspécifique, nous permettra de mieux comprendre les modes de séparation (avec ou sans flux de gènes) entre ces espèces.

Etant donné la faible différenciation estimée entre les cinq espèces en utilisant une trentaine de marqueurs, les approches génomiques constituent une alternative prometteuse afin d'améliorer notre capacité à distinguer différents scénarios de divergence interspécifique. Cela permettrait également d'explorer si certaines zones génomiques demeurent plus ou moins perméables aux introgressions afin de tester l'existence d'« îlots de divergence », c'est-à-dire de zones génomiques plus différenciées que le reste du génome car portant des gènes sous sélection diversifiante. Chez de nombreux modèles biologiques, des gènes impliqués dans la divergence ont été découverts par cette approche, mais des questions se posent quant au rôle de l'organisation de ces gènes dans le génome des populations en divergence permettant la différenciation malgré l'existence de flux de gènes (Feder *et al.*, 2012; Kulathinal *et al.*, 2009; Nosil, 2008; Renaut *et al.*, 2012). Dans ce contexte, l'hypothèse de l'existence d' « îlots de divergence » postule que seules de petites régions génomiques impliquées dans la divergence écologique ou comportementale se différencieraient au début entre espèces, le reste du génome restant perméable aux flux de gènes (Feder *et al.*, 2012). La spéciation avec flux de gènes impliquerait quatre phases, correspondant à des étapes avec des régions de divergence au sein du génome de tailles croissantes au cours du temps, malgré l'existence de flux de gènes, et appelées métaphoriquement des « îlots » puis des « continents » génomiques de divergence. L'un des exemples bien documentés de l'existence de ces « îlots » de divergence concerne deux sous-espèces d'anophèles sympatriques (*Anopheles gambiae,* formes M et S) présentant trois régions génomiques avec une forte différenciation génétique par rapport au reste du génome (Turner, Hahn, 2007; Turner *et al.*, 2005). Malgré d'autres exemples ayant étayé cette hypothèse d'« îlots de divergence » (Ellison *et al.*, 2011; Harr, 2006; Michel *et al.*, 2010; Neafsey *et al.*, 2008; Neafsey *et al.*, 2010), elle reste encore très controversée, même chez *Anopheles* (White *et al.*, 2010). En effet, la détection de ces îlots se base sur l'observation d'une différenciation (F_{ST}) plus élevée de ces régions par rapport au reste du génome. Cependant, d'autres explications alternatives peuvent expliquer ces hétérogénéités de différenciation génomique, impliquant de potentielles mauvaises interprétations sur la signification et l'origine évolutive de ces îlots, en particulier par l'utilisation des mesures de

différenciation ou divergence nucléotidique (F_{ST} ou *Da*) (Noor, Bennett, 2009). Par exemple, l'effet d'une recombinaison restreinte dans certaines parties du génome pourrait biaiser l'estimation de ces paramètres.

Une faible différenciation génétique, de forts effectifs de population et des flux de gènes entre espèces apparentées font de nos cinq espèces de pommiers des modèles idéaux pour étudier la génomique de la divergence chez les arbres. En particulier, ce complexe d'espèces offre l'opportunité de tester les hypothèses d'îlots génomiques de divergence, question encore largement débattue, et encore jamais abordée chez les arbres. De futures études génomiques devraient donc être entreprises dans cette direction.

2. Domestication du pommier cultivé à partir de *Malus sieversii* et contribution majeure inattendue du pommier sauvage européen (*Malus sylvestris*)

D'après des études antérieures utilisant des séquences nucléaires et chloroplastiques, le pommier sauvage d'Asie Centrale, *M. sieversii,* serait l'ancêtre sauvage du pommier cultivé (Harris *et al.*, 2002; Velasco *et al.*, 2010). Une contribution du pommier sauvage européen *M. sylvestris* avait été cependant suggérée (Forsline *et al.*, 2002), en particulier suite à la découverte d'haplotypes partagés au niveau de séquences nucléaires, mais cette contribution restait contestée (Coart *et al.*, 2006; Harrison, Harrison, 2011; Micheletti *et al.*, 2011). En effet, seul un faible nombre de variétés et d'individus sauvages avaient pu être analysés. Dans le manuscrit A, j'ai montré l'existence d'une contribution majeure, par introgressions récentes, du pommier sauvage européen, *M. sylvestris,* dans le génome du pommier cultivé. J'ai aussi montré que *M. sieversii* est l'espèce la plus proche génétiquement du pommier cultivé, si les individus cultivés présentant de forts signaux récents d'introgression avec le pommier sauvage européen sont retirés des analyses. *Malus sylvestris* a ainsi joué un rôle dans la diversification post-domestication du pommier cultivé plutôt que dans sa domestication originelle en Asie Centrale. On peut d'ailleurs se demander si nos résultats sont réellement en accord avec le statut de *M. sieversii* en tant qu'ancêtre du pommier cultivé. En effet, il est possible que l'entité génétique *M. domestica*, différenciée des autres espèces, panmictique et avec une forte variabilité génétique, représente une

espèce à part entière, existant ou ayant existé à l'état sauvage, au même titre que *M. sieversii* dans le Tian Shan, mais que les populations sauvages de cette espèce putative n'aient toujours pas été identifiées et échantillonnées. Cette espèce aurait ensuite été dispersée et aurait subi des croisements avec d'autres espèces présentes le long de la route de la Soie. *M. sieversii* elle-même n'a été découverte qu'en 1929 par le botaniste Russe Nikolai Vavilov. Malheureusement pour les pommiers Kazakhs, Vavilov fût envoyé en prison où il mourut en 1943. Ce n'est qu'au début des années 1990 que *M. sieversii* fut ainsi « redécouverte » grâce à un étudiant Kazakh de Vavilov, Aimak Djangaliev (Morgan, Richards, 2003).

Etonnamment, aucun goulet d'étranglement n'a été détecté chez le pommier cultivé, qui forme par ailleurs une entité génétique panmictique indépendante. On aurait pu pourtant s'attendre à ce que la technique de propagation clonale par greffage utilisée pour cultiver le pommier ait créé une structuration génétique clonale chez l'espèce cultivée, limitant ainsi la différenciation génétique avec son ancêtre sauvage, s'il s'agit bien de *M. sieversii*. L'absence de structure clonale (quand on ne considère qu'un réplicat par variété dans les analyses) peut être expliquée par les techniques de création de cultivars. Le temps de génération élevé des pommiers et leur système de reproduction allogame (à cause d'un système d'auto-incompatibilité) rend la sélection de certains individus «élites» et l'amélioration par croisement longues et laborieuses. Les techniques historiques de création des variétés de pommiers étaient donc de simples « allo-pollinisations au hasard» suivies du choix des meilleurs individus dans la descendance, qui étaient ensuite propagés et maintenus par clonage. Cela a eu pour conséquence le maintien d'une grande diversité génétique au sein du pommier cultivé. Ces méthodes de sélection et de culture sont aussi très utilisées chez d'autres fruitiers comme la vigne, qui présente d'ailleurs une histoire évolutive très proche de celle du pommier cultivé (Myles *et al.*, 2011). En effet, chez la vigne *V. v. vinifera*, les cultivars d'Europe de l'Ouest, originaires des populations sauvages de *V. vinifera sylvestris* dans le Proche Orient, ont subi des introgressions par les populations sauvages de *V. vinifera sylvestris* d'Europe de l'Ouest lors de l'introduction de la vigne en Europe il y a 2800 ans. Bien que les méthodes de croisements entre cultivars élites (intra-*V. v. vinifera*) aient favorisé la coancestralité de la majorité des variétés de vignes créées, ces méthodes de croisements inter-variétaux et inter-populationnels avec *V. v. sylvestris*

localement en Europe de l'Ouest ont permis le maintien d'une forte diversité génétique. Ainsi, les méthodes de sélection ont favorisé les hybridations interspécifiques et intraspécifiques dans l'espèce cultivée, ce qui a même pu éliminer les traces de propagation clonale et d'un éventuel goulet d'étranglement initial. La propagation végétative et ces hybridations ont donc eu un rôle primordial dans l'évolution du pommier cultivé, et probablement aussi chez d'autres fruitiers ligneux.

La vitesse de la domestication est déterminée par la combinaison de l'intensité de la sélection sur les traits phénotypiques - qui dépend à la fois de l'architecture génétique des caractères sélectionnés (mono- *versus* polygéniques) et de la pré-existence ou non de variation génétique chez le taxon sauvage contribuant au phénotype du taxon domestiqué (Doebley *et al.*, 2006) - de la sévérité du goulet d'étranglement lors de la domestication, des tailles efficaces et du système de reproduction (Glémin, Bataillon, 2009; Tenaillon, Manicacci, 2011). Chez les arbres, le mode de propagation et le long temps de génération jouent aussi sur la durée du processus de domestication : à durée égale, le nombre de générations sur laquelle la sélection et les effets de dérive peuvent agir est plus limité que chez les espèces annuelles. La domestication des arbres a été aussi plus tardive dans l'histoire que celle des céréales (Zohary, Hopf, 2000; Zohary, Spiegel-Roy, 1975). Globalement, les exemples de détection de forts goulets d'étranglement sont rares chez les arbres fruitiers. Chez l'abricoter (*Prunus armeniaca)*, une baisse de diversité génétique a été détectée dans une population supposée être dans son centre d'origine secondaire (Bourguiba *et al.*, 2012). De la même manière, les populations cultivées de *Spondias purpurea***,** arbre fruitier en Mésoamérique, présentent une diversité génétique légèrement inférieure à celle des populations sauvages (Miller, Schaal, 2006) ; on peut aussi retrouver une baisse de diversité génétique entre les populations cultivées et sauvages de café (*Coffea arabica*) (Anthony et al., 2002). Cependant, ces baisses de diversité génétique sont bien moindres que celles observées chez certaines plantes annuelles comme le maïs (Tenaillon *et al.*, 2004), le soja (Hyten *et al.*, 2006), le tournesol (Liu, Burke, 2006) ou le blé (Haudry *et al.*, 2007). La plupart des espèces d'arbres fruitiers présentent en effet les mêmes niveaux de diversité génétique que leurs apparentés sauvages, comme par exemple chez l'olivier, le figuier, le pommier ou l'amandier (Besnard *et al.*, 2011; Cornille *et al.*, 2012b; Delplancke *et al.*, 2011; Miller, Gross, 2011; Miller, Schaal, 2006). D'autre part, les syndromes de

domestication sont souvent peu marqués, même si certaines modifications phénotypiques peuvent exister concernant la taille et la teneur en sucre (grenadier, figuier, dattier, vigne et ananas) ou en huile (olivier) du fruit (Zohary, Spiegel-Roy, 1975). En particulier, le pommier cultivé ne présente pas de syndrome de domestication marqué à proprement parler : la qualité du fruit (calibre, couleurs vives et variées) est retrouvée de manière naturelle chez certains arbres dans les forêts du Tian Shan en Asie Centrale, et aucun trait associé à de meilleures conditions de récolte ni de stockage des fruits par exemple n'est présent chez le pommier cultivé (Forsline *et al.*, 2002; Juniper, Mabberley, 2006; Luby *et al.*, 2001). Chez les arbres, les phénotypes des taxons domestiqués sont ainsi souvent très proches des apparentés sauvages, provoquant de nombreux débats sur le statut taxonomique des taxons domestiqués et sauvages (Juniper, Mabberley, 2006; Lumaret, Ouazzani, 2001; Zohary, Spiegel-Roy, 1975). Une explication à ce syndrome de domestication peu marqué peut être que la plupart des traits d'importance agronomique ont un contrôle polygénique, ce qui a rendu, en plus des longs temps de génération, l'amélioration variétale difficile (Neale, 2007). Ce n'est que depuis une petite dizaine d'années, suite à l'avènement des approches de génomique d'association, que les composantes génétiques de ces caractères commencent à être révélées pour certaines espèces d'arbres, en particulier celles utilisées pour la production de bois (Neale, 2007).

Ainsi, les traits d'histoire de vie (larges tailles efficaces et longs temps de génération) et les difficultés d'amélioration inhérentes au cycle de vie des arbres fruitiers (long temps de génération et système de reproduction allogame avec système d'auto-incompatibilité) ont entrainé des différences phénotypiques peu évidentes entre les arbres domestiqués et leurs ancêtres sauvages. Ces observations nous amènent, non pas à se poser des questions sur la réelle domestication des arbres fruitiers depuis le Néolithique, mais à remettre en question l'utilisation chez les arbres d'une définition de domestication basée sur l'existence d'un syndrome de domestication marqué, telle qu'elle est communément utilisée chez les plantes herbacées. Chez les arbres en général certains auteurs font référence à une « domestication en cours » ou « semi-domestication » comme par exemple chez les arbres fruitiers amazoniens comme *Stenocereuse stellatus* (Clement *et al.*, 2010; Pickersgill, 2007). Dans le cas du pommier, les pommes produites par *M. sieversii* ne sont pas toutes phénotypiquement proches de *M. domestica,* suggérant une potentille sélection par

l'homme des phénotypes avantageux chez *M. domestica.* Il est d'ailleurs impossible d'avoir un verger de pommes de bon calibre toutes identiques (clonales) sans l'action de l'homme. Ainsi, le pommier apparaît bien comme une plante domestiquée formant un groupe panmictique différencié de ses apparentés sauvages, mais sans syndrome de domestication extrêmement marqué.

De nouvelles approches en termes de marqueurs génétiques utilisés et de nouveaux efforts d'échantillonnage permettraient certainement d'approfondir nos connaissances sur l'histoire évolutive du pommier cultivé et de ses apparentés sauvages. L'échantillonnage devrait se focaliser sur des régions encore non explorées comme l'Afghanistan, l'Iran, l'Iraq, la Turquie et le sous-continent Indien, ainsi que dans les régions présentant la plus grande diversité spécifique de pommiers, le Sichuan, le Guizhou et le Yunnan en Chine. Ces échantillonnages nous apporteraient de nouveaux éléments sur les relations de parenté entre le pommier domestique et des espèces sauvages encore jamais étudiées, et ainsi révéler d'autres contributions potentielles au génome du pommier domestique. D'autre part, de la même manière que nous avons montré dans le manuscrit A que certains cultivars locaux du Bassin Méditerranéen étaient introgressés par le pommier sauvage caucasien (*M. orientalis*), l'accès à d'autres accessions de cultivars locaux dans le Caucase, le Moyen-Orient, la Chine, l'Afrique du Nord et dans des régions localisées d'Europe, pourraient nous apporter de nouveaux éléments sur des contributions d'autres espèces sauvages à des cultivars locaux le long de la route de la Soie (Juniper, Mabberley, 2006). Il serait aussi intéressant d'établir les relations de parenté entre *M. domestica* et *M. asiatica*, une autre espèce exclusivement cultivée en Asie pour la production de pommes jusqu'à l'introduction, il y un siècle, de *M. domestica* comme principal producteur de pommes commerciales.

Les techniques de séquençage haut débit, rapides et à faibles prix permettraient de déterminer si certaines zones du génome ont été ou non soumises à sélection lors du processus de domestication. Chez des taxons domestiqués présentant des syndromes de domestication marqués, comme par exemple le poulet, des balayages sélectifs ont été mis en évidence sur des gènes codant pour certains caractères phénotypiques d'intérêt agronomique (Rubin *et al.*, 2010). Etant donné le syndrome de domestication peu marqué chez le pommier cultivé, peut-on s'attendre à l'absence de balayage sélectif chez des gènes impliqués dans les caractères phénotypiques comme la couleur, la taille, le goût de la

pomme ? Même si aucun balayage sélectif n'est détecté, observerait-on tout de même des traces de sélection chez ces gènes ou dans d'autres régions du génome ? Ces données génomiques constitueraient une information très utile d'un point de vue fondamental mais aussi pour les programmes d'amélioration variétale (Hamblin *et al.*, 2011; Neale, 2007; Neale, Kremer, 2011).

3. Flux de gènes dans le genre *Malus* : fortes capacités de dispersion et hybridations interspécifiques et intraspécifiques

L'importance des hybridations dans l'évolution des plantes est largement reconnue (Abbott, 1992; Baack, Rieseberg, 2007; Schaal *et al.*, 1998), en particulier chez les plantes pérennes (Petit, Hampe, 2006; Rieseberg *et al.*, 2000). Les introgressions constituent un phénomène clé dans l'évolution des populations car elles jouent sur la spéciation, la diversification et l'adaptation à de nouveaux environnements (Baack, Rieseberg, 2007). Nos études à différentes échelles évolutives des processus de diversification nous ont permis d'estimer les flux de gènes entre les cinq espèces de pommiers et au sein de ces espèces. Jusqu'alors, les études génétiques entre les cinq espèces de pommiers étaient basées sur des séquences nucléaires et chloroplastiques peu variables (Juniper, Mabberley, 2006; Robinson *et al.*, 2001) ou sur un nombre limité d'échantillons de chacune des cinq espèces (Coart *et al.*, 2006; Coart *et al.*, 2003; Larsen *et al.*, 2006). La combinaison de marqueurs hypervariables et d'un échantillonnage représentatif de la distribution géographique de chaque espèce nous a permis d'estimer l'importance des hybridations interspécifiques entre pommiers sauvages et cultivés (Manuscrits A et D), des hybridations intraspécifiques lors des processus de recolonisation (Manuscrit B), et d'estimer les flux de gènes historiques chez les espèces sauvages (Manuscrits B et D). Nous avons exclu *M. baccata* de ces analyses puisque nous n'avions détecté que très peu d'hybrides (Manuscrit A), et l'échantillonnage ne nous permettait pas d'estimer les flux de gènes de manière spatialisée.

