Effets de la sélection naturelle et de l'histoire démographique sur les patrons de polymorphisme nucléaire

—

Comparaisons interspécifiques chez *Arabidopsis halleri* et *A. lyrata* entre le fond génomique et deux régions cibles de la sélection

THÈSE

présentée et soutenue publiquement le 17 Décembre 2010

pour l'obtention du

Doctorat de l'Université de Lille 1 – Sciences et Technologies

(spécialité Biologie Évolutive et Écologie)

par

Camille Roux

Composition du jury

Président :	Pascal Touzet, Professeur	GEPV, Université de Lille 1
Rapporteurs :	Nicolas Bierne, Chargé de Recherche CNRS	Institut des Sciences de l'Evolution, Sète
	Olivier François, Professeur	TIMC, Grenoble INP
Examinateur :	Frantz Depaulis, Chargé de Recherche CNRS	Ecole Normale Supérieure, Paris
Directeurs :	Xavier Vekemans, Professeur	GEPV, Université de Lille 1
	Vincent Castric, Chargé de Recherche CNRS	GEPV, Université de Lille 1

UNIVERSITÉ DE LILLE 1 – SCIENCES ET TECHNOLOGIES
ÉCOLE DOCTORALE SCIENCES DE LA MATIÈRE, DU RAYONNEMENT ET DE L'ENVIRONNEMENT
Laboratoire de Génétique et Evolution des Populations Végétales — UMR CNRS 8016
U.F.R. de Biologie – Bât. SN2 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 20 43 40 24 – Télécopie : +33 (0)3 20 43 69 79

Table des matières

Table des matières

Remerciements

Par ordre chronologique, j'ai une pensée :

pour mes parents qui ont toujours respecté une grande liberté de choix d'orientation de vie chez leurs enfants, malgré le manque parfois inquiétant de communication de ma part.

pour ma soeur Charlotte et son compagnon Antoine qui ont embelli mes années à Orsay, et celles qui ont suivi.

pour Solène qui est un exemple d'indépendance et d'exigence pour moi (L'Anar-cheese vaincra un jour).

pour les diverses rencontres de collègues de fac et de cité-u qui ont été marquantes (particulièrement Tristan, Manu, Linda, Coeur, Chouchou, Laurent, Fany, Nolwenn, Benjamin, et Laurine).

pour Johnny.H, de m'avoir condamné à raconter 1,000 fois la même annecdote.

pour M.G qui n'a put finir sa thèse.

pour Jean Vidal et Graham Noctor qui en licence m'ont fait préférer la recherche dans le végétal, sans aucun regret lorsque je repense à mes choix d'orientation.

pour Patrice Meimoun, maître de stage en M1, qui m'a décoincé à la paillasse (ouille !).

pour Sébastien Thomine, maître de stage en M2, qui a été honnête sur les défauts que j'ai à corriger. ça m'a accompagné pendant toute ma thèse.

pour Viviane Lanquar et son "le mieux est l'ennemi du bien" que j'essaie d'appliquer depuis 4 ans, ce qui est plus compliqué qu'en apparence.

pour l'équipe Auto-incompatibilité du GEPV de m'avoir donné ma chance alors que bon...ce fut surtout une énorme chance pour moi de découvrir une autre biologie que celle que je connaissais. C'était donc une opportunité révélatrice/déterminante/marquante/décisive pour moi.

particulière pour Xavier qui me fait penser que j'ai encore beaucoup de progrès à faire avant de prétendre à un poste de titulaire un jour.

pour le GEPV dans son ensemble qui forme un cadre agréable où pratiquer de la recherche.

pour les réunions du GDR, remotivantes.

pour les thésards du GEPV (Isa, J.B, Camille-bis, Benon et évidemment AUDE DARRACQ qui comme chacun sait, adore qu'on la mette en avant. J'aurai voulu mettre fin à la malédiction de Aude pour les remerciements foireux mais aargh....A cette liste j'intègre volontier Lucy, l'altérophyle de l'endive, et Fafa-Gaga que j'ai du mal à prendre pour un non-doctorant).

pour Marine, et pour sa patience.

Remerciements

Introduction

Le processus de spéciation a été représenté pendant tout le XXème siècle comme une succession de séparations dichotomiques de lignées évolutives. Cette vision dichotomique de l'évolution émerge dès la publication du livre de Charles Darwin "on the origin of species" illustré par son unique figure (Figure 1) [Darwin, 1859].

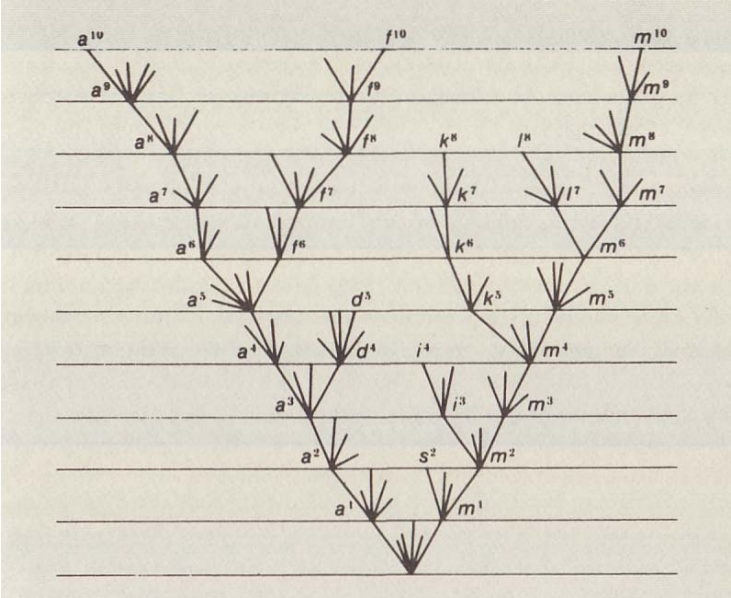

FIG. 1 – Vision dichotomique de la diversité au sein d'un même genre selon [Darwin, 1859].

Si un tel patron respecte les relations évolutives entre les genres du vivant, l'accumulation sans précédent de données de polymorphisme de séquences d'ADN pour des espèces proches a révélé des relations phylogénétiques plus complexes entre les génomes de ces espèces. D'un point de vue génomique, la vision dichotomique de l'évolution signifierait que les espèces seraient des populations isolées de génomes au sein desquels les séquences orthologues à chaque locus seraient phylogénétiquement plus proches au sein de la même population que de celles présentes dans une autre population. Sans les travaux du dernier siècle et demi qui ont conduit au développement de nouvelles voies de recherches s'intéressant aux structures des génomes et aux comportements des mutations au sein des populations, il aurait été difficile d'imaginer la spéciation autrement que par ce processus de différenciation homogène, affectant à l'identique l'ensemble des génomes. La possibilité récente de comparer les génomes nucléaires de deux espèces différentes marquera d'abord tout observateur par les variations génomiques locales des niveaux de divergence. C'est ainsi que les publications successives des génomes de

l'Homme [Consortium, 2001, Venter et al., 2001, Sequencing, 2005] ont permis de révéler une hétérogénéité des niveaux de divergence entre ces deux espèces à l'échelle génomique (Figure 2). Cette hétérogénéité du niveau de divergence observé le long des génomes est la résultante de nombreux facteurs et de nombreuses forces évolutives impliquant à la fois des processus neutralistes et des processus adaptatifs.

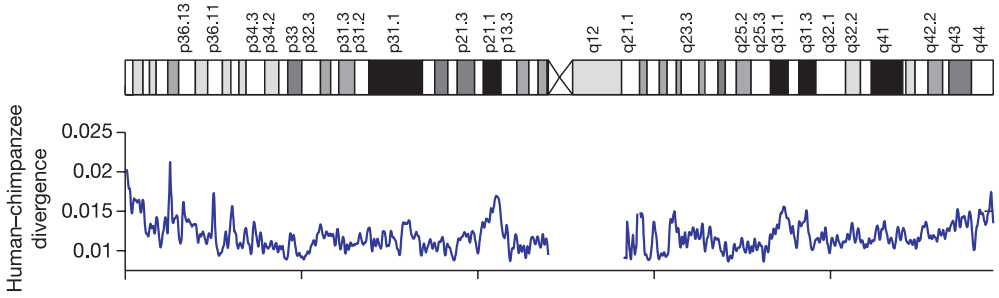

FIG. 2 – Variations des taux de divergence entre l'Homme et le chimpanzé le long du chromosome 1, tiré de [Sequencing, 2005].

1 Effectifs efficaces des populations et évolution de la complexité génomique.

Depuis les travaux de Kimura dans les années 60 sur la théorie neutraliste de l'évolution, nous avons beaucoup appris sur l'importance des effets de l'effectif efficace des populations (N) sur le comportement des patrons de polymorphisme et sur l'évolution des structures génomiques. Comme pour l'émergence de chaque nouvelle voie de recherche, celle de la théorie neutraliste de l'évolution fut immédiatement accompagnée de débats vigoureux opposant les chantres du "tout adaptation" à ceux du "tout neutralisme". Débats excessifs car les importances relatives de la sélection naturelle et de la dérive génétique dépendent étroitement d'un facteur central qui est la valeur de N. En effet, ces forces n'influent pas sur l'évolution des fréquences alléliques au sein des populations avec les mêmes intensités pour toutes les valeurs que N peut prendre. A partir du principe où l'influence des effets stochastiques dans une population est proportionnelle à $1/(2N)$, il est intuitif de penser que la dérive prédomine sur la sélection naturelle au sein des populations de faibles effectifs efficaces. Réciproquement, la sélection naturelle influera plus sur les variations de fréquence des variants pour des populations caractérisées par de fortes valeurs de N. Si l'histoire démographique d'une population, son système de reproduction ainsi que la fraction de la population réellement impliquée dans l'effort reproductif sont des éléments qui déterminent la valeur de N, Lynch et Conery [Lynch and Conery, 2003] ont montré qu'une forte diminution des valeurs de N a accompagné l'évolution des eucaryotes pluricellulaires depuis les procaryotes. La complexification des génomes qui a suivi cette évolution a été décrite comme une conséquence de la réduction des effectifs des populations. Ici, une complexification génomique sous-entend une augmentation de la taille des génomes par une augmentation du nombre de gènes ou de copies de gènes, par une plus grande quantité d'ADN intronique, et par l'accumulation d'éléments mobiles transposables. Ainsi, l'accumulation d'ADN intronique et d'éléments transposables dans les génomes apparait comme une conséquence directe de fixations par dérive

génétique de mutations faiblement délétères dans des populations à faibles effectifs, plutôt qu'à des processus adaptatifs. Bien que les introns présentent des signes de contraintes évolutives plus fortes que les positions synonymes dans les séquences codantes [Andolfatto, 2005], l'apparition de nouveaux introns dans un gène entraine un désavantage sélectif menant à une élimination quasi systématique chez les procaryotes. Les variations de taille des génomes par amplification de familles multigéniques est également expliquée par la probabilité de fixation plus élevée de gènes dupliqués pour de faibles valeurs de N [Force et al., 1999]. Par redondance fonctionnelle, une partie des paralogues néoformés peut accumuler des mutations inhibant leur fonction initiale, puis évoluer soit vers une pseudo-fonctionnalisation, soit vers une sous-fonctionnalisation. Finalement, il est impressionnant de voir l'écart qu'il y a entre la vision finaliste par laquelle la complexification des génomes m'a été enseignée lors de mes premières années universitaires ("une grande complexité des génomes est avantageuse car permet une régulation adaptative plus fine"), et l'importance des facteurs stochastiques en liaison avec la valeur du paramètre N dans l'évolution de cette complexité.

2 Effectifs efficaces des populations et discordances phylogénétiques.

Placé dans un contexte de spéciation récente, la dérive agit sur l'hétérogénéité de la divergence le long des génomes en fixant aléatoirement des allèles qui étaient en ségrégation dans une population ancestrale. Quand une espèce avec un niveau de polymorphisme donné est séparée en deux sous-populations, une partie du polymorphisme neutre ancestral est transmis aux deux populations filles. Ainsi, à la génération qui a immédiatement suivi l'évènement de spéciation, le polymorphisme ancestral est présent à son maximum chez les deux nouvelles populations. A cette étape de la spéciation, le polymorphisme ancestral est donc principalement à l'origine du partage de polymorphisme entre les deux populations. Ici, nous parlons de polymorphisme partagé lorsque deux espèces partagent le même état bi-allélique à la même position génomique au niveau nucléotidique. En supposant un modèle d'isolement strict, la rétention du polymorphisme ancestral chez les deux populations filles diminuera au cours du temps par effet de la dérive génétique. Si au sein de chaque lignée il est possible de calculer la distribution du temps de fixation pour un allèle neutre [Kimura, 1955], le même modèle appliqué à deux lignées en cours de divergence permet d'estimer la distribution du temps de fixation d'un allèle dans l'une ou l'autre des deux lignées, conduisant ainsi à la perte de ce polymorphisme partagé (Figure 3) [Clark, 1997].

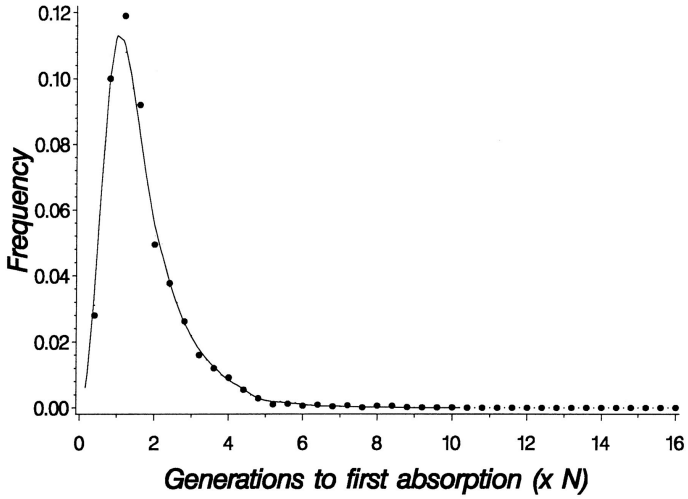

FIG. 3 – Densité de probabilité du temps de perte du polymorphisme partagé neutre par effet de la dérive génétique. La ligne représente la résolution numérique à partir de la densité du temps d'absorption intégrée par Kimura, tiré de [Clark, 1997].

Cette probabilité de perte du polymorphisme ancestral est une fonction du temps depuis l'évènement de spéciation exprimée en N générations où N est l'effectif efficace des populations. Dans un modèle de Wright et Fisher où la population ancestrale et les deux populations filles ont la même valeur du paramètre N, le temps moyen de perte du polymorphisme partagé est relativement court, environ 1.7 N générations. Cependant, la distribution du temps de perte du polymorphisme partagé est très étiré vers la droite. Ainsi, si 50% des locus échantillonnés dans les deux populations de génomes sont réciproquement monophylétiques au bout de 4 à 7 N générations après le début de l'isolement des deux lignées en moyenne, 5% des locus n'auront pas encore atteint le statut d' "espèces phylogénétiques" pour des temps plus anciens de 9 à 12 N générations [Hudson and Coyne, 2002]. C'est principalement par ce processus de tri incomplet de lignées que s'explique l'hétérogénéité des relations phylogénétiques le long des génomes de l'humain, du chimpanzé et du gorille (Figure 4).

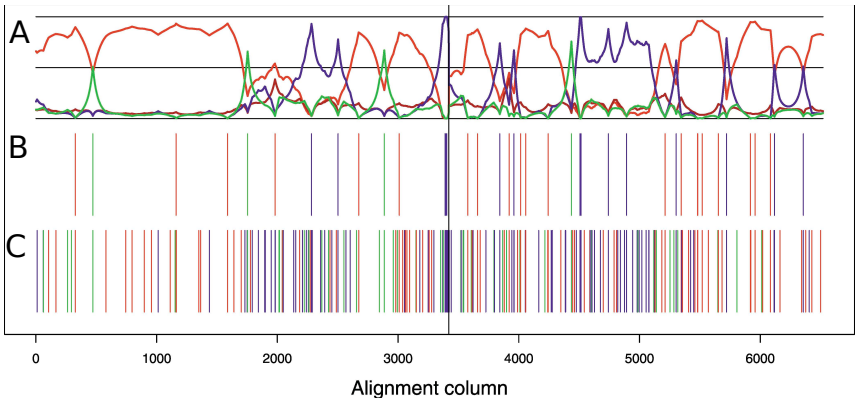

FIG. 4 – Hétérogénéité des généalogies le long du génome humain.
(A) Probabilités a posteriori des généalogies : Homme-chimpanzé (rouge), Homme-gorille (bleu) et chimpanzé-gorille (vert) sur une région de 100kb.
(B) Succession des généalogies fortement supportées en utilisant un groupe externe, coloré comme en (A).
(C) Succession des généalogies sans groupe externe. D'après [Hobolth et al., 2007].

Une analyse de plusieurs fragments génomiques séquencés chez l'Homme, le chimpanzé et le gorille a en effet révélé qu'une fraction de ces génomes n'était pas en cohérence avec la phylogénie des espèces [Patterson et al., 2006]. Puisque la période séparant les deux évènements de spéciation successifs chez l'Homme, le chimpanzé et le gorille est relativement courte par rapport aux effectifs efficaces des populations, toutes les régions génomiques chez l'humain n'ont pas eu le temps de coalescer avec celles du chimpanzé avant la spéciation du gorille. C'est par cette stochasticité des processus de coalescence que certaines régions génomiques chez l'Homme (ou chez le chimpanzé) sont plus proches du gorille que du chimpanzé (ou de l'Homme) (Figure 5).

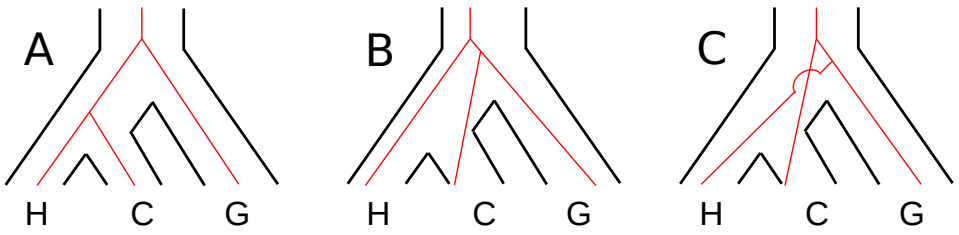

FIG. 5 – Conséquences de la stochasticité de la dérive sur les phylogénies de gènes menant à : (A) des généalogies cohérentes ou (B et C) incohérentes avec la généalogie des espèces.

Une telle discordance de relations phylogénétiques est liée à la recombinaison qui a pour effet de casser la corrélation entre les histoires évolutives le long des génomes [Hobolth et al., 2007].

3 Dérive et isolement-reproducteur.

Si nous avons exposé précédemment les effets de la stochasticité des processus de coalescence couplée à ceux de la recombinaison génomique sur l'hétérogénéité locale des niveaux de divergence entre génomes, la divergence génomique moyenne tend elle à augmenter linéairement avec le

temps. Son augmentation est la résultante de la fixation différentielle d'allèles ancestraux par dérive génétique, de l'apparition aléatoire de mutations spécifiques dans les lignées suivies dans certains cas de la fixation des allèles dérivés dans l'une des deux lignées (Figure 6).

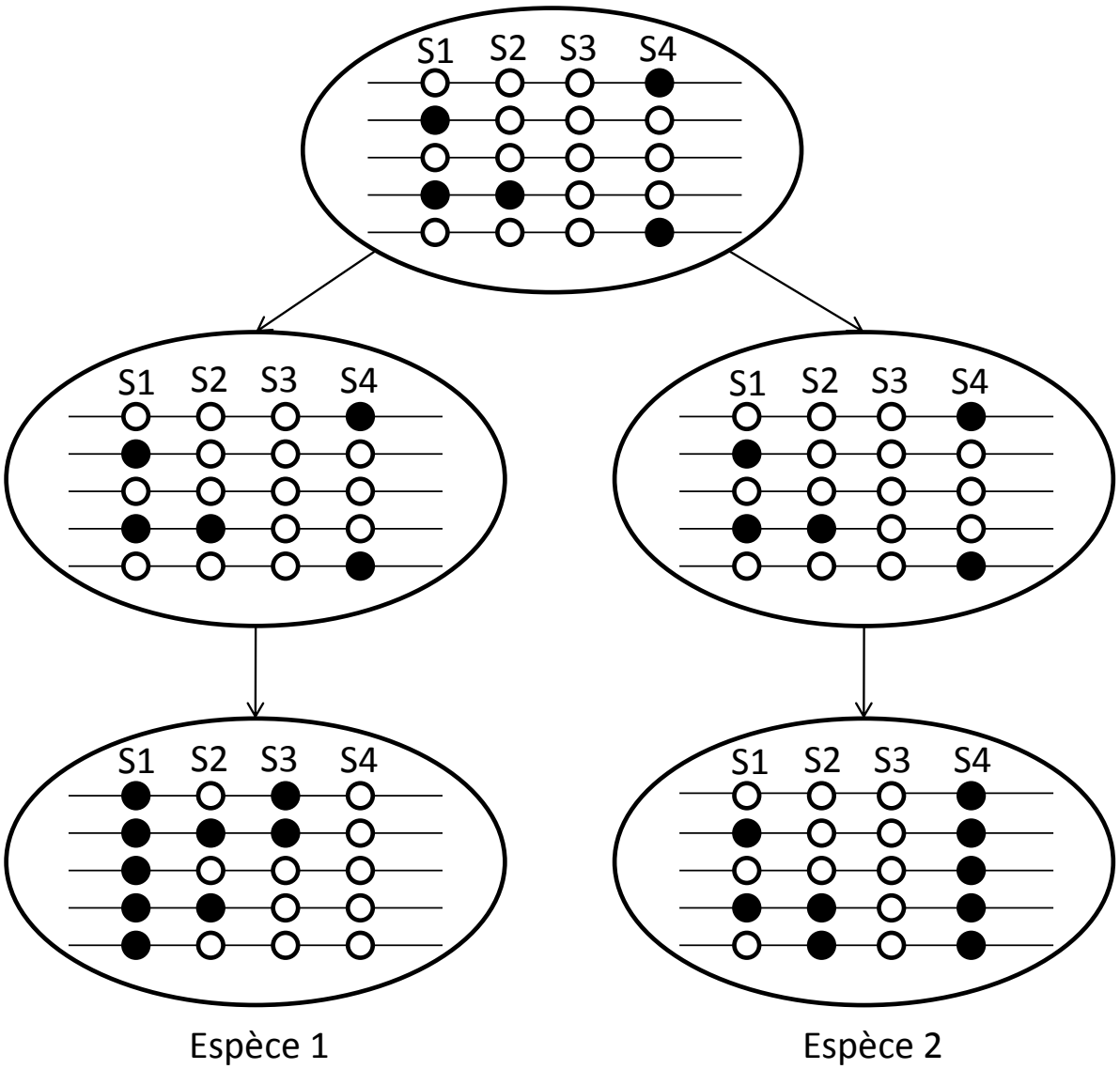

FIG. 6 – Evolution des patrons de polymorphisme neutre depuis la subdivision d'une population ancestrale en deux populations filles isolées.
Les ronds blancs et noirs représentent respectivement les allèles ancestraux et les allèles dérivés. Avant l'évènement de spéciation un état bi-allélique est observé au niveau des positions S1, S2 et S4. A la génération qui suit immédiatement l'évènement de spéciation, le polymorphisme ancestral est redistribué dans les nouvelles sous-populations. Par dérive génétique, l'allèle dérivé en ségrégation dans la population ancestrale peut être fixé chez l'espèce 1 mais rester en ségrégation chez l'espèce 2 (S1). Le polymorphisme ancestral peut être maintenu chez les deux espèces formant du polymorphisme partagé (S2). Avec le temps, des mutations exclusives à chacune des espèces apparaissent aléatoirement dans les populations (S3). Fixation de deux allèles différents au niveau d'une même position à partir d'un polymorphisme ancestral (S4), cette différence fixée peut être observée pour une mutation exclusive fixée par dérive.

En supposant que chaque substitution différentiant deux lignées pourrait contribuer à la diminution de la valeur sélective des hybrides à partir des deux lignées, Allen Orr a modélisé l'accumulation des incompatibilités de Dobzhansky-Muller au cours du temps (Figure 7).