Des flux de gènes très importants ont été détectés chez les espèces de pommiers analysées. En effet, l'analyse des populations du pommier sauvage européen (*M. sylvestris*) montre une faible structuration génétique à l'échelle de l'Europe à l'intérieur des groupes génétiques résultant de la recolonisation des refuges glaciaires. De plus, les paramètres de

dispersion estimés (*Sp* et $m_{x\text{-}y}$) chez cette espèce, ainsi que chez *M. orientalis* et *M. sieversii*, dans la limite des données collectées, nous suggèrent de fortes capacités de dispersion, comparables à celles estimées chez certains arbres tropicaux, présentant la caractéristique d'être en faible densité (Vekemans, Hardy, 2004), tout comme les pommiers sauvages (exceptée *M. sieversii* qui forme des forêts denses). Ce résultat suggère que la dispersion des fruits et du pollen par les animaux, connue pour limiter les distances de dispersion et augmenter la structuration génétique spatiale chez d'autres espèces (Vekemans, Hardy, 2004), ne semble pas limiter la flux de gènes chez les pommiers. L'utilisation de marqueurs microsatellites à hérédité bi-parentale ne nous permet pas de distinguer les flux de gènes imputables à la dispersion du pollen ou des graines. Il est donc encore difficile d'apprécier les rôles respectifs des disperseurs de fruits et des pollinisateurs dans les processus de dispersion. Chez les pommiers, les ours, les chèvres, les bovidés sauvages et les oiseaux se nourrissant de pommes, sont des vecteurs de dispersion des graines capables de parcourir de grandes distances (Juniper, Mabberley, 2006; Khoshbakht, Hammer, 2006; Larsen *et al.*, 2006). Les pollinisateurs du pommier sont majoritairement généralistes : abeilles, bourdons, abeilles solitaires et abeilles maçonnes et peuvent butiner sur de longues distances (Juniper, Mabberley, 2006). L'homme a aussi dû et doit encore jouer un rôle essentiel dans la dispersion de ces espèces, comme c'est le cas pour d'autres espèces d'arbres, en particulier beaucoup d'espèces d'arbres invasives telles que *Ginkgo biloba* ou *Quercus robra* (Petit *et al.*, 2004a). En effet, l'homme à travers les échanges et le commerce notamment, a joué un rôle majeur lors de la diffusion du pommier cultivé à travers l'Eurasie. Malgré le peu d'indices à ce jour, la littérature suggère que la combinaison de différents vecteurs de dispersion relativement extensifs, comme chez les pommiers, expliquerait l'observation de flux de gènes aussi importants (Pasquet *et al.*, 2008). D'autre part, les forts flux de gènes observés chez *M. sieversii,* qui elle présente des densités de populations très fortes dans les Montagnes du Tian Shan en Asie Centrale, laissent penser que d'autres facteurs que la densité jouent un rôle dans la dispersion de ces pommiers. Ces facteurs pourraient être liés à des écologies ou des histoires de colonisation différentes entre les pommiers d'Asie de l'Ouest (*M. orientalis* et *M. sylvestris*) et ceux d'Asie Centrale (*M. sieversii*).

Nous avons vu dans le paragraphe précédant que les hybridations interspécifiques avaient joué un rôle inattendu dans l'évolution du pommier cultivé (Manuscrit A), en

particulier dans la diversification post-domestication. Les résultats présentés dans le manuscrit D montrent aussi l'importance des flux de gènes du pommier cultivé vers les pommiers sauvages. Les études précédentes s'intéressant à ces questions n'avaient détecté que très peu de ces évènements d'hybridation, concluant qu'ils étaient rares (Coart *et al.*, 2006; Coart *et al.*, 2003; Larsen *et al.*, 2006). Notre étude (Manuscrit D) montre qu'en réalité ces hybridations sont fréquentes. Cependant, il est difficile de déterminer si l'apparente contradiction de nos résultats avec ceux des études précédantes est liée à des effets confondants tels que des échantillonnages plus exhaustifs, la proximité de zones plus urbanisées ou plus agricoles (par exemple plus proches de vergers à pommiers cultivés) ou l'histoire des zones échantillonnées (anciennes parcelles agricoles). Dans une optique de conservation des ressources génétiques, ces questions sont primordiales. En effet, l'identification de populations de pommiers sauvages non introgressées par le pommier cultivé pourrait permettre de guider les mises en place de vergers à graines de pommiers sauvages « purs » (non introgréssés par le pommier cultivé). Ils auraient un double intérêt. D'une part, *M. sylvestris*, *M. sieversii* et *M. orientalis* représentent des ressources génétiques sauvages importantes, au service de l'amélioration variétale, puisque contributrices au génome du pommier cultivé. D'autre part, *M. sylvestris* à l'heure actuelle (mais aussi dans le futur *M. orientalis* et *M. sieversii*), est de plus en plus utilisée dans les projets agro-forestiers visant au retour de l'arbre dans les agrosystèmes. Dans ces projets, les agroforestiers souhaitent utiliser des individus sauvages dont ils connaissent les caractéristiques génétiques, par exemple des individus non introgressés par le pommier cultivé et correspondant au groupe génétique de la région où ils seront replantés.

Dans mes travaux, j'ai estimé les flux de gènes sur de larges échelles et de manière globale. L'échantillonnage était peu adapté pour estimer plus finement les processus de dispersion: les échantillons pour chacune des espèces étaient concentrés en populations espacées, avec souvent l'absence de coordonnées géographiques individuelles. Pour caractériser la structure génétique spatiale des populations de manière optimale, l'échelle spatiale d'échantillonnage doit être maximisée et relativement continue, par exemple en échantillonnant le long d'un transect, à une échelle correspondant à celle à laquelle les processus de dispersion se déroulent (Vekemans, Hardy, 2004). Ce type d'échantillonnage est donc difficile à mettre en oeuvre sans informations préliminaires sur les capacités de

dispersion des espèces, données que nous avons obtenues dans le manuscrit D. L'estimation fine des capacités de dispersion via le pollen et les graines nous permettrait de mieux comprendre les phénomènes d'hybridations interspécifiques entre le pommier cultivé et le pommier sauvage européen. Les estimations des distances de dispersion nous permettront en effet de cibler les populations de pommiers les plus sujettes à long terme à des introgressions par le pommier domestiqué, selon la distance aux zones géographiques présentant une forte densité de vergers ou de pommiers domestiqués « marrons » (c'est-à-dire « échappés » de cultures domestiques mais poussant à l'état sauvage). Ces perspectives seront étudiées dans le cadre d'un projet que j'ai rédigé, et qui a été financé par la région Ile-de-France dans le cadre de l'appel d'offre PICRI pour 2012-2014 (Annexe 1). D'un point de vue plus fondamental, le processus de colonisation de l'espèce cultivée de l'Asie vers l'Europe pourrait être interprété comme l'invasion d'une espèce (*M. domestica)* en Europe et utilisé comme modèle pour comprendre les conséquences génétiques des hybridations interspécifiques dans les expansions et les invasions (Excoffier *et al.*, 2009; Petit *et al.*, 2004b) chez les espèces domestiquées et leurs apparentés sauvages.

4. *Approximate Bayesian computation* : avantages et inconvénients dans le cadre de l'étude de la diversification dans le genre *Malus*

Afin d'estimer les paramètres démographiques et génétiques des populations de pommiers, je me suis basée sur une approche de plus en plus utilisée : l'*approximate Bayesian computation* (ABC). Les raisons de son utilisation dans ma thèse sont avant tout pragmatiques. En effet, mes jeux de données (5 espèces, 1200 individus, 26 marqueurs microsatellites) ne me permettaient pas d'inférer des paramètres démographiques en un temps raisonnable avec des méthodes basées sur des approches Bayesiennes standardes intégrant des MCMC et des simulations de coalescence comme *IMa* (Hey, 2006; Hey, Nielsen, 2004; Hey, Nielsen, 2007). A titre illustratif, avec de telles approches, la simulation de deux populations de 168 et 140 individus respectivement, avec 26 marqueurs microsatellites m'a demandé deux mois de calcul pour seulement 250000 étapes de « pré-chauffe ». Une autre raison du choix des approches de type ABC vient de l'adéquation des

modèles simulés par ces approches avec les processus évolutifs mis en jeu lors des processus de diversification que je souhaitais étudier durant ma thèse. En effet, les méthodes d'inférences démographiques Bayesiennes utilisant des modèles d'« isolement avec migration » n'étaient pas forcément pertinentes pour tester des hypothèses démographiques complexes caractéristiques des processus de diversification. Les scénarios simples des modèles Bayesiens existant, type IMa (Hey, 2006; Hey, Nielsen, 2004; Hey, Nielsen, 2007), n'intégrent que deux ou quelques populations. Les méthodes ABC permettent donc l'exploration d'un très grand nombre de modèles différents (nombre et valeurs de paramètres utilisés pour les simulations). Ceci leur confèrent un avantage certain lors du choix de scénario évolutifs comme par exemple dans le cas présent où nous avons testé l'existence de flux de gènes lors des processus de diversification dans le genre *Malus* ou les différentes modalités de domestication du pommier cultivé.

Les résultats issus de l'ABC dans l'étude de l'histoire de la domestication du pommier cultivé (Manuscrit A) nous ont permis d'avancer des arguments solides sur la contribution du pommier sauvage européen (*M. sylvestris*). Nous avons choisi d'utiliser *DIYABC* (Cornuet *et al.*, 2008), logiciel qui fait l'hypothèse que les introgressions de chaque espèce ont lieu à un instant *t* (documenté pour chaque espèce sauvage par des données historiques) plutôt qu'avec des flux de gènes à chaque génération depuis un temps *t*. Les résultats ont confirmé ceux de *STRUCTURE* (Pritchard *et al.*, 2000) - indiquant des introgressions récentes sur quelques générations - démontrant en plus qu'il y avait eu des introgressions plus anciennes lors de l'introduction de *M. domestica* en Europe il y a 3000 ans.

Nous avons également utilisé des approches ABC afin de retracer l'histoire de la recolonisation post-glaciaire du pommier sauvage (*M. sylvestris*). Nous avons simulé des coalescents en utilisant *SimCoal2* (Laval, Excoffier, 2004). Ce logiciel permet de simuler des flux de gènes constants entre populations en divergence à chaque génération à partir d'un instant *t.* Nous avons ainsi fait l'hypothèse que, durant les processus de recolonisation, la remise en contact secondaire de populations après isolement implique des flux de gènes à chaque génération, et non pas une introgression à un temps *t* (comme implémenté dans *DIYABC*). Les simulations d'ABC analysées avec *ABCtoolBox* (Wegmann *et al.*, 2010) nous ont permis de suggérer l'existence d'un troisième refuge glaciaire dans les Carpates chez le pommier sauvage européen durant la dernière glaciation ainsi que l'existence de flux de

gènes entre populations recolonisantes, qui avaient été isolées dans les refuges glaciaires durant le Pléistocène (Manuscrit B).

L'approche ABC a été très puissante dans le choix des scénarios avec des supports supérieurs à 70% pour les meilleurs scénarios (Manuscrits A et B). Cependant, l'estimation des paramètres a été plus difficile dans le cas de l'étude de la recolonisation post-glaciaire du pommier sauvage européen. Lors de la reconstruction du processus de domestication du pommier cultivé, les estimations de paramètres présentaient des intervalles de confiance raisonnables et les données observées s'ajustaient bien au modèle choisi (« *Model checking* » : les statistiques résumées observées ne sont pas significativement différentes des statistiques résumées simulées avec le modèle choisi). Au contraire, lors de l'étude de la recolonisation post-glaciaire du pommier sauvage, les estimations des paramètres ont un faible support, avec des intervalles de confiance particulièrement larges. En effet, même si le scénario choisi est bien significativement meilleur que les autres (Manuscrit B), les valeurs des statistiques résumées observées sont significativement différentes de celles obtenues avec le meilleur modèle. Une des possibilités pour vérifier la validité des estimations est de comparer les courbes de distributions postérieures des paramètres à leurs courbes de rejet (Encadré 1). La courbe de rejet est directement liée à l'adéquation du modèle testé aux données observées : quand les données observées sont significativement différentes des données simulées, alors les courbes de rejet et les courbes de distribution postérieures diffèrent. Etant donnés les estimations et les intervalles de confiance obtenus, ainsi que la proximité des données observées avec celles simulées pour la domestication du pommier (*i.e.*, avec *DIYABC)*, nous n'avons réalisé cette comparaison *post-hoc* que lors de l'étude de l'histoire de recolonisation du pommier sauvage européen (*i.e.,* avec *SimCoal* et *ABCtoolBox*). Dans ce cas, pour la majorité des paramètres estimés les courbes de rejet et les courbes de distribution postérieures sont décalées. Ce résultat suppose des problèmes d'inférences des paramètres.

Plusieurs explications peuvent être avancées afin d'expliquer les problèmes rencontrés. En premier lieu, les processus de recolonisation sont plus complexes que les scénarios testés dans le manuscrit B. Les mauvaises estimations seraient donc directement liées au modèle lui même. Une autre explication pourrait être liée aux statistiques résumées utilisées pour choisir les simulations dont les paramètres contribueront à la construction de

la distribution posterieure des paramètres. Le manuscrit B présente une ACP sur les paramètres estimés et les statistiques résumées issues de 3000 simulations pour chaque scénario démographique (Figure 3). Cette figure montre que certains paramètres (en rouge) ne sont « expliqués » par quasiment aucune des statistiques résumées (en bleu). Dans le cas présent, les taux de migration ($m_{x\text{-}y}$) et de croissance des populations (Gr_x) semblent difficilement estimables sur la base des statistiques résumées choisies pour l'inférence. Cette approche descriptive fournit un bon diagnostic de la pertinence du choix des statistiques résumées et donc de nos capacités d'estimation.

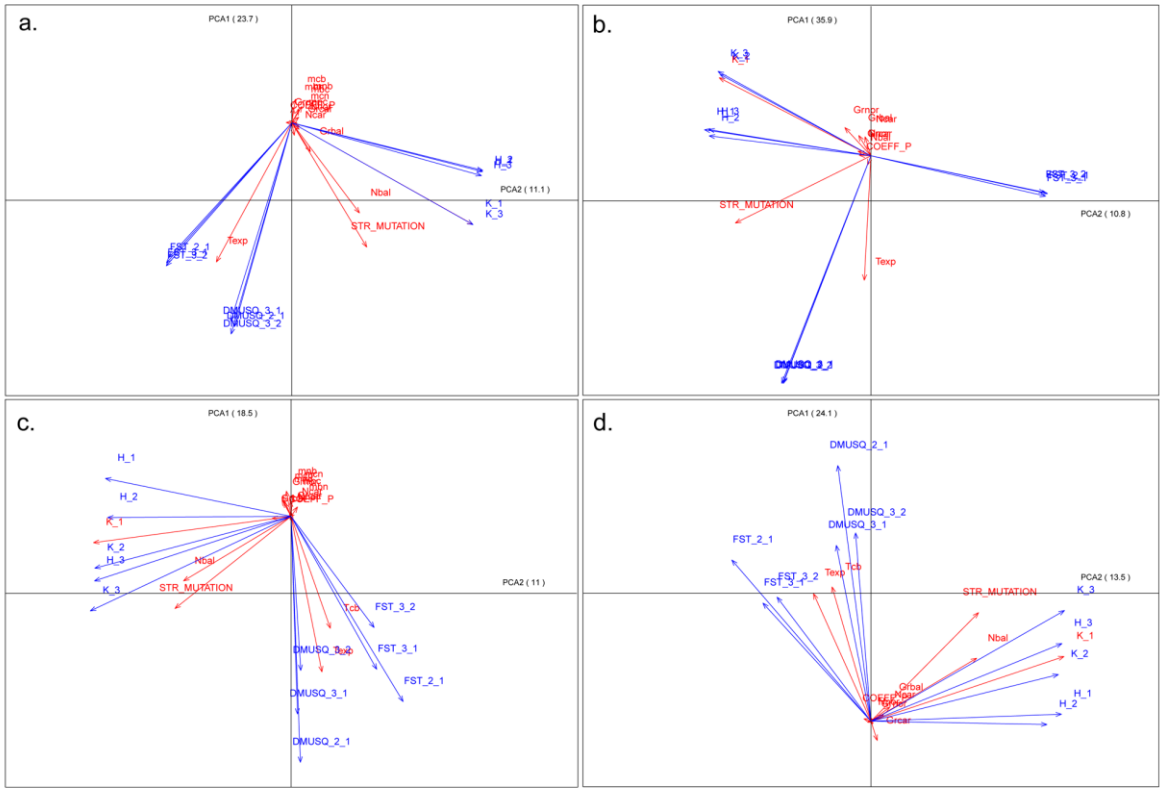


Figure 3. ACP sur 3000 simulations chez *Malus sylvestris*. Les axes 1 et 2 sont représentés pour les modèles : *a* (a), *b* (b) *c* (c) et *d* (d) (voir détails des modèles dans le manuscrit B). Le pourcentage de variance totale expliquée est indiqué sur chaque axe. En bleu, les statistiques résumées : F_{ST}*_x_y*: différenciation génétique entre les populations x et y, *K_x* : nombre d'allèle moyenné sur l'ensemble des loci dans la population x, *H_x* : hétérozygotie moyennée sur l'ensemble des loci dans la population x, *DMUSQ_x* : $\delta\mu^2$ (Goldstein *et al.*, 1995) moyenné sur l'ensemble des loci dans la population x. En rouge, les paramètres du modèle : $m_{x\text{-}y}$: taux de migration par génération de la population x à la population y, *STR_MUTATION* : taux de mutation des microsatellites, N_x : taille efficace de la population x, Gr_x : taux de croissance de la population x, $T_{x\text{-}y}$: temps de divergence entre la population x et y, *COEFF_P* : paramètre géométrique de la distribution du modèle mutationnel.

Finalement, la dernière explication possible concernant les problèmes d'estimations de nos modèles est liée à l'utilisation de marqueurs microsatellites. En particulier, le manque de statistiques résumées disponibles pour ces marqueurs dans les logiciels « clé en main » comme *ABCtoolBox*, ainsi que le manque de connaissance de leur modèle mutationnel pourraient expliquer que les données simulées avec un bon scénario s'ajustent mal aux données observées. Alors que les taux de mutation et modèles mutationnels sont relativement bien connus pour les données de séquences, dans nos simulations nous avons mis des *priors* sur les paramètres du taux de mutation ainsi qu'un *hyperprior* (*i.e., prior* sur les *priors*) sur la forme de la distribution gamma. Ces *priors* et *hyperpriors* peuvent provoquer des bruits dans l'estimation des paramètres. A ma connaissance aucune étude n'a comparé la puissance d'inférence des paramètres démographiques et génétiques à partir de marqueurs microsatellites à celle obtenue à partir de séquences. Cette problématique fera l'objet d'une collaboration avec Aurélien Tellier (Université Munich, Allemagne) lors de mon année de post-doctorat financée par le projet PICRI de la région Ile-de-France. Nous utiliseront pour cela les jeux de données microsatellites et des séquences nucléaires chez les cinq espèces de pommier du genre *Malus*.

5. Co-divergence hôte-pathogène lors des processus de domestication

Comprendre les mécanismes d'émergence de pathogènes chez les plantes domestiquées et leur dispersion dans les agrosystèmes est un enjeu essentiel si l'on veut mieux les contrôler dans le futur (Anderson, May, 1982). *Venturia inaequalis* est un champignon pathogène responsable d'une maladie appelée la tavelure, affectant le pommier cultivé et causant de sérieuses baisses de qualité dans la production de pommes (Gladieux *et al.*, 2008) (Figure 4). Des résultats antérieurs ont retracé l'histoire évolutive de ce pathogène depuis le centre d'origine du pommier cultivé (*M. domestica*) en Asie Centrale jusqu'en Europe, montrant que ce pathogène était originaire d'Asie Centrale (Gladieux *et al.*, 2008) et pointant *M. sieversii* comme son hôte d'origine (Gladieux *et al.*, 2010b). Ces résultats sont conformes avec l'hypothèse d'un phénomène d'*host-tracking*[5] dans lequel *V.*

[5] Ce terme se réfère à un des mécanismes à l'origine des agents pathogènes dans les agrosystèmes : la co-divergence hôte/pathogène. Cette co-divergence réfère à l'évolution d'un agent pathogène lors

inaequalis se serait dispersé jusqu'en Europe en suivant son hôte *M. domestica* lors de sa domestication. Il aurait ensuite colonisé l'Europe par expansion démographique, en s'établissant sur *M. sylvestris*, hôte qui ne présentait jusqu'alors pas de traces de ce pathogène. D'autre part, la domestication du pommier a été associée à une différenciation génétique chez le pathogène (Gladieux *et al.*, 2010b).

Figure 4. Tavelure de pommier (*Venturia inaequalis*) sur des feuilles de *Malus sylvestris* (Photo : Bruno LeCam).

De la même manière qu'on observe des syndromes de domestication chez les plantes cultivées, des changements de traits d'histoire de vie des pathogènes s'adaptant à leur nouvel hôte domestiqué et aux agrosystèmes associés peuvent apparaitre. Une étude à laquelle j'ai collaboré (Manuscrit E, Annexe 2, Le Van *et al* 2011) a ainsi testé par inoculations croisées les changements de pathogénicité de *V. inaequalis* associés à la domestication du pommier cultivé et à sa dispersion jusqu'en Europe. Les trois espèces hôtes *M. domestica*, *M. sylvestris* et *M. sieversii* ne sont pas faciles à discriminer sur la base de caractères morphologiques, j'ai donc génotypé les arbres inoculés pour vérifier leur statut

---

du processus de domestication de son hôte (*host-tracking*). Dans ce modèle, l'agent pathogène et l'hôte partageraient le même centre d'origine.

taxonomique. En résumé, les résultats montrent une évolution de la pathogénicité des populations de *V. inaequalis* lors de l'adaptation à son hôte domestiqué à travers l'Eurasie.

Les études sur les processus de domestication des micro-organismes commencent à émerger (Libkind *et al.*, 2011; Liti *et al.*, 2009), en particulier l'étude des processus de « domestication » (dans le sens « changement évolutif dû à l'homme, mais par forcément intentionnellement ») des pathogènes associés à la domestication de leur hôte (Stukenbrock *et al.*, 2007; Zaffarano *et al.*, 2008). Les résultats sur *V. inaequalis* (Gladieux *et al.*, 2008; Gladieux *et al.*, 2010b; Lê Van *et al.*, 2012) montrent l'impact de la domestication d'un arbre cultivé sur l'évolution de son pathogène à la fois en termes de différenciation génétique et d'évolution de traits d'histoire de vie. L'interaction *V. inaequalis* - *M. domestica* représente ainsi un modèle idéal afin d'étudier la coévolution hôte-pathogène lors des processus de domestication. Le développement des approches de génomique rapides et peu coûteuses, permettront d'obtenir les génomes de pommiers cultivés, d'Asie Centrale et d'Europe, ainsi que les génomes des souches de *V. inaequalis* associées à ces individus hôtes afin de tester l'existence de balayage sélectifs liés à l'adaptation à l'hôte domestiqué. En particulier, il serait intéréssant d'étudier les gènes impliqués dans la pathogénicité, et en parallèle d'établir la co-structuration hôte-pathogène liée à *l'host-tracking*.

Références bibliographiques

Références bibliographiques

Abbott RJ (1992) Plant invasions, interspecific hybridization and the evolution of new plant taxa. *Trends in Ecology & Evolution* **7**, 401-405.

Allaby R, Peterson G, Merriwether D, Fu Y-B (2005) Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the *sad2* locus. *TAG Theoretical and Applied Genetics* **112**, 58-65.

Allaby RG, Fuller DQ, Brown TA (2008) The genetic expectations of a protracted model for the origins of domesticated crops. *PNAS* **105**, 13982-13986.

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43**, 1223-1232.

Alves I, Šrámková Hanulová A, Foll M, Excoffier L (2012) Genomic data reveal a complex making of humans. *PLoS Genet* **8**, e1002837.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217-1229.

Anderson LL, Hu FS, Nelson DM, Petit RJ, Paige KN (2006) Ice-age endurance: DNA evidence of a white spruce refugium in Alaska. *Proceedings of the National Academy of Sciences* **103**, 12447-12450.

Anderson RM, May aRM (1982) Coevolution of hosts and parasites. *Parasitology* **85** 411-426

Anthony FA, Combes MC, Astorga CA*, et al.* (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *TAG Theoretical and Applied Genetics* **104**, 894-900.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33**, 1677-1688.

Araújo MB, New M (2007) Ensemble forecasting of species distributions. *Trends in Ecology &; Evolution* **22**, 42-47.

Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 1565-1571.

Arnold ML (2004) Natural hybridization and the evolution of domesticated, pest and disease organisms. *Molecular Ecology* **13**, 997-1007.

Arrigo N, Guadagnuolo R, Lappe S*, et al.* (2011) Gene flow between wheat and wild relatives: empirical evidence from *Aegilops geniculata*, *Ae. neglecta* and *Ae. triuncialis*. *Evolutionary Applications* **4**, 685-695.

Atkinson RG, Cipriani G, Whittaker DJ, Gardner RC (1997) The allopolyploid origin of kiwifruit, *Actinidia deliciosa* (*Actinidiaceae*). *Plant Systematics and Evolution* **205**, 111-124.

Austerlitz F, Mariette S, Machon N, Gouyon PH, Godelle B (2000) Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics* **154**, 1309-1321.

Avise JC (2000) *Phylogeography: the history and formation of species* Harvard University Press, Cambridge, Massachusetts.

Avise JC (2009) Phylogeography: retrospect and prospect. *Journal of Biogeography* **36**, 3-15.

Baack EJ, Rieseberg LH (2007) A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics & Development* **17**, 513-518.

Badr A, M K, Sch R*, et al.* (2000) On the origin and domestication history of barley (*Hordeum vulgare*). *Molecular Biology and Evolution* **17**, 499-510.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729-744.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* **19**, 2609-2625.

Besnard G, Hernandez P, Khadari B, Dorado G, Savolainen V (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biology* **11**, 80.

Besnard G, Rubio de Casas R, Vargas P (2007) Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). *J Biogeogr* **34**, 736-752.

Blackman BK, Scascitelli M, Kane NC*, et al.* (2011) Sunflower domestication alleles support single domestication center in eastern North America. *PNAS* **108**, 14360-14365.

Boré JM, Fleckinger J (1997) *Pommiers à cidre: variétés de France* INRA éditions, Paris, FRANCE (1997) (Monographie).

Bourguiba H, Audergon J-M, Krichen L*, et al.* (2012) Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. *BMC Plant Biology* **12**, 49.

Brewer S, Cheddadi R, de Beaulieu JL, Reille M (2002) The spread of deciduous *Quercus* throughout Europe since the last glacial period. *Forest Ecology and Management* **156**, 27-48.

Brown JL, Knowles LL (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of Pleistocene glaciation and recent climate change on the American Pika. *Molecular Ecology* **21**, 3757-3775.

Brown TA, Jones MK, Powell W, Allaby RG (2009) The complex origins of domesticated crops in the Fertile Crescent. *Trends Ecol. Evol.* **24**, 103-109.

Bruford MW, Bradley DG, Luikart G (2003) DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet* **4**, 900-910.

Caicedo AL, Williamson SH, Hernandez RD*, et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3**, 1745-1756.

Carling MD, Lovette IJ, Brumfield RT (2010) Historical divergence and gene flow: coalescent analyses of mitochondrial, autosomal and sex-linked loci in *Passerina* Buntings *Evolution* **64**, 1762-1772.

Carstens BC, Richards CL (2007) Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* **61**, 1439-1454.

Caswell JL, Mallick S, Richter DJ*, et al.* (2008) Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet* **4**, e1000057.

Cheddadi R, Vendramin GG, Litt T*, et al.* (2006) Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. *Global Ecology and Biogeography* **15**, 271-282.

Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**, 747-756.

Chen H, Morrell PL, Ashworth VETM, de la Cruz M, Clegg MT (2009a) Tracing the geographic origins of major avocado cultivars. *J. Hered.* **100**, 56-65.

Chen J, Kallman T, Gyllenstrand N, Lascoux M (2009b) New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity* **104**, 3-14.

Chessa B, Pereira F, Arnaud F*, et al.* (2009) Revealing the history of sheep domestication using retrovirus integrations. *Science* **324**, 532-536.

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427.

Clement C, De Cristo-Araùjo M, D'Eeckenbrugge GC, Alves Pereira A, Picanço-Rodrigues D (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72-106.

Clotault J, Thuillet A-C, Buiron M*, et al.* (2011) Evolutionary history of pearl millet (*Pennisetum glaucum* [L.] R. Br.) and selection on flowering genes since its domestication. *Molecular Biology and Evolution*, First published online.

Coart E, Van Glabeke S, De Loose M, Larsen AS, Roldán-Ruiz I (2006) Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus*

sylvestris (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Molecular Ecology* **15**, 2171-2182.

Coart E, Vekemans X, Smulders MJM*, et al.* (2003) Genetic variation in the endangered wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment length polymorphism and microsatellite markers. *Molecular Ecology* **12**, 845-857.

Collins WD, Bitz CM, Blackmon ML*, et al.* (2006) The Community Climate System Model Version 3 (CCSM3). *Journal of Climate* **19**, 2122-2143.

Cornille A, Giraud T, Bellard C*, et al.* (2012a) Post-glacial recolonization history of the European crabapple (Malus sylvestris Mill.), a wild contributor to the domesticated apple. *Molecular Ecology* **in revision**.

Cornille A, Gladieux P, Smulders MJM*, et al.* (2012b) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* **8**, e1002703.

Cornuet J-M, Ravigne V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics* **11**, 401.

Cornuet J-M, Santos F, Beaumont MA*, et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**, 2713-2719.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001-2014.

Coyne JA, Orr HA (2004) *Speciation* Sunderland (MA): Sinauer Associates.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410-418.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. *Science* **292**, 673-679.

De Andrés MT, Benito A, Pérez-Rivera G*, et al.* (2012) Genetic diversity of wild grapevine populations in Spain and their genetic relationships with cultivated grapevines. *Molecular Ecology* **21**, 800-816.

Delplancke M, Alvarez N, Espíndola A*, et al.* (2011) Gene flow among wild and domesticated almond species: insights from chloroplast and nuclear markers. *Evolutionary Applications* **5**, 317-329.

Diamond J (1997) *Guns, Germs, and Steel: The Fates of Human Societies* Norton, W. W. & Company, Inc. Sales.

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700-707.

Diego M, Michela T, Francesco S*, et al.* (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1044-1045.

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* **127**, 1309-1321.

Dzhangaliev AD (2003) The wild apple tree of Kazakhstan. In: *Hortic Rev* pp. 63-303. John Wiley & Sons, Inc.

Ellison CE, Hall C, Kowbel D*, et al.* (2011) Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences*.

Ellstrand NC (1992) Gene flow by pollen: implications for plant conservation genetics. *Oikos* **63**, 77-86.

Ellstrand NC (2003) Current knowledge of gene flow in plants: implications for transgene flow. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **358**, 1163-1170.

Ellstrand NC (2005) *Dangerous Liaisons?: When Cultivated Plants Mate with Their Wild Relatives* Johns Hopkins University Press.

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **Vol. 30**, 539-563.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11**, 1591-1604.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

Excoffier L, Foll M, Petit RJ (2009) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40**, 481-501.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

Fagundes NJR, Ray N, Beaumont M*, et al.* (2007) Statistical evaluation of alternative models of human evolution. *PNAS* **104**, 17614-17619.

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.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics* **28**, 342-350.

Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. *Trends Genet* **24**, 24-32.

Fieldings AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**, 38-49

Fitzpatrick BM, Fordyce JA, Gavrilets S (2009) Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology* **22**, 2342-2347.

Forsline PL, Aldwinckle HS, Dickson EE, Luby JJ, Hokanson SC (2002) Collection, maintenance, characterization and utilization of wild apples of Central Asia. In: *Hortic Rev* pp. 1-61. John Wiley & Sons, Inc.

François O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet* **4**, e1000075.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Fuller DQ, Willcox G, Allaby RG (2012) Early agricultural pathways: moving outside the 'core area' hypothesis in Southwest Asia. *Journal of Experimental Botany* **63**, 617-633.

Gao L-z, Innan H (2008) Nonindependent domestication of the two rice subspecies, *Oryza sativa ssp. indica* and *ssp. japonica*, demonstrated by multilocus microsatellites. *Genetics* **179**, 965-976.

Gardiner SE, Bus VGM, Rusholme RL*, et al.* (2007) Fruits and Nuts: Apple (ed. Kole C), pp. 1-62. Springer Berlin Heidelberg.

Gepts P, Papa R (2003) Possible effects of (trans)gene flow from crops on the genetic diversity from landraces and wild relatives. *Environmental Biosafety Research* **2**, 89-103.

Geraldes A, Basset P, Gibson B*, et al.* (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology* **17**, 5349-5363.

Gharghani A, Zamani Z, Talaie A*, et al.* (2009) Genetic identity and relationships of Iranian apple (*Malus domestica* Borkh.) cultivars and landraces, wild *Malus* species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genet Resour Crop Ev* **56**, 829-842.

Gianfranceschi L, Seglias N, Tarchini R, Komjanc M, Gessler C (1998) Simple sequence repeats for the genetic analysis of apple. *Theoritical Applied Genetics* **96**, 1069-1076.

Giovannoni J (2010) Harvesting the apple genome. *Nat Genet* **42**, 822-823.

Gladieux P, Vercken E, Fontaine MC*, et al.* (2010a) Maintenance of fungal pathogen species that are specialized to different hosts: allopatric divergence and introgression through secondary contact. *Molecular Biology and Evolution* **28**, 459-471.

Gladieux P, Zhang X-G, Afoufa-Bastien D*, et al.* (2008) On the origin and spread of the scab disease of apple: out of Central Asia. *PLoS ONE* **3**, e1455.

Gladieux P, Zhang XG, Roldàn-Ruiz I*, et al.* (2010b) Evolution of the population structure of *Venturia inaequalis*, the apple scab fungus, associated with the domestication of its host. *Molecular Ecology* **19**, 658-674.

Glémin S, Bataillon T (2009) A comparative view of the evolution of grasses under domestication. *New Phytol.* **183**, 273-290.