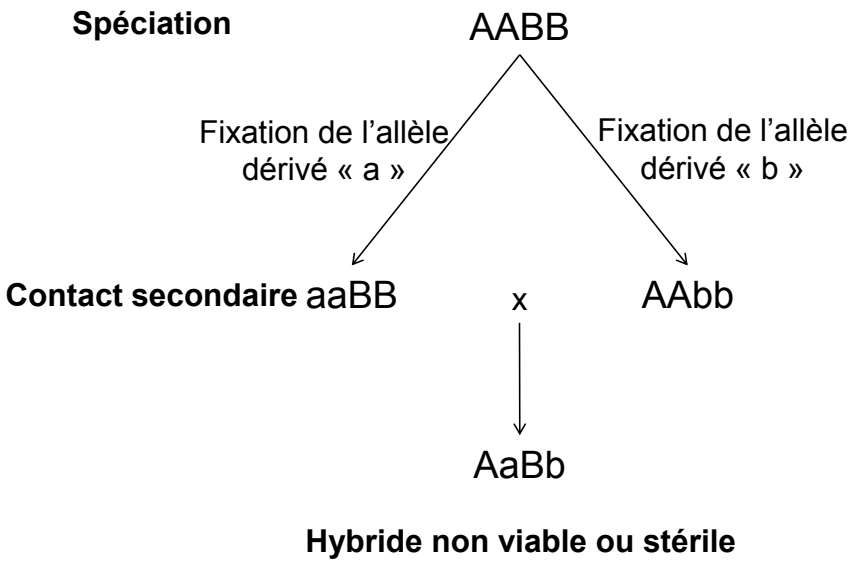

FIG. 7 – Incompatibilités de Dobzhansky-Muller : après la séparation d'une population ancestrale en deux populations filles, deux gènes peuvent évoluer indépendamment jusqu'à la fixation d'allèles dérivés dans une des deux populations au niveau de chacun des gènes. Suite à un contact secondaire suivi d'hybridations entre individus des deux espèces filles, de nouvelles interactions épistatiques apparaissent chez les hybrides et peuvent résulter soit à une stérilité des hybrides, soit à leur létalité.

Plus le temps depuis la séparation des lignées isolées est grand, plus la probabilité que des évènements de mutation affectent différents gènes augmente et de ce fait, un plus grand nombre de combinaisons d'allèles est attendu. Comme le nombre de combinaisons alléliques augmente après chaque mutation dans les deux lignées, le nombre d'incompatibilités contribuant à l'isolement post-zygotique peut augmenter aussi rapidement que le carré du temps de spéciation, contrastant avec la divergence génomique moyenne qui évolue plutôt de façon linéaire avec le temps [Orr, 1995]. Ainsi, des processus stochastiques liés à l'accumulation de substitutions peuvent générer des barrières d'isolement reproducteur. Puisque l'accumulation de différences nucléotidiques fixées entre les espèces peut être traduite en unités temporelles, il est permis de suggérer une hétérogénéité génomique pour la densité en incompatibilités de type Dobzhansky-Muller si elle est effectivement corrélée avec le temps de divergence (Figure 8) sans que ces incompatibilités aient directement un effet sur le niveau de divergence nucléotidique entre deux populations allopatriques.

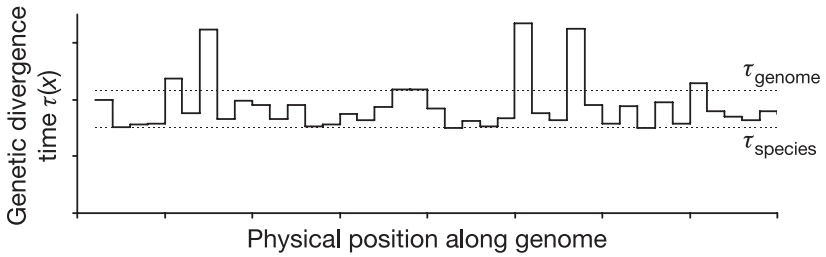

FIG. 8 – Temps de divergence génétique $\tau(x)$ en fonction de la position génomique. $\tau_{species}$ est le temps de spéciation entre deux espèces. τ_{genome} est la moyenne génomique pour la divergence, elle excède toujours tspecies. Tiré de [Patterson et al., 2006].

4 Migration et divergence.

Au cours des premières générations qui suivent la séparation de deux lignées évolutives, il est attendu que les barrières reproductives ne soient pas suffisamment accumulées dans les deux populations de génomes pour contre-sélectionner systématiquement les hybrides potentiels (sauf si évidemment la mise en place soudaine de telles barrières soit elle-même à l'origine de la spéciation). Ainsi, si les aires de distributions géographiques des lignées se chevauchent au moins partiellement, et si les périodes reproductives sont synchronisées, alors des flux de gènes peuvent avoir lieu avec comme conséquence la formation d'hybrides entre les deux lignées. L'histoire démographique des flux de gènes joue un rôle important sur la direction et l'efficacité de la sélection naturelle ; en effet, des taux de migration élevés vont diminuer le potentiel pour l'adaptation locale [Ronce and Kirkpatrick, 2001]. En absence complète d'incompatibilité de Dobzhansky-Muller, les évènements de migrations vont tendre à empêcher la fixation de différences nucléotidiques entre deux populations de génomes. A une position génomique donnée, le transfert d'un allèle ancestral d'une population vers une population ayant fixé un allèle dérivé convertira une différence fixée entre deux populations en un site exclusivement polymorphe dans la seconde population (Figure 9).

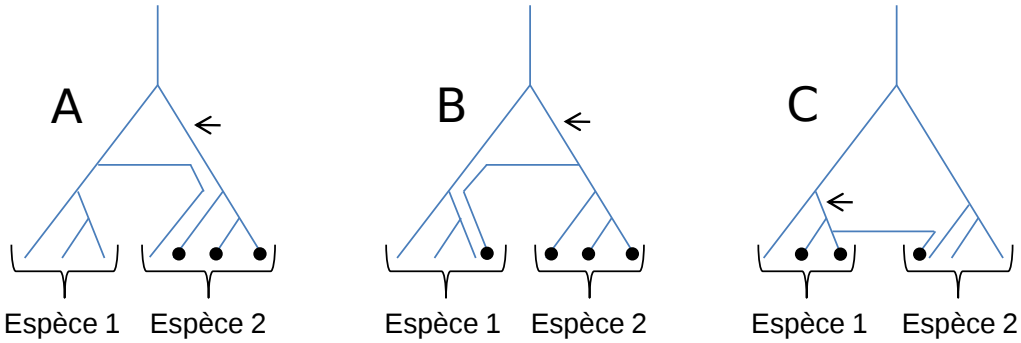

FIG. 9 – Effets sur le patron de polymorphisme de la migration entre deux espèces en divergence. La flèche indique une branche de l'arbre le long de laquelle apparait une mutation conduisant à un allèle dérivé (rond noir).
(A) La mutation a lieu le long d'une branche menant à une différence fixée entre les deux espèces. La migration transfert un allèle ancestrale depuis l'espèce 1 vers l'espèce 2 et convertie une différence fixée en un polymorphisme exclusif à l'espèce 2.
(B) La mutation a lieu le long d'une branche menant à une différence fixée entre les deux espèces. La migration transfert un allèle dérivé depuis l'espèce 2 vers l'espèce 1 et convertie une différence fixée en un polymorphisme d'origine ancestral chez l'espèce 1.
(C) La mutation a lieu le long d'une branche menant à du polymorphisme exclusif à l'espèce 1. La migration transfert un allèle dérivé depuis l'espèce 1 vers l'espèce 2 et convertie un polymorphisme exclusif à l'espèce 1 en un polymorphisme partagé entre les deux espèces.

La migration génère donc du polymorphisme exclusif au détriment des différences fixées, mais peut également générer du polymorphisme partagé à partir des allèles dérivés en ségrégation dans une population Figure (9-C). En revanche, les flux de gènes seront contre-sélectionnés au niveau des régions génomiques liées aux incompatibilités de Dobzhansky-Muller sans que ces dernières soient forcément issues d'un processus adaptatif. Ce patron hétérogène des échanges de gènes établit la nature mosaïque des génomes en cours de divergence, avec des régions qui maintiennent du partage de polymorphisme et d'autres qui présentent des différences fixées dans chacune des populations [I. Wu, 2001].

5 La mutation ponctuelle et l'hétérogénéité de la divergence.

Il a été décrit chez les mammifères que les taux de mutation nucléotidique varient entre les chromosomes [Gaffney and Keightley, 2005] (Figure 10). De manière générale, les erreurs de

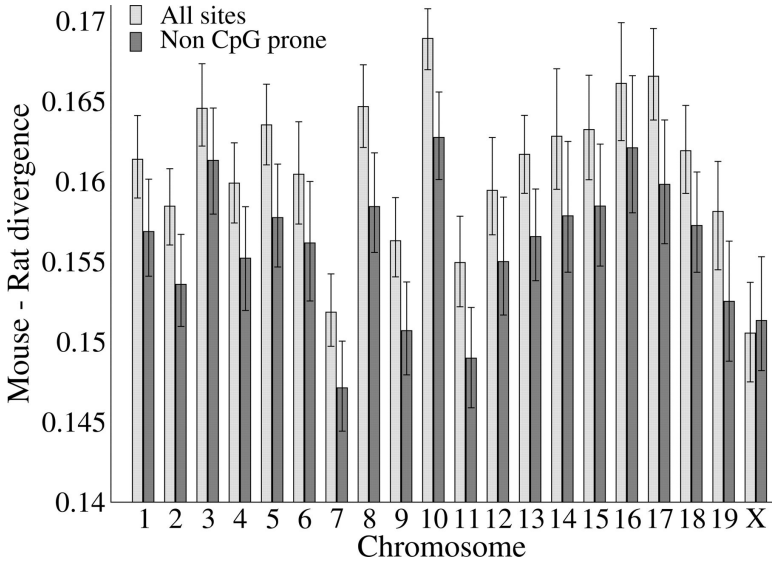

FIG. 10 – Moyennes estimées des taux de substitution pour l'ensemble des sites et pour les sites non-CpG de chaque chromosomes de la souris. Tiré de [Gaffney and Keightley, 2005].

réplication de l'ADN sont connues pour former la principale source de mutations, mais l'identification des mécanismes expliquant la variation des taux de mutation chromosomique reste encore anecdotique. Par exemple, les taux de mutations des chromosomes Y des mammifères sont plus élevés que ceux des chromosomes X du fait du plus grand nombre de cycles cellulaires dans les lignées germinales chez les mâles [Makova and Li, 2002]. Mais les facteurs influençant les taux de mutation des chromosomes autosomaux sont encore trop méconnus pour déterminer l'impact qu'auraient de telles différences sur l'hétérogénéité des niveaux de divergence entre lignées évolutives. Pour le moment nous avons plus de connaissances sur l'implication du contexte génomique local dans les variations du taux de mutation. Le facteur le mieux identifié à ce jour est le rôle des îlots CpG ("p" désigne le pont phosphate entre une cytosine et une guanine). Brièvement, les îlots CpG sont souvent les cibles de méthyl-transférases pour réguler l'expression des gènes [Merlo et al., 1995]. La conversion d'une cytosine en 5-méthyle-cytosine stimule localement l'activité des histones-désacétylases et des histones-méthyle-transférases qui vont respectivement retirer un groupement acétyle puis transférer un groupement méthyle en position N-terminale des histones. L'effet d'un groupement acétyle en position N-terminale d'une protéine histone est d'empêcher l'interaction de cette histone avec l'ADN. L'acétylation des histones entraine donc la décompactation locale de l'ADN, libérant ainsi l'accès vers les sites de régulation dans la région promotrice d'un gène par le complexe d'initiation de la transcription. La méthylation des îlots CpG dans les régions promotrices des gènes aura l'effet inverse sur la compactation de l'ADN en éteignant localement la transcription [Yang and Seto, 2007]. Si la méthylation de l'ADN au niveau des îlots CpG a un rôle important dans la régulation de l'ex-

pression des gènes, la très grande instabilité des cytosines méthylées a pour effet d'augmenter localement les taux de transitions nucléotidiques de C vers T et de G vers A qui apparaissent comme des sites hypermutatoires [Arndt et al., 2003]. Cependant, il semble qu'il existe un niveau de complexité dans les déterminants pour les taux de mutations qui ne dépende pas des séquences [Hodgkinson et al., 2009]. En alignant 309,158 séquences de 81 paires de bases centrées sur un SNP obtenues chez le chimpanzé, les auteurs ont montré que le nombre de mutations récurrentes chez l'humain est environ trois fois plus élevé que l'attendu selon l'hypothèse nulle que les mutations affecteraient avec la même probabilité les 81 nucléotides de la fenêtre d'observation (Figure 11). L'origine de cet excès de mutations récurrentes affectant les génomes est encore inconnu, nous savons seulement à ce jour qu'il n'est pas la conséquence de l'hypermutabilité des îlots CpG, ni celle de la rétention du polymorphisme ancestral, ni enfin causée par des effets sélectifs.

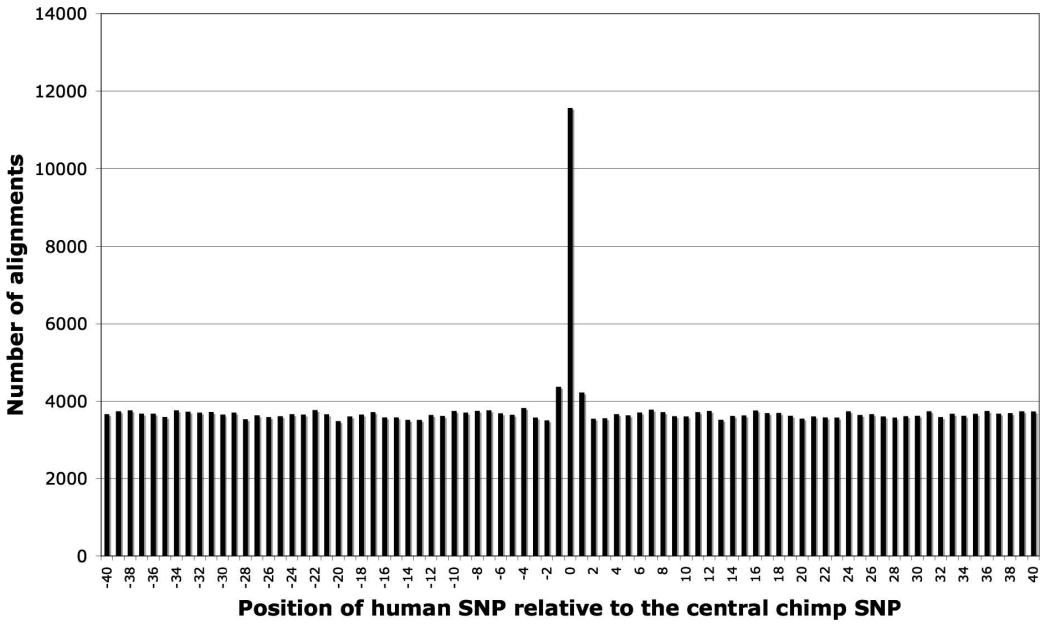

FIG. 11 – Nombre de SNPs humains pour chaque site des alignements Homme-chimpanzé utilisés. D'après [Hodgkinson et al., 2009].

6 La sélection directionnelle.

Lorsque deux populations isolées évoluent dans des environnements écologiques contrastés, la sélection peut agir dans des directions différentes chez les deux populations, favorisant des phénotypes opposés voire extrêmes. Cette sélection divergente peut provenir de différences environnementales telles que des différences dans les climats, dans les ressources disponibles, ou bien dans les prédateurs ainsi que les compétiteurs présents dans chacun des deux milieux. Durant les premières générations qui suivent la séparation de deux lignées dans deux environnements écologiquement différents, les régions génomiques contenant les QTL majeurs pour les traits différemment sélectionnés deviennent fixées dans chacune des deux populations plus rapidement qu'attendu sous la neutralité [Lewontin and Krakauer, 1973]. La sélection agit de manière "locus spécifique" en entrainant une augmentation de la fréquence de l'allèle favorable dans un environnement donné jusqu'à fixation, ainsi qu'une augmenta-

tion en fréquence des allèles neutres qui lui sont physiquement liés (phénomène d'"autostop moléculaire") [Maynard Smith and Haigh, 1974]. Une telle purge du polymorphisme ancestral au niveau de ces régions génomiques a pour conséquence une réduction accélérée du nombre de sites polymorphes partagés entre les deux lignées évolutives, et de ce fait, augmente localement le niveau de différenciation par rapport au fond génomique. De plus, les changements génétiques qui augmentent l'adaptation peuvent indirectement accélérer l'évolution de l'isolement reproductif entre les deux populations [Coyne and Orr, 2004]. Puisque les QTL majeurs impliqués dans un trait clé divergent rapidement sous l'effet de la sélection, ils deviennent résistants aux flux de gènes. Ainsi, les régions en forts déséquilibres de liaisons avec de tels QTL peuvent accumuler des facteurs d'isolement reproductifs comme des sous-produits de l'interruption de flux de gènes au niveau génomique local. Ce développement indirect de l'isolement reproducteur s'effectue plus rapidement que dans les régions non liées à ces mêmes QTL où les flux de gènes ne sont pas initialement contre sélectionnés [Via and West, 2008]. Comme dans le cas des incompatibilités de Dobzhansky-Muller, les flux de gènes vont entrainer une hétérogénéité génomique de la divergence entre les espèces naissantes, avec pour différence que les régions fortement divergentes résultent ici d'un effet d'entrainement moléculaire par des allèles positivement sélectionnés. Mais la sélection peut également faciliter l'introgression de certaines régions génomiques si elles sont avantageuses pour l'espèce receveuse. De nombreux cas d'introgressions asymétriques depuis des plantes cultivées vers des fonds génomiques sauvages ont ainsi été observés pour des transgènes [Stewart et al., 2003]. Par exemple, la résistance à un herbicide ou à un pathogène conférée par un transgène à une plante cultivée peut en effet faciliter l'introgression de ce transgène chez une espèce proche sauvage, en fonction de la pression de sélection dans l'environnement naturel à proximité de la zone cultivée. Ainsi, les régions génomiques impliquées dans l'introgression adaptative peuvent présenter des niveaux de divergence inférieurs au fond génomique moyen et générer de l'hétérogénéité dans la divergence génomique entre espèces.

7 La sélection balancée.

En absence de sélection naturelle, le polymorphisme observé à une position génomique donnée finira par disparaitre comme la conséquence de la fluctuation des fréquences alléliques au cours du temps sous l'effet de la dérive génétique, jusqu'à la fixation d'un allèle dans une population. Si le temps moyen de fixation est de $4.N$ générations pour un locus neutre, la sélection naturelle peut dans certains cas agir pour maintenir la diversité allélique à un locus donné, les allèles fonctionnellement différents persistant ainsi plus longtemps dans une population. Cette sélection, dite "balancée" résulte principalement de trois processus différents identifiés : la superdominance, la sélection hétérogène dans l'espace et la sélection fréquence-dépendante. La superdominance décrit une situation où les génotypes hétérozygotes ont un avantage sélectif sur les homozygotes. L'hypothèse de la superdominance est souvent invoquée pour expliquer les niveaux élevés de polymorphisme observés pour les gènes du complexe majeur d'histocompatibilité (MHC), impliqués dans l'initiation de la réponse immunitaire chez les vertébrés [Doherty and Zinkernagel, 1975]. Selon ce modèle, les individus hétérozygotes ont un avantage sélectif sur les individus homozygotes en ayant un spectre d'identification de pathogènes et de parasites conférés par deux allèles différents plus large que chez les individus homozygotes [Kekalainen et al., 2009, Penn et al., 2002, Oliver et al., 2009]. La sélection balancée de type fréquence-dépendante maintient la diversité observée au locus sélectionné en conférant un avantage sélectif aux génotypes les plus rares. La diversité allélique peut être maintenue

sur de longues périodes évolutives, et l'âge des allèles peut dépasser l'âge d'espèces proches [Muirhead et al., 2002]. Ainsi, il est plus probable d'observer du polymorphisme partagé entre deux espèces différentes au niveau d'un locus soumis à une telle sélection balancée qu'au niveau du fond génomique moyen. Compte tenu de l'étendue des généalogies des locus soumis à sélection balancée [Takahata, 1990], la divergence interspécifique mesurée au niveau de ces locus ne dépend plus essentiellement du temps depuis l'évènement de spéciation mais des classes alléliques échantillonnées pour mesurer la divergence nucléotidique. La divergence interspécifique au sein d'une même classe allélique peut en effet être inférieure à la divergence mesurée entre deux allèles fonctionnellement différents au niveau intraspécifique.

8 Prise en compte de l'histoire démographique des populations.

Lors des premières études de génomique des populations examinant un locus ou un faible nombre de locus, les observations de patrons de polymorphisme qui étaient incompatibles avec les attendus selon le modèle standard neutre (population panmictique ayant une taille constante) étaient souvent interprétées comme des signes de la sélection naturelle plutôt que comme des écarts aux hypothèses du modèle démographique par la population considérée [Wright and Gaut, 2005]. Les effets de la subdivision des populations ainsi que ceux des changements de tailles sont désormais connus pour modifier les patrons de polymorphisme neutres et entrainer des interprétations erronées des tests de neutralité couramment utilisés. De tels tests pratiqués au sein d'une population échangeant des gènes avec une population non échantillonnée peuvent être biaisés par une élévation de la variance des niveaux de polymorphisme entre les locus causée par la migration [Stadler et al., 2008]. Cette variance entre les gènes dans les niveaux de polymorphismes peut également être la conséquence de changement d'effectif des populations et en particulier celle d'un goulot d'étranglement. Le risque d'une variance interlocus élevée est d'entrainer une élévation du nombre de faux positifs à des tests de neutralités supposant le modèle standard neutre [Andolfatto, 2008, Wright and Gaut, 2005]. Ainsi, les événements démographiques incluant la croissance des populations, la présence d'un goulot d'étranglement dans l'histoire d'une espèce ainsi que les effets des flux de gènes sur les patrons de polymorphisme peuvent générer des signatures moléculaires mimant celles de la sélection naturelle lorsque le modèle standard neutre est considéré. Il est donc nécessaire de connaitre au mieux l'histoire démographique d'un modèle biologique pour étudier l'action de la sélection naturelle sur des régions génomiques particulières. Les progrès récents dans les méthodes d'inférences des histoires démographiques permettent aujourd'hui de tester la pertinence du modèle standard neutre contre des modèles démographiques alternatifs pour un échantillon de données multilocus [Csilléry et al., 2010]. En supposant qu'un sous-ensemble de locus anonymes soit représentatif de l'histoire démographique expérimentée par le fond génomique moyen, l'utilisation de telles approches statistiques renseigne sur la chronologie de la mise en place de l'isolement reproducteur entre différentes populations issues d'une population ancestrale. Ces approches permettent en effet de sélectionner quel modèle démographique explique le mieux les données observées parmi un ensemble de modèles qui correspondent chacun à différentes hypothèses biologiques d'évolution divergente entre les populations. Une meilleure connaissance de l'histoire démographique permise par ces approches permet ensuite d'obtenir par simulations des attendus théoriques selon l'hypothèse de neutralité sélective correspondant mieux au modèle biologique étudié que le modèle standard.

9 Prise en compte des résultats de l'analyse fonctionnelle des gènes.

Parallèlement à ces avancées récentes dans les méthodes d'analyses des histoires démographiques, les connaissances fondamentales en physiologie ont fortement progressé, notamment avec l'essor de la physiologie moléculaire étudiant les fonctions des gènes. Cette discipline permet aujourd'hui de confirmer l'implication d'un gène dans un trait phénotypique particulier, préalablement proposé comme gène candidat par des approches de génétique directe ou indirecte. La validation fonctionnelle d'un gène candidat permet de définir physiquement une région génomique sur laquelle il est possible d'étudier les importances relatives de différents processus impliqués dans son évolution moléculaire. De nombreux travaux ont été réalisés sur l'identification des gènes impliqués dans les processus d'adaptation et d'isolement reproducteur entre espèces proches [Ting et al., 2000].

10 Objectifs de la thèse.

Dans mon travail de thèse, je me suis intéressé à la question des différences entre l'histoire démographique expérimentée par un fond génomique moyen et l'évolution de régions appartenant à ce fond génomique mais soumises à différents processus de sélection naturelle. Pour cela j'ai travaillé sur le couple d'espèces allogames *Arabidopsis halleri* et *A. lyrata*, proches de la plante modèle des biologistes moléculaires *A. thaliana* [Clauss and Koch, 2006]. Dans le genre Arabidopsis, *A. halleri* se singularise en étant la seule espèce tolérante et hyperaccumulatrice de métaux lourds, notamment du Zinc et du Cadmium (Figure 12) [Mitchell-Olds, 2001, Becher et al., 2004, Pauwels et al., 2008].