Goedbloed DJ, Megens HJ, Van Hooft P*, et al.* (2012) Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. *Molecular Ecology*, in press.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences* **92**, 6723-6727.

Gross BL, Olsen KM (2009) Genetic perspectives on crop domestication. *Trends Plant Sci.* **15**, 529-537.

Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* **156**, 1-13.

Hamblin MT, Buckler ES, Jannink J-L (2011) Population genetics of genomics-based crop improvement methods. *Trends in genetics : TIG* **27**, 98-106.

Hampe A, Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters* **8**, 461–467.

Hardy OJ, Maggia L, Bandou E*, et al.* (2006) Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. *Molecular Ecology* **15**, 559-571.

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620.

Harr B (2006) Genomic islands of differentiation between house mouse subspecies. *Genome Research* **16**, 730-737.

Harris SA, Robinson JP, Juniper BE (2002) Genetic clues to the origin of the apple. *Trends Genet* **18**, 426-430.

Harrison N, Harrison RJ (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1043-1044.

Harter AV, Gardner KA, Falush D*, et al.* (2004) Origin of extant domesticated sunflowers in eastern North America. *Nature* **430**, 201-205.

Hartman Y, Hooftman DAP, Uwimana B*, et al.* (2012) Genomic regions in crop–wild hybrids of lettuce are affected differently in different environments: implications for crop breeding. *Evolutionary Applications*, in press.

Hasumi H, Emori S (2004) K-1 coupled GCM (MIROC) description (ed. Tokyo: Center for Climate System Research UoT).

Haudry A, Cenci A, Ravel C*, et al.* (2007) Grinding up wheat: A massive loss of nucleotide diversity since domestication. *Molecular Biology and Evolution* **24**, 1506-1517.

Heuertz M, Carnevale S, Fineschi S*, et al.* (2006) Chloroplast DNA phylogeography of European ashes, *Fraxinus* sp. (Oleaceae): roles of hybridization and life history traits. *Molecular Ecology* **15**, 2131-2140.

Heuertz M, Hausman J-F, Hardy OJ*, et al.* (2004) Nuclear microsatellites reveal contrasting patterns of genetic structure between Western and Southeastern European populations of the common ash (*Fraxinus excelsior* L.) *Evolution* **58**, 976-988.

Heun M, Abbo S, Lev-Yadun S, Gopher A (2012) A critical review of the protracted domestication model for Near-Eastern founder crops: linear regression, long-distance gene flow, archaeological, and archaeobotanical evidence. *Journal of Experimental Botany* **63**, 4333-4341.

Heun M, Schäfer-Pregl R, Klawan D*, et al.* (1997) Site of einkorn wheat domestication identified by DNA fingerprinting. *Science* **278**, 1312-1314.

Hewitt GM (1990) Divergence and speciation as viewed from an insect hybrid zone. *Canadian Journal of Zoology* **68**, 1701-1715.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58**, 247-276.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-917.

Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 183-195.

Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development* **16**, 592-596.

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.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences* **104**, 2785-2790.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-1978.

Hu FS, Hampe A, Petit RJ (2008) Paleoecology meets genetics: deciphering past vegetational dynamics. *Frontiers in Ecology and the Environment* **7**, 371-379.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* **9**, 1322-1332.

Hübner S, Günther T, Flavell A*, et al.* (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Molecular Ecology* **21**, 1115-1129.

Hufford MB, Xu X, van Heerwaarden J*, et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nat Genet* **44**, 808-811.

Hyten DL, Song Q, Zhu Y*, et al.* (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences* **103**, 16666-16671.

Jackson ST, Weng C (1999) Late Quaternary extinction of a tree species in eastern North America. *Proceedings of the National Academy of Sciences* **96**, 13847-13852.

Jacques D, Vandermijnsbrugge K, Lemaire S, Antofie A, Lateur M (2009) Natural distribution and variability of the wild apple (*Malus sylvestris*) in Belgium *Belgian Journal of Botany* **142**, 39-49

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Janick J (2005) The origins of fruits, fruit growing, and fruit breeding. In: *Plant Breeding Reviews*, pp. 255-321. John Wiley & Sons, Inc.

Jay F, Manel S, Alvarez N*, et al.* (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology* **21**, 2354-2368.

Jezkova T, Olah-Hemmings V, Riddle BR (2011) Niche shifting in response to warming climate after the last glacial maximum: inference from genetic data and niche assessments in the chisel-toothed kangaroo rat (*Dipodomys microps*). *Global Change Biology* **17**, 3486-3502.

Jolivet C, Degen B (2011) Spatial genetic structure in wild cherry (*Prunus avium* L.): II. Effect of density and clonal propagation on spatial genetic structure based on simulation studies. *Tree Genetics & Genomes* **7**, 541-552.

Joy DA, Feng X, Mu J*, et al.* (2003) Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300**, 318-321.

Juniper BE, Mabberley DJ (2006) *The Story of the Apple* Imber Press, Inc.

Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* **106**, 625-632.

Kareiva P, Watts S, McDonald R, Boucher T (2007) Domesticated nature: shaping landscapes and ecosystems for human welfare. *Science* **316**, 1866-1869.

Katoh K, Kuma K-i, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511-518.

Khoshbakht K, Hammer K (2006) Savadkouh (Iran) – an evolutionary centre for fruit trees and shrubs. *Genetic Resources and Crop Evolution* **53**, 641-651.

Kilian B, Özkan H, Walther A*, et al.* (2007) Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during *Triticum monococcum* (Einkorn) domestication: implications for the origin of agriculture. *Mol. Biol. Evol.* **24**, 2657-2668.

King AR, Feris C (1998) Chloroplast DNA phylogeography of *Alnus glutinosa* (L.) Gaertn. *Molecular Ecology* **7**, 1151-1161.

Kisel Y, Barraclough TG (2010) Speciation has a spatial scale that depends on levels of gene flow

The American Naturalist **175,** , 316-334.

Kliman RM, Andolfatto P, Coyne JA*, et al.* (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**, 1913-1931.

Koopman WJM, Li Y, Coart E*, et al.* (2007) Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. *Mol Ecol* **16**, 243-256.

Kovach MJ, Sweeney MT, McCouch SR (2007) New insights into the history of rice domestication. *Trends Genet.* **23**, 578-587.

Kremer A, Ronce O, Robledo-Arnuncio JJ*, et al.* (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters* **15**, 378-392.

Kronforst MR, Young LG, Kapan DD*, et al.* (2006) Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences* **103**, 6575-6580.

Krutovsky KV, Burczyk J, Chybicki I*, et al.* (2012) Gene flow, spatial structure, local adaptation, and assisted migration in trees

Genomics of Tree Crops, pp. 71-116. Springer New York.

Kulathinal RJ, Stevison LS, Noor MAF (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet* **5**, e1000550.

Larsen A, Asmussen C, Coart E, Olrik D, Kjær E (2006) Hybridization and genetic variation in Danish populations of European crab apple (*Malus sylvestris*). *Tree Genetics & Genomes* **2**, 86-97.

Larsen A, Kjær E (2009) Pollen mediated gene flow in a native population of *Malus sylvestris* and its implications for contemporary gene conservation management. *Conservation Genetics* **10**, 1637-1646.

Larson G, Karlsson EK, Perri A*, et al.* (2012) Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences* **109**, 8878-8883.

Lascoux M, Palmé AE, Cheddadi R, Latta RG (2004) Impact of Ice Ages on the genetic structure of trees and shrubs. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 197-207.

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20**, 2485-2487.

Le Normand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution* **17**, 183-189.

Lê Van A, Gladieux P, Lemaire C*, et al.* (2012) Evolution of pathogenicity traits in the apple scab fungal pathogen in response to the domestication of its host. *Evolutionary Applications*, no-no.

Lea AGH, Piggott JR (2003) *Fermented Beverage Production* Kluwer Academic/Plenum.

Lenne JM, Wood D (1991) Plant diseases and the use of wild germplasm. *Annual Review of Phytopathology* **29**, 35-63.

Leuenberger C, Wegmann D (2009) Bayesian computation and model selection without likelihoods. *Genetics* **184**, 243-252.

Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. *Genetics* **184**, 243-252.

Lexer C, Widmer A (2008) The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 3023-3036.

Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2**, e166.

Li Y, Stocks M, Hemmilä S*, et al.* (2010) Demographic histories of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution* **27**, 1001-1014.

Li Z, Zou J, Mao K*, et al.* (2011) Population genetic evidence for complex evolutionary histories of four high altitude *Juniper* species in the Qinghai-Tibetan plateau. *Evolution*, no-no.

Libkind D, Hittinger CT, Valério E*, et al.* (2011) Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences* **108**, 14539-14544.

Liebhard R, Gianfranceschi L, Koller B*, et al.* (2002) Development and characterisation of 140 new microsatellites in apple (*Malus* x *domestica* Borkh.). *Molecular Breeding* **10**, 217-241.

Liepelt S, Cheddadi R, de Beaulieu J-L*, et al.* (2009) Postglacial range expansion and its genetic imprints in *Abies alba* (Mill.) A synthesis from palaeobotanic and genetic data. *Review of Palaeobotany and Palynology* **153**, 139-149.

Liti G, Carter DM, Moses AM*, et al.* (2009) Population genomics of domestic and wild yeasts. *Nature* **458**, 337-341.

Liu A, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* **173**, 321-330.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17**, 145-151.

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P (1994) Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences* **91**, 2757-2761.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **Vol. 82**, 1420-1425

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006) Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences* **103**, 9578-9583.

Lu G, Basley DJ, Bernatchez L (2001) Contrasting patterns of mitochondrial DNA and microsatellite introgressive hybridization between lineages of lake whitefish (*Coregonus clupeaformis*); relevance for speciation. *Molecular Ecology* **10**, 965-985.

Luby JJ, Alspach PA, Bus VGM, Oraguzie NC (2001) Field resistance to fire blight in a diverse apple (*Malus* sp.) germplasm collection. *J Am Soc Hortic Sci* **127**, 245–253.

Luikart G, Gielly L, Excoffier L*, et al.* (2001) Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proceedings of the National Academy of Sciences* **98**, 5927-5932.

Lumaret R, Ouazzani N (2001) Plant genetics: Ancient wild olives in Mediterranean forests. *Nature* **413**, 700-700.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753-1766.

Mabberley DJ, Jarvis CE, Juniper BE (2001) The name of the apple. *Telopea* **9**, 2001.

Magri D, Vendramin GG, Comps B*, et al.* (2006) A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist* **171**, 199-221.

Mallet J (2007) Hybrid speciation. *Nature* **446**, 279-283.

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

Manel Sp, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**, 189-197.

Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* **15**, 59-69.

Matsuoka Y, Vigouroux Y, Goodman MM*, et al.* (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *PNAS* **99**, 6080-6084.

McKey D, Elias M, Pujol B, Duputié A (2010) The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist* **186**, 318-332.

Meirmans PG, Van Tienderen PH (2004) Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes* **4**, 792-794.

Michel AP, Sim S, Powell THQ*, et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences* **107**, 9724-9729.

Micheletti D, Troggio M, Salamini F*, et al.* (2011) On the evolutionary history of the domesticated apple. *Nat Genet* **43**, 1044-1045.

Miller A, Gross BL (2011) From forest to field: perennial fruit crops domestication. *Am. J. Bot.* **98**, 1389-1414.

Miller A, Schaal B (2005) Domestication of a Mesoamerican cultivated fruit tree, *Spondias purpurea*. *PNAS* **102**, 12801-12806.

Miller AJ, Schaal BA (2006) Domestication and the distribution of genetic variation in wild and cultivated populations of the Mesoamerican fruit tree *Spondias purpurea* L. (Anacardiaceae). *Mol. Ecol.* **15**, 1467-1480.

Molina J, Sikora M, Garud N*, et al.* (2011) Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences*.

Monserud RA, Leemans R (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling* **62**, 275-293.

Morgan J, Richards A (2003) *The New Book of Apples* Brogdale Horticultural Trust, Ebury Press.

Myles S, Boyko AR, Owens CL*, et al.* (2011) Genetic structure and domestication history of the grape. *PNAS* **108**, 3530-3535.

Nadachowska K, Babik W (2009) Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). *Molecular Biology and Evolution* **26**, 829-841.

Neafsey D, Schaffner S, Volkman S*, et al.* (2008) Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence. *Genome Biology* **9**, R171.

Neafsey DE, Barker BM, Sharpton TJ*, et al.* (2010) Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. *Genome Research* **20**, 938-946.

Neale DB (2007) Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development* **17**, 539-544.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* **12**, 111-122.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583-590.

Network TMCS (2011) What do we need to know about speciation? *Trends in ecology & evolution (Personal edition)* **27**, 27-39.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: A Markov Chain Monte Carlo Approach. *Genetics* **158**, 885-896.

Niemiller ML, Fitzpatrick BM, Miller BT (2008) Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: *Gyrinophilus*) inferred from gene genealogies. *Molecular Ecology* **17**, 2258-2275.

Nogués-Bravo D, Rodríguez J, Hortal J, Batra P, Araújo MB (2008) Climate Change, Humans, and the Extinction of the Woolly Mammoth. *PLoS Biology* **6**, e79.

Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* **103**, 439-444.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology* **17**, 2103-2106.

Oard J, Cohn MA, Linscombe S, Gealy D, Gravois K (2000) Field evaluation of seed production, shattering, and dormancy in hybrid populations of transgenic rice (*Oryza sativa*) and the weed, red rice (*Oryza sativa*). *Plant Science* **157**, 13-22.

Oddou-Muratorio S, Demesure-Musch B, Pelissier R, Gouyon PH (2004) Impacts of gene flow and logging history on the local genetic structure of a scattered tree species, *Sorbus torminalis* L. Crantz. *Molecular Ecology* **13**, 3689-3702.

Oddou-Muratorio S, Klein EK (2008) Comparing direct vs. indirect estimates of gene flow within a population of a scattered tree species. *Molecular Ecology* **17**, 2743-2754.

Olsen KM, Gross BL (2008) Detecting multiple origins of domesticated crops. *PNAS* **105**, 13701-13702.

Olsen KM, Schaal BA (2001) Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *American Journal of Botany* **88**, 131-142.

Orton V (1973) *The American cider Book: The story of America's natural beverage* Farrar, Straus and Giroux edn. North Point Press.

Oumar I, Mariac C, Pham J-L, Vigouroux Y (2008) Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theor. Appl. Genet.* **117**, 489-497.

Özkan H, Brandolini A, Schäfer-Pregl R, Salamini F (2002) AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in Southeast Turkey. *Molecular Biology and Evolution* **19**, 1797-1801.

Palmé AE, Vendramin GG (2002) Chloroplast DNA variation, postglacial recolonization and hybridization in hazel, *Corylus avellana*. *Molecular Ecology* **11**, 1769-1779.

Papa R (2005) Gene flow and introgression between domesticated crops and their wild relatives. In: *Proceedings of the International Workshop on the Role of Biotechnology for the Characterisation and Conservation of Crop, Forestry, Animal and Fishery Genetic Resources*, Turin, Italy.

Parducci L, Jørgensen T, Tollefsrud MM*, et al.* (2012) Glacial survival of boreal trees in Northern Scandinavia. *Science* **335**, 1083-1086.

Pasquet RmS, Peltier A, Hufford MB*, et al.* (2008) Long-distance pollen flow assessment through evaluation of pollinator foraging range suggests transgene escape distances. *Proceedings of the National Academy of Sciences* **105**, 13456-13461.

Patin E, Laval G, Barreiro LB*, et al.* (2009) Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* **5**, e1000448.

Patocchi A, Fernàndez-Fernàndez F, Evans K*, et al.* (2009) Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. *Tree Genetics Genomes* **5**, 211-223.

Pautasso M (2009) Geographical genetics and the conservation of forest trees. *Perspectives in Plant Ecology, Evolution and Systematics* **11**, 157-189.

Pauwels M, Vekemans X, Godé C*, et al.* (2012) Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, *Arabidopsis halleri* (Brassicaceae). *New Phytologist* **193**, 916-928.

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* **12**, 361-371.

Pearson RG, Thuiller W, Araújo MB*, et al.* (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography* **33**, 1704-1711.

Pereira-Lorenzo S, Ramos-Cabrer AM, Fischer M (2009) Breeding Apple (*Malus x Domestica* Borkh). Breeding Plantation Tree Crops: Temperate Species, pp. 33-81. Springer New York.

Perrier X, De Langhe E, Donohue M*, et al.* (2011) Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proceedings of the National Academy of Sciences* **108**, 11311-11318.

Petit RJ, Aguinagalde I, de Beaulieu J-L*, et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**, 1563-1565.

Petit RJ, Bialozyt R, Pauline G-G, Hampe A (2004a) 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 (2004b) Hybridization as a mechanism of invasion in oaks. *New Phytologist* **161**, 151-164.

Petit RJ, Csaikl UM, Bordás Sn*, et al.* (2002) Chloroplast DNA variation in European white oaks: Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management* **156**, 5-26.

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends in ecology & evolution (Personal edition)* **24**, 386-393.

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37**, 187-214.

Petit RJ, Hu FS, Dick CW (2008) Forests of the past: a window to future changes. *Science* **320**, 1450-1452.

Pickersgill B (2007) Domestication of plants in the Americas: insights from mendelian and molecular genetics. *Ann Bot* **100**, 925-940.

Pinho C, Hey J (2010) Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics* **41**, 215-230.

Piry S, Luikart G, Cornuet JM (1999) Computer note. BOTTLENECK: a computer program for detecting recent reductions in the effective size using allele frequency data. *Journal of Heredity* **90**, 502-503.

Ponomarenko V (1991) On a little known species *Malus* x *asiatica* (Rosaceae). *Bot Zhurn* **76**, 715-720.

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

Provan J, Bennett KD (2008) Phylogeographic insights into cryptic glacial refugia. *Trends in ecology & evolution* **23**, 564-571.

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. *Nature* **457**, 843-848.

R (2011) Development Core Team.

Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics* **3**, 29-43.

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

Rehder A (1940) *Manual of cultivated trees and shrubs*, 2 edn. Macmillan, New York.

Renaut S, Maillet N, Normandeau E*, et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 354-363.

Richards C, Volk G, Reilley A*, et al.* (2009) Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. *Tree Genetics & Genomes* **5**, 339-347.

Richards CL, Carstens BC, Lacey Knowles L (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34**, 1833-1845.

Rieseberg LH, Baird SJE, Gardner KA (2000) Hybridization, introgression, and linkage evolution. *Plant Molecular Biology* **42**, 205-224.

Robinson JP, Harris SA, Juniper BE (2001) Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. *Plant Syst. Evol.* **226**, 35-58.

Ross-Ibarra J, Gaut BS (2008) Multiple domestications do not appear monophyletic. *Proceedings of the National Academy of Sciences* **105**, E105.

Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the Genus *Zea*. *Genetics* **181**, 1399-1413.

Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Row JR, Brooks RJ, MacKinnon CA*, et al.* (2011) Approximate Bayesian computation reveals the factors that influence genetic diversity and population structure of foxsnakes. *Journal of Evolutionary Biology* **24**, 2364-2377.

Rozas J, SÃ¡nchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.

Rubin C-J, Zody MC, Eriksson J*, et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-591.

Russell J, Dawson IK, Flavell AJ*, et al.* (2011) Analysis of >1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol.* **191**, 564-578.

Sagnard F, Deu M, Dembélé D*, et al.* (2011) Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region. *TAG Theoretical and Applied Genetics* **123**, 1231-1246.

Salazar C, Jiggins C, Taylor J, Kronforst M, Linares M (2008) Gene flow and the genealogical history of *Heliconius heurippa*. *BMC Evolutionary Biology* **8**, 132.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. *Current Opinion Plant Biology* **10**, 162-167.

Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, and Systematics* **38**, 595-619.

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: problems and prospects. *Molecular Ecology* **7**, 465-474.

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology* **4**, 11.

Schuster M, Büttner R (1995) Chromosome numbers in the *Malus* wild species collection of the genebank Dresden-Pillnitz. *Genet Resour Crop Ev* **42**, 353-361.

Schwacke L, Schwacke J, Rosel P (2005) RE-RAT: relatedness estimation and rarefaction analysis tool. http://people.musc.edu/~schwaclh/.

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography* **31**, 1555-1568.

Silfverberg-Dilworth E, Matasci C, Van de Weg W*, et al.* (2006) Microsatellite markers spanning the apple (*Malus* x *domestica* Borkh.) genome. *Tree Genetics Genomes* **2**, 202-224.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology* **20**, 5123-5140.

Spooner DM, McLean K, Ramsay G, Waugh R, Bryan GJ (2005) A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14694-14699.

St. Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE (2011) Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Molecular Ecology* **20**, 3306-3320.

Städler T, Arunyawat U, Stephan W (2008) Population genetics of speciation in two cosely related wild tomatoes (Solanum Section Lycopersicon). *Genetics* **178**, 339-350.

Städler T, Roselius K, Stephan W (2005) Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow *Evolution* **59**, 1268-1279.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* **68**, 978-989.

Stewart JR, Lister AM, Barnes I, Dalén L (2009) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**, 661-671.

Stewart JR, Lister AM, Barnes I, Dalén L (2010) Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**, 661-671.

Strauss (2011) Transgenic biotechnology in forestry: what a long strange trip it's been. *BMC Proceedings* **5**.

Stukenbrock EH, Banke Sr, Javan-Nikkhah M, McDonald BA (2007) Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Molecular Biology and Evolution* **24**, 398-411.

Svenning J-C, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand S (2011) Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews* **30**, 2930-2947.

Svenning J-C, Normand S, Kageyama M (2008) Glacial refugia of temperate trees in Europe: insights from species distribution modelling. *Journal of Ecology* **96**, 1117-1127.

Svenning J-C, Skov F (2007) Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology Letters* **10**, 453-460.

Szpiech ZA, Jakobsson M, Rosenberg NA (2008) ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* **24**, 2498-2504.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations *Genetics* **105**, 437-460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Tanno K-i, Willcox G (2006) How Fast Was Wild Wheat Domesticated? *Science* **311**, 1886.

Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences* **108**, 17052-17057.

Tenaillon MI, Manicacci D (2011) Maize origins: an old question under the spotlights. In: *Advances in Maize (Essential Reviews in Experimental Biology)* (eds. Prioul J-L, Thévenot C, Molnar T), pp. 89-110. The Society for Experimental Biology.

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214-1225.

Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**, 1607-1619.

Thuiller W, Brotons L, Araújo MB, Lavorel S (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165-172.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369-373.

Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* **13**, 397-405.

Turner TL, Hahn MW (2007) Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae*. *Molecular Biology and Evolution* **24**, 2132-2138.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol* **3**, e285.

Tyson RC, Wilson JB, Lane WD (2011) A mechanistic model to predict transgenic seed contamination in bee-pollinated crops validated in an apple orchard. *Ecological Modelling* **222**, 2084-2092.

Uwimana B, D'Andrea L, Felber F*, et al.* (2012) A Bayesian analysis of gene flow from crops to their wild relatives: cultivated (*Lactuca sativa* L.) and prickly lettuce (*L. serriola* L.) and the recent expansion of *L. serriola* in Europe. *Molecular Ecology* **21**, 2640-2654.

van Heerwaarden J, Doebley J, Briggs WH*, et al.* (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *PNAS* **108**, 1088-1092.

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.

Vavilov NI (1926) Studies on the origin of cultivated plants. *Trudy Byuro. Prikl. Bot.* **16**, 139–245.

Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921-935.

Velasco R, Zharkikh A, Affourtit J*, et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics* **42**, 833-839.

Vercken E, Fontaine MC, Gladieux P*, et al.* (2010) Glacial refugia in pathogens: European genetic structure of anther smut pathogens on *Silene latifolia* and *Silene dioica*. *PLoS Pathogens* **6**, e1001229.

Vilà C, Leonard JA, Götherström A*, et al.* (2001) Widespread origins of domestic horse lineages. *Science* **291**, 474-477.

Volk GM, Richards CM, Reilley AA*, et al.* (2008) Genetic diversity and disease resistance of wild *Malus orientalis* from Turkey and Southern Russia. *Journal of the American Society for Horticultural Science* **133**, 383-389.

Wachowiak W, PalmÉ AE, Savolainen O (2011) Speciation history of three closely related pines *Pinus mugo* (T.), *P. uliginosa* (N.) and *P. sylvestris* (L.). *Molecular Ecology* **20**, 1729-1743.

Wagner I, Weeden NF (2000) Isozyme in *Malus sylvestris*, *Malus* x *domestica* and in related *Malus* species. *Acta Horticulturae* **538**, 51-56.

Wakeley J (2008) *Coalescent Theory: An Introduction.* Roberts & Company Publishers, Greenwood Village, Colorado.

Waltari E, Hijmans RJ, Peterson AT*, et al.* (2007) Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**, e563.

Wang C, Chen J, Zhi H*, et al.* (2010) Population genetics of foxtail millet and its wild ancestor. *BMC Genet.* **11**, 90.

Wang Y, Hey J (2011) Estimating divergence parameters with small samples from a large number of loci. *Genetics* **184**, 363-379.

Warmuth V, Eriksson A, Bower MA*, et al.* (2012) Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proceedings of the National Academy of Sciences*.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256-276.

Wegmann D, Excoffier L (2010) Bayesian Inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* **27**, 1425-1435.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Weiss E, Kislev ME, Hartmann A (2006) Autonomous cultivation before domestication. *Science* **312**, 1608-1610.

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. In: *Advances in Agronomy*, pp. 139-186. Academic Press.

Westman AL, Medel S, Spira TP*, et al.* (2004) Molecular genetic assessment of the potential for gene escape in strawberry, a model perennial study crop. In: *Introgression from genetically modified plants into wild relatives*, pp. 75-88. H. C. M.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology* **19**, 925-939.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* **22**, 506-519.

Zaffarano PL, McDonald BA, Linde CC (2008) Rapid speciation following recent host shift in the plant pathogenic fungus *Rhynchosporium*. *Evolution* **62**, 1418-1436.

Zeder MA, Emshwiller E, Smith BD, Bradley DG (2006) Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* **22**, 139-155.

Zhang C, Chen X, He T*, et al.* (2007) Genetic structure of *Malus sieversii* population from Xinjiang, China, revealed by SSR markers. *Journal of Genetics and Genomics* **34**, 947-955.

Zheng X-M, Ge S (2010) Ecological divergence in the presence of gene flow in two closely related Oryza species (*Oryza rufipogon* and *O. nivara*). *Molecular Ecology* **19**, 2439-2454.

Zhou R, Zeng K, Wu W*, et al.* (2007) Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in Mangroves. *Molecular Biology and Evolution* **24**, 2746-2754.

Zohary D (2004) Unconscious selection and the evolution of domesticated plants. *Econ Bot* **58**, 5-10.

Zohary D, Hopf M (2000) *Domestication of plants in the Old World*, 3 edn. New York: Oxford University Press.

Zohary D, Spiegel-Roy P (1975) Beginnings of Fruit Growing in the Old World. *Science*, 319-327.

Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561-577.

Annexes

Annexe 1: ***Conserver et utiliser les ressources génétiques du pommier sauvage européen (Malus sylvestris)***

Cornille Amandine, Giraud Tatiana, Collin Eric
(2012) Forêt-entreprise, 205: 40-41.

Conserver et utiliser les ressources génétiques du pommier sauvage

Amandine Cornille*, Tatiana Giraud* , Éric Collin**

Les hybridations entre pommiers sauvages et arboricoles sont étudiées afin de mieux préciser la provenance et la qualité des plants réintroduits en agroforesterie.

L'intensification de l'agriculture depuis la révolution verte jusqu'à nos jours s'est accompagnée d'une perte de biodiversité considérable dans les agrosystèmes. Il faut désormais revisiter nos pratiques agricoles, non seulement pour une meilleure efficacité agronomique mais aussi pour la qualité de notre environnement. L'aménagement de parcelles agroforestières favorise la protection intégrée des productions agricoles en hébergeant une faune auxiliaire abondante et variée. Une part importante des agriculteurs (30 %[1]) se montre intéressée par la mise en place d'une ou plusieurs parcelles agroforestières dans leur exploitation. Il est donc pertinent et primordial que les gestionnaires du territoire, en particulier les agriculteurs et les collectivités, puissent disposer d'outils simples, lisibles et efficaces pour un développement plus intense et mieux raisonné de la réimplantation de l'arbre champêtre. Cependant, ces mesures, si elles se veulent raisonnées et durables sur le long terme, dépendent directement de la connaissance de l'identité génétique des arbres plantés, ainsi que de leur diversité génétique locale, mais aussi de la diversité génétique existant à plus grande échelle.

© Duchemin

© Laboratoire Ecologie Systématique et Évolution

Pommier sauvage : Malus sylvestris, *Ardèche, 2008.*

Le pommier sauvage, une espèce rare qu'on apprend à connaître

Les agriculteurs et les forestiers désireux de réintroduire des pommiers sauvages dans le paysage bocager ou forestier sont confrontés à un double problème : comment se procurer des plants, comment être sûr qu'il s'agit bien de pommiers sauvages et non de résultats de croisements avec des variétés fruitières ?

© Laboratoire Écologie Systématique et Évolution

Pommes de Malus sylvestris, *Orléans 2009.*

Les instances en charge de la conservation des ressources génétiques des arbres forestiers en France (CRGF[2]) et en Europe (EUFORGEN[3]) s'inquiètent des conséquences de plantations forestières réalisées avec du matériel de provenance inconnue ou à base génétique trop étroite (ex : graines récoltées sur un seul arbre). Une thèse[4] en cours d'achèvement et un projet (PICRI[5]) qui la prolonge vont heureusement permettre d'apporter des réponses à ces questions et des bases pour combattre cette cause d'inquiétude.

Sur le plan fondamental, cette thèse met en lumière les mécanismes génétiques qui ont conduit à la différenciation des espèces de pommiers en Asie et en Europe ces 60 000 dernières années et qui ont accompagné la domestication du pommier par l'homme il y a moins de 10 000 ans. De manière plus appliquée, elle fournit des indications pour la conservation et l'utilisation des ressources génétiques des pommiers sauvages de France.

Une recommandation serait d'éviter les transferts de graines et plants entre l'est et l'ouest du pays, dont les populations actuelles de pommier semblent génétiquement différenciées, car issues de refuges glaciaires géographiquement isolés. De plus, les travaux de thèse ont montré que 50 % de pommiers sauvages échantillonnés à travers l'Europe (800 individus), identifiés sur la base de critères morphologiques, se sont révélés être des hybrides avec le pommier cultivé. Les analyses de l'ADN conduites sur 1 200 échantillons de pommiers forestiers et fruitiers de différents pays d'Europe ou d'Asie ont également permis de sélectionner des marqueurs moléculaires utilisables en routine à moindre coût pour des études plus appliquées, comme celles qui seront mises en œuvre dans le projet qui vient de débuter.

Un vaste échantillonnage en cours

Le projet a pour but premier d'étudier les hybridations entre pommiers sauvages et cultivés en Île-de-France, mais son champ s'étend naturellement à l'ensemble du territoire national, notamment pour estimer la diversité et la spécificité éventuelle des ressources génétiques de pommiers sauvages. Grâce au vaste échantillonnage[6] engagé par des bénévoles et des organismes forestiers, Ce projet permettra au laboratoire Écologie, Systématique et Évolution de l'Université Paris-Sud et CNRS, de mesurer la diversité génétique à l'intérieur de populations des pommiers sauvages, la différenciation entre populations et les hybridations entre espèces sauvage et cultivé.

Cette étude permettra à l'association française d'Agroforesterie (AFAF) d'estimer quelles sont les meilleures populations sauvages candidates à choisir pour la réimplantation des arbres en milieu agricole. Leur but final est de mettre en place des politiques locales de réintroduction du pommier sauvage dans les plus brefs délais et diffuser les informations acquises le plus largement possible vers les agriculteurs et les régions. L'association a aussi pour but de généraliser cette démarche et de la pérenniser à long terme sur les autres espèces fruitières comme le poirier, le cormier, le merisier, le chêne, et l'érable.

Les données recueillies permettront aussi de préciser les précautions à respecter en matière de récolte et de transfert géographique des graines et plants destinés aux plantations. Elles permettront aussi de repérer les pommiers faussement forestiers, issus en fait de pépins de variétés fruitières ou de croisements sauvage-cultivé. Des récoltes de greffons pourront être réalisées sur une cinquantaine de pommiers sauvages pour constituer un (des) verger(s) à graines conservatoire(s) régionalisé(s), ce qui permettrait de résoudre, pour la (les) région(s) concernée(s), les problèmes d'utilisation et de conservation évoqués ci-dessus. Plus généralement, les résultats de la thèse et du projet permettront à la CRGF de jeter les bases d'une stratégie de conservation du pommier sauvage en France et de faciliter son élargissement à l'échelle paneuropéenne dans le cadre du réseau EUFORGEN. ■

Amandine Cornille, Tatiana Giraud* : Laboratoire Écologie, Systématique et Évolution, Université Paris-Sud, 91405 Orsay ; amandine.cornille@u-psud.fr*
*Éric Collin** : UR Ecosystèmes Forestiers, Irstea, 45290 Nogent-sur-Vernisson ; eric.collin@irstea.fr ; secrétaire de la CRGF*

(1) **Hamon X., (2007).** *Test de la faisabilité de l'agroforesterie dans l'Hérault.* Revue d'agroforesterie, 1, 36 p.
(2) Commission des Ressources Génétiques Forestières ; http://agriculture.gouv.fr/conservation-des-ressources.
(3) European Forest Genetic Resources programme *;* http://www.euforgen.org/.
(4) **Cornille A.** *Histoire de spéciations, de domestications, phylogéographie et hybridations dans le genre Malus.*
(5) Projet d'Initiative Citoyenne et de Recherche de la Région Île-de-France, (2011-2013).
(6) http://www.tela-botanica.org/actu/article4867.html

Annexe 2: ***Evolution of pathogenicity traits in the apple scab fungal pathogen in response to the domestication of its host*** **(Manuscrit E).**

Amandine Lê Van, Pierre Gladieux, Christophe Lemaire, Amandine Cornille, Tatiana Giraud, Charles-Eric Durel, Valérie Caffier and Bruno Le Cam (2012) Evolutionary Applications, in press.