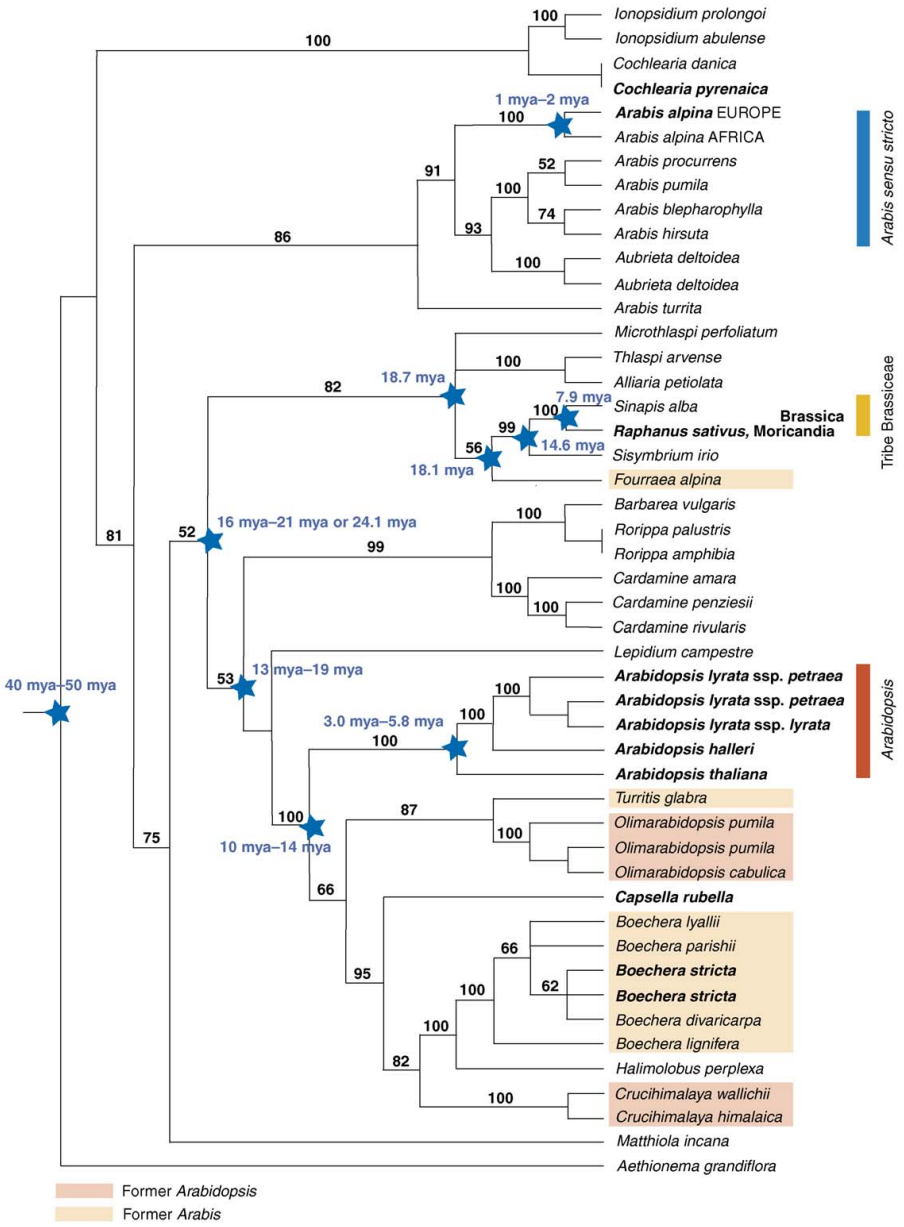

FIG. 12 – Phylogénie de 48 espèces parmi les Brassicaceae d'après [Clauss and Koch, 2006]. Les étoiles désignent les nœuds datés en million d'années. Une carte génétique est disponible pour les espèces indiquées en gras.

L'accumulation des métaux consiste en leur concentration dans les organes aériens à partir de ceux puisés dans le sol par les racines. Une plante est considérée comme hyperaccumulatrice lorsque plus de 0.1% de sa masse sèche est composée d'un élément de type Nickel, Cobalt, Cuivre ou Plomb, ou plus de 1% de Zinc [Clemens, 2001]. Sans structure physiologique adaptée, une concentration cytoplasmique en métaux élevée peut entrainer des dommages au niveau cellulaire. En effet, si les propriétés oxydo-réductrices des métaux sont utilisées par de nombreuses protéines nécessaires au métabolisme, notamment par les coenzymes impliquées dans les chaines de transfert d'électrons, un excès de métaux risque de produire des doses toxiques de dérivés réactifs de l'oxygène, pouvant désorganiser les membranes cellulaires et interférer avec les structures fonctionnelles des protéines [Prasad, 2009]. De par ses spécificités phénotypiques, *A. halleri* est un modèle biologique intéressant pour étudier les architectures génétiques impliquées dans l'accumulation des métaux ainsi que celles impliquées dans les mécanismes d'évitement du stress oxydant. Au sein du genre Arabidopsis, la distinction entre les espèces non-tolérantes et les espèces dites tolérantes peut se faire sans ambiguïté car ces deux groupes sont phénotypiquement bien différenciés, sans continuum intermédiaire observé à ce jour [Pauwels et al., 2006] (Figure 13). Si l'ensemble des populations connues d'*A. lyrata* et d'*A. thaliana* ne montrent pas des niveaux

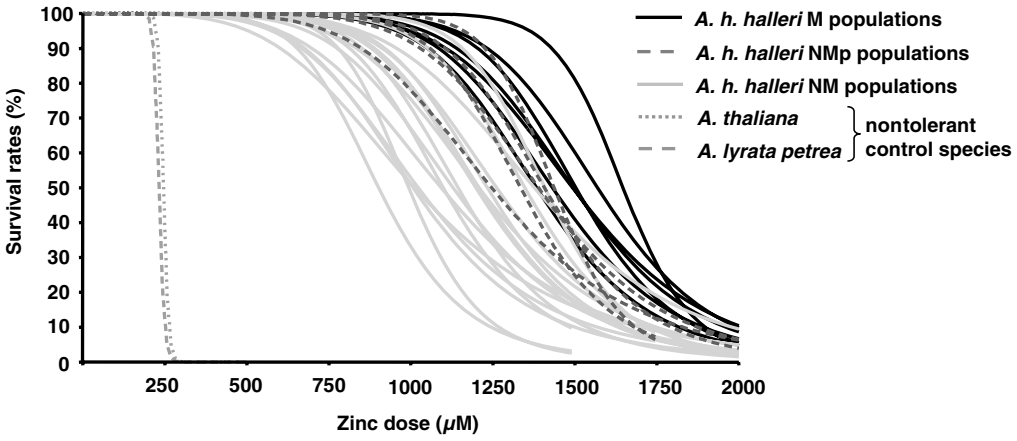

FIG. 13 – Courbes de survies mesurées chez différentes populations *A. halleri*, *A. thaliana* et *A. lyrata* face à différentes intensités de stress oxydant par le zinc.

de tolérance particulièrement élevés pour les métaux, aussi bien les populations d'*A. halleri* trouvées sur des sites pollués aux métaux lourds que les populations naturelles trouvées sur des sites non-pollués montrent de forts niveaux de tolérance. Ce caractère permettant à *A. halleri* de se développer sur des sols pollués mais constitutif d'une espèce distribuée en Europe Centrale et en Asie [Al-Shehbaz and O'Kane, 2002] soulève la question des conditions écologiques qui ont pu sélectionner une telle architecture génétique. Une première hypothèse est que l'hyperaccumulation et la tolérance des métaux lourds spécifiques à *A. halleri* seraient des traits phénotypiques dérivés, apparus dans la lignée *A. halleri* après sa spéciation avec *A. lyrata*. Cette hypothèse est associée à une sélection des génotypes tolérants par les milieux fortement anthropisés, suivie de colonisations vers des milieux non pollués, expliquant le caractère constitutif de la tolérance chez cette espèce. Une hypothèse alternative serait que la sélection des génotypes hyperaccumulateurs et donc tolérants soit l'évènement fondateur de la spéciation vers *A. halleri*, ainsi, l'adaptation à des sols pollués serait un sous-produit de cette sélection ancestrale. L'objectif du chapitre 2

de ce manuscrit de thèse, intitulé "Demographic history and adaptation genomics in the Arabidopsis genus", est d'approcher la question de la relation entre l'évolution de la tolérance et de l'hyperaccumulation des métaux chez *A. halleri* et l'histoire démographique de cette espèce. Ma démarche fut d'étudier d'une part l'histoire d'*HMA4*, un QTL validé chez *A. halleri* pour être impliqué dans la tolérance et l'hyperaccumulation des métaux, puis d'étudier parallèlement l'histoire démographique du couple d'espèces *A. halleri* et *A. lyrata* en utilisant des simulations de scénarios alternatifs.

Dans le chapitre 3 intitulé "Extent of linkage to a locus under balancing selection", je mesure chez *A. halleri* et *A. lyrata* l'étendue de la région génomique locale subissant l'effet de la liaison génétique au locus S, sous la forme d'une élévation le niveau de polymorphisme neutre de cette région génomique. En effet, le locus-S est soumis à une sélection balancée de type fréquence-dépendante, c'est-à-dire que la valeur sélective des individus possédant des allèles rares est supérieure à celle des individus dont les allèles du locus-S sont plus fréquents dans la population. Puisque les études sur *HMA4* et sur le locus-S impliquent une bonne connaissance de l'histoire démographique du modèle biologique utilisé, le premier chapitre intitulé "interprétation des inférences démographiques de divergence : le chant des sirènes" est un chapitre exploratoire visant à évaluer les limites des méthodes existantes pour étudier des scénarios démographiques.

Chapitre 1

Interprétation des inférences de modèles démographiques de divergence : le chant des sirènes

1.1 Résumé.

Au cours de la dernière décennie, l'accumulation de jeux de données de polymorphismes nucléotidiques pour des paires de taxons proches a permis l'essor de la "génomique des populations en divergence", discipline visant à étudier l'histoire démographique de populations phylogénétiquement proches. Plus précisément, la génomique des populations en divergence cherche à caractériser la chronologie des évènements conduisant aux patrons de polymorphisme neutre observés chez ces espèces. Les avancées théoriques et statistiques associées à la génomique des populations en divergence visent à expliquer les signatures laissées par la démographie sur les patrons de polymorphisme en inférant l'importance relative de processus tels que la dérive génétique, le flux génique, et même les fluctuations des effectifs des populations. Différentes motivations peuvent mener à l'étude de l'histoire démographique d'un modèle biologique, allant de la reconstruction du processus de spéciation expérimenté par un ensemble d'espèces, à l'étude d'une région génomique candidate, susceptible d'avoir expérimenté une histoire particulière par rapport au reste du génome. Plusieurs outils disponibles à ce jour permettent d'aborder ce problème, en utilisant soit des méthodes dites "clefs en main" avec notamment les programmes "IM", "IMa" et "MIMAR" qui explorent l'espace des paramètres au moyen d'une approche MCMC s'appuyant sur un unique modèle de divergence [Hey and Nielsen, 2004, Hey and Nielsen, 2007, Becquet and Przeworski, 2007], soit des méthodes présentant un cadre théorique plus flexible connu sous la dénomination ABC (Approximate Bayesian Computation, [Tavare et al., 1997, Beaumont et al., 2002]). La flexibilité de l'ABC s'explique par la possibilité d'étudier des modèles adaptés par l'expérimentateur à un modèle d'étude donné, d'évaluer statistiquement des modèles alternatifs d'évolution puis d'estimer les paramètres des modèles considérés. Actuellement, les méthodes MCMC sont préférées aux approches ABC avec environ 250 articles inventoriés en Aout 2009 [Pinho and Hey, 2010] publiant des estimations effectuées avec les logiciels IM et IMa contre environ 20 articles publiant des estimations de paramètres démographiques dans un cadre ABC. Cependant, des études par simulations ont montré que ces méthodes MCMC produisent des estimations erronées lorsque le modèle démographique s'éloigne du modèle implémenté dans la méthode

[Becquet and Przeworski, 2009]. En conséquence, il apparait que la première étape de toute inférence démographique à partir de jeux de données réels doit être l'étude des modèles démographiques alternatifs, analyse possible dans le cadre de l'ABC.

Dans ce chapitre de thèse, je propose d'aborder par simulations de coalescence l'efficacité d'une approche ABC pour différencier correctement quatre modèles de divergence des populations contrastés, différents par leur patron temporel de migration. Les simulations montrent que la sélection de modèle dans notre approche permet de différencier avec une grande confiance deux catégories de modèles se distinguant par l'intégration ou non d'évènements récents de migration. En revanche il s'avère plus difficile de trancher au sein de ces deux catégories. Cependant, les simulations montrent également qu'une sélection erronée de modèle au sein de la catégorie sans migration récente biaise peu les interprétations biologiques tirées à partir des estimations de paramètres contrairement au biais entrainé sur les estimations avec IM et MIMAR. En analysant par ABC des jeux de données dont les résultats ont été publiés avec MIMAR, nous montrons que le modèle implémenté dans MIMAR n'est jamais soutenu comme étant le modèle le plus pertinent, particulièrement à cause de la violation de l'hypothèse d'un effectif constant des populations depuis l'évènement de spéciation.

Ce premier chapitre forme une ébauche d'un manuscrit visant à produire une discussion sur l'emploi des méthodes d'inférence démographique existantes. Il est issu du constat personnel que les outils "clefs en main" IM et MIMAR s'appuyant sur le modèle "isolement avec migration" sont généralement utilisés sans avoir préalablement testé la pertinence de ce modèle. Or dans certains cas les résultats sont utilisés pour appuyer de "belles histoires", notamment en rapport avec le processus de spéciation sympatrique. Les analyses à suivre viseront à étudier dans quelle mesure il est actuellement possible de complexifier utilement un modèle démographique.

1.2 Introduction.

Les travaux sur les distributions géographiques des espèces effectués par les biogéographes entre la fin du XIX^e et le début du XX^e siècle ont lié la question de la spéciation à celle de l'étude des échanges d'individus entre les populations [Coyne and Orr, 2004]. En observant que les espèces les plus proches sont presque toutes séparées par des barrières géographiques, Wagner écrivit en 1873 [Wagner, 1873] que de telles barrières seraient nécessaires au processus de spéciation. Cette vision est en contradiction avec la théorie Darwinienne où l'origine des espèces serait une conséquence directe de la sélection naturelle [Darwin, 1859]. Avec l'avènement du concept biologique de l'espèce proposé par Dobzhansky [Dobzhansky, 1935] puis affiné par Mayr [Mayr, 1942], la relation entre l'étude de la spéciation et l'étude de la géographie des espèces a glissé vers la compréhension des effets réciproques de la migration et de l'évolution de l'isolement reproducteur. Pour un couple d'espèces proches étudiées, décrire la chronologie des évènements de migration aiderait à identifier les régions génomiques impliquées dans les premières étapes d'une évolution divergente [Ting et al., 2000]. En plaçant ainsi les patrons temporels de migration au centre de la description des histoires démographiques sans *a priori* sur la cause de l'isolement reproducteur nous pouvons distinguer quatre catégories de scénarios différents.

Dans le scénario de spéciation le plus simple (Figure 1.1), une espèce ancestrale a été soudainement subdivisée en deux populations filles qui ont évolué en complet isolement (scénario nommé ici SI. pour "Strict Isolation"). Chaque lignée évolutive a pu accumuler des incompatibilités de type Dobzhansky-Muller [Dobzhansky, 1937, Muller, 1942] comme des sous-produits de la divergence entre les deux sous-populations [Coyne and Orr, 2004]. Une accumulation d'in-

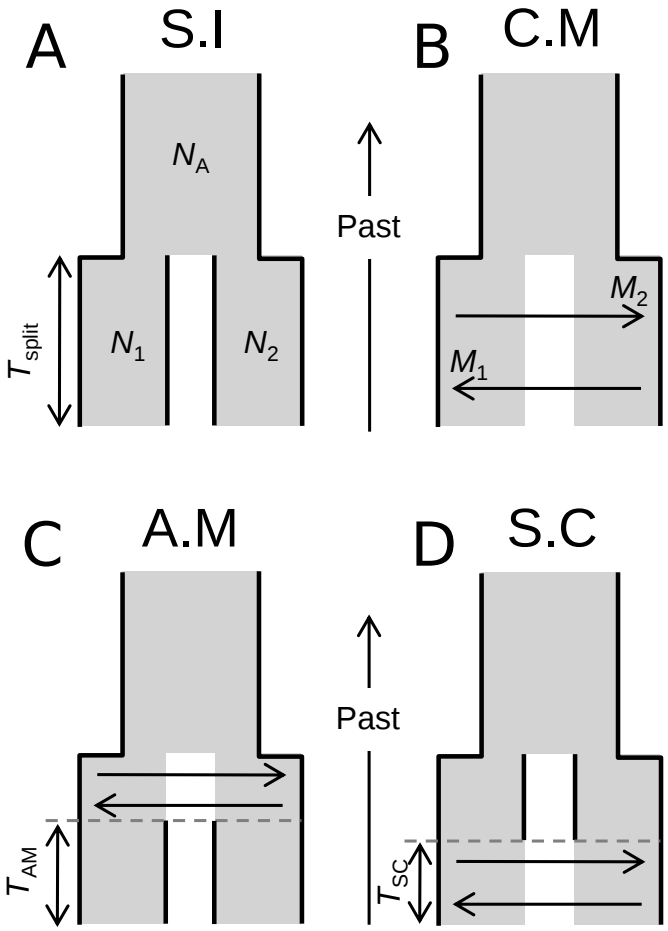

FIG. 1.1 – Description des quatre modèles démographiques étudiés. (A) Modèle d'isolement strict (SI) où une population panmictique ancestrale de taille N_A est subdivisée en deux populations filles de tailles N_1 et N_2 depuis T_{split} générations dans le passé. (B) Modèle de migration constante (CM) depuis la séparation des deux lignées où les populations échangent des migrants à des taux M_1 (depuis la population 2 vers la population 1) et M_2 (depuis la population 1 vers la population 2). (C) Modèle d'isolement ancien (AM) où les flux de gènes entre les lignées séparées ne s'effectuent qu'au cours des premiers stades de la spéciation jusqu'au temps T_{AM} dans le passé. (D) Modèle de divergence en isolement suivi d'un contact secondaire (SC) au cours du quel les deux lignées échangent des gènes à partir de T_{SC} générations dans le passé

compatibilités de Dobzhansky-Muller à l'échelle du génome tend à diminuer la viabilité et/ou la fertilité d'hybrides potentiels entre les populations. Alternativement, les deux populations filles peuvent n'être que partiellement isolées génétiquement durant la période de spéciation (Figure 1.1-B et Figure 1.1-C). L'effet des flux de gènes entre deux populations est d'inhiber le processus de spéciation en ralentissant les évolutions génétiques indépendantes au sein des différentes populations. Ainsi, pour que la spéciation soit possible dans un scénario avec migration (spéciation

sympatrique ou parapatrique), il est nécessaire que les forces évolutives à l'origine de la divergence des populations soient plus fortes que les flux de gènes qui tendent à homogénéiser ces populations [Slatkin, 1985, Gavrilets, 2006]. Un tel scénario implique la fixation rapide dans les deux populations d'allèles alternatifs dans certaines régions génomiques qui ne peuvent pas être introgressés dans le génome de la population pour des raisons intrinsèques (incompatibilités de Dobzhansky-Muller) ou extrinsèques (divergence pour des QTLs clefs impliqués dans des adaptations locales) [I. Wu, 2001]. Les échanges de gènes entre populations peuvent s'effectuer dans les régions non liées aux allèles réduisant la valeur sélective des hybrides, créant ainsi un patron génomique en mosaïque de différenciation qui caractérise le scénario de spéciation en présence de flux de gènes [Via and West, 2008]. Dans le cas ou le flux de gène est encore maintenu actuellement, le scénario sera nommé CM (pour "Constant Migration") dans ce document, ce scénario étant par ailleurs plus connu sous le nom de I.M (pour "Isolation with Migration") terme que nous éviterons ici afin d'éviter la confusion avec un logiciel d'analyse couramment utilisé [Hey and Nielsen, 2004]. Le scénario AM (pour "Ancient Migration") quant à lui décrit deux espèces actuellement isolées, mais ayant connu une période de flux de gènes lors des premières étapes qui ont suivi leur séparation (Figure 1.1-C) [Via, 2001]. Un dernier scénario démographique, SC, envisage la possibilité d'une introgression récente lors d'un contact secondaire précédé d'une longue période évolutive sans échanges de gènes entre les deux populations (Figure 1.1-D).

La génétique des populations de la divergence [Kliman et al., 2000] est une discipline récente qui s'intéresse aux échanges de gènes entre génomes en cours de différenciation. Un de ses principaux objectifs consiste à permettre de différencier les effets de l'introgression de ceux de la rétention du polymorphisme ancestral qui tous les deux convergent vers le maintien de polymorphisme partagé entre deux populations. L'outil originel développé par Wakeley et Hey [Wakeley and Hey, 1997] permet d'estimer les quatre paramètres qui décrivent le modèle de divergence de deux populations en isolement strict. Sa limite d'utilisation imposée par la simplicité du modèle démographique considéré a motivé le développement d'outils intégrant la possibilité des échanges géniques. Ainsi en 2001, Nielsen et Wakeley [Nielsen and Wakeley, 2001] proposèrent une méthode plus générale qui permet d'estimer des taux de migration efficace entre deux lignées en divergence. Le progrès méthodologique substantiel apporté par l'intégration possible de la migration s'est fait au détriment de l'hypothèse contraignante d'une absence de recombinaison intragénique pour les locus analysés. Les améliorations apportées par Hey et Nielsen à la méthode IM [Hey and Nielsen, 2004, Hey and Nielsen, 2007] permettent d'utiliser les informations multilocus pour inférer les taux de migrations ainsi que la distribution de l'âge de ces évènements de migrations, mais toujours à partir de données haplotypiques. L'approche méthodologique du logiciel IM et ses dérivés (IMa,...) est une approche par chaine de Markov (MCMC) explorant de manière combinée l'espace des paramètres du modèle CM (distributions *a priori* des paramètres) et l'espace des généalogies d'haplotypes, utilisant à chaque étape de la chaine un calcul exact de la vraisemblance des généalogies compte tenu des données de séquences nucléotidiques. L'inférence des taux de migrations dans le modèle CM a récemment été généralisée à des jeux de donnés multilocus recombinants (permettant la recombinaison intragénique pour chaque locus) par le développement du logiciel MIMAR [Becquet and Przeworski, 2007]. Celui-ci utilise également une approche MCMC explorant de manière combinée l'espace des paramètres et des généalogies, mais l'évaluation de la vraisemblance à chaque étape de la chaine est réalisée par comparaison d'un jeu de statistiques résumées calculées sur la généalogie simulée et sur les données. Cependant, si des outils tels que IM et MIMAR peuvent être efficaces pour estimer les paramètres d'un modèle de séparation avec migra-

tion continue, les estimations qu'ils produisent peuvent être biaisées, parfois fortement, si les patrons temporels de migrations ne correspondent pas au modèle [Becquet and Przeworski, 2009]. Malgré les hypothèses très fortes liées à l'adhésion à un modèle démographique unique (scénario CM), l'utilisation de ces méthodes est devenu pratiquement systématique, alors que des méthodologies plus flexibles comme l'"ABC" (pour "Approximate Bayesian Computation", [Tavare et al., 1997, Beaumont et al., 2002]) offrent la possibilité de tester divers scénarios avec ou sans flux de gènes et de répondre ainsi à des questions biologiques pertinentes [Tavare et al., 1997, Pritchard et al., 1999, Beaumont et al., 2002, Fagundes et al., 2007]. Ceci illustre parfaitement un problème répandu parmi les biologistes empiristes et souligné par [Beaumont et al., 2010] : "in evolutionary biology there remains a tendency to treat statistical procedures uncritically as "black boxes", and to accept apparently easy solutions, especially those that fit with common-sense nostrums", bien que cette remarque ne fut pas directement adressée aux méthodes "clefs en mains" IM et MIMAR j'estime qu'elle est de circonstance ici. Les approches ABC fournissent des outils d'analyses qui se placent dans un cadre statistique d'évaluation des scénarios démographiques alternatifs. Ainsi, plusieurs scénarios alternatifs concernant l'origine de l'Homme moderne ont été débattus au cours des deux dernières décennies, chacun tentant d'expliquer certains aspects des patrons de distribution des polymorphismes observés [Templeton, 2002, Cavalli-Sforza and Feldman, 2003, Krings, 1997, Liu et al., 2006]. Ce débat a souffert du manque de cadre statistique pour confronter les différents scénarios proposés. Fagundes & al [Fagundes et al., 2007] ont exploité la souplesse de l'approche ABC pour calculer les probabilités *a posteriori* des différents modèles concurrents. Cette avancée méthodologique a permis d'apporter des éléments de réponses objectifs menant au rejet des scénarios les moins probables. En simulant de manière explicite les différents patrons de migration et de changement démographique, il a été montré que le scénario de remplacement récent des populations eurasiennes par une population d'origine africaine sans introgression avec les populations locales (le modèle "Out-of-africa" équivalent à SI) était supérieur aux différents scenarios incluant une divergence ancienne de populations humaines connectées par migration (les modèles "Multiregionaux" équivalents à CM). De même chez le maïs, Ross-Ibarra & al [Ross-Ibarra et al., 2009] ont utilisé ce cadre d'analyse pour inférer plus finement les patrons temporels de migration suivant les scénarios SI, CM, AM et SC. Ils ont ainsi identifié une hétérogénéité dans les processus de divergence au sein du genre Zéa à partir de données de séquences obtenues chez *Z. mays mays*, *Z. m. mexicana*, *Z. m. parviglumis et Z. luxurians*.

Au coeur de nombreuses questions évolutives en lien avec le processus de spéciation se trouve la distinction *a posteriori* entre scénarios démographiques de divergence tels que SI, CM, AM et SC. Au vu de l'accélération récente de la production de données de polymorphisme et de divergence nucléotidiques entre espèces proches, il apparait que les tentatives de confrontation de scénarios de spéciation et d'estimation de leurs paramètres vont proliférer dans les prochaines années. Il n'existe cependant pas encore d'analyses exploratoires à partir de simulations pour tester l'efficacité de l'approche ABC à distinguer respectivement ces scénarios et à estimer leurs paramètres dans différentes situations biologiques.