Evolutionary Applications
Open Access
Evolutionary Applications ISSN 1752-4571

ORIGINAL ARTICLE

Evolution of pathogenicity traits in the apple scab fungal pathogen in response to the domestication of its host

Amandine Lê Van,[1] Pierre Gladieux,[2,3] Christophe Lemaire,[4] Amandine Cornille,[2,3] Tatiana Giraud,[2,3] Charles-Eric Durel,[1] Valérie Caffier[1] and Bruno Le Cam[1]

1 INRA, UMR1345, IRHS (INRA, Agrocampus-Ouest, Université d'Angers), SFR QUASAV, Beaucouzé, France
2 CNRS, UMR 8079, Ecologie, Systématique et Evolution, Univ. Paris-Sud, Orsay, France
3 AgroParisTech, Orsay, France
4 Université d'Angers, UMR1345 IRHS (INRA, Agrocampus-Ouest, Université d'Angers), SFR QUASAV, Angers, France

Keywords
apple scab, coevolution, disease emergence, plant–microbe interactions, wild crop relatives.

Correspondence
Bruno Le Cam, UMR1345 IRHS (INRA, Agrocampus-Ouest, Université d'Angers), SFR QUASAV, F-49071 Beaucouzé, France.
Tel.: +33 241 225 735;
fax: + 33 241 225 705;
e-mail: bruno.lecam@angers.inra.fr

Received: 23 December 2011
Accepted: 9 January 2012

doi:10.1111/j.1752-4571.2012.00246.x

Abstract

Understanding how pathogens emerge is essential to bring disease-causing agents under durable human control. Here, we used cross-pathogenicity tests to investigate the changes in life-history traits of the fungal pathogen *Venturia inaequalis* associated with host-tracking during the domestication of apple and subsequent host-range expansion on the wild European crabapple (*Malus sylvestris*). Pathogenicity of 40 isolates collected in wild and domesticated ecosystems was assessed on the domesticated apple, its Central Asian main progenitor (*M. sieversii*) and *M. sylvestris*. Isolates from wild habitats in the centre of origin of the crop were not pathogenic on the domesticated apple and less aggressive than other isolates on their host of origin. Isolates from the agro-ecosystem in Central Asia infected a higher proportion of plants with higher aggressiveness, on both the domesticated host and its progenitor. Isolates from the European crabapple were still able to cause disease on other species but were less aggressive and less frequently virulent on these hosts than their endemic populations. Our results suggest that the domestication of apple was associated with the acquisition of virulence in the pathogen following host-tracking. The spread of the disease in the agro-ecosystem would also have been accompanied by an increase in overall pathogenicity.

Introduction

The domestication of plants, expanding global trade and agriculture are major drivers of plant pathogen emergence (Pysek et al. 2010). Understanding how pathogens emerge on domesticated plants and spread in the agro-ecosystem is essential to bring disease-causing agents under durable human control (Anderson et al. 2004). Population genetic analyses on samples from domesticated and wild hosts have provided important insights into the origins of fungal pathogens (Gomez-Alpizar et al. 2007; Frenkel et al. 2010; Stukenbrock et al. 2011; Torriani et al. 2011). The emergence of pathogens on domesticated hosts can result from the colonization of a novel host or from a process of host-tracking where the pathogen simply follows its current host along the domestication process (e.g. Munkacsi et al. 2008; Gladieux et al. 2010; Giraud et al. 2010; Hansen 1987; Stukenbrock and McDonald 2008). Colonization of a novel host can be due to host-range expansion (i.e. colonization of a new host species while remaining pathogenic on the ancestral host) or host-shift (i.e. colonization of a new host species associated with the loss of the ability to infect the ancestral host). By definition, host-shifts involve very strong (qualitative) host specialization and can lead to rapid speciation through the evolution of an association between genes involved in host adaptation and genes conferring reproductive isolation (Couch et al. 2005; Giraud et al. 2010). In the case of host-range expansion or host-tracking, specialization is not an obligate outcome but quantitative specialization to

the new host can nonetheless be observed, with a lower performance on the ancestral host of pathogen populations from the new host compared to native populations (Fry 2003; Sicard et al. 2007). In fungal plant pathogens, specialization is favoured by some particular life-history traits including a large number of spores, mating within the host and high selection coefficients on a limited number of genes (Giraud et al. 2010).

In addition to the divergent selective pressures caused by the genetic differences among ancestral and domesticated hosts, pathogens are also exposed to a novel habitat, the agro-ecosystem, drastically different from natural ecosystems. The particular features of the human-engineered environment are thought to further enhance pathogen specialization to the domesticated hosts (Stukenbrock and McDonald 2008). Unlike natural plant populations, the high density and genetic uniformity of cultivated plant populations are highly conducive to pathogen transmission between infected and noninfected hosts, which favours more aggressive pathogens (Anderson and May 1982; Hochberg 2000; Thrall et al. 2007). Moreover, cultivated crops represent large targets for pathogens shifting from other hosts, while the large and widely connected pathogen populations associated with the agro-ecosystem are potential reservoirs for novel epidemics on naïve hosts that do not have evolved defence mechanisms (Couch et al. 2005). The vast-scale and homogenous availability of nutrients in the agro-ecosystem is also expected to enable the development of very large populations of pathogens, thereby increasing the efficiency of selection and accelerating adaptation (Karasov et al. 2010).

In analogy with the common suite of morphological and physiological traits that distinguish crops from their wild ancestors (Doebley et al. 2006; Zeder et al. 2006), the changes in life-history traits of pathogens adapting to domesticated hosts and to the agro-ecosystem can be regarded as a 'domestication syndrome'. Unlike plants, however, for which the domestication syndrome has been extensively investigated, the study of the evolutionary changes in pathogens associated with the domestication of plants and ecosystems is still in its infancy. This might be related to the lack of archaeological data on pathogens and to the difficulty in identifying, getting access and collecting samples in the centre of origins of diseases. Studies on the rice blast pathogen, *Magnaporthe oryzae*, have demonstrated differences in pathogenicity traits between populations infecting domesticated rice and the ancestral host of the pathogen, Setaria millet. Isolates from Setaria millet were either not virulent on rice or less aggressive than isolates from rice (Couch et al. 2005). Similarly, the pathogen *Rhynchosporium* shifted from an unidentified ancestor to cultivated barley and rye and speciated after adaptation to its new hosts at the time of domestication of cereals in the Fertile Crescent (Zaffarano et al. 2008).

The pathosystem *Malus* spp.-*Venturia inaequalis* (apple scab) is an excellent system to investigate the changes in life-history traits of pathogens adapting to domesticated hosts and to the agro-ecosystem. First, the life-history traits of *V. inaequalis* confer to this ascomycete an important evolutionary potential (McDonald and Linde 2002). The fungus reproduces both asexually during spring and summer (epidemic phase) and sexually during winter (saprotrophic phase). Reproduction occurs between strains of opposite mating types that have infected the same leaf. This reproductive system, where mating occurs between individuals that were able to grow on the same host genotype, is highly conducive to rapid ecological divergence (Giraud et al. 2010; Gladieux et al. 2011). Second, the story of the apple domestication is well documented. Historical information (Juniper and Mabberley 2006) and (partial) genetic evidence (Harris et al. 2002; Velasco et al. 2010) suggested that the centre of origin of the cultivated apple (*M.* × *domestica*) was Central Asia, where the wild apple *M. sieversii*, its main progenitor, forms forests (Harris et al. 2002). From Central Asia, the domesticated apple was moved westward to Europe and eastward to China following the Silk Road (Juniper and Mabberley 2006). During the spread of apple cultivation, several other *Malus* species may have contributed to the gene pool of *M.* × *domestica*. While the domesticated varieties appear closely related to *M. sieversii* (Velasco et al. 2010), a possible contribution from the European crabapple *M. sylvestris* to the domesticated apple gene pool is still debated (Coart et al. 2006; Harrison and Harrison 2011; Micheletti et al. 2011; A Cornille, P. Gladieux, I. Roldán-Ruiz, F. Laurens, B. Le Cam, M. J. M. Smulders, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X. G. Zhang, M. I. Tenaillon, T. Giraud, unpublished manuscript). The speciation between European and Central Asian wild apples likely occurred during Pleistocene repeated glaciations owing to the retreat and fragmentation of an ancient corridor of Tertiary temperate forests that ranged from the Atlantic Ocean to Bering (Juniper and Mabberley 2006). *Malus sylvestris* is now considered as an endangered tree species in some European regions, with a very scattered distribution (Stephan et al. 2003; A Cornille, P. Gladieux, I. Roldán-Ruiz, F. Laurens, B. Le Cam, M. J. M. Smulders, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X. G. Zhang, M. I. Tenaillon, T. Giraud, unpublished manuscript). Third, population genetics studies on *V. inaequalis* provided important clues about the evolutionary history of this pathogen. A previous study on the population structure of *V. inaequalis* showed that the pathogen shared a common origin with its host in Central Asia

(Gladieux et al. 2008). A subsequent study on populations of *V. inaequalis* infecting the wild apples *M. sieversii* and *M. sylvestris* pinpointed *M. sieversii* as the host of origin of the fungus (Gladieux et al. 2010). Results were consistent with a host-tracking scenario in which *V. inaequalis* spread into Europe together with the domesticated apple and subsequently expanded its range to *M. sylvestris*, previously free of apple scab. Population genetic analyses indicated that apple domestication had a strong impact on the population structure of the pathogen: apple domestication was associated with significant changes in the genetic differentiation of *V. inaequalis* populations in their centre of origin but had little impact on historical demography and mating system of the fungus (Gladieux et al. 2010). Three distinct gene pools were indeed identified based on microsatellite data by Gladieux et al. (2010): a population geographically restricted to the south-eastern mountains of Kazakhstan parasitizing *M. sieversii* (CAM population), an Asian population infecting *M.* × *domestica* and *M. sieversii* in peri-urban or agricultural areas (CAP population) and a European population present on *M.* × *domestica* and *M. sylvestris* (EU population) (Gladieux et al. 2010). Gladieux et al. (2010) hypothesized that the mountain population (CAM) could represent a relict of the ancestral populations that infected *M. sieversii* before apple domestication and from which the other populations would have diverged following domestication. The CAM population would represent an undisturbed population of the pathogen from natural ecosystems, while the CAP and EU populations would be evolved populations in contact with the agro-ecosystem.

Here, we used cross-inoculation tests to investigate the changes in pathogenicity traits of the apple scab fungus *V. inaequalis* associated with the domestication and spread of its host. Pathogenicity of 40 isolates collected in wild and domesticated ecosystems was assessed on the domesticated apple (*Malus* × *domestica*), its Central Asian main progenitor (*M. sieversii*) and the wild European crabapple (*M. sylvestris*). Two components of pathogenicity were analysed: virulence, that is, the ability to infect a given host genotype, and aggressiveness, that is, the severity of the disease in successful infections. We tested the hypotheses that agro-ecosystem features such as high host density favoured pathogen specialization on *M.* × *domestica* as well as an increase in aggressiveness, while the features of the European forest ecosystem with a very scattered host distribution did not lead to the specialization on *M. sylvestris*. We tested more specifically the following hypotheses: (i) host-tracking of *V. inaequalis* from the wild ancestor to the cultivated apple has been associated with a gain in virulence; evidence would be that isolates from the wild Asian progenitor are unable to cause disease on *M.* × *domestica*; (ii) adaptation to the cultivated apple has been associated with increased overall pathogenicity; evidence would be that isolates from domesticated apple trees are more aggressive or more frequently virulent on the wild Asian progenitor than the endemic isolates; (iii) the emergence of apple scab on wild crabapple populations in Europe was due to a host-range expansion and not a host-shift of *V. inaequalis* populations from agro-ecosystems; evidence would be that isolates from crabapple trees are still able to cause disease on the domesticated trees; iv) the host-range expansion on *M. sylvestris*, nevertheless, resulted in a certain degree of specialization; evidence would be that crabapple isolates show lower aggressiveness or lower frequency of virulence on domesticated trees than isolates from *M.* × *domestica*.

Materials and methods

Fungal isolates

This study was based on a total of 40 isolates of *V. inaequalis* (Table 1) sampled on *M. sieversii*, *M.* × *domestica* and *M. sylvestris*. Three core collections of *V. inaequalis* were previously constructed, one per *Malus* species of origin, as maximizing neutral genetic diversity among isolates genotyped with 12 SSR loci (Lê Van et al. 2011). Each core collection was constituted by 15 isolates except the '*M.* × *domestica* core collection' constituted by 10 isolates. The isolates originating from CAM, CAP or EU populations (Gladieux et al. 2010) were classified into five pools labelled after their geographic origin (Asia or Europe), the environment of origin (wild or agro-ecosystem) and the host of origin (*M. sieversii*, *M.* × *domestica* or *M. sylvestris*): 'WildAsiaSiev', 'AgroAsiaSiev', 'AgroAsiaDom', 'AgroEuDom' and 'WildEuSylv' (Table 1).

Plant material

Two cultivars of *M.* × *domestica*, three accessions of *M. sieversii* (GMAL 3619.b, PI 633797.d and PI 633799.e) and four accessions of *M. sylvestris* (X 9650, X 9651, X 9653 and X 9654) were used in this study. The two cultivars, Gala and Top Red Delicious (latter on called 'Top Red'), are extensively cultivated worldwide. *Malus sieversii* accessions were collected in Kazakhstan in two different localities (Figure S1). GMAL 3619.b and PI 633799.e were collected from the Tarbagatai mountain range, and PI 633797.d was collected from the Djungarsky mountain range (Forsline et al. 2010; USDA website: http://www.ars-grin.gov/cgi-bin/npgs/html/search.pl). *Malus sylvestris* accessions were collected in the French forest of Rambouillet. Budwoods were grafted on 'MM106' apple rootstocks and then maintained in a greenhouse. The plants of *M. sylvestris* used in pathogenicity tests were

Table 1. Description of the *Venturia inaequalis* isolates used in this study.

Isolate	Country of origin	Sampled year	Population name	*Malus* host (Cultivar)
2217	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2219	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2220	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2221	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2222	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2223	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2224	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2225	Kazakhstan	2006	WildAsiaSiev†	*M. sieversii*
2227	Kazakhstan	2006	AgroAsiaSiev‡	*M. sieversii*
2228	Kazakhstan	2006	AgroAsiaSiev‡	*M. sieversii*
2229	Kazakhstan	2006	AgroAsiaSiev‡	*M. sieversii*
2230	Kazakhstan	2006	AgroAsiaSiev‡	*M. sieversii*
2231	Kazakhstan	2006	AgroAsiaSiev‡	*M. sieversii*
2233	China	2005	AgroAsiaSiev‡	*M. sieversii*
2234	China	2005	AgroAsiaSiev‡	*M. sieversii*
2278	China	2005	AgroAsiaDom‡	*M.* × *domestica* (Golden Delicious)
2279	China	2005	AgroAsiaDom‡	*M.* × *domestica* (Gala)
2281	China	2005	AgroAsiaDom‡	*M.* × *domestica* (New Century)
2284*	China	2005	AgroAsiaDom‡	*M.* × *domestica* (New Century)
2285	China	2005	AgroAsiaDom‡	*M.* × *domestica* (Gala)
2286	France	2005	AgroEuDom§	*M.* × *domestica* (Mutsu)
2288	France	2005	AgroEuDom§	*M.* × *domestica* (Mutsu)
2289	France	2005	AgroEuDom§	*M.* × *domestica* (Mutsu)
2291	Spain	2005	AgroEuDom§	*M.* × *domestica* (Wellspur)
EU-D-16	Germany	1999	AgroEuDom§	*M.* × *domestica* (Coop 9)
2237	France	2005	WildEuSylv§	*M. sylvestris*
2238*	France	2005	WildEuSylv§	*M. sylvestris*
2239*	France	2005	WildEuSylv§	*M. sylvestris*
2240*	France	2005	WildEuSylv§	*M. sylvestris*
2241*	France	2005	WildEuSylv§	*M. sylvestris*
2245	France	2005	WildEuSylv§	*M. sylvestris*
2246*	France	2005	WildEuSylv§	*M. sylvestris*
2247	France	2005	WildEuSylv§	*M. sylvestris*
2248*	France	2005	WildEuSylv§	*M. sylvestris*
2249*	France	2005	WildEuSylv§	*M. sylvestris*
2251*	France	2005	WildEuSylv§	*M. sylvestris*
2252*	France	2005	WildEuSylv§	*M. sylvestris*
2254*	France	2005	WildEuSylv§	*M. sylvestris*
2255*	France	2005	WildEuSylv§	*M. sylvestris*
2256	France	2005	WildEuSylv§	*M. sylvestris*

*These isolates were inoculated onto *M.* × *domestica* and *M. sylvestris* but not onto *M. sieversii*.
†This population belongs to the previously identified CAM (Central Asian Mountains) population (Gladieux et al. 2010).
‡This population belongs to the previously identified CAP (Central Asian Plains) population.
§This population belongs to the previously identified EU (European) population.

genotyped using microsatellite markers, and their assignment to the gene pool of their putative species of origin was checked using a reference data set (Data S1).

Cross-pathogenicity tests

The 40 isolates were inoculated onto all host genotypes during three rounds of experiments per host species. For each experiment, one to six isolates of each fungal population were inoculated onto one *Malus* species. For inoculations onto *M. sieversii*, 28 of 40 isolates were inoculated because of limited plant material available (Table 1). Only actively growing plants with uniform growth were chosen and transferred to a quarantine-controlled climate chamber for subsequent inoculations. We used a quarantine room because of the exotic origin of numerous isolates whose unknown virulence may present a risk. Inocula were obtained by growing monoconidial isolates of the

fungus on cellophane sheets deposited onto malt agar medium (Bus et al. 2005). Monoconidial suspensions of each isolate were adjusted to a concentration of 1.5×10^5 conidia/mL. Germination rates were assessed for each monoconidial suspension on malt agar plates to check for the viability of conidia. Germination rates ranged from 34% to 95% depending on the isolate and the experiment, with more than 64% reached in 75% of cases. Each isolate was sprayed using an air pressure hand-sprayer on four to five replicates of each host genotype. All leaves were inoculated. For the first 48 h after inoculation, the plants were kept in darkness with humidity maintained at 100% and temperature at 18°C to allow conidia germination and fungal infection. Humidity was then reduced to 80% with 16-h light per day. Percentage of each leaf showing sporulation was scored visually at 14, 21 and 28 days after inoculation (dai) on an ordinal scale, ranging from 0 (no sporulation) to 8 (100% of leaf area showing sporulation) (Lê Van et al. 2011). For virulence, each replicate was either scored as infected or not (no visible sporulating symptoms at 28 days).

Data analyses

Analyses of virulence

Virulence was defined as the ability of an isolate to produce sporulating symptoms on a host genotype. A Pearson's chi-squared test was performed on contingency tables to test for independence between virulence and population of origin. Because of expected cell count below five, *P*-values were computed for a Monte Carlo test (Hope 1968) with 1×10^5 replicates. When virulence and the population of origin were not independent, multiple comparisons were made using Pearson's chi-squared tests on two-by-two contingency tables using the Bonferroni correction. A Pearson's chi-squared test was also conducted to test for independence between virulence and *Malus* species tested.

Analyses of aggressiveness

Aggressiveness was measured as the area under the disease progress curve (AUDPC) calculated on the sporulation percentage of the most diseased leaf of each replicate. The 'AUDPC' variable was analysed using a linear mixed-effect model (LME). The 'isolate' was treated as random factor and nested in the population of origin. The cultivars were treated as fixed factors. A variance function was used for modelling the within-group heteroscedasticity. Each factor (isolate, population of origin, *Malus* species of origin, tested cultivar and round of experiment) was included in the model based on an ascendant selection using BIC (Bayesian Information Criterion) to select the best model (Pinheiro and Bates 2000). The model was fitted by maximizing the log-likelihood. All statistical analyses (virulence and aggressiveness) were performed using the 'nlme' package (Pinheiro et al., 2008) in R version 2.10.1 (R-Development-Core-team, 2008).

Results

Virulence

Malus sieversii plants were significantly more often infected by isolates from Central Asia sampled in the agro-ecosystem area, either from *M. sieversii* (AgroAsiaSiev population) or from *M.* × *domestica* (AgroAsiaDom population), than by isolates sampled in the wild area in Central Asia (WildAsiaSiev population) or isolates from Europe, either from wild or from agro-ecosystem (Pearson's chi-squared tests on two-by-two contingency tables; Fig. 1). Isolates belonging to the Asian population from the agro-ecosystem were able to infect up to three *M. sieversii* accessions, whereas isolates from Europe were able to infect up to two accessions. Only one of the three accessions was infected by isolates from the WildAsiaSiev population (Table 2). The higher frequency of virulence on *M. sieversii* in fungal populations from Asian agro-ecosystems suggests that the spread of the disease on the domesticated apple has been associated with an increase in pathogenicity of *V. inaequalis*.

None of the *M.* × *domestica* cultivars was infected by the WildAsiaSiev population. *Malus* × *domestica* cultivars were significantly less frequently infected by the WildAsiaSiev population than by other pathogen populations (Pearson's chi-squared tests on two-by-two contingency tables; Fig. 1). All other populations were able to infect both *M* × *domestica* cultivars (Fig. 1). The lack of pathogenicity of isolates from wild habitats of the centre of origin of the crop when inoculated on the domesticated apple indicates that the host-tracking of *V. inaequalis* from the wild ancestor to the cultivated apple has demanded acquisition of new virulence.

Although isolates from both *M.* × domestica and *M. sylvestris* were able to infect *M.* × *domestica* cultivars, the number of infected plants was significantly lower when challenged with isolates from *M. sylvestris* (WildEuSylv population) than with isolates from *M.* × *domestica* either collected in Asia or in Europe ($\chi^2 = 9.44$; $P < 0.01$ and $\chi^2 = 6.82$; $P < 0.01$, respectively). This suggests a certain degree of specialization by isolates parasitizing the European crabapple. No significant differences were observed between isolates from *M.* × *domestica* from either Asia or Europe ($P = 1$) (Fig. 1).

Malus sylvestris plants were significantly more frequently diseased after inoculation by isolates from *M. sylvestris* than by isolates from other *Malus* species (Pearson's chi-squared tests on two-by-two contingency

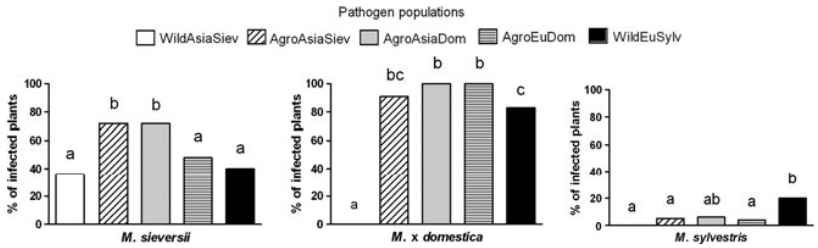


Figure 1 Percentage of plants of three *Malus* species infected by five *Venturia inaequalis* populations. Note that an empty place indicates zero infected plant for the corresponding pathogen population. Virulence was measured on three accessions for *M. sieversii*, two cultivars for *M.* × *domestica* and four accessions for *M. sylvestris*. Different letters indicate significant differences between populations ($P < 0.05$) (Pearson's chi-squared tests with Bonferroni correction on two-by-two contingency tables).

Table 2. Number of virulent isolates of *Venturia inaequalis* out of the number of isolates tested for each genotype of three *Malus* species.

	M. sieversii accessions			*M.* × *domestica* cultivars		*M. sylvestris* accessions			
Pathogen population	GMAL 3619.b	PI 633 797.d	PI 633 799.e	Gala	Top Red	X9650	X9651	X9653	X9654
WildAsiaSiev	8/8	0/8	0/7	0/8	0/7	0/8	0/8	0/8	0/8
AgroAsiaSiev	7/7	3/7	5/7	6/7	7/7	0/7	0/7	2/7	0/6
AgroAsiaDom	4/4	3/4	3/4	5/5	5/5	0/5	0/5	1/4	0/4
AgroEuDom	5/5	1/5	1/5	5/5	5/5	0/5	0/5	0/5	1/5
WildEuSylv	4/4	0/4	1/4	13/15	12/15	1/14	2/15	4/12	6/14

tables; Fig. 1), except for AgroAsiaDom. Moreover, some isolates from *M. sylvestris* were able to infect up to three different *M. sylvestris* accessions, whereas isolates from other populations were able to infect a single accession. This suggests that the colonization of *M. sylvestris* has demanded a certain degree of adaptation.

The number of diseased plants was not significantly different when challenged with isolates from the agro-ecosystem in Asia sampled on *M.* × *domestica* or on *M. sieversii* ($\chi^2 = 0.03$; $P = 1$). Similar to the pattern of virulence observed on *M.* × *domestica*, the eight isolates from WildAsiaSiev were not able to infect *M. sylvestris* accessions (Fig. 1). *Malus sylvestris* accessions were resistant to a significantly higher number of isolates than *M.* × *domestica* or *M. sieversii* accessions ($\chi^2 = 356.28$; $P < 0.0001$).

Aggressiveness

Data of all experiments conducted on the same *Malus* species were pooled. The experiment factor had no significant effect and did not improve LME models. As a consequence, it was not eventually included in the models. The within-group heteroscedasticity was modelled as a power function of mean fitted values.

Because a single *M. sieversii* accession was susceptible to all pathogen populations, statistical analyses were conducted only for this accession (GMAL 3619.b). The factor 'population of origin' explained the area under the disease progress curve (AUDPC) variance (BIC = 3158) and significantly improved the null model (BIC = 3176; $P < 0.0001$). Isolates from Asia collected in the agro-ecosystem, either from *M. sieversii* or from *M.* × *domestica*, were significantly more aggressive than isolates from other populations on the *M. sieversii* accession (Fig. 2). The AgroAsiaSiev and the AgroAsiaDom populations were on average fivefold and fourfold, respectively, more aggressive than the WildAsiaSiev population. This higher aggressiveness of populations from agro-ecosystems on *M. sieversii* is another evidence indicating an increase in overall pathogenicity following the spread of the pathogen on the domesticated apple.

Aggressiveness of isolates from WildAsiaSiev was significantly lower than aggressiveness of isolates from AgroAsiaSiev, but not significantly different from aggressiveness of European isolates, regardless of their host of origin (*M.* × *domestica* or *M. sylvestris*). The response of isolates from *M.* × *domestica* was different across populations. Isolates from *M.* × *domestica* collected in Asia (AgroAsiaDom) were significantly more aggressive

than isolates from Europe (AgroEuDom) on *M. sieversii*. The lower aggressiveness and the lower frequency of virulent isolates in European population from the agro-ecosystem support the view that, unlike in Central Asia, pathogen populations in Europe evolved a quantitative specialization to the apple-based agro-ecosystem.

For tests on *M.* × *domestica* accessions, the best model included the factor '*Malus* species of origin' (BIC = 3709), significantly improving the null model (BIC = 3715; $P < 0.001$). There was no significant effect of the cultivar, cv. Gala and cv. Top Red having similar responses. Thus, adding this trait to the model did not significantly improve the BIC score (BIC = 3721). The most aggressive isolates were those from *M.* × *domestica* regardless of their geographic origin (Asia or Europe) (Fig. 2). However, isolates from *M.* × *domestica* were not significantly more aggressive than isolates from AgroAsiaSiev ($P = 0.81$). WildEuSylv was the least aggressive population, suggesting that the emergence of the disease on *M. sylvestris* was followed by the evolution of quantitative specialization in pathogen populations.

The number of isolates virulent on *M. sylvestris* accessions was low (Table 2). Furthermore, the severity of the disease caused by virulent isolates was weak (Fig. 2). As a consequence, the statistical analysis of aggressiveness had too low a power to infer any reliable conclusion and is therefore not presented.

Discussion

Previous studies exploiting population genetic inference revealed a marked impact of domestication on fungal pathogen population structure, leading in some cases to the emergence of novel pathogen species (Couch et al. 2005; Munkacsi et al. 2007; Stukenbrock et al. 2007). However, the impact of domestication on pathogenicity traits has rarely been investigated, despite being of major importance to understand the consequences of modern human activities on disease emergence. We used cross-pathogenicity tests in controlled quarantine conditions to investigate the changes in pathogenicity traits of the apple scab fungus, *V. inaequalis*, associated with the domestication and spread of its host. Our main findings were that host-tracking was associated with a change in virulence and an increase in aggressiveness of pathogen populations from the agro-ecosystem. Our results suggested that the transition from wild to apple-based agro-ecosystem did not promote the evolution of specialized populations of *V. inaequalis*, as they were still able to infect the ancestral host plant. In contrast, host-range expansion from the domesticated apple to the European wild apple was associated with a certain degree of host specialization, as populations of *V. inaequalis* from *M. sylvestris* caused a less severe disease on *M.* × *domestica*.

Changes in pathogenicity traits associated with host-tracking

The pathogenicity of isolates collected in natural ecosystems on the wild apple *M. sieversii* (WildAsiaSiev) can be compared to more recently founded pathogen populations from the agro-ecosystem to draw inferences on the evolutionary changes associated with the emergence and spread of *V. inaequalis* on the domesticated apple. The lack of pathogenicity of isolates from the WildAsiaSiev population when inoculated onto the domesticated apple suggests a strong ecological differentiation between the WildAsiaSiev population and the populations from the agro-ecosystem. A similar pattern of pathogenicity was

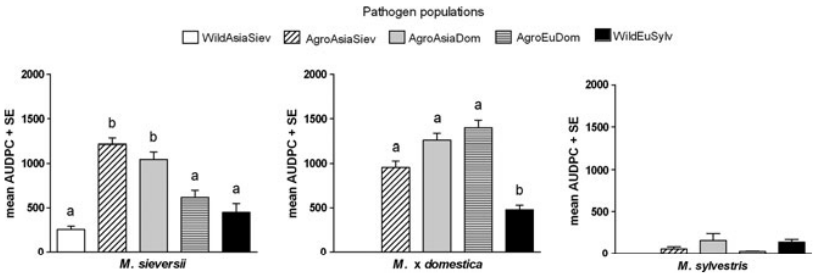


Figure 2 Mean area under the disease progress curve (AUDPC) (+SE) of five *Venturia inaequalis* populations inoculated onto three *Malus* species. AUDPC was measured on one accession for *Malus sieversii*, two cultivars for *M.* × *domestica* and two accessions for *M. sylvestris*. Different letters indicate significant differences between populations ($P < 0.05$), parameters being estimated by the maximum-likelihood algorithm in the linear mixed-effect model.

observed in a study on the rice blast fungus where isolates of *M. oryzae* from ancestral undomesticated hosts (Setaria millet) were not able to infect or were less aggressive on domesticated rice (Couch et al. 2005). Unlike *M. oryzae*, however, *V. inaequalis* did not emerge on the domesticated crop following a host-shift but through a more continuous process of host-tracking. It could be hypothesized that the disruptive change during domestication corresponded to input of new resistance genes in the domesticated apple species. Indeed, along the process of domestication, *M.* × *domestica* might have hybridized with several wild species of *Malus* such as *M. sieversii*, *M. baccata*, *M. kirghisorumi*, *M. orientalis* (Coart et al. 2006; Forsline et al. 2010; A Cornille, P. Gladieux, I. Roldán-Ruiz, F. Laurens, B. Le Cam, M. J. M. Smulders, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X. G. Zhang, M. I. Tenaillon, T. Giraud, unpublished manuscript) from which new resistant genes could have been introgressed. Resistance genes from *M. sieversii* might also be unconsciously selected during the course of domestication. A recent study also suggests that novel resistant alleles may also have been created in crops during domestication (Zhai et al. 2011).

Additional insights into the changes in pathogenicity associated with the apple domestication can be gained by comparing the frequency of virulence and aggressiveness of isolates collected in populations from natural and agro-ecosystems. The different populations could only be compared on *M. sieversii*, as isolates from the WildAsiaSiev population were avirulent on *M.* × *domestica*. Isolates from the WildAsiaSiev population were less aggressive and less frequently virulent on *M. sieversii* than isolates from the population representing the Central Asian agro-ecosystem (AgroAsiaSiev and AgroAsiaDom populations), suggesting that spread of the disease on the domesticated apple may have been associated with a quantitative increase in pathogenicity. These differences in overall pathogenicity (both virulence and aggressiveness) may be related to the contrasted ecological properties of the two environments. Higher density and homogeneity of the agro-ecosystem would have promoted higher pathogenicity of populations infecting domesticated hosts, while the patchy and geographically structured populations of *M. sieversii* (Richards et al. 2009) would have impeded the evolution of high pathogenicity in populations from natural ecosystems. The lower overall pathogenicity of WildAsiaSiev could also be explained by differences in resistance to *V. inaequalis* between *M. sieversii* populations from Central Asian mountains and plains. A study of local adaptation using *M. sieversii* genotypes from which WildAsiaSiev and AgroAsiaSiev populations were sampled would be interesting for assessing to what extent WildAsiaSiev and AgroAsiaSiev populations are adapted to their host, taking into account the potential diversity in resistance to *V. inaequalis* existing within *M. sieversii*.

Apple-based agro-ecosystem did not promote the evolution of specialized populations of *V. inaequalis*

The higher level of environmental homogeneity of the agro-ecosystem is thought to promote the ecological specialization of pathogens associated with cultivated species (Stukenbrock and McDonald 2008). Following host-tracking, *V. inaequalis* populations on *M.* × *domestica* did not loose their capacity to infect their wild native host *M. sieversii*. In Central Asia, specialization of the *V. inaequalis* population from *M.* × *domestica* agro-ecosystem might have been impeded by recurrent gene flow between populations infecting domesticated apples and neighbouring populations from *M. sieversii* in human-managed habitats. In Europe, pathogenicity experiments supported quantitative specialization of the *V. inaequalis* population from the agro-ecosystem. This population was less frequently virulent and less aggressive than Asian populations from the agro-ecosystem on *M. sieversii*. Allelic combinations providing higher pathogenic fitness on *M. sieversii* would have been unnecessary in European populations and would thus have been progressively lost through genetic drift. The observed quantitative specialization in Europe could therefore be due to geographic distance rather than to environmental differences between wild and agricultural ecosystems.