Dans cette étude nous nous intéressons aux limites de l'approche dans une optique d'application à des jeux de données issus de re-séquençage multilocus en populations naturelles. Nous testons dans un premier temps la robustesse d'une approche ABC pour distinguer les scénarios sans migration récente (SI et AM) des scénarios avec migration récente (CM et SC, Figure 1.1). Nous déterminerons ensuite dans quelle mesure il est possible de différencier des scénarios emboités. En effet, le scénario SI apparaît comme un cas particulier du scénario AM où les effets de la migration ancestrale sont restreints dans le temps. De la même manière, le scénario CM

impliquant de la migration en continu depuis la séparation des lignées est un cas extrême du scénario SC où la période intermédiaire au cours de laquelle les deux lignées évoluent isolément serait trop courte pour avoir des effets significatifs sur les patrons de diversité et de divergence. Ensuite, nous testons les écarts possibles au scénario standard CM en estimant les probabilités *a posteriori* des principaux scénarios à partir de jeux de données dont les résultats d'analyses effectuées avec MIMAR ont été publiés. Finalement, nous étudions l'effet de ces écarts sur les estimations des paramètres estimés par le logiciel MIMAR.

1.3 Matériels et méthodes.

1.3.1 Scénarios démographiques étudiés.

Nous avons simulé par coalescence quatre scénarios démographiques (Figure 1.1) qui ont pour point commun de décrire la séparation d'une population panmictique ancestrale de diversité θ_A ($= 4.N_A.\mu$ où N_A représente la taille de la population en nombre efficace d'individus et μ le taux de mutation) au temps T_{split} en deux populations filles panmictiques, de diversités θ_1 ($= 4.N_1.\mu$) et θ_2 ($= 4.N_2.\mu$). Ces quatre scénarios se distinguent par différents patrons temporels de migration. Le scénario nul est un scénario avec une subdivision de la population ancestrale au temps T_{split} suivi d'un isolement strict (SI pour Strict Isolation) entre les deux populations en divergence. Un tel scénario est défini par les quatre paramètres θ_1, θ_2, θ_A et T_{split}. Le scénario CM pour "Constant Migration" désigne un scénario où les deux populations divergent en échangeant réciproquement et de façon continue des migrants aux taux constants M_1 ($= 4.N_1.m_1$, où m_1 est la proportion de N_1 composée à chaque génération d'individus provenant de la population 2) et M_2. Les scénarios AM (Ancient Migration) et SC (Secondary Contact) ont la particularité de présenter une alternance entre une période de migration et une période d'isolement. Dans le cas du scénario AM, la séparation de la population ancestrale est suivie d'échanges de gènes entre les deux nouvelles lignées à des taux M_1 et M_2, puis d'une période d'isolement à partir du temps T_{AM} jusqu'au temps présent. Le scénario SC décrit une séparation au temps T_{split} avec un isolement strict entre les deux populations jusqu'au temps T_{SC} dans le passé, à partir duquel les deux populations rentrent en contact secondaire au travers d'évènements de migrations à des taux M_1 et M_2. Lors des ré-analyses des jeux de donnés publiés nous avons étudié pour chacun des quatre modèles principaux deux modèles alternatifs selon le comportement des tailles des populations filles. Ainsi, nous avons confronté au sein de chacun des scénarios SI, CM, AM et SC un modèle où les tailles efficaces des populations filles sont restées constantes depuis l'évènement de spéciation à un modèle où les tailles de populations croissaient exponentiellement. Selon ce dernier modèle, la taille $N(T_{split})$ au moment de la séparation des populations est égale à $N(T_{split}) = N_0 e^{-\alpha T_{split}}$ où N_0 est une constante de référence exprimée en nombre efficace d'individus, α est le taux de croissance de la population et T_{split} est le nombre de générations depuis la séparation.

1.3.2 Simulations de coalescents.

Les scénarios SI, CM, AM et SC ont été simulés sans croissance exponentielle des populations filles pour s'intéresser dans un premier temps à l'influence seule de la migration sur les analyses faites en ABC. Bien que nous n'ayons pas encore exploré la précision de cette approche pour les inférences de changements de tailles des populations, nous avons tout de même intégré ces modèles dans les analyses faites à partir de jeux de données publiés. Les simulations de coalescents ont été pratiquées en utilisant le programme msnsam [Ross-Ibarra et al., 2008] qui est une version

modifiée du générateur de patrons de polymorphismes ms [Hudson, 2002], et qui permet de faire varier le nombre d'individus échantillonné pour chaque locus, contrairement à la version initiale. Chaque réplica de simulation est réalisé en tirant indépendamment chacun des paramètres du modèle démographique dans une distribution a priori, à l'aide d'une version modifiée du logiciel Priorgen [Ross-Ibarra et al., 2008], qui permet par ailleurs pour chaque locus d'utiliser des valeurs différentes des paramètres de mutation et de recombinaison. Les lignes de commandes utilisées pour simuler les différents scénarios sont placées en annexe.

Les modifications apportées au logiciel Priorgen [Ross-Ibarra et al., 2008] permettent d'une part de ne pas contraindre un des deux effectifs efficaces des populations filles comme dans la version originale, et d'autres part de faire varier ces effectifs au cours du temps (pour modèles à croissance exponentielle). Techniquement, chaque locus est simulé par msnsam en lui fournissant une valeur θ_{ref} ($= 4.N_{ref}.\mu.L$ où L est la longueur du locus simulé en nombre de paires de bases et μ le taux de mutation au locus considéré) qui est le paramètre de mutation populationnelle de la population de référence. Les valeurs réellement utilisées pour chaque population fille et pour la population ancestrale sont obtenues par tirage des rapports θ_1/θ_{ref}, θ_2/θ_{ref} et θ_A/θ_{ref} dans une distribution *a priori* uniforme bornée entre 0 et 10, correspondant aux bornes 0 et 1,000, 000 pour les tailles N_1, N_2 et N_A. De la valeur θ_{ref} initialement attribuée par l'expérimentateur vont dépendre les valeurs des temps où se sont déroulés les différents évènements démographiques $T/4.N_{ref}$ (T est exprimé en nombre de générations). La valeur de N_{ref} a été ici arbitrairement fixée à 100,000 individus. Les taux de migrations M_1 et M_2 sont tirés aléatoirement entre 0 et 2 individus passant d'une population à l'autre à chaque génération. Le paramètre $T_{split}/4.N_{ref}$ est tiré dans une distribution bornée par 0 et 25, correspondant à un tirage de T_{split} dans un intervalle compris entre 0 et 10,000,000 générations. Puisque les paramètres T_{AM} et T_{SC} correspondent à des évènements démographiques logiquement plus récents que la séparation des deux lignées, leurs valeurs ont été uniformément tirées dans un intervalle compris entre 0 et la valeur préalablement tirée pour T_{split}. 5,000,000 de jeux de données multilocus ont été ainsi simulés pour chacun des scénarios considérés. Les jeux de données ont été simulés de façon à s'approcher des études empiriques récentes et comprennent chacun 30 locus de 500 paires de bases échantillonnés chez 80 individus répartis équitablement entre les deux espèces filles considérées. Les taux de mutation et de recombinaison sont identiques entre les locus, et affectent tous les deux chaque paire de base avec une probabilité de 2.10^{-9} par génération.

1.3.3 Calcul de statistiques résumées pour décrire les jeux de données simulés.

Nous avons retenu un ensemble de 34 statistiques résumées : le nombre total de sites polymorphes, le nombre de sites appartenant à chacune des 7 catégories (S_s, S_{x1}, S_{x2}, S_{f1}, S_{f2}, S_{x1f2} et S_{x2f1}) décrites dans la figure 1.2, les estimateurs de diversités θ_{Wat} (estimateur θ de Watterson) et π (diversité nucléotidique) pour chacune des deux populations, le D de Tajima par population, la statistique F_{ST} interspécifique (calculée comme étant égale à $1-\pi_s/\pi_{tot}$, où π_s représente la moyenne des valeurs de π par population et π_{tot} la valeur de π mesurée sur l'ensemble de l'échantillon), la divergence brute entre les deux populations et la divergence nette. Comme statistiques résumées, nous avons utilisé la moyenne et l'écart type entre locus de chacune de ces statistiques, calculés par un programme écrit en C par Xavier Vekemans. Plusieurs combinaisons de ces statistiques ont été utilisées dans des études précédentes pour inférer des paramètres ancestraux [Wakeley and Hey, 1997, Fagundes et al., 2007, Becquet and Przeworski, 2007], puis nous les avons complété avec un ensemble de statistiques que nous suspections d'être informa-

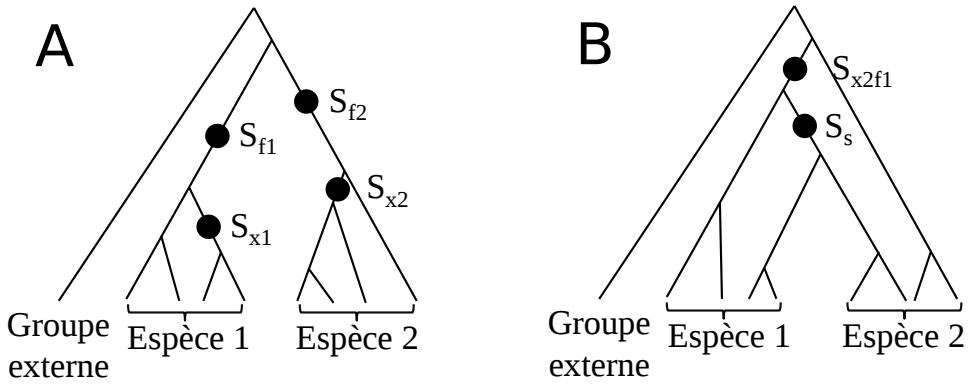

FIG. 1.2 – Extension par Ramos-Onsins et al. [Ramos-Onsins et al., 2004] des quatre classes de sites polymorphes (S_{x1}, S_{x2}, S_f et S_s) décries par Wakeley et Hey [Wakeley and Hey, 1997]. Chaque cercle noir symbolise la position dans l'arbre coalescent d'une mutation affectant un allèle ancestral formant un allèle dérivé. En utilisant un groupe externe et en supposant un modèle à nombre infinie de sites, il est possible d'orienter la mutation en cas de différence fixée entre les deux espèces. Dans un tel cas de figure, on considère que la mutation a eu lieu dans la lignée évolutive menant à l'espèce où le nucléotide diffère de celui du groupe externe.
S_{x1} : site polymorphe spécifique à l'espèce 1.
S_{x2} : site polymorphe spécifique à l'espèce 2.
S_{f1} : différence fixée où tous les individus de l'espèce 1 possèdent le même allèle dérivé à cette position et où tous les individus de l'espèce 2 possèdent l'allèle ancestral.
S_{f2} : différence fixée où tous les individus de l'espèce 2 possèdent le même allèle dérivé à cette position et où tous les individus de l'espèce 1 possèdent l'allèle ancestral.
S_s : position où un état bi-allélique ancestral/dérivé se retrouve à la même position génomique chez les deux espèces.
S_{x1f2} : non représenté sur la figure. Site où l'allèle dérivé est fixé chez l'espèce 2 mais qui reste polymorphe avec l'allèle ancestral chez l'espèce 1.
S_{x2f1} : site où l'allèle dérivé est fixé chez l'espèce 1 mais qui reste polymorphe avec l'allèle ancestral chez l'espèce 2

tives (moyenne et variance du D de Tajima, moyenne et variance de θ_{Wat}, moyenne et variance des divergences brutes et nettes, moyenne et variance de S_{x1f2} et de S_{x2f1}).

1.3.4 Procédure de sélection de modèles appliquée aux données simulées.

Nous avons simulé 1,000 jeux de données et calculé les statistiques résumées pour chacun des quatre modèles SI, CM, AM et SC sans croissance exponentielle des populations filles. Ensuite, une approche de sélection de modèle par ABC a été réalisée sur chacun des 4,000 jeux de statistiques résumées obtenus afin de déterminer les probabilités relatives d'ajustement à chacun des 4 modèles testés (SI, CM, AM et SC). Pour l'approche ABC, chaque modèle démographique a fait l'objet de 5,000,000 de simulations en utilisant les procédures décrites plus haut. Pour chaque réplica, le jeu de 34 statistiques résumées décrit à la section précédente a été calculé. Ensuite une mesure de l'écart entre chacun des 4 x 5,000,000 de réplicas ABC et chacun des 4,000 jeux simulés est calculée au moyen d'une distance euclidienne sur l'ensemble des 34 statis-

tiques résumées. La procédure choisie pour la sélection de modèle est celle de la méthode de rejet simple, consistant à sélectionner les 10,000 meilleurs réplicas ABC sur l'ensemble, et à calculer la proportion de ces réplicas correspondant à chacun des 4 modèles démographiques simulés, le meilleur modèle étant celui conduisant à la proportion la plus élevée. Cette procédure permet finalement de calculer pour l'ensemble des 1,000 jeux simulés avec chacun des modèles démographiques, la proportion de ces jeux qui conduisent à une détermination correcte du meilleur modèle démographique (Figure 1.3). Le choix de la méthode de rejet simple a été

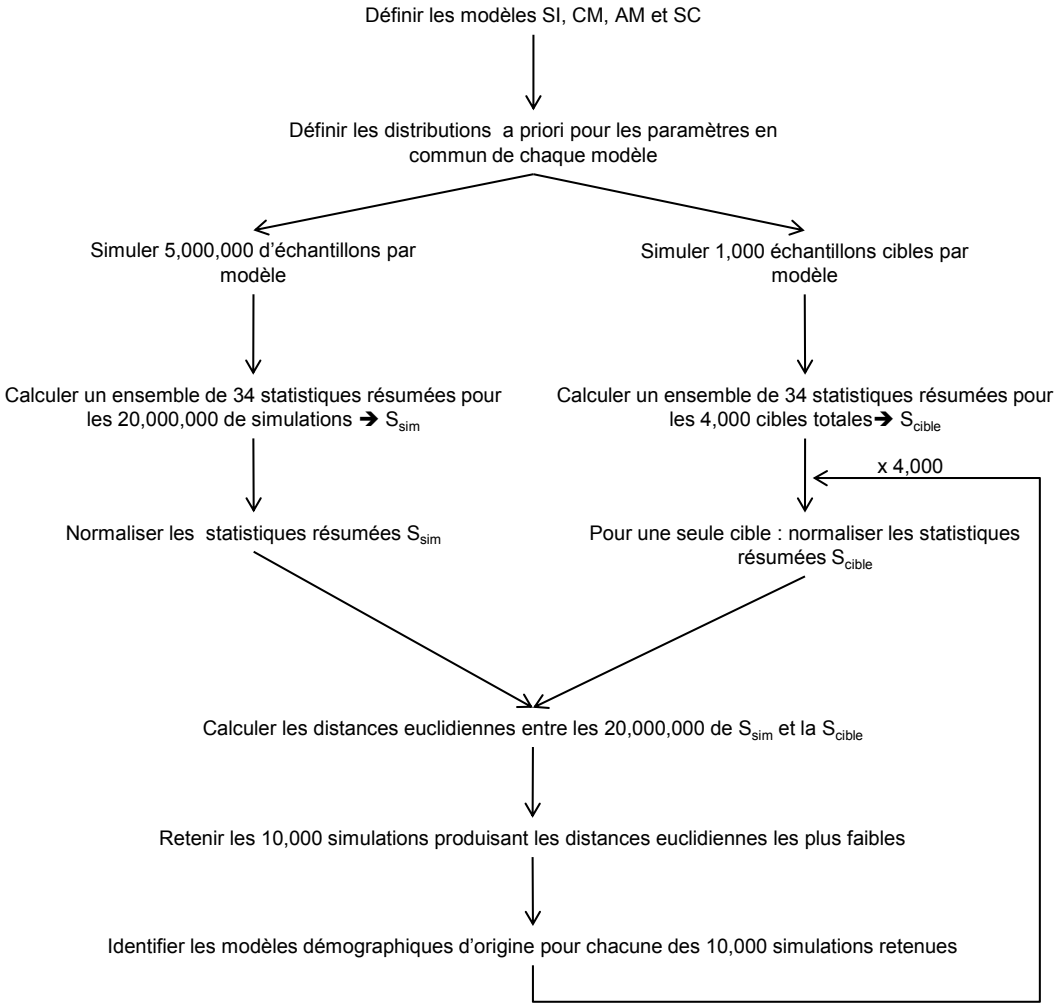

FIG. 1.3 – Procédure de sélection de modèle appliquée à 4,000 jeux de données simulés en utilisant un algorithme de simple rejet.

motivé par la raison suivante : dans un premier temps, nous étions intéressés par la confrontation des 4,000 jeux simulés avec les réplicas ABC à l'aide d'une approche de régression non linéaire multivariée avec la librarie R "abc" [Csillery et al., 2010] utilisant le package "nnet" [Blum and François, 2010, Venables and Ripley, 2002, R Development Core Team, 2008], mais après 8 jours de traitement, seulement 300 jeux simulés sur les 4,000 avaient été analysés. En

constatant que l'utilisation d'une simple méthode de rejet sur ces 300 analyses préliminaires avait donné des estimations convergentes à celles fournies par le modèle plus flexible de Blum et François, mais pour des durées d'analyses 30 fois plus courtes que cette dernière, nous avons décidé d'utiliser la méthode de rejet pour effectuer la comparaison de modèles sur l'ensemble des jeux simulés (1.3).

1.3.5 Application à des jeux de données de séquences nucléotidiques publiées.

Nous avons analysé les séquences disponibles pour les couples d'espèces décrits dans le Tableau 1.1 et publiés dans [Foxe et al., 2009, Li et al., 2010, Thalmann et al., 2007].

Paires d'espèces étudiées	Référence	Meilleur modèle estimé en ABC
Capsella grandiflora vs. C. rubella	Foxe & al. 2009	CM + croissance exponentielle
Picea likiangensis vs. P. schrenkiana	Li & al. 2010	AM + croissance exponentielle
Picea purpurea vs. P. schrenkiana	Li & al. 2010	SI + taille constante
Picea wilsonii vs. P. schrenkiana	Li & al. 2010	AM + croissance exponentielle
Pan paniscus vs. P. t. troglodytes	Fischer & al. 2006 Becquet & Przeworski 2007	SI + croissance exponentielle
P. t. schweinfurthii vs. P. t. verus	Fischer & al. 2006 Becquet & Przeworski 2007	CM + croissance exponentielle
Gorilla gorilla vs. G. beringei	Thalmann & al. 2007 Becquet & Przeworski 2007	AM + croissance exponentielle

TAB. 1.1 – Paires d'espèces pour lesquelles des analyses avec MIMAR ont été publiées précédemment, puis réanalysées ici dans un cadre ABC.

Ces jeux de données ont fait l'objet d'analyses avec le logiciel MIMAR dont les résultats ont été publiés. Bien que les jeux de données concernant les primates aient également été analysés par le logiciel IM [Won and Hey, 2005], nous nous intéressons uniquement ici aux analyses effectuées avec MIMAR [Becquet and Przeworski, 2007]. Pour tous les jeux de données, nous avons utilisé les mêmes valeurs de taux de mutations et de taux de recombinaison que celles rapportées par les auteurs. Chacun des jeux de données a été téléchargé à partir du site du NCBI, aligné avec la version 3.8.31 de MUSCLE [Edgar, 2004], puis corrigé manuellement avec MEGAv4 [Tamura et al., 2007]. Afin d'utiliser la même information génétique que les auteurs des analyses d'origines, les statistiques résumées décrites précédemment ont été calculées pour l'ensemble des nucléotides des données "épicéa" et "grands singes", et pour uniquement les positions synonymes dans le cas du jeu de données "*Capsella*". Une approche ABC a été menée pour chacun des 7 jeux de données, avec un nombre de réplicas de simulation égal à 5,000,000 pour chaque modèle démographique (les 4 modèles avec taille constante des populations filles, et les 4 modèles avec croissance exponentielle). Les distributions *a priori* sont spécifiques de chacune des 7 analyses et ont été arbitrairement choisies pour être bien plus larges que les intervalles de confiance rapportés par les auteurs (à 90% ou 95% selon les études). La procédure utilisée pour la sélection de modèle est celle utilisant un réseau de neurones pour effectuer une régression non-linéaire multivariée [Blum and François, 2010]. Dans un premier temps, nous avons utilisé la fonction R "model selection abc nnet" développée par Blum & François [Blum and François, 2010] pour évaluer au sein de chaque scénario SI, CM, AM et SC la possibilité que les populations filles soient en croissance exponentielle. Puis dans un second temps, avec la même fonction R, nous avons comparé entre eux les modèles les mieux soutenus au sein de chacun des quatre scénarios. Pour chaque analyse, 0.05% des points simulés ont été conservés, le

nombre de réseaux de neurones à été fixé à 50, et le nombre de réseaux cachés est égal au double du nombre de modèles comparés. Pour effectuer les sélections de modèles, toutes les statistiques sauf les moyennes et les écarts types de θ_{Wat}, de la divergence et de la divergence nette, ont été conservées, soit 26 statistiques résumées connues pour capturer une grande partie de l'information génétique [Wakeley and Hey, 1997, Fagundes et al., 2007, Becquet and Przeworski, 2007]. Toutes les statistiques résumées calculées n'ont pas été utilisées afin de pouvoir pratiquer plus tard un test d'ajustement à partir de statistiques non-impliquées dans les inférences. L'estimation des paramètres a été faite en utilisant la fonction R "abc nnet multivar" de Blum & François [Blum and François, 2010]. Nous avons repris les 26 statistiques utilisées précédemment sans les moyennes et les variances du D de Tajima pour les deux populations. Le D de Tajima, sensible aux changements de tailles des populations, était uniquement conservé précédemment pour apporter une information liée à une possible croissance de tailles des populations depuis la séparation des lignées. Pour chaque modèle concerné, nous avons retenu les 5,000 simulations les plus proches des valeurs observées dans un total de 5,000,000 de simulations. Pour que les estimations ne dépassent pas les distributions a priori, les statistiques résumées ont été transformées avant de procéder à la régression comme selon Hamilton [Hamilton et al., 2005] suivant $y = -ln(tan(\frac{x-min}{max-min}\frac{\pi}{2})^{-1})$ où "min" et "max" sont les limites inférieures et supérieures de la distribution *a priori*. Pour ces analyses, le nombre de réseaux de neurones est également fixé à 50, avec un nombre de réseaux cachés égal au double du nombre de paramètres à estimer.

1.3.6 Test d'ajustement aux données pour valider les estimations (goodness of fit test).

Le test d'ajustement permet de vérifier si le modèle estimé par MIMAR ou par l'ABC fournit des échantillons proches des données observées. Pour chacun des jeux de données empiriques publiés, les attendus neutres sous les modèles démographiques estimés par le logiciel MIMAR et par l'ABC ont été générés à partir de 2,000 simulations de coalescents en utilisant le logiciel msnsam. Puisque nous ne disposons pas des distributions *a posteriori* obtenues par MIMAR, toutes les simulations ont été effectuées en reprenant uniquement les valeurs des modes estimés pour chaque paramètre. Cette différence est de prime importance car selon les études trouvées dans la bibliographie, les auteurs rapportent comme estimations ponctuelle soit le mode soit la médiane de la distribution *a posteriori*. De plus, les goodness-of-fit publiés dans la bibliographie sont pratiqués à partir de l'ensemble des points des distributions *a posteriori*, élargissant la distribution attendue des statistiques résumées. Nous risquons donc ici d'avoir un excès de faux négatifs en utilisant uniquement une seule combinaison de paramètres correspondant aux points estimés rapportés. Les P-values ont été calculées ici comme la probabilité par statistique résumée que la valeur obtenue par simulation selon le modèle considéré soit égale à la valeur observée ou qu'elle prenne une valeur plus extrême [Becquet and Przeworski, 2009].

1.4 Résultats

1.4.1 Sensibilité de l'approche ABC aux évènements migratoires récents.

Nous avons étudié dans un premier temps l'efficacité de l'approche ABC pour différencier les scénarios intégrant de la migration récente (CM+SC) des scénarios décrivant une histoire démographique sans migration récente (SI+AM). Ainsi, nous avons empiriquement estimé à partir des 2,000 simulations réalisées suivant les modèles CM et SC la proportion d'analyses

pour lesquelles l'approche ABC identifie un cas de figure opposé. De la même manière nous avons estimé la proportion parmi 2,000 jeux de données simulés selon un modèle SI et AM où l'approche ABC estime correctement le modèle d'origine. La figure 1.4 montre que la procédure différencie efficacement les jeux de données simulés dans les scénarios faisant intervenir la migration récente de ceux qui ne le font pas. Nous observons 1.4% de faux négatifs, c'est-à-dire de jeux de données simulés suivant un modèle avec migration récente non retrouvés comme tels par l'ABC. Ce taux de faux négatifs est plus important dans le sens opposé. En effet, la méthode utilisée conclut à l'existence de migration récente pour 6.4% des jeux de données simulés en l'absence de migration récente.

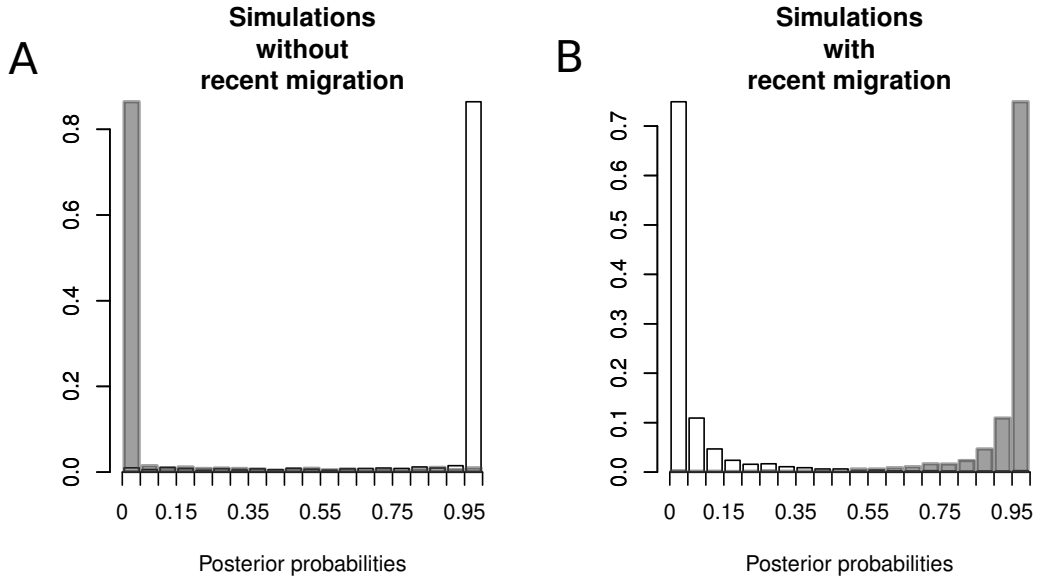

FIG. 1.4 – Distribution des probabilités *aposteriori* que des jeux de données simulés (A) sans migration récente ou (B) avec migration récente soient attribués à un modèle sans migration récente (barres blanches) et à un modèle avec migration récente (barres grises) par la procédure de sélection de modèle.

1.4.2 Efficacité pour détecter une absence de migration ou de la migration ancienne.

En comparant les différentes probabilités d'assignation à l'un des quatre modèles sachant que le modèle sans migration (SI) est le bon modèle, nous observons que la procédure de sélection de modèle retrouve SI dans 67.6% des cas (Figure 1.5A Tableau 1.2).

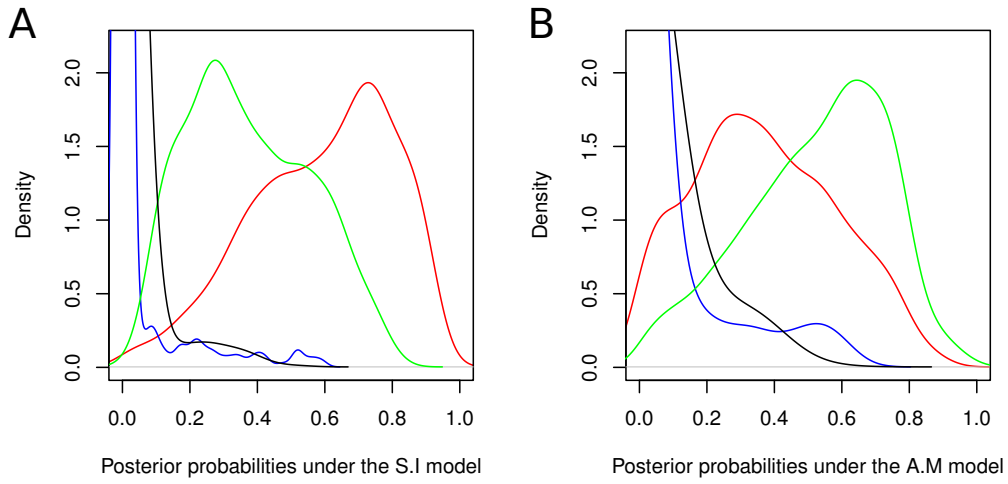

FIG. 1.5 – Distributions empiriques des probabilités relatives d'être associé aux modèles SI (rouge), CM (bleu), AM (vert) et SC (noir) quand (A) le modèle SI est le bon modèle et quand (B) AM est le bon modèle.
Ces distributions sont issues des simulations de 1,000 jeux de données multilocus sous le modèle SI et de 1,000 jeux de données sous le modèle AM. Pour chacun des jeux de données simulés nous avons estimé les probabilités relatives de chaque modèle selon la procédure décrite figure 1.3.

X	$P(X > 0.5\|SI)$	$P(X = \mathbf{max}\|SI)$	$P(X > 0.5\|AM)$	$P(X = \mathbf{max}\|AM)$
SI	0.668	0.676	0.286	0.288
CM	0.009	0.018	0.041	0.085
AM	0.284	0.302	0.592	0.621
SC	0.001	0.004	0.002	0.006

TAB. 1.2 – Proportions parmi 1,000 simulations SI et 1,000 simulations AM de simulations attribuées aux modèles X par la procédure de sélection de modèle (décrite figure 1.3). Les modèles sont soutenus soit en ayant seulement la plus forte probabilité relative, soit en ayant une probabilité *a posteriori* supérieure à 0.5.

Nous observons également que le modèle SI est directement en compétition avec le modèle AM. Ainsi, un modèle réellement SI est identifié erronément comme étant un modèle AM dans 30.2% des cas. En revanche, dans seulement 2.2% des analyses, la comparaison de modèle identifie un modèle SI comme étant un des modèles qui intègrent de la migration récente. Afin de mieux comprendre les causes de la confusion observée entre les modèles par l'approche ABC, nous avons examiné la relation entre les valeurs de paramètres des simulations effectuées selon le modèle S.I. et le support par la procédure ABC pour l'un des quatre modèles (Figure 1.6). La probabilité d'identifier correctement un modèle SI dépend essentiellement de la valeur du paramètre T_{split} (Figure 1.6-D), avec une probabilité d'autant plus grande que le temps depuis la séparation des deux espèces est important.

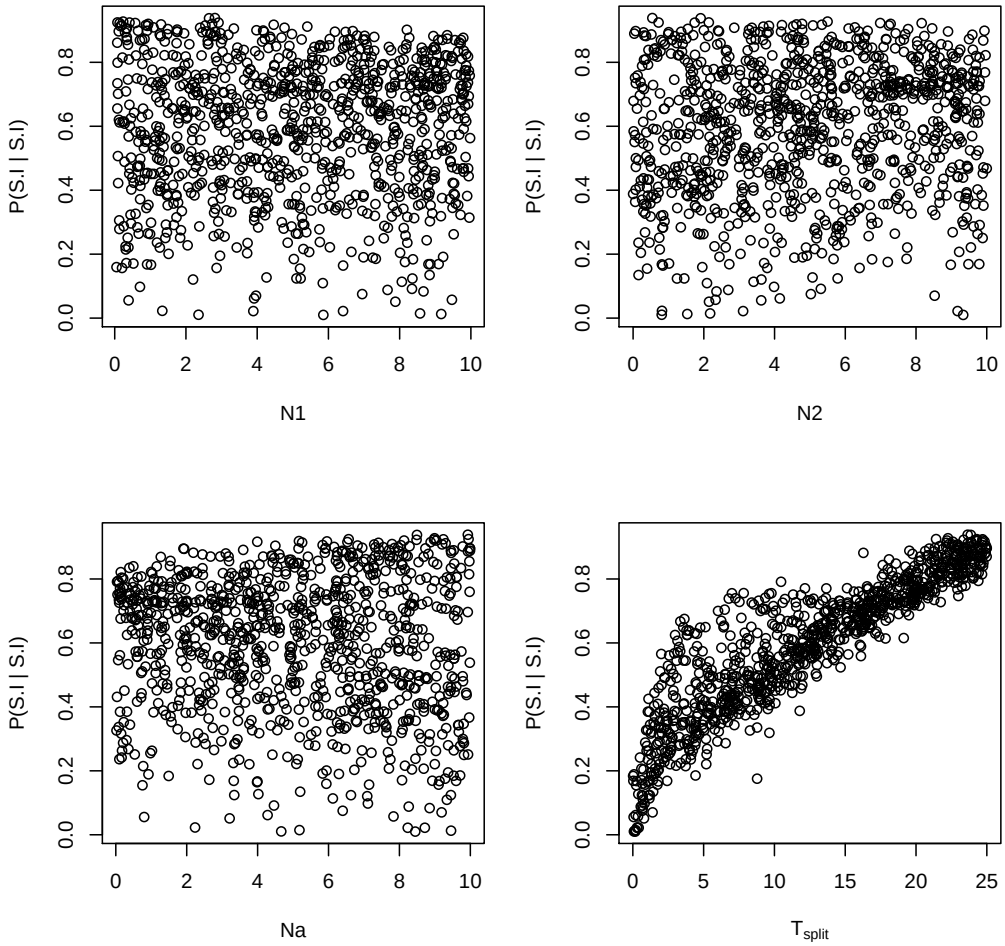

FIG. 1.6 – Probabilité $P(SI|SI)$ calculée pour 1,000 jeux de données cibles simulés selon un modèle SI en fonction de : (A) La taille efficace N_1 de la population 1. (B) La taille efficace N_2 de la population 2. (C) La taille efficace N_A de la population ancestrale. (D) Le temps T_{split} où les lignées ont évolué en isolement complet.

De la même manière, 62.1% des jeux de données simulés suivant un modèle avec migration ancienne (AM) sont correctement identifiés dans notre approche ABC (Figure 1.5-B Tableau 1.2). Les faux négatifs sont dus à une compétition avec le modèle SI (28.8%) puis dans une moindre mesure avec les modèles intégrant de la migration récente (9.1%). L'analyse de l'influence des valeurs de paramètres sur les résultats d'assignation des simulations sous le modèle AM montre que pour des séparations très anciennes de lignées suivies par peu de migration (Figure 1.7-B) la comparaison de modèles tend à attribuer un modèle SI au lieu du modèle AM. A l'inverse, des simulations de données dans un modèle AM où les valeurs de paramètres décrivent une divergence récente suivie d'échanges de gènes soutenus dans le temps vont être attribuées comme appartenant aux modèles CM ou SC (Figure 1.7-C).

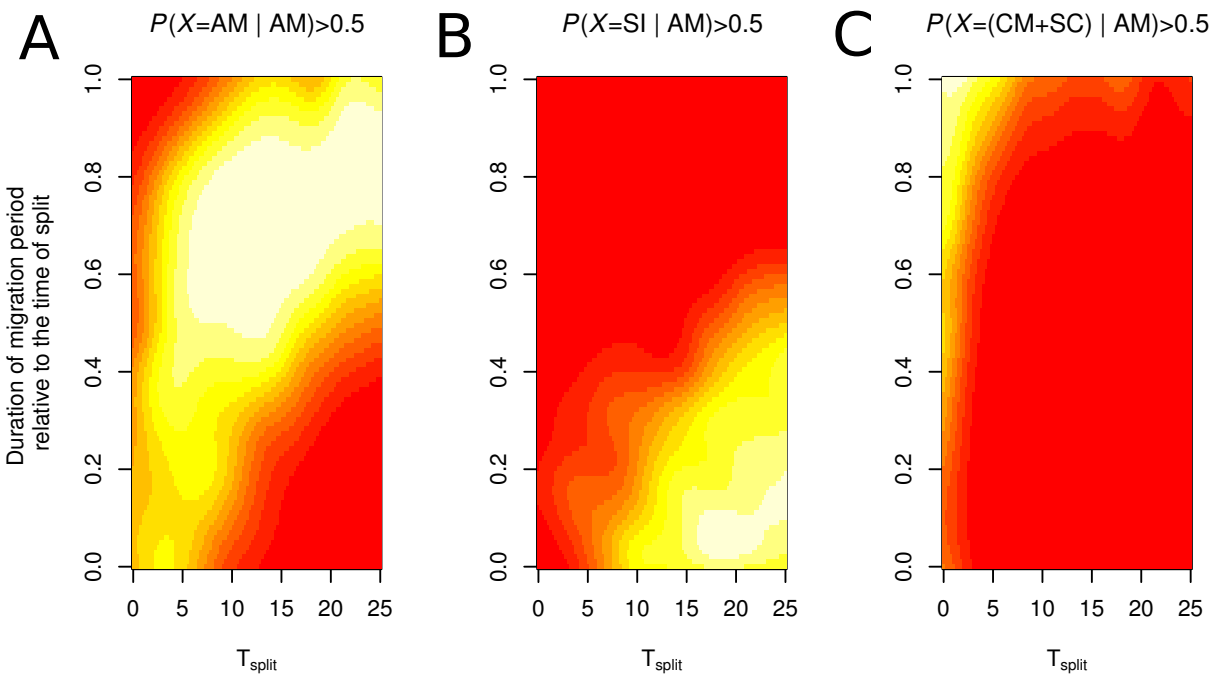

FIG. 1.7 – Distributions jointes des valeurs des paramètres T_{split} et $D(=(T_{split}-T_{AM})/T_{split})$ pour des simulations effectuées selon le modèle AM puis qui ont été assimilées par la méthode de sélection de modèles comme étant : (A) AM ($P(X=AM|AM)>0.5$), (B) SI ($P(X=SI|AM)>0.5$) ou (C) un modèle avec de la migration récente ($P(X=(CM+SC)|AM)>0.5$). Les valeurs les plus claires représentent les combinaisons de paramètres T_{split} et D les plus représentées.

1.4.3 Efficacité pour détecter de la migration récente.

L'efficacité de l'approche ABC pour identifier un modèle CM sachant que le jeu de données simulé résulte réellement d'un modèle CM est montré au Tableau 1.3 et à la Figure 1.8-A .

X	$P(X > 0.5\|CM)$	$P(X =$**max**$\|CM)$	$P(X > 0.5\|SC)$	$P(X =$**max**$\|SC)$
SI	0	0	0.001	0.002
CM	0.824	0.903	0.57	0.658
AM	0.006	0.017	0.007	0.019
SC	0.052	0.08	0.279	0.321

TAB. 1.3 – Proportions parmi 1,000 simulations CM et 1,000 simulations SC qui ont été attribuées aux modèles X par la procédure de sélection de modèle (décrite figure 1.3). Les modèles sont soutenus soit en ayant seulement la plus forte probabilité relative, soit en ayant une probabilité *a posteriori* supérieure à 0.5.

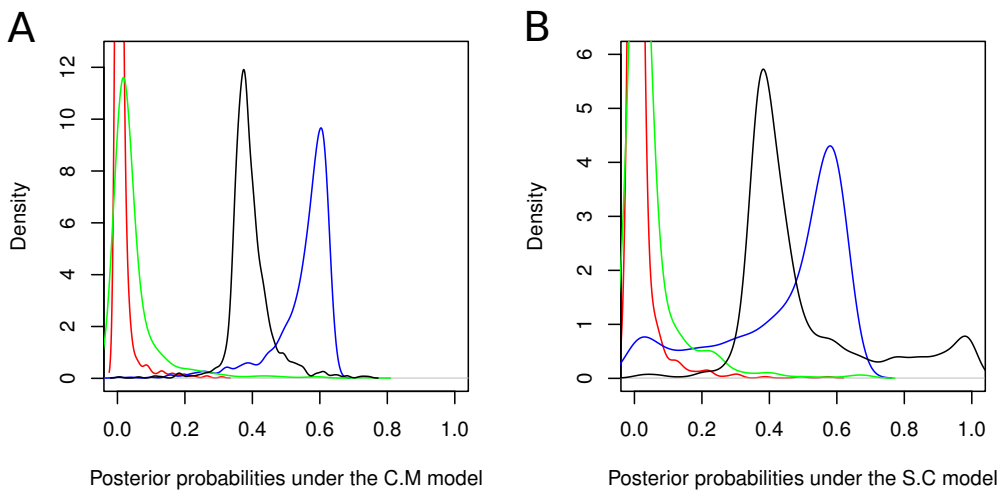

FIG. 1.8 – Distributions empiriques des probabilités relatives d'être associé aux modèles SI (rouge), CM (bleu), AM (vert) et SC (noir) quand (A) le modèle CM est le bon modèle et quand (B) SC est le bon modèle.
Ces distributions sont issues des simulations de 1,000 jeux de données multilocus sous le modèle CM et de 1,000 jeux de données sous le modèle SC. Pour chacun des jeux de données simulés nous avons estimé les probabilités relatives de chaque modèle selon la procédure décrite dans le matériel et méthode.

Une proportion de 90.3% des jeux de données est correctement soutenue par la classification de modèles. Les 9.7% de faux négatifs sont dus pour l'essentiel à une compétition avec le modèle SC (8%) puis, dans une moindre mesure, avec le modèle AM (1.7%). Le modèle SI n'est jamais désigné par la procédure de sélection de modèle quand le modèle simulé est CM. La même procédure de comparaison de modèles semble moins précise pour retrouver le modèle SC lorsqu'un jeu de données est effectivement simulé suivant ce modèle (Figure 1.8-B Tableau 1.3). Seulement 32.1% des comparaisons de modèles estiment correctement le modèle SC, alors que

65.8% des analyses produisent des faux négatifs en faveur du modèle CM et 2.1% des comparaisons supportent les modèles avec absence de migration récente. La figure 1.9 décrit l'influence des valeurs de paramètres (temps depuis la spéciation et durée relative de la période d'isolement précédent le contact secondaire) sur les résultats d'assignation des simulations sous modèle SC. Une assignation correcte n'est obtenue que dans les cas où le contact secondaire est précédé d'une période d'isolement prolongée et uniquement si l'évènement de spéciation est ancien (Figure 1.9-A). En revanche, pour des valeurs de durée d'isolement inférieures à la moitié du temps depuis la spéciation, l'approche ABC conduit à identifier un modèle CM plutôt que SC (Figure 1.9-B). Enfin, si la période de contact secondaire est très courte mais que le temps depuis la spéciation est lui-même relativement court, les échantillons simulés sont assignés aux modèles SI ou AM (Figure 1.9-C).

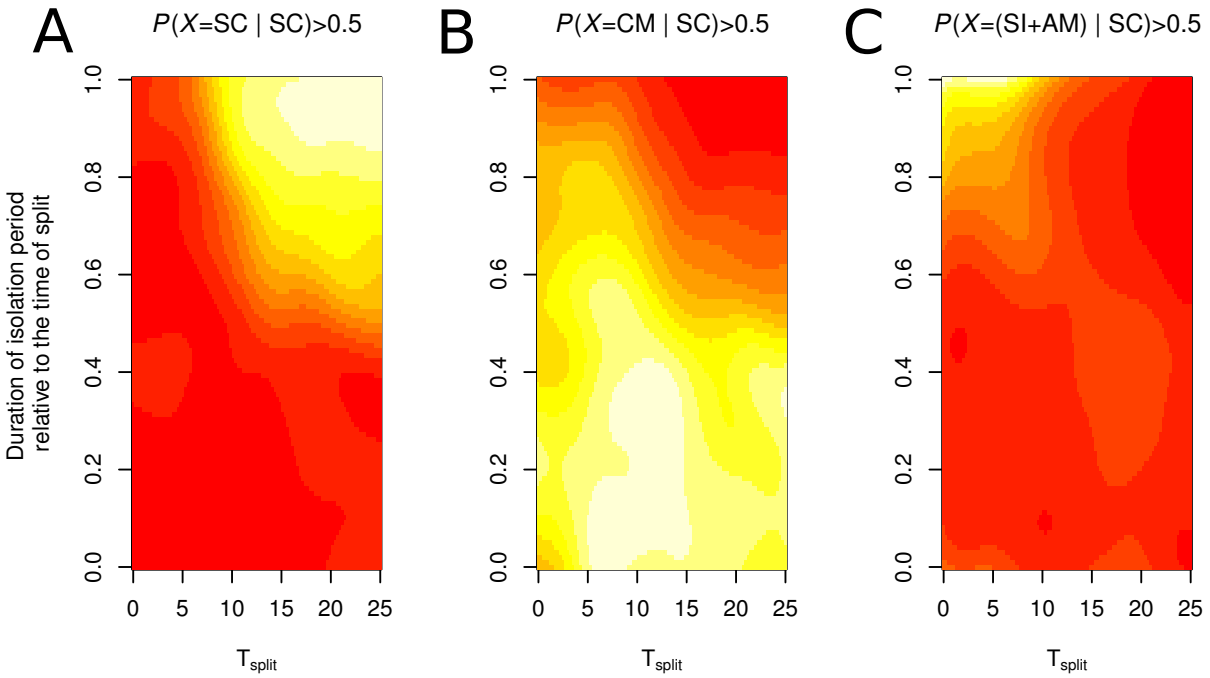

FIG. 1.9 – Distributions jointes des valeurs des paramètres T_{split} et $D(=(T_{split}-T_{SC})/T_{split})$ pour des simulations effectuées selon le modèle SC et qui ont été assimilées par la méthode de sélection de modèles comme étant : (A) SC ($P(X = SC|SC) > 0.5$), (B) CM ($P(X = CM|SC) > 0.5$) ou (C) un modèle sans migration récente ($P(X = (SI + AM)|SC) > 0.5$). Les valeurs les plus claires représentent les combinaisons de paramètres T_{split} et D les plus représentées.

1.4.4 Patrons canoniques de polymorphisme pour chaque scénario.

Pour illustrer les patrons de polymorphismes typiquement observés selon chacun des quatre scénarios démographiques, nous représentons sur la Figure 1.10 les spectres joints des fréquences des SNPs dérivés chez deux espèces en divergence suivant les scénarios SI, CM, AM et SC. Les fréquences des allèles dérivés pour chaque SNP ont été obtenues par scénario à partir des 10 jeux de données les mieux soutenus par l'approche ABC comme appartenant au bon scénario. En ce qui concerne le scénario SI, plus de 65% des sites polymorphes observés dans les données

de simulation sont des différences fixées entre les deux espèces. Les SNPs restant représentent des polymorphismes exclusifs de l'une des espèces (Figure 1.10-A Tableau 1.4).

Catégories de sites polymorphes	**Modèles**			
	SI	**AM**	**CM**	**SC**
S_x	34.05	84.82	46.52	32.63
S_s	0	9.95	51.12	65.24
S_f	65.95	1.78	0	0
S_{xf}	0	3.45	2.36	2.13

TAB. 1.4 – Proportions moyennes de chaque catégories de sites polymorphes mesurées sur les 10 jeux de données simulés par modèle qui ont été correctement soutenus avec la plus forte probabilié *a posteriori* parmi 1,000 simulations.

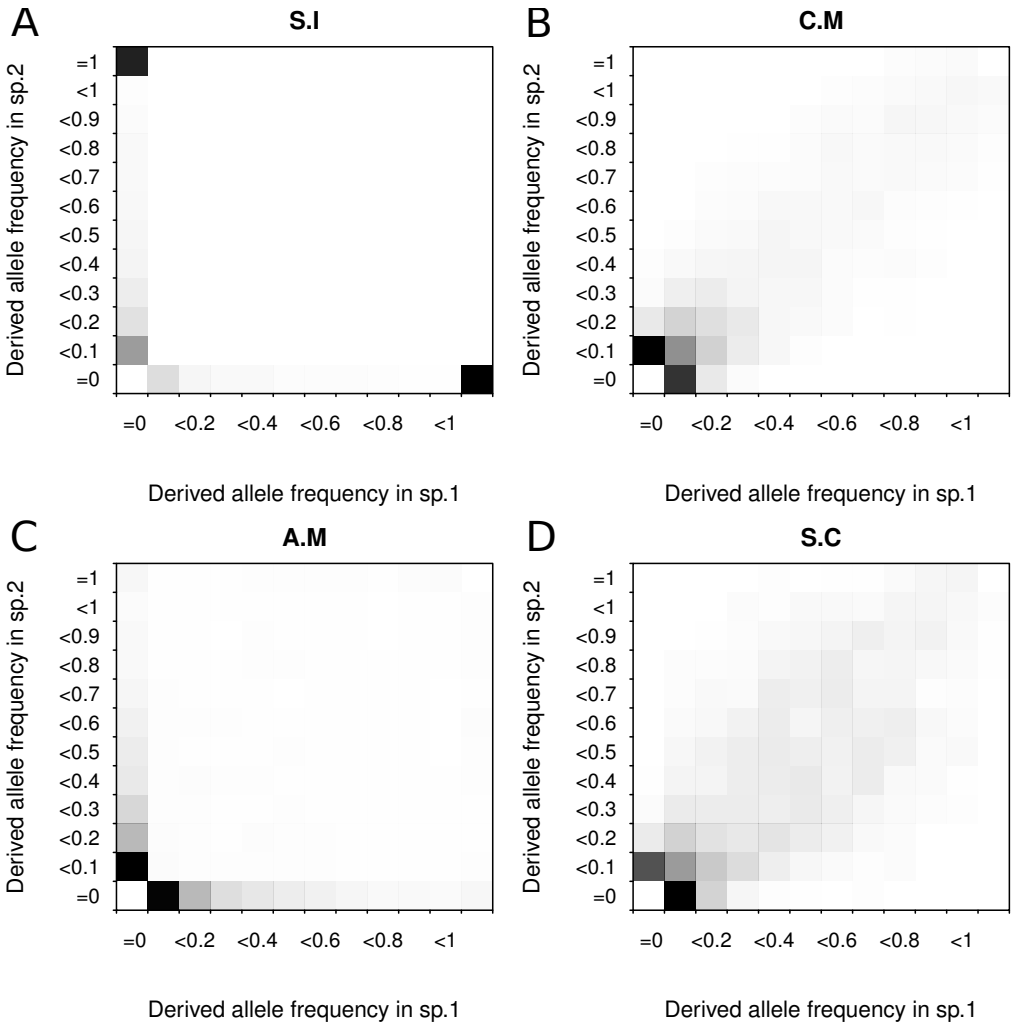

FIG. 1.10 – Spectres joints des fréquences des allèles dérivés pour l'ensemble des SNPs trouvés chez deux espèces différentes. Les graphiques ont été construits pour les modèles SI, CM, AM et SC en calculant les fréquences des allèles dérivées issues des 10 simulations les mieux estimées comme étant le bon modèle parmi les 1,000 cibles, suivant l'approche décrite figure 1.3. Pour chaque graphique, la catégorie représentée en noire désigne la classe la plus représentée. L'intensité de la coloration des autres classes dépend de leurs importances relatives à la classe la plus représentée. La figure 1.2 peut être traduite sur ce graphique de la façon suivante :
$S_{x1} : f2(d) = 0$ & $0 < f1(d) < 1$, où $f1(d)$ et $f2(d)$ représentent la fréquence de l'allèle dérivé chez les espèces 1 et 2 respectivement.
$S_{x2} : f1(d) = 0$ & $0 < f2(d) < 1$.
$S_{f1} : f1(d) = 1$ & $f2 = 0$.
$S_{f2} : f2(d) = 1$ & $f1 = 0$.
$S_s : 0 < f1(d) < 1$ & $0 < f2(d) < 1$.
$S_{x1f2} : 0 < f1(d) < 1$ & $f2(d) = 1$.
$S_{x2f1} : 0 < f2(d) < 1$ & $f2(d) = 0$.

Il n'y a donc aucun SNP partagé entre les deux espèces, ceci découle du fait que les valeurs de paramètres canoniques pour le scénario SI correspondent à des temps de divergence très élevés, nettement supérieurs aux temps de coalescence des généalogies de gènes au sein de chaque espèce fille. Le patron typique de séquences simulées suivant un scénario AM représente toutes les classes de sites polymorphes (Tableau 1.4), dont la plus représentée (85%) est celle des sites polymorphes exclusifs à l'une des deux espèces. Pour ces sites, l'allèle dérivé est majoritairement présent en très faible fréquence (Figure 1.10-C). L'action de la migration récente spécifique aux scénarios CM et SC est caractérisée dans des cas extrêmes par l'absence totale de différences fixées entre les deux espèces (Tableau 1.4), et par une prépondérance des polymorphismes partagés entre les deux espèces. Les spectres joints de fréquences d'allèles dérivés montrent des patrons très similaires entre les scénarios CM et SC. (Figure 1.10-B et Figure 1.10-D), ce qui confirme les difficultés d'assignation relative entre ces deux scénarios.

1.4.5 Performance de l'estimation des paramètres des scénarios SI, AM et CM.

Nous nous intéressons dans cette partie aux estimations des paramètres décrivant les scénarios SI, CM et AM pour dix jeux de données simulés selon chacun de ces trois scénarios et étant correctement assignés par la procédure de sélection de modèles. A ce stade, le faible nombre de jeux de données analysés ne permet qu'une vision grossière du comportement de l'approche sans réellement l'interpréter. Il est prévu d'étendre ultérieurement cette analyse aux 4,000 cibles précédemment utilisées pour le test de performance de la comparaison de scénarios. La figure 1.11 présente les valeurs réelles ainsi que les estimations fournies par l'ABC des quatre paramètres qui décrivent un scénario SI pour dix jeux de données effectivement simulés suivant un scénario SI.

Les distributions *a posteriori* estimées par l'ABC sont centrées autour des valeurs réelles avec relativement peu de variance pour les paramètres liés aux tailles des populations filles ainsi que pour le temps de spéciation. Les statistiques résumées sélectionnées pour les inférences bayésiennes contiennent donc l'information suffisante pour estimer avec précision et avec une bonne définition les paramètres N_1, N_2 et T_{split} (Figure 1.11-A, 1.11-B et 1.11-D). En revanche, les distributions *a posteriori* peuvent être plus larges pour le paramètre lié à la taille de la population ancestrale (Figure 1.11-C), mais contiennent toujours la valeur réelle. Becquet et Przeworski [Becquet and Przeworski, 2007] ont décrit le même manque de précision dans l'estimation de la taille ancestrale en SI. Ils ont montré par ailleurs que la variance de l'estimation de N_A diminue quand la taille d'échantillonnage en termes de nombre de locus étudiés augmente.

La précision des estimations de paramètres sous un scénario CM est illustrée sur la Figure 1.12. La précision des estimations des effectifs actuels des populations filles n'apparait pas être affectée par les évènements de migration (Figure 1.12-A et 1.12-B). Les distributions *a posteriori* des estimations des tailles ancestrales et des temps de spéciation sont plus étendues que celles obtenues en absence complète de migration (Figure 1.12-C et 1.12-D), et de ce fait contiennent la valeur réelle mais avec une faible précision de l'estimation, rendant difficile toute interprétation biologique. Cependant, dans les rares cas observés où les distributions *a posteriori* de N_A et/ou de T_{split} sont étroites (réplicas 6 et 7, Figure 1.12), on observe que ces distributions sont bien centrées sur la vraie valeur du paramètre et donc permettent une interprétation biologique pertinente. Les estimations des taux de migrations M_1 et M_2 sont généralement centrées sur la vraie valeur du paramètre, mais l'étendue de la distribution a posteriori est parfois très large (Figure 1.12-E et 1.12-F).

La figure 1.13 montre les inégalités dans les qualités d'estimations des sept paramètres décrivant le scénario AM. Les effectifs des populations filles sont correctement estimés avec une faible variance (Figure 1.13-A et 1.13-B). A l'inverse, les variances des estimations de l'effectif de la population ancestrale sont souvent trop grandes pour permettre une interprétation biologique pertinente, sauf dans un cas sur les dix simulations (réplica 10, Figure 1.13-C). Des analyses supplémentaires doivent être pratiquées pour mieux décrire la qualité de l'estimation de l'effectif ancestral dans un scénario AM, et en particulier préciser dans quelles situations les estimations sont particulièrement peu efficaces. En ce qui concerne l'estimation du temps de séparation des deux espèces, les résultats sont particulièrement hétérogènes avec même un cas dramatique montrant une distribution a posteriori étroite mais incorrecte car non centrée sur la valeur réelle du paramètre (réplica 4, Figure 1.13-D). Les estimations des taux de migrations présentent de larges variances ne permettant pas d'interprétation biologique pertinente (Figure 1.13-E et 1.13-F). En revanche, nous observons une grande précision dans l'estimation du temps d'installation de la période d'isolement des deux espèces après migration ancienne (Figure 1.13-G).

1.4.6 Effets de la violation de l'hypothèse de non-migration sur les estimations des paramètres.

Les estimations des paramètres du scénario SI par une approche ABC sont fortement affectées lorsque les données ont été simulées suivant un scénario CM (Figure 1.14). Mis à part pour deux analyses, les estimations des effectifs de la population ancestrale ne retrouvent pas les valeurs réelles malgré de faibles variances des distributions *a posteriori* (Figure 1.14-C). Les temps de spéciation sont systématiquement et largement sous-estimés (Figure 1.14-D). Ceci souligne l'importance d'une comparaison préalable de scénarios démographiques et notamment la mise en compétition de modèles avec ou sans migration récente entre les deux populations.

A l'inverse, nous avons testé l'efficacité pour l'estimation des 6 paramètres d'un scénario CM par une approche ABC sur des données simulées selon un modèle SI. Dans ce cas, les paramètres liés aux effectifs des populations filles sont estimés avec peu de variance et les distributions sont centrées sur les valeurs réelles (Figure 1.15-A 1.15-B). En outre, les estimations des effectifs de la population ancestrale (Figure 1.15-C) ainsi que celles des temps de spéciation (Figure 1.15-D) ne sont pas affectées par l'inclusion d'une migration potentielle dans le modèle utilisé pour l'ajustement (par comparaison à la Figure 1.10, car les jeux de données simulés sont identiques dans les deux cas). Enfin, en accord avec le modèle simulé (absence de migration), les estimations des taux de migration sont très proches de zéro et les variances des distributions a posteriori sont très faibles (Figure 1.15-E 1.15-F).

1.4.7 Conséquences des erreurs de la phase de sélection de modèles sur l'estimation de paramètres.

Dans un premier temps, nous nous intéressons aux estimations de paramètres pour des jeux de données simulés sous AM et ayant été assignés par la procédure de sélection de modèles à un scénario SI (Figure 1.16). Comme précédemment, les effectifs des populations filles sont estimés avec une grande fiabilité, à l'inverse des estimations de l'effectif de la population ancestrale (Figure 1.16-A, 1.16-C). Il a été montré précédemment que l'assignation erronée à un scénario SI d'un jeu de données simulé sous AM a lieu principalement lorsque la période de migration ancienne est restreinte. Bien que les intervalles de confiance des estimations pour T_{split} soient étroits, ils contiennent dans la majorité des cas les valeurs réelles à la fois pour le temps de

séparation et le temps d'arrêt de la migration ancienne (Figure 1.16-D). Pour les deux cas observés où la période de migration ancienne est relativement longue (réplicas 6 et 8), l'approche ABC a considéré le temps d'arrêt du palier de migration comme étant le temps de séparation des deux lignées selon un modèle SI, menant donc à une interprétation erronée pour le temps de spéciation, mais permet des interprétations biologiquement pertinente sur le temps d'arrêt d'échanges géniques entre les deux populations. La figure 1.17 montre les estimations de paramètres pour des jeux de données simulés sous S.I. ayant été assignés à un scénario AM par la procédure de sélection de modèle. Les effectifs des populations filles sont à nouveau estimés avec fiabilité et précision (Figure 1.17-A, 1.17-B). Les estimations de l'effectif de la population ancestrale sont variables en précision, mais en aucun cas ne mènent à des interprétations erronées (Figure 1.17-C). De la même manière, les estimations du temps d'arrêt de la migration (Figure 1.17-G) et du temps de spéciation (Figure 1.17-D) convergent et leurs distributions sont approximativement centrées sur les valeurs réelles de T_{split}, ce qui biologiquement se rapproche du scénario SI simulé au départ.

1.4.8 Sélection de modèle et estimation de paramètres par ABC pour des jeux de données publiés.

Nous utilisons des données de re-séquençage obtenues chez quatre espèces d'épicéa, deux espèces de capselles et six espèces de primates pour lesquelles des résultats issus d'analyses effectuées avec le logiciel MIMAR ont été publiés (Tableau 1.1).

Etonnamment, l'application de la procédure ABC pour la sélection de modèles indique que les modèles avec migration récente (et donc le modèle CM implémenté dans MIMAR) sont rejetés pour cinq des sept jeux de données publiés avec des résultats utilisant ce logiciel (Tableau 1.1). En effet, seuls les jeux de données *Capsella grandiflora - C. rubella* et *Pan. t. schweinfurthii - P. t. verus* soutiendraient les hypothèses du scénario CM. De plus, les modèles assignés par cette procédure incorporent une croissance exponentielle des effectifs dans tous les cas à l'exception de la comparaison *Picea purpurea - P. schrenkiana* (Tableau 1.1).

En analysant les jeux de données par l'approche ABC sous l'hypothèse du même modèle démographique que celui utilisé par le logiciel MIMAR (CM + effectifs constants depuis la séparation des deux espèces filles), nous obtenons des estimations de effectifs des populations filles globalement très proches de celles estimées par MIMAR (Tableaux 1.5 1.6 Figure 1.18-A 1.18-B).

Couple d'espèces	Auteurs	N_1	N_2	N_A
Capsella grandiflora vs. C. rubella	Foxe &al	463,600 (352,400-602,200)	4,700 (2,000-17,500)	463,600 (352,400-602,200)
	C. Roux	430,026 (332,534-670,543)	6,270 (3,270-17,801)	713,782 (231,652-1,144,745)
Picea likiangensis vs. P. schrenkiana	Li &al	155,000	33,000	10,000
	C. Roux	134,685.9 (103,873-205,858)	34,760 (23,924-56,392)	47,009 (14,815-243,937)
Picea purpurea vs. P. schrenkiana	Li &al	206,000	32,000	151,000
	C. Roux	284,275 (183,763-455,729)	35,127 (18,750-57,597)	28,237 (6,738-127,932)
Picea wilsonii vs. P. schrenkiana	Li &al	140,000	20,000	178,000
	C. Roux	174,738.1 (107,047-276,858)	24,090 (14,638-42,774)	39,656 (9,795-302,180)
Pan paniscus vs. P. t. troglodytes	Becquet & al	11,500 (9,150-15,200)	19,900 (15,300-25,600)	31,600 (22,200-48,700)
	C. Roux	12,432 (7,384-16,781)	21,984 (15,318-36,211)	4,820 (942.7078-49,656)
P. t. schweinfurthii vs. P. t. verus	Becquet & al	24,700 (18,600-71,800)	10,800 (8,040-21,100)	11,000 (2,270-21,900)
	C. Roux	20,547 (16,367-32,222)	9,857 (6,212-17,732)	4,286 (661-38,950)
Gorilla gorilla vs. G. beringei	Becquet & al	9,130 (5,090-14,100)	8,140 (3,570-18,100)	26,400 (5,990-49,100)
	C. Roux	7,423 (4,111-17,545)	2,634 (1,273-13,817)	43,130 (13,818-96,829)

TAB. 1.5 – Comparaison entre les estimations des paramètres N_1, N_2 et N_A de 7 jeux de données effectuées avec MIMAR et par ABC pour le modèle CM.

Les estimations faites par l'approche ABC et par MIMAR diffèrent de façon plus importante pour l'effectif de la population ancestrale mais sans biais systématique (Tableaux 1.5 et

Couple d'espèces	Auteurs	M_1	M_2	T_{split} (années)
Capsella grandiflora vs. C. rubella	Foxe &al	1.9 (0.007-4.2)	1.9 (0.007-4.2)	15,000 (800-3,546,800)
	C. Roux	0.292 (0.125-0.788)	2.129 (1.197-6.200)	3,621,068 (434,472-4,702,637)
Picea likiangensis vs. P. schrenkiana	Li &al	0.94	0.33	10,900,000
	C. Roux	0.086 (0.040-0.452)	0.090 (0.033-0.992)	15,877,080 (8,130,493-47,487,818)
Picea purpurea vs. P. schrenkiana	Li &al	0.48	0.5	3,200,000
	C. Roux	0.106 (0.045-0.404)	0.186 (0.081-1.009)	12,809,922 (7,316,389-38,571,544)
Picea wilsonii vs. P. schrenkiana	Li &al	2.12	0.06	5,100,000
	C. Roux	0.088 (0.037-0.556)	0.392 (0.207-1.411)	13,168,828 (5,120,687-91,381,910)
Pan paniscus vs. P. t. troglodytes	Becquet & al	0.062 (0.001-0.1)	0.062 (0.001-0.1)	785,000 (616,000-1,350,000)
	C. Roux	0.047 (0.022-0.308)	0.053 (0.018-0.276)	2,251,473 (1,484,487-2,954,076)
P. t. schweinfurthii vs. P. t. verus	Becquet & al	0.425 (0.143-2.622)	0.425 (0.143-2.622)	282,000 (230,000-1,210,000)
	C. Roux	0.487 (0.213-1.144)	0.071 (0.025-0.603)	1,135,732 (640,984-1,950,379)
Gorilla gorilla vs. G. beringei	Becquet & al	0.867 (0.282-2.059)	0.867 (0.282-2.059)	91,500 (84,300-1,440,000)
	C. Roux	0.424 (0.080-5.986)	0.864 (0.132-7.622)	646,950 (118,903-1,703,858)

TAB. 1.6 – Comparaison entre les estimations des paramètres M_1, M_2 et T_{split} de 7 jeux de données effectuées avec MIMAR et par ABC pour le modèle CM.

1.6, Figure 1.18-C). En revanche, les temps de spéciation estimés par l'approche ABC sont systématiquement plus élevés que ceux estimés par MIMAR (Tableaux 1.5 et 1.6, Figure 1.18-D).

1.5 Discussion et perspectives.

Nous avons testé ici les possibilités d'utiliser efficacement une approche ABC pour différencier quatre modèles démographiques de divergence sur la base des patrons de flux de gènes historiques. Ces quatre modèles ont été proposés pour simuler des scénarios de spéciation contrastés [Hey, 2006, Ross-Ibarra et al., 2009] tout en gardant à l'esprit qu'ils représentent de fortes simplifications de la réalité biologique comme le soulignent Box & Draper [Box and Draper, 1987] : "Remember that all models are wrong ; the practical question is how wrong do they have to be to not be useful" (cités par [Beaumont et al., 2010]). Nos résultats montrent que la procédure de sélection de modèles utilisée permet de distinguer certains scénarios ou groupes de scénarios, mais la fiabilité de l'approche dépend fortement des valeurs des paramètres considérées. Nos résultats montrent également que l'estimation des valeurs de paramètres peut être fortement altérée par l'utilisation d'un modèle de divergence non approprié. Ceci concorde avec les résultats de Becquet et Przeworski [Becquet and Przeworski, 2009] qui ont étudié le comportement des logiciels IM et MIMAR (faisant l'hypothèse d'un scénario CM) face à un biais dans le modèle démographique. En effet, les deux méthodes sont fortement affectées pour l'ensemble des paramètres lorsque les données correspondent à un processus de contact secondaire entre les espèces. L'effet parfois spectaculaire de cet écart au modèle risque de conduire les expérimentateurs à des conclusions erronées si ces méthodes sont appliquées à des situations biologiques qui s'écartent fortement du modèle considéré par IM et MIMAR. Cet effet justifie donc pleinement le recours à des procédures élaborées de comparaison de modèles alternatifs [Csilléry et al., 2010].

Nos résultats indiquent que l'approche ABC pour la sélection de modèles détecte efficacement des évènements de migration récents dans un scénario de divergence entre deux populations. Une telle approche permet de soutenir cette affirmation avec des arguments statistiques, contrairement à l'utilisation des approches IM et MIMAR qui appliquent un scénario CM ce qui nécessite le choix d'un seuil arbitraire de taux de migration en-dessous duquel l'observateur conclut à une absence de migration. Nos résultats indiquent également que s'il est possible de distinguer dans une importante majorité de cas les modèles d'isolement (SI) de ceux avec de la migration ancienne (AM), les expérimentateurs utilisant cette approche doivent être conscient du fort taux de faux négatifs. Ainsi, la probabilité d'assigner correctement un jeu de données à un scénario SI augmente si les deux espèces étudiées sont isolées depuis des temps de spéciation anciens.

Ceci est lié à l'augmentation de la proportion de généalogies de gènes montrant une coalescence complète au sein de chaque espèce. Ce signal spécifique d'un isolement strict diminue pour des temps de séparation plus courts. Dans un tel cas, nous nous attendons à observer une plus grande proportion de polymorphismes d'origine ancestrale chez les deux populations filles, une proportion plus faible de différences fixées, et également une plus grande proportion d'allèles dérivés exclusifs présents en faibles fréquences. C'est ainsi que des données non affectées par de la migration depuis la séparation des lignées peuvent être néanmoins assignées à un scénario comportant une signature de migration ancienne (AM). Dans de telles situations biologiques, l'effet de la rétention du polymorphisme ancestral peut exceptionnellement être confondu avec celui d'une migration récente.

Bien que des données simulées suivant un modèle AM ne soient pas correctement soutenues de façon systématique par la méthode de sélection de modèle, le biais observé conserve un sens biologique. En effet, si nous définissons les modèles comme étant quatre catégories discrètes, certaines combinaisons de paramètres démographiques peuvent converger vers des interprétations biologiquement semblables. C'est donc le devoir de l'expérimentateur d'estimer au préalable dans quelle mesure il doit distinguer précisément un modèle avec séparation lointaine sans migration d'un modèle avec séparation aussi ancienne mais suivit d'une très courte période d'échanges de gènes. Si confusion entre les modèles SI et AM il y a, son incidence sur l'estimation des paramètres est identifiable dans le cadre de l'ABC contrairement à ce qu'il a été observé pour les logiciels IM et MIMAR [Becquet and Przeworski, 2009]. En effet, l'utilisation de ces méthodes MCMC pour des jeux de données AM tend à fournir des valeurs sous-estimées du temps de spéciation, sans aucun moyen de vérifier de quel scénario se rapproche le jeu de données étudié. Ainsi, en cas de choix erroné de modèle démographique en ABC, un jeu de données avec pour histoire la subdivision d'une population ancestrale au temps T_{split} en deux populations filles échangeant des migrants jusqu'au temps T_{AM} sera interprété comme une spéciation à un temps T_{split} proche de la valeur de T_{AM}. Réciproquement, les estimations des paramètres du modèle AM sur des données ayant réellement eu pour histoire la séparation au temps T_{split} de deux lignées isolées sera interprétée comme un modèle AM où T_{split} et T_{AM} seront proches de la vraie valeur de T_{split}. Si cela entraine une perte d'information quant à la vraie date de séparation des deux lignées ou quant aux taux de migration ancienne, les expérimentateurs peuvent néanmoins avoir une grande confiance dans les dates estimées à partir desquelles les deux lignées ont arrêté d'échanger des gènes contrairement avec IM et MIMAR.

Pour les modèles intégrant de la migration récente, la confusion possible entre les deux scénarios CM et SC se présente de manière différente que dans le cas des modèles SI et AM. En effet, un jeu de données aléatoirement simulé suivant un modèle CM aura effectivement 90% de chances d'être correctement soutenu par la comparaison de modèles et souffrira peu d'être affecté par de faux négatifs en faveur du modèle SC. A l'inverse, la majorité des jeux de données suivant un modèle SC seront soutenus comme appartenant au modèle CM. Heureusement, le support d'un modèle SC par la sélection de modèles est rarement un faux positif. Pour qu'un contact secondaire soit identifié comme tel, les deux lignées doivent avoir évoluées en isolement complet pendant une période suffisamment longue pour que les allèles dérivés exclusifs atteignent des fréquences intermédiaires avant le contact secondaire. Si l'impact de la présence ou non de migration ancienne n'a pas d'incidence majeure sur l'estimation du temps à partir duquel deux espèces évoluent en isolement, des analyses futures devront être effectuées pour mesurer l'impact d'une confusion entre CM et SC par la méthode de choix de modèle sur les estimations de paramètres.

De plus, un expérimentateur qui restreint son analyse à l'estimation de paramètres

démographiques doit avoir conscience que la ré-analyse par deux méthodes de jeux de données publiés utilisant les mêmes alignements et le même modèle démographique peuvent diverger dans les estimations de certains paramètres. Ceci a une incidence majeure sur la manière d'interpréter des jeux de données, selon le paramètre sur lequel se repose l'argumentation. Ainsi, en estimant des temps de spéciation d'environ 15,000 ans pour trois des quatre configurations d'analyses publiées avec MIMAR (la quatrième estimant un temps de séparation ancien de $1.4\text{x}10^6$ années), Foxe &al intitulent leur article "Recent Speciation associated with the evolution of selfing in Capsella". L'analyse du même modèle démographique effectuée en utilisant une approche ABC estime au contraire une spéciation plus ancienne de $3.6\text{x}10^6$ ans, mieux supportée par une analyse d'ajustement des données (Tableau 1.7). De même, l'hypothèse de taille constante des populations filles depuis la spéciation est souvent rejetée au profit des modèles intégrant une croissance exponentielle. Même si l'effet d'un tel écart réaliste au modèle sur les estimations fournies par MIMAR peut être testé, la détection d'une expansion de population ou d'un effet fondateur est une information clée sur le mode de spéciation que seul l'ABC authorise.

A partir de ce constat, une réflexion devient nécessaire sur la place dans la bibliographie que doivent avoir les argumentations basées sur les estimations d'un seul paramètre, par une seule méthode et pour un seul modèle démographique. Réflexion qui doit être effectuée paramètre par paramètre et modèle par modèle pour l'ensemble des méthodes existantes, avec les données déjà disponibles, pour savoir dans quelle mesure les méthodes MCMC existantes sont des boites à fantasmes ou non. En parallèle doit être poursuivi un effort exploratoire pour mieux comprendre les limites de l'ABC, ainsi que pour optimiser la manière de résumer les données génétiques dans les inférences. Une voie à explorer est l'utilisation d'une description des patrons de polymorphismes au moyen des spectres joints de fréquences des allèles dérivés observés pour chaque couple d'espèce (Figure 1.19), et de comparer avec les résultats obtenues en résumant les données par la moyenne et l'écart type du nombre de site dans chaque catégorie.

Analysis		Foxe &al (MIMAR)	Roux (ABC)
$Bi-allelicpositions$	**Mean**	0.171	0.303
	SD	0.1685	0.0895
S_f	**Mean**	0.068	0.3175
	SD	0.1085	0.1765
$S_{x-grandiflora}$	**Mean**	0.0985	0.4275
	SD	0.224	0.061
$S_{x-rubella}$	**Mean**	0.066	0.2695
	SD	0.117	0.1165
S_s	**Mean**	0.2145	0.229
	SD	0.3025	0.346
$\pi_{grandiflora}$	**Mean**	0.121	0.1965
	SD	0.4465	0.424
$\pi_{rubella}$	**Mean**	0.282	0.1335
	SD	0.409	0.24
$\theta_{grandiflora}$	**Mean**	0.199	0.2755
	SD	0.17	0.1215
$\theta_{rubella}$	**Mean**	0.215	0.216
	SD	0.3415	0.462
$Taj\ D_{grandiflora}$	**Mean**	0.3275	0.3925
	SD	0.125	0.1205
$Taj\ D_{rubella}$	**Mean**	0.147	0.4245
	SD	0.385	0.223
$GrossDivergence$	**Mean**	**0.031**	0.263
	SD	0.265	0.0755
$NetDivergence$	**Mean**	**0.0065**	0.422
	SD	0.338	0.1135
F_{ST}	**Mean**	**0.0045**	0.469
	SD	**0.0315**	**0.0405**

TAB. 1.7 – Test d'ajustement des estimations des paramètres du modèle CM effectuée par MIMAR et par ABC au jeu de données observées pour le couple *Capsella*. Les valeurs en gras indiquent des valeurs P inférieures à 5%.

1.6 Annexes

Les lignes de commandes utilisées pour simuler des échantillons sous les scénarios SI, CM, AM et SC, avec effectifs constants dans le temps sont respectivement :
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 0 -m 2 1 0 -n 1 tbs -n 2 tbs -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 tbs -m 2 1 tbs -n 1 tbs -n 2 tbs -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 0 -m 2 1 0 -n 1 tbs -n 2 tbs -ema tbs 2 0 tbs tbs 0 -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 tbs -m 2 1 tbs -n 1 tbs -n 2 tbs -eM tbs 0 -ej tbs 2 1 -eN tbs tbs

Les lignes de commandes pour les modèles alternatifs des scénarios SI, CM, AM et SC intégrant des croissances exponentielles pour les deux populations filles sont :
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 0 -m 2 1 0 -g 1 tbs -n 1 tbs -eg tbs 1 0 -g 2 tbs -n 2 tbs -eg tbs 2 0 -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 tbs -m 2 1 tbs -g 1 tbs -n 1 tbs -eg tbs 1 0 -g 2 tbs -n 2 tbs -eg tbs 2 0 -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 0 -m 2 1 0 -g 1 tbs -n 1 tbs -eg tbs 1 0 -g 2 tbs -n 2 tbs -eg tbs 2 0 -ema tbs 2 0 tbs tbs 0 -ej tbs 2 1 -eN tbs tbs
./msnsam tbs niter -t tbs -r tbs tbs -I 2 tbs tbs 0 -m 1 2 tbs -m 2 1 tbs -g 1 tbs -n 1 tbs -eg tbs 1 0 -g 2 tbs -n 2 tbs -eg tbs 2 0 -eM tbs 0 -ej tbs 2 1 -eN tbs tbs

Le logiciel msnsam va attribuer aux arguments associés à un élément tbs (To Be Specified) une valeur puisée dans la distribution a priori.

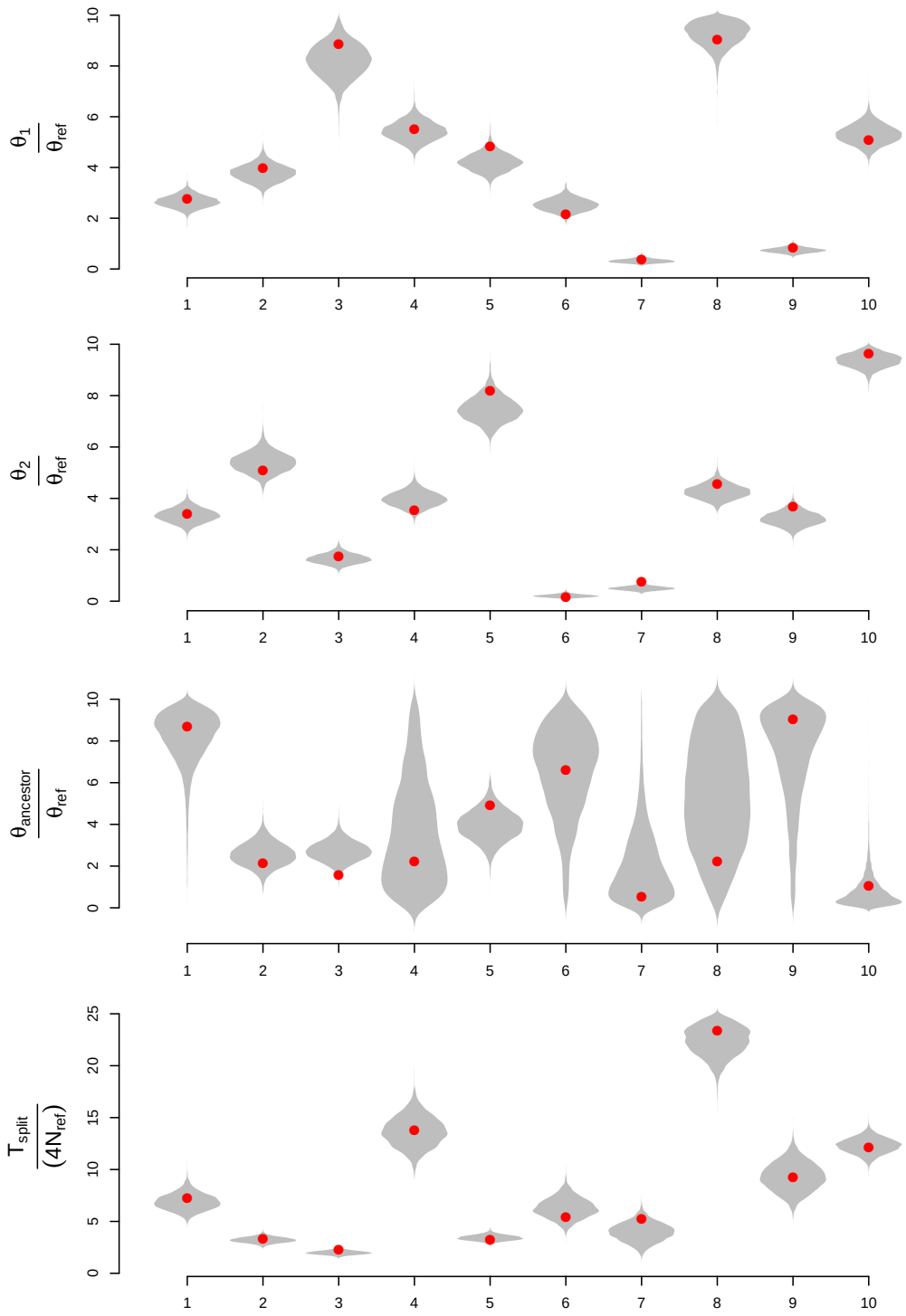

FIG. 1.11 – Estimations des quatre paramètres θ_1, θ_2, θ_A, et T_{split} du scénario SI pour 10 jeux de données simulés puis correctement soutenus par la procédure de sélection de modèles. Les valeurs réelles aléatoirement choisies sont représentées par des points rouges. Les régions grisées correspondent aux distributions *aposteriori*.

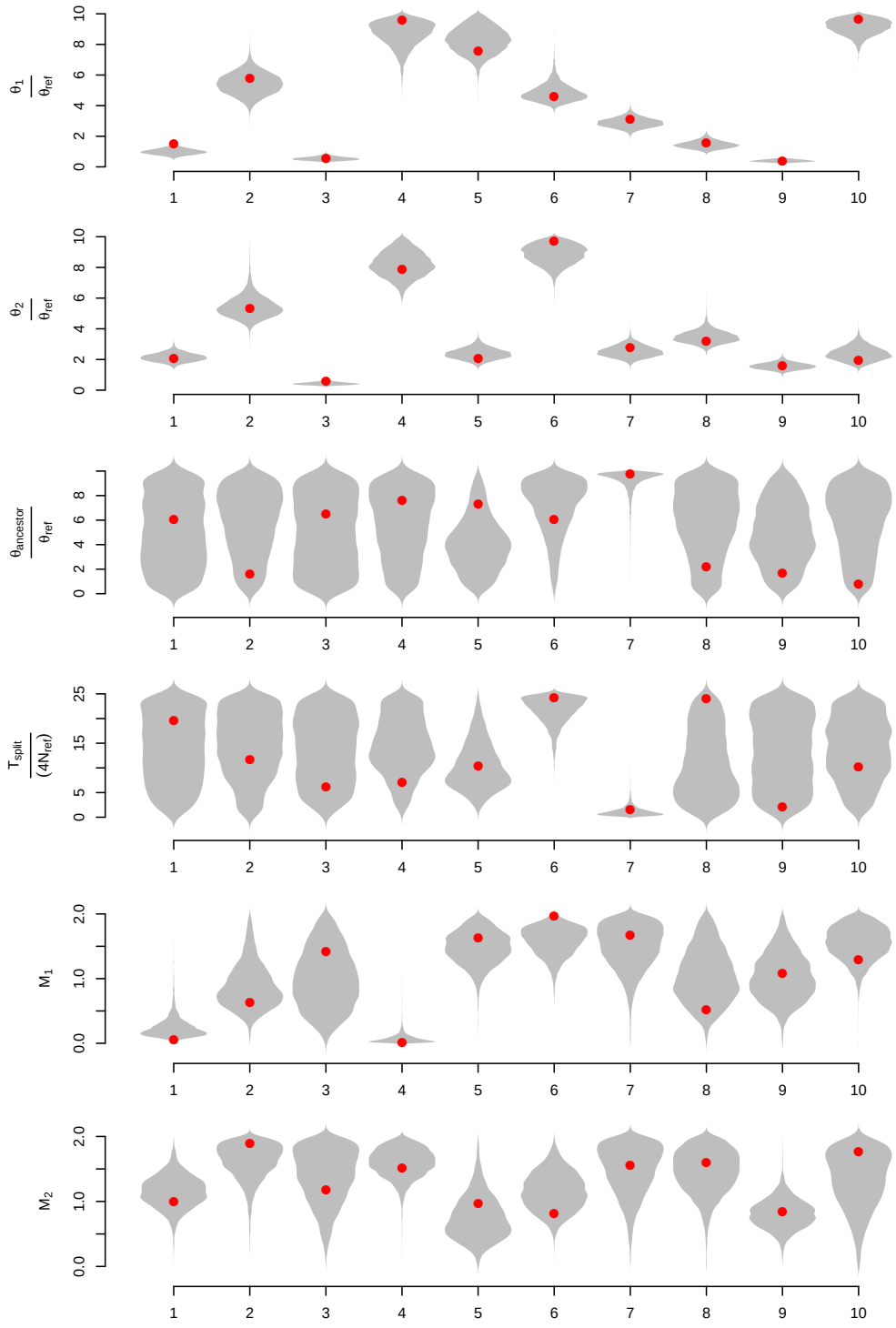

FIG. 1.12 – Estimations des 6 paramètres θ_1, θ_2, θ_A, T_{split}, M_1 et M_2 du scénario CM pour 10 jeux de données simulés puis correctement soutenus par la procédure de sélection de modèles. Les valeurs réelles aléatoirement choisies sont représentées par des points rouges. Les régions grisées correspondent aux distributions *aposteriori*.

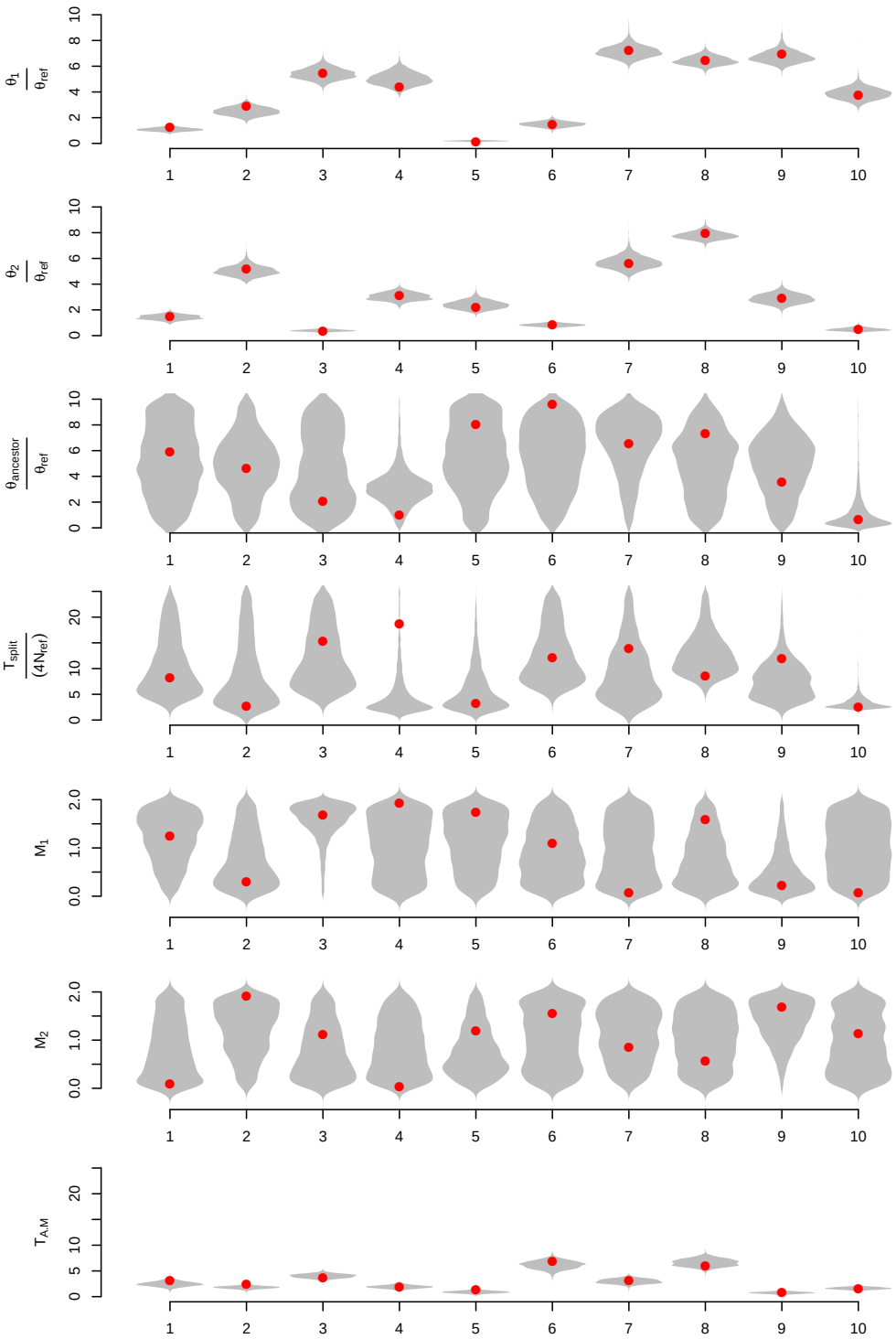

FIG. 1.13 – Estimations des 7 paramètres θ_1, θ_2, θ_A, T_{split}, M_1, M_2 et T_{AM} du scénario AM pour 10 jeux de données simulés puis correctement soutenus par la procédure de sélection de modèles. Les valeurs réelles aléatoirement choisies sont représentées par des points rouges. Les régions grisées correspondent aux distributions *aposteriori*.

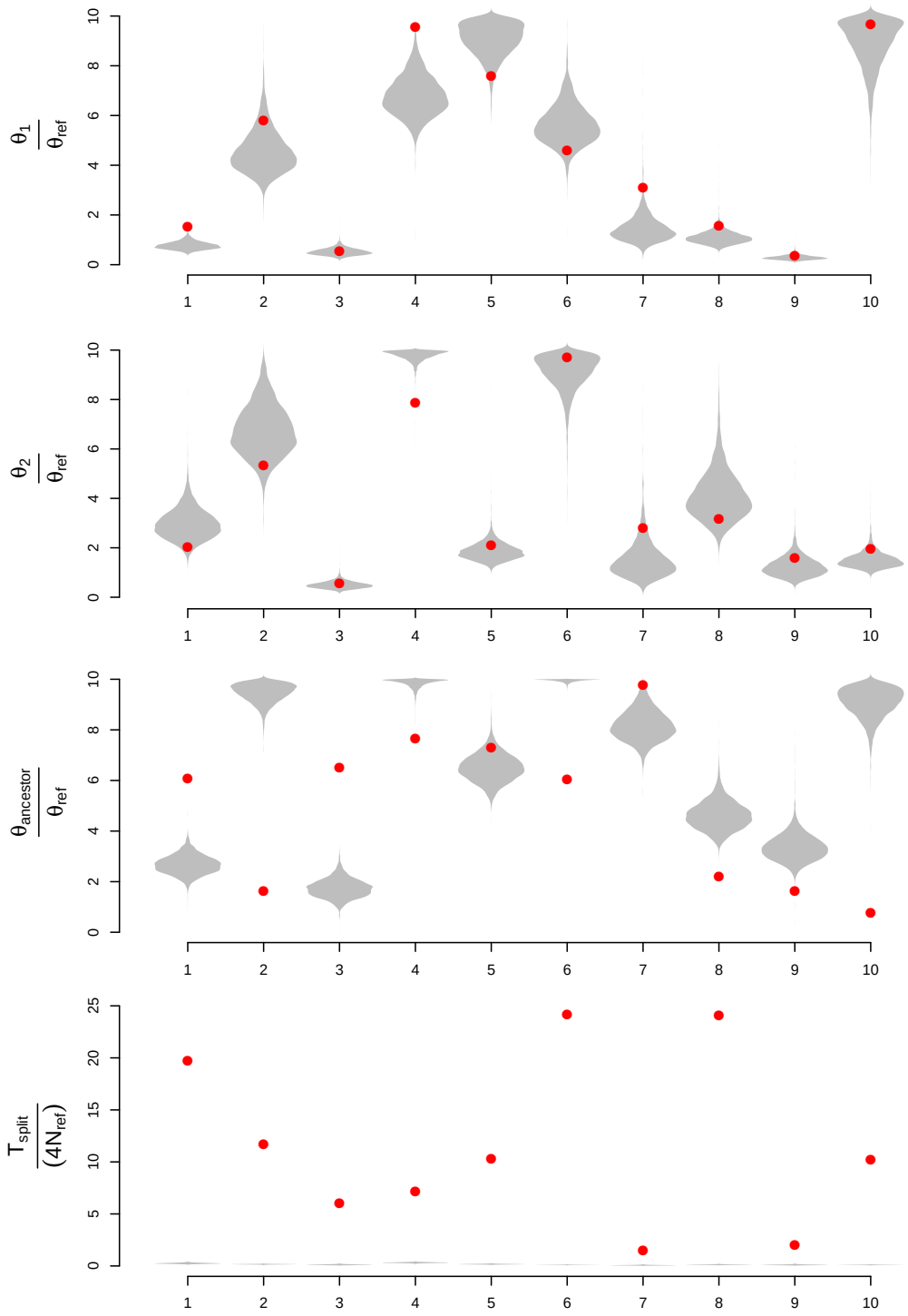

FIG. 1.14 – Estimations des quatre paramètres θ_1, θ_2, θ_A et T_{split} du scénario SI pour 10 jeux de données simulés selon un modèle CM. Les valeurs réelles aléatoirement choisies sont représentées par des points rouges. Les régions grisées correspondent aux distributions *aposteriori*.

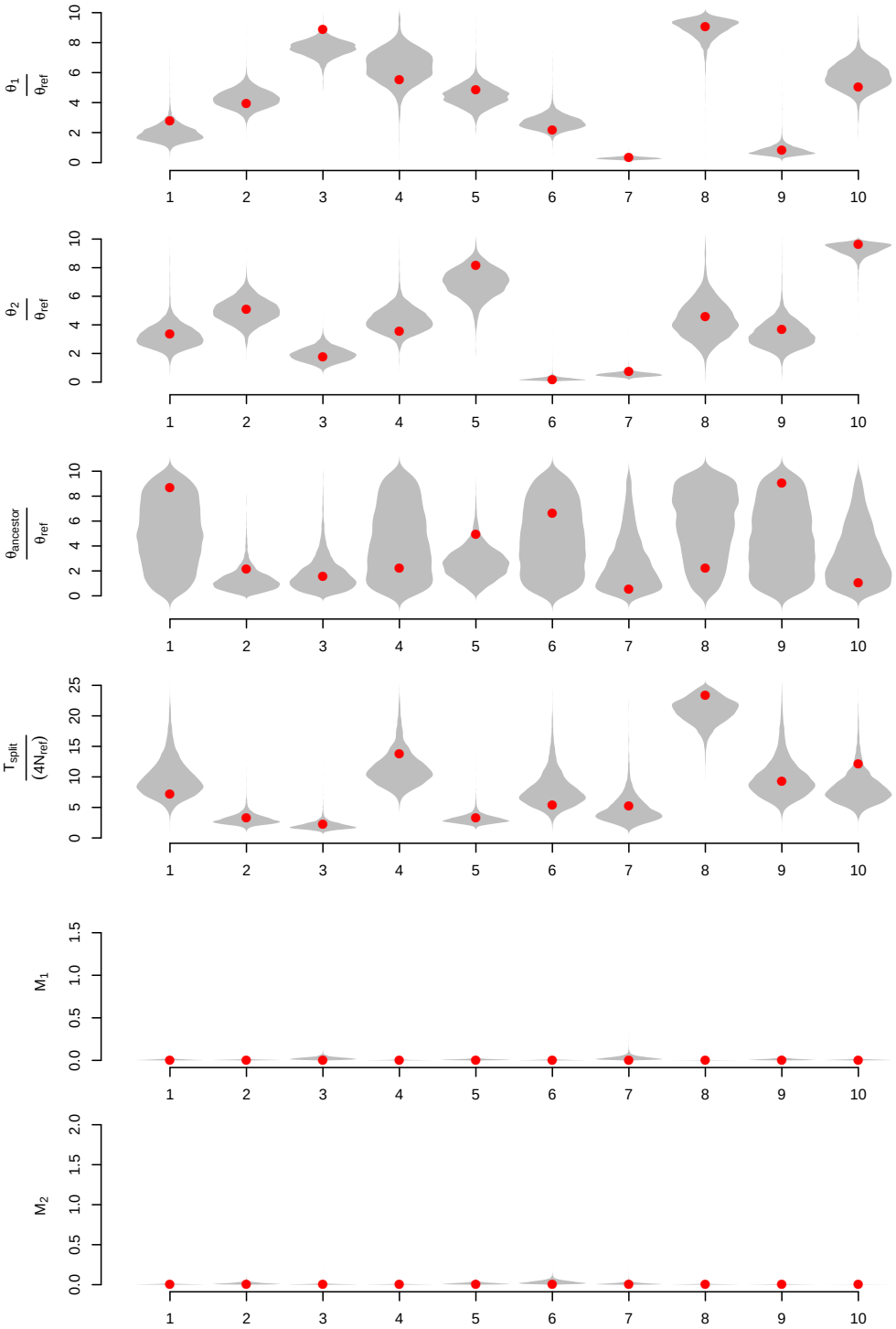

FIG. 1.15 – Estimations des 6 paramètres θ_1, θ_2, θ_A, T_{split}, M_1, et M_2 du scénario CM pour 10 jeux de données simulés selon un modèle SI. Les valeurs réelles aléatoirement choisies sont représentées par des points rouges. Les régions grisées correspondent aux distributions *aposteriori*.

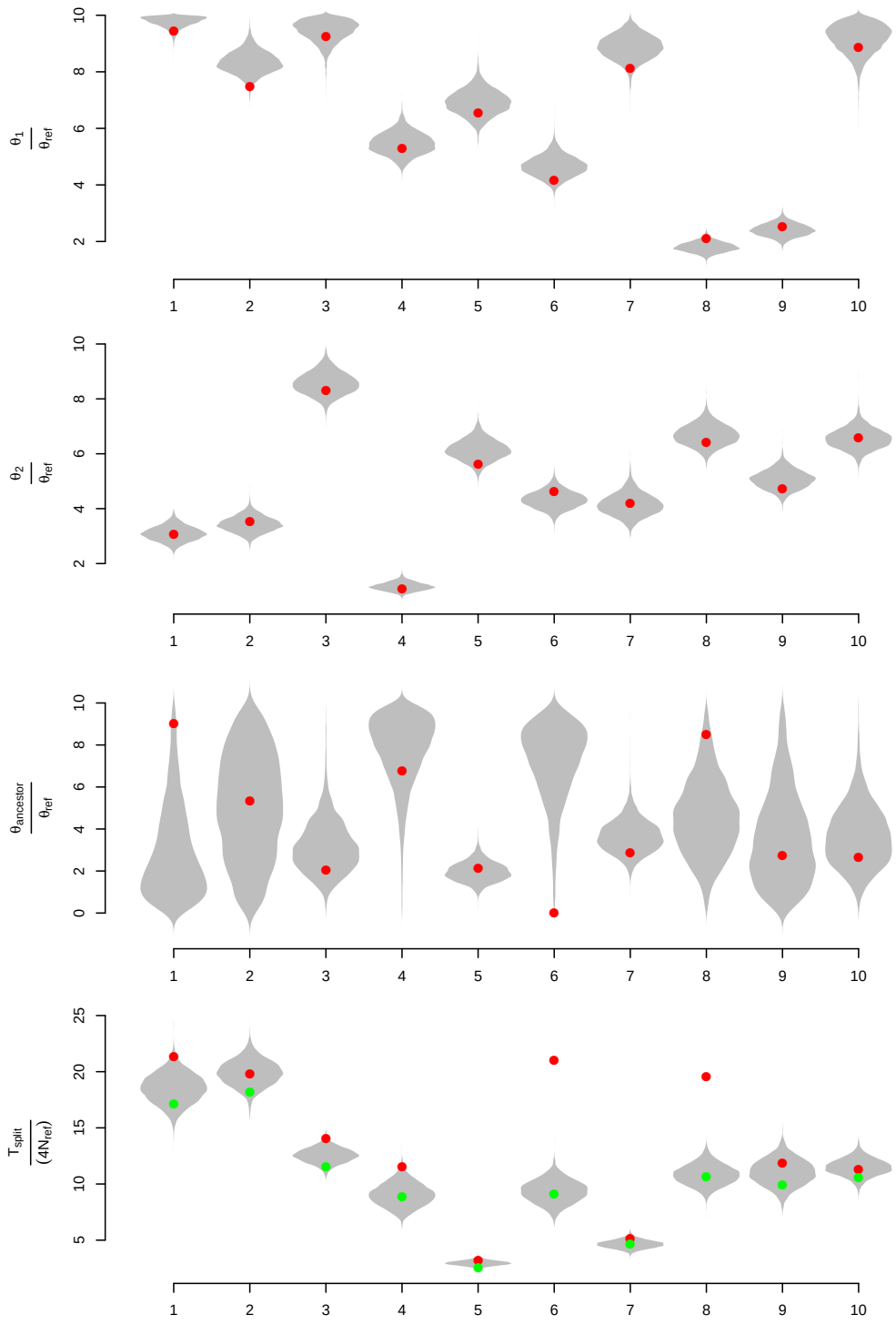

FIG. 1.16 – Estimations des paramètres d'un scénario SI pour 10 jeux de données issus du scénario AM mais dont une probabilité *aposteriori* supérieure à 0.5 d'appartenir au scénario SI fut attribuée par la procédure de sélection de modèles. Les valeurs réelles, aléatoirement choisies, correspondent aux points rouges, à l'exception de la valeur réelle du temps T_{AM} d'arrêt de la migration ancienne représenté en vert. Les régions grisées correspondent aux distributions *aposteriori* pour chacun des quatre paramètres et pour chacune des 10 analyses.

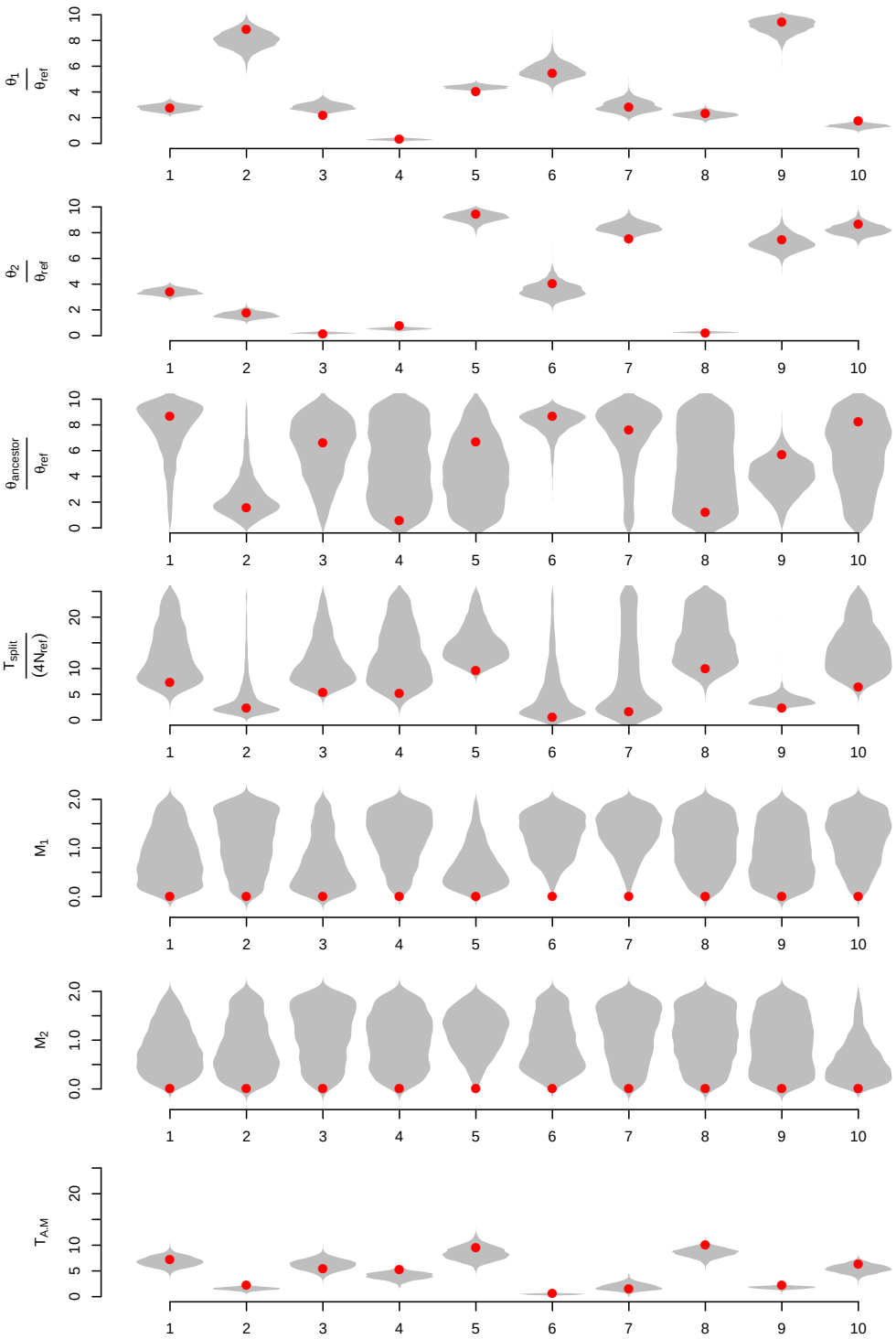

FIG. 1.17 – Estimations des paramètres AM pour 10 jeux de données issus du scénario SI mais dont une probabilité *aposteriori* supérieure à 0.5 d'appartenir au scénario AM fut attribuée par la procédure de sélection de modèles. Les valeurs réelles, aléatoirement choisies, correspondent aux points rouges. Les valeurs réelles pour les taux de migrations est de zéro.

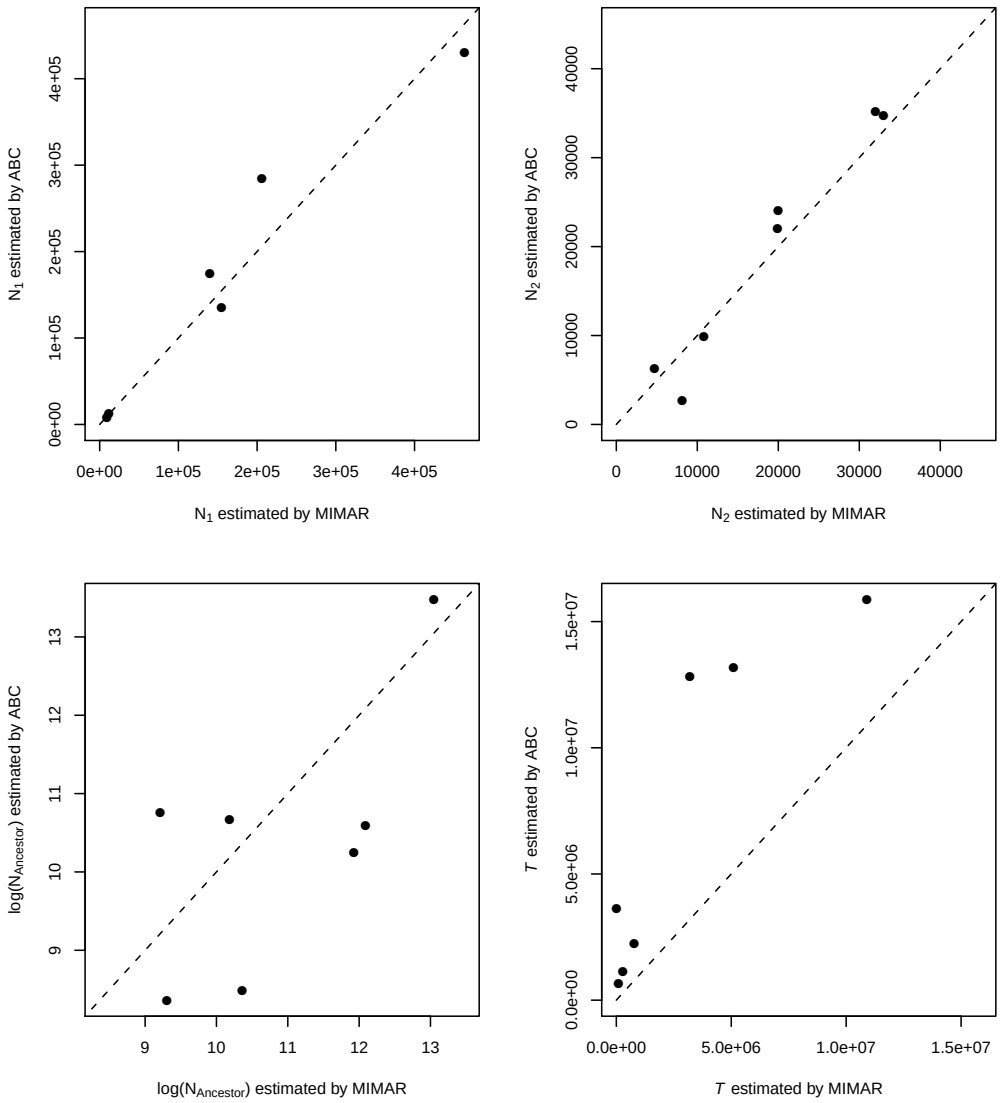

FIG. 1.18 – Estimations des paramètres θ_1, θ_2, θ_A et T_{split} du modèle CM obtenues par MIMAR et par une approche ABC à partir de jeux de données similaires. Pour un paramètre donné, un point éloigné au dessus de la diagonale en pointillée indique une valeur estimée de ce paramètre plus élevée en ABC que par MIMAR. Un point éloigné en dessous de la diagonale indique une valeur d'estimation par MIMAR plus petite que par une approche ABC.

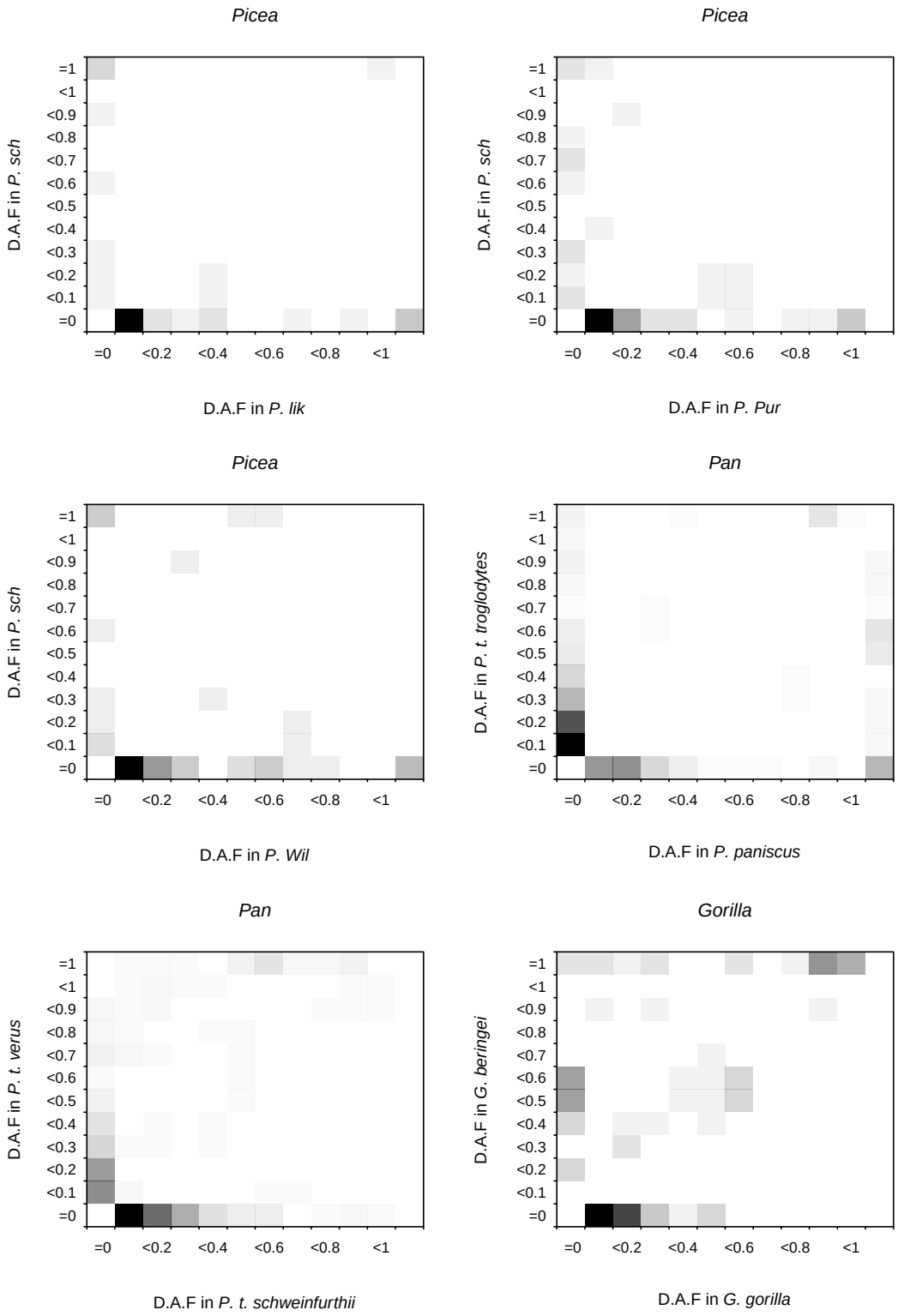

FIG. 1.19 – Spectres joints de fréquences des allèles dérivés (Derived Allele Frequency) calculés pour 6 jeux de données publiés. L'allèle dérivé est déterminé en utilisant des séquences orthologues chez un groupe externe proche. L'absence de tel groupe externe pour le couple *Capsella* nous empêche de représenter fidèlement le spectre. Dans le papier d'origine, Foxe &al ont utilisé des séquences obtenues chez *A. thaliana*.

Chapitre 2

Demographic history and adaptation genomics in the Arabidopsis genus

2.1 Résumé.

Au sein du genre végétal Arabidopsis, l'espèce *A. halleri* est la seule à posséder une architecture génomique impliquée dans une concentration surélevée dans les parties aériennes des cations divalents Zinc et Cadmium puisés dans le sol. Ce trait phénotypique appelé hyperaccumulation des métaux est la résultante fonctionnelle d'un ensemble de protéines impliquées dans l'assimilation des métaux à partir du sol, de leur entrée dans le cortex racinaire, de leur passage dans la stèle racinaire, du transfert dans le xylème, de leur remobilisation vers les tissus aériens à partir du xylème, de leur entrée dans les cellules puis dans la compartimentation sub-cellulaire dans la vacuole et les organites (plastes et mitochondries). Bien que les métaux soient nécessaires à la conformation optimale de nombreuses protéines et bien que leurs propriétés oxydo-réductrices soient utilisées comme catalyseurs enzymatiques pour réaliser des transferts d'électrons, un excès en métaux entraine un risque de toxicité pour la cellule en créant des espèces réactives en oxygène désorganisant des structures protéiques, interagissant négativement avec la membrane lipidique et ayant des effets mutateurs de l'ADN. Claude Bernard a défini cette régulation physiologique fine d'un élément essentiel par le développement du concept d'homéostasie. Ainsi, le risque de toxicité lié à l'hyperaccumulation de métaux implique des mécanismes permettant le maintien de l'homéostasie métallique dans les cellules végétales. Ce double caractère hyperaccumulateur et hypertolérant est rencontré chez toutes les populations naturelles européennes et asiatiques d'*A. halleri* étudiées. Les populations d'*A. halleri* en Europe se rencontrent à la fois sur des sites métallifères et sur des sites non contaminés par les métaux, qui diffèrent ainsi par une moins forte pression de sélection exercée par la concentration en métaux dans le sol. La fixation de ce trait phénotypique dans l'ensemble d'une espèce non endémique des sols pollués soulève la question de l'évolution d'une telle architecture génétique dans la lignée *A. halleri*. Ce caractère constitutif à *A. halleri* pourrait être expliqué par des processus récents et indépendants de sélection des génotypes hypertolérants à partir de la "standing variation" en conséquence de l'anthropisation importante de certains milieux, suivies de colonisations vers des milieux non pollués. L'hypothèse opposée faisant également intervenir la sélection naturelle serait la sélection du caractère hyperaccumulateur au moment de la séparation des lignées menant vers *A. halleri* et *A. lyrata* par une pression de sélection indéterminée, en rapport direct ou non avec les concentrations en métaux lourds rencontrées dans leur environnement historique.

Dans ce chapitre de thèse intitulé "Demographic history and adaptation genomics in the Arabidopsis genus" j'étudie dans un premier temps l'histoire démographique du couple d'espèces proches *A. halleri* et *A. lyrata* en comparant huit scénarios démographiques différents par leur patrons temporels de flux géniques et par la contrainte sur l'effectif efficace des populations. Il apparait que ces deux lignées évolutives se sont séparées sans avoir échangé de flux de gènes depuis environ 330,000 ans et sans avoir expérimenté de fluctuation importante dans les effectifs efficaces. Dans un second temps, j'étudie la triplication qui a spécifiquement affecté le gène *HMA4* chez *A. halleri*. Ce gène identifié par une approche QTL comme étant un gène majeur impliqué dans l'hyperaccumulation et l'hypertolérance des métaux code pour un transporteur de zinc exprimé dans les racines. Des études de physiologie moléculaire ont montré que son action contribue fortement au flux de zinc depuis les racines vers les parties aériennes. Il a également été montré que l'augmentation de ce flux de zinc chez *A. halleri* par rapport à ses plus proches congénères est une conséquence de l'amplification du nombre de transcrits *HMA4*, amplification conditionnée à la fois par des mutations dans la région promotrice de *HMA4* et par la présence spécifique dans le génome d'*A. halleri* de 3 copies *HMA4* en tandem. En constatant la relation forte entre l'hyperaccumulation, le nombre de transcrits du gène *HMA4* et l'effet additif de la triplication de *HMA4* sur le nombre de transcrits, j'ai daté les évènements de duplication en tandem de *HMA4* pour situer ces étapes importantes de l'évolution de l'hyperaccumulation des métaux dans l'histoire d'*A. halleri*. Nos résultats suggèrent que la première duplication eue lieu il y a environ 390,000 ans, supportant l'hypothèse que l'évolution de l'hyperaccumulation des métaux serait intimement liée à la spéciation d'*A. halleri*.

2.2 Introduction.

Since Darwin [Darwin, 1859] introduced the idea that natural selection may be the driving force for the origin of species, the role of adaptive processes at play during speciation have remained controversial. A first approach has tried to catch speciation "in flagrante delicto" by focusing on partially reproductively isolated ecotypes or races, asking how ecology and genetics interact to cause the evolution of reproductive barriers [Schluter, 2001, Via, 2009]. While this approach is well suited to investigate the modes of speciation, and in particular to reveal the ecological speciation process, its validity has been questioned because there is no guarantee that the studied ecotypes or races will ever attain species status. Hence, a second "retrospective" approach has considered well-established species among which reproductive isolation has been completed. Such studies are well suited to determine the genetics of reproductive barriers and hybrid sterility, but the modes of speciation, and in particular the roles of divergent selection are notoriously difficult to infer a posteriori [Via, 2009].

Recent developments in population genomic tools have brought new prospects for the retrospective approach [Hey and Nielsen, 2004, Becquet and Przeworski, 2007, Hey, 2006]. Indeed, these developments now make it possible to study the divergence process a posteriori by estimating parameters under simple demographic models of speciation. More recently, the Approximate Bayesian Computation approach (ABC) has provided a framework for testing among alternative demographic models [Fagundes et al., 2007, Blum and François, 2010], and also allowed more flexibility in the type of models that can be compared. In particular, models with gene flow in the first stage of speciation (sympatric or parapatric models) can now be explicitly tested against allopatric speciation models with or without secondary contact. In parallel, recent advances in population and ecological genomics have also lead to the identification of ma-

jor genes driving ecological specialization within or between species (e.g. [Linnen et al., 2009]. In particular, the availability of genomic tools in model species along with population genomic and candidate gene approaches [Tenaillon and Tiffin, 2008] have resulted in the identification of the genetic basis and molecular processes responsible for key ecological shifts or adaptations [Bradshaw and Schemske, 2003]. This identification work aims to ultimately help to understand the chronology of genetic mechanisms underlying responses of organisms to their natural environment. Strikingly, these two lines of advances have rarely been integrated, and the demographical and historical context of most documented ecological adaptations remains poorly documented. In particular, it remains largely unknown whether key divergent ecological adaptations were indeed associated with speciation events or have evolved secondarily and independently within sister species after the split.

Here, we gain insight into the ecological speciation process using a retrospective approach by combining demographic inference on the timing of speciation with studies on the molecular targets of adaptation. Our approach is to take advantage of candidate genes of ecological relevance that have been functionally validated to test whether major adaptive molecular changes at these genes were contemporary with important stages of the speciation process. We focus on the pair of plant species *Arabidopsis halleri* and *A. lyrata* (Brassicaceae), two close relatives of the model species *A. thaliana* from which they diverged about 5 million years ago [Koch and Matschinger, 2007]. *A. halleri* is mainly distributed in Continental Europe, although a subspecies (*A. halleri ssp. gemmifera*) with a disjunct distribution is occurring in Eastern Eurasia [Al-Shehbaz and O'Kane, 2002]. In comparison, *A. lyrata* has a circumboreal distribution but also occurs in Western and Central Europe [Al-Shehbaz and O'Kane, 2002]. The two species differ in an important ecological trait. *A. halleri* is a pseudometallophyte species occurring on metalliferous and non-metalliferous sites but exhibiting constitutive tolerance to high soil concentrations in zinc and cadmium and hyperaccumulation of these metals in aerial parts [Pauwels et al., 2006, Kashem et al., 2010, Zhao et al., 2000]. *A. lyrata*, together with the outgroup *A. thaliana*, are both non-accumulator and sensitive to zinc and cadmium, which strongly suggests that heavy-metal tolerance and hyperaccumulation in *A. halleri* are derived ecological traits. Recently, several quantitative trait loci (QTL) regions involved in tolerance to zinc and cadmium in *A. halleri* have been revealed in a backcross family with *A. lyrata* [Willems et al., 2007]. A candidate gene in one of these QTLs was functionally characterized : *AhHMA4* (Heavy Metal ATPase 4) is a metal pump controlling root to shoot Zn transport by loading Zn into xylem vessels [Hanikenne et al., 2008]. [Hanikenne et al., 2008] demonstrated that this gene has a strikingly high transcript level in *A. halleri*, as a result of *cis*-regulatory changes and a tandem triplication. RNA silencing of *AhHMA4* in *A. halleri* strongly supported that this gene played a major role in the acquisition of Zn and Cd tolerance and hyperaccumulation in *A. halleri*.

In this paper, we first characterize patterns of genetic variation across the genomic backgrounds of *A. halleri* and *A. lyrata* to investigate their evolutionary history using an ABC framework and evaluate alternative demographic models of speciation. We then estimate the timing of the duplication of *AhHMA4*, specifically conferring heavy metal tolerance and hyperaccumulation to the *A. halleri* lineage. Our analysis suggests that the historical split between *A. halleri* and *A. lyrata* closely coincides with the evolution of heavy metal tolerance and hyperaccumulation in the *A. halleri* lineage, in accordance with the ecological speciation scenario. It also clearly indicates that heavy-metal tolerance evolved in *A. halleri* well before the spread of zinc and cadmium polluted areas by industrial activities.

2.3 Methods

2.3.1 Plant material.

For *A. halleri*, we sampled 31 diploid individuals from 6 populations scattered throughout the European distribution of the species (F1, France ($N = 6$) ; I5, Italy ($N = 5$) ; D13, Germany ($N = 5$) ; SLO5, Slovenia ($N = 5$) ; PL1, Poland ($N = 5$) ; and CZ8 ($N = 5$). Precise locations of these sites are given in Pauwels et al. (submitted). Leaves were collected in the field, dried and used for DNA extraction as described in [Pauwels et al., 2006]. For *A. lyrata*, we used published sequences from four populations [Ross-Ibarra et al., 2008] : the Plech reference population in Germany ($N = 12$) that has been identified as part of the center of diversity of the species [Clauss and Mitchell-Olds, 2006, Ross-Ibarra et al., 2008], Sweden ($N = 9$), Iceland ($N = 12$) and Russia ($N = 15$). We report analyses performed either with a pooled sample of all *A. lyrata* populations ("Pool" analyses), or with a sample comprising only the reference *A. lyrata* German population ("Plech" analyses).

2.3.2 DNA sequencing.

Single large exons at 29 independent loci in *A. halleri* (Table 2.1) were amplified [30x(30" at 95C, 45" at 55C, 