The emergence of apple scab on the European wild apple resulted from a host-range expansion associated with quantitative host specialization

The pathogenicity of isolates from European crabapple on the domesticated trees indicated that the emergence of apple scab on *M. sylvestris* resulted from a host-range expansion and not a host-shift. Analyses further revealed that the host-range expansion from *M.* × *domestica* to *M. sylvestris*, nevertheless, led to a quantitative specialization with a lower frequency of virulence and lower aggressiveness of populations from *M. sylvestris* on the native host *M.* × *domestica*, suggesting a trade-off between adaptations onto these two hosts. The lower phylogenetic distance between *M. sieversii* and *M.* × *domestica* (Velasco et al. 2010) compared with that between *M.* × *domestica* and *M. sylvestris* could explain the less pronounced trade-off between adaptations to these two former hosts. Stronger trade-offs are indeed expected for host-range expansion involving phylogenetically distant hosts than for host-tracking. Isolates from *M. sylvestris* were not more aggressive on *M. sylvestris* than isolates sampled on other hosts. The highly scattered distribution of *M. sylvestris*

across Europe and within forests may not promote evolution towards higher aggressiveness (Gilbert 2002). Environmental context could thus be of primary importance for the evolution of pathogenicity and aggressiveness.

Concluding remarks

We have investigated the changes in pathogenicity in populations of the pathogen *V. inaequalis*, associated with the domestication of their host. The emergence of *V. inaequalis* on domesticated hosts was associated with a gain in virulence and a subsequent increase in aggressiveness on *M. sieversii* trees in contact with the agro-ecosystem. These results are expected characteristics of the domestication syndrome in pathogens. We hypothesize that pathogen populations did not specialize on the domesticated host because of the close relatedness between their original wild host and *M.* × *domestica* and the persistence of gene flow between pathogen populations from *M. sieversii* and *M.* × *domestica* in Central Asia. A decrease in the levels of gene flow in Europe would have led to a decrease in aggressiveness of the European population from *M.* × *domestica* but without a loss of its capacity to infect its original wild host. The introduction of *V. inaequalis* in Europe was followed by a host-range expansion from *M.* × *domestica* to the more phylogenetically distant *M. sylvestris*. The existence of efficient resistance traits in this wild species, in association with its very low density in forests, may have limited the increase in aggressiveness of the corresponding pathogen population.

Our findings have important implications regarding the assessment of risk for the emergence of highly aggressive pathogens in wild and agricultural ecosystems. We show here that regulation agencies, policy makers, as well as plant breeders, should consider very carefully the risk of host-tracking by pathogens onto domesticated species. The favourable environment provided by the agro-ecosystem can foster the emergence of new pathogens with increased virulence and aggressiveness. While many plant and animal species are still under domestication (Diamond 2002), our results point to the considerable risk that potentially unnoticed pathogens can adapt via host-tracking and subsequently spread across continents. Moreover, the finding that a pathogen having emerged in the agro-ecosystem has increased its aggressiveness without loosing its ability to infect its original host also suggests that such pathogens can subsequently pose serious threats to wild crop relatives. The policy regarding cultivated areas should therefore take into account the surrounding wild ecosystem to prevent a 'boomerang' effect, that is, the return of more aggressive pathogens back on wild original hosts.

Acknowledgements

We are greatly indebted to all the people that helped with sample collection: Catherine Peix, Xiu Guo Zhang, François Laurens, Laurent Brun, Marta Pujol, Brigitte Musch, Laurent Levèque, M. Rigoleur, M. Arigoni, M. Ihl and M. Le Valégan. We are also grateful to Pascal Heitzler for providing *M. sieversii* budwoods and Yves Le Vallegan and M. Arigoni for providing *M. sylvestris* budwoods. We thank Pauline Lasserre-Zuber and Caroline Denancé for complementary *Malus* genotyping, Pascale Expert for her help in pathological tests and Frédérique Didelot and Frédéric Fabre for advices on statistical analyses. Amandine Lê Van was supported by a fellowship from INRA, Department SPE and GAP and the Region Pays de La Loire. This work was funded by an Agence Nationale de la Recherche grant ANR 07-BDIV-003 and by the CO-SAVE program (Région Pays de La Loire).

Data archiving statement

Data deposited in the Dryad repository: doi:10.5061/dryad.6bp470fn

Literature cited

Anderson, R. M., and R. M. May 1982. Coevolution of hosts and parasites. Parasitology **85**:411–426.

Anderson, P. K., A. A. Cunningham, N. G. Patel, F. J. Morales, P. R. Epstein, and P. Daszak 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in Ecology & Evolution **19**:535–544.

Bus, V. G. M., F. N. D. Laurens, W. E. van de Weg, R. L. Rusholme, E. H. A. Rikkerink, S. E. Gardiner, H. C. M. Bassett *et al.* 2005. The *Vh8* locus of a new gene-for-gene interaction between *Venturia inaequalis* and the wild apple *Malus sieversii* is closely linked to the *Vh2* locus in *Malus pumila* R12740-7A. New Phytologist **166**:1035–1049.

Coart, E., S. Van Glabeke, M. De Loose, A. S. Larsen, and I. Roldan-Ruiz 2006. Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). Molecular Ecology **15**:2171–2182.

Couch, B. C., I. Fudal, M. H. Lebrun, D. Tharreau, B. Valent, P. van Kim, J. L. Notteghem *et al.* 2005. Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. Genetics **170**:613–630.

Diamond, J. 2002. Evolution, consequences and future of plant and animal domestication. Nature **418**:700–707.

Doebley, J. F., B. S. Gaut, and B. D. Smith 2006. The molecular genetics of crop domestication. Cell **127**:1309–1321.

Forsline, P. L., H. S. Aldwinckle, E. E. Dickson, J. J. Luby, and S. C. Hokanson 2010. Collection, maintenance, characterization, and utilization of wild apples of Central Asia. In: J. Janick, ed. Horticultural Reviews: Wild Apple and Fruit Trees of Central Asia, Vol. **29**, pp. 1–61. John Wiley & Sons, Inc, Oxford, UK.

Frenkel, O., T. L. Peever, M. I. Chilvers, H. Ozkilinc, C. Can, S. Abbo, D. Shtienberg *et al.* 2010. Ecological genetic divergence of the fungal pathogen *Didymella rabiei* on sympatric wild and domesticated *Cicer* spp (Chickpea). Applied and Environmental Microbiology **76**:30–39.

Fry, J. D. 2003. Detecting ecological trade-offs using selection experiments. Ecology **84**:1672–1678.

Gilbert, G. S. 2002. Evolutionary ecology of plant diseases in natural ecosystems. Annual Review of Phytopathology **40**:13–43.

Giraud, T., P. Gladieux, and S. Gavrilets 2010. Linking the emergence of fungal plant diseases with ecological speciation. Trends in Ecology & Evolution **25**:387–395.

Gladieux, P., X. G. Zhang, D. Afoufa-Bastien, R. M. V. Sanhueza, M. Sbaghi, and B. Le Cam 2008. On the origin and spread of the scab disease of apple: out of Central Asia. PLoS One **3**:e1455.

Gladieux, P., X. G. Zhang, I. Roldan-Ruiz, V. Caffier, T. Leroy, M. Devaux, S. Van Glabeke *et al.* 2010. Evolution of the population structure of *Venturia inaequalis*, the apple scab fungus, associated with the domestication of its host. Molecular Ecology **19**:658–674.

Gladieux, P., F. Guérin, T. Giraud, V. Caffier, L. Parisi, F. Didelot, C. Lemaire *et al.* 2011. Emergence of novel fungal pathogens by ecological speciation: importance of the reduced viability of immigrants. Molecular Ecology **20**:4521–4532.

Gomez-Alpizar, L., I. Carbone, and J. B. Ristaino 2007. An Andean origin of *Phytophthora infestans* inferred from mitochondrial and nuclear gene genealogies. Proceedings of the National Academy of Sciences of the United States of America **104**:3306–3311.

Hansen, E. M. 1987. Speciation in plant pathogenic fungi: the influence of agricultural practice. Canadian Journal of Plant Pathology **9**:403–410.

Harris, S. A., J. P. Robinson, and B. E. Juniper 2002. Genetic clues to the origin of the apple. Trends in Genetics **18**:426–430.

Harrison, N., and R. J. Harrison 2011. On the evolutionary history of the domesticated apple. Nature Genetics **43**:1043–1044.

Hochberg, M. E. 2000. Evidence that specialists are special. Trends in Ecology & Evolution **15**:490.

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society **30**:582–598.

Juniper, B. E., and D. J. Mabberley 2006. The Story of the Apple. Timber Pressm, Portland.

Karasov, T., P. W. Messer, and D. A. Petrov 2010. Evidence that adaptation in Drosophila is not limited by mutation at single sites. Plos Genetics **6**:e1000924.

Lê Van, A., C. E. Durel, B. Le Cam, and V. Caffier 2011. The threat of wild habitat to scab resistant apple cultivars. Plant Pathology **60**:621–630.

McDonald, B. A., and C. Linde 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annual Review of Phytopathology **40**:349–379.

Micheletti, D., M. Troggio, F. Salamini, R. Viola, R. Velasco, and S. Salvi 2011. On the evolutionary history of the domesticated apple. Nature Genetics **43**:1044–1045.

Munkacsi, A. B., S. Stoxen, and G. May 2007. Domestication of maize, sorghum, and sugarcane did not drive the divergence of their smut pathogens. Evolution **61**:388–403.

Munkacsi, A. B., S. Stoxen, and G. May 2008. *Ustilago maydis* populations tracked maize through domestication and cultivation in the Americas. Proceedings of the Royal Society B-Biological Sciences **275**:1037–1046.

Pinheiro, J. C., and D. M. Bates 2000. Mixed-Effects Models in S and S-PLUS, Statistics and Computing. Springer-Verlag, New York, USA.

Pinheiro, J. C., D. M. Bates, S. DebRoy, and D. Sarkar; R-Core-team 2008. Nlme: Linear and Nonlinear Mixed Effects Models. R Package (version 3). http://www.R-project.org (accessed on 1 December 2008).

Pysek, P., V. Jarosik, P. E. Hulme, I. Kuhn, J. Wild, M. Arianoutsou, S. Bacher *et al.* 2010. Disentangling the role of environmental and human pressures on biological invasions across Europe. Proceedings of the National Academy of Sciences of the United States of America **107**:12157–12162.

R-Development-Core-team 2008. R: A Language and Environment for Statistical Computing. Vol. ISBN 3-900051-07-0. R foundation for Statistical Computing, Vienna, Austria [http://www.R-project.org] (accessed on 14 December 2009).

Richards, C. M., G. M. Volk, A. A. Reilley, A. D. Henk, D. R. Lockwood, P. A. Reeves, and P. L. Forsline 2009. Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. Tree Genetics & Genomes **5**:339–347.

Sicard, D., P. S. Pennings, C. Grandclement, J. Acosta, O. Kaltz, and J. A. Shykoff 2007. Specialization and local adaptation of a fungal parasite on two host plant species as revealed by two fitness traits. Evolution **61**:27–41.

Stephan, B. R., I. Wagner, and J. Kleinschmit 2003. EUFORGEN Technical Guidelines for Genetic Conservation and Use for Wild Apple and Pear (*Malus sylvestris* and *Pyrus pyraster*). International Plant Genetic Resources Institute, Rome, Italy, 6.

Stukenbrock, E. H., and B. A. McDonald 2008. The origins of plant pathogens in agro-ecosystems. Annual Review of Phytopathology **46**:75–100.

Stukenbrock, E. H., S. Banke, M. Javan-Nikkhah, and B. A. McDonald 2007. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. Molecular Biology and Evolution **24**:398–411.

Stukenbrock, E. H., T. Bataillon, J. Y. Dutheil, T. T. Hansen, R. Li, M. Zala, B. A. McDonald *et al.* 2011. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Research **21**:2157–2166.

Thrall, P. H., M. E. Hochberg, J. J. Burdon, and J. D. Bever 2007. Coevolution of symbiotic mutualists and parasites in a community context. Trends in Ecology & Evolution **22**:120–126.

Torriani, S. F. F., P. C. Brunner, and B. A. McDonald 2011. Evolutionary history of the mitochondrial genome in *Mycosphaerella* populations infecting bread wheat, durum wheat and wild grasses. Molecular Phylogenetics and Evolution **58**:192–197.

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana *et al.* 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nature Genetics **42**:833–839.

Zaffarano, P. L., B. A. McDonald, and C. C. Linde 2008. Rapid speciation following recent host shifts in the plant pathogenic fungus *Rhynchosporium*. Evolution **62**:1418–1436.

Zeder, M. A., E. Emshwiller, B. D. Smith, and D. G. Bradley 2006. Documenting domestication: the intersection of genetics and archaeology. Trends in Genetics **22**:139–155.

Zhai, C., F. Lin, Z. Dong, X. He, B. Yuan, X. Zeng, L. Wang *et al.* 2011. The isolation and characterization of *Pik*, a rice blast resistance gene which emerged after rice domestication. New Phytologist **189**:321–334.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Geographic location of strains (circles) and hosts (squares) sampling sites. Populations of strains are identified by different colours: WildAsiaSiev (red), AgroAsiaSiev (orange), AgroAsiaDom (purple), AgroEuDom (yellow) and WildEuSylv (green). Red squares represent the three accessions of *M. sieversii* from Kazakhstan. Accessions of *M. sylvestris* are located in France in the Rambouillet forest (next to Paris). Maps are provided by Google Earth©.

Data S1. *M. sylvestris* accessions were genotyped using microsatellite markers, and their assignment to the gene pool of their putative species of origin was checked using a reference data set.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Résumé : Malgré son importance économique, culturelle et historique, l'histoire évolutive du pommier cultivé (*Malus domestica*) ainsi que celle de ses apparentés sauvages supposés, restaient encore très peu connues. En s'appuyant sur les nouvelles approches de génétique des populations (*approximate Bayesian computation*) avec l'utilisation de marqueurs microsatellites et de séquences nucléaires, cette thèse a eu pour objectif d'étudier, à différentes échelles évolutives (phylogéographie, spéciation, domestication), les mécanismes de diversification naturelle et artificielle dans le genre *Malus*. Mes travaux ont porté sur quatre espèces de pommiers sauvages distribuées à travers l'Eurasie (*Malus orientalis* (Caucase), *Malus sieversii* (Asie Centrale), *Malus sylvestris* (Europe), et *Malus baccata* (Sibérie)) et sur la seule espèce domestiquée du genre, *Malus domestica*. Cette thèse s'est articulée en quatre parties visant respectivement à inférer : (*i*) l'histoire de la domestication du pommier cultivé depuis son centre d'origine en Asie Centrale, (*ii*) l'histoire de la recolonisation post-glaciaire du pommier sauvage Européen (*M. sylvestris*), (*iii*) les histoires de spéciation entre les cinq espèces de *Malus,* (*iv*) les hybridations interspécifiques et les capacités de dispersion des trois principaux contributeurs (*M. sylvestris*, *M. sieversii* et *M. orientalis*) au génome du pommier cultivé. L'étude des mécanismes de diversification artificielle montre que les processus de domestication sont originaux chez cet arbre fruitier, de par l'absence de goulet d'étranglement et l'existence d'introgressions post-domestication fréquentes par une autre espèce sauvage (*M. sylvestris*) que l'espèce ancestrale (*M. sieversii)*. L'étude des processus de diversification naturelle (phylogéographie, spéciation et structure des populations) révèlent de grandes tailles de populations, de forts flux de gènes et de faibles structures génétiques spatiales chez chacune des espèces. Cette thèse a aussi révélé de forts taux d'hybridations interspécifiques, en particulier de fortes introgressions des espèces de pommiers sauvages par le pommier cultivé en Europe et en Asie Centrale. Cette étude a permis l'amélioration des connaissances de la structuration des populations de pommiers sauvages ayant contribué au génome du pommier cultivé ainsi que de l'étendue des hybridations du pommier cultivé avec les espèces sauvages. Ces travaux revêtent une grande importance autant pour la conservation des pommiers sauvages, pour le maintien de leur intégrité dans des habitats fragmentés que pour l'amélioration variétale du pommier domestiqué.

Abstract: Despite its economic, cultural and historical importance, few studies have investigated the evolutionary history of the domesticated apple (*Malus domestica*) as well as those of its wild relatives. Using new population genetic approaches (*approximate Bayesian computation*) with microsatellites and nuclear sequences, this thesis aimed at unravelling, at different evolutionary scales (phylogeography, speciation, domestication), the natural and artificial diversification processes at play in the *Malus* genus. My research focused on the four wild apple species distributed across Eurasia (*Malus orientalis* (Caucasus), *Malus sieversii* (Central Asia), *Malus sylvestris* (Europe), and *Malus baccata* (Siberia)) and on the single domesticated apple species in the genus, *Malus domestica*. This thesis was divided into four parts: (*i*) domestication history of the cultivated apple, from its origin in Central Asia to Europe, (*ii*) post-glacial recolonization history of the European crabapple (*M. sylvestris*), (*iii*) the history of speciation among the five *Malus* species*,* (*iv*) crop-to-wild gene flow and dispersal capacities of the closest wild relative species (*M. sylvestris*, *M. sieversii* and *M. orientalis*). By investigating artificial diversification, we evidenced unique processes of domestication in this fruit tree, with no bottleneck and with extensive post-domestication introgressions by another wild species (*M. sylvestris*) than the ancestral progenitor (*M. sieversii)*. Natural diversification patterns (phylogeography, speciation and population structure) revealed large effective population sizes, high dispersal capacities and weak spatial genetic structures. This thesis also revealed high levels of interspecific hybridizations, particularly high level of crop-to-wild gene flow in Europe and Central Asia. This study extended our knowledge about population structures for wild species that contributed to the cultivated apple genome, as well as the extent of hybridization rates. This work is essential for the conservation of wild apple populations, the integrity maintenance of wild species facing fragmentation and future breeding programs concerning the domesticated apple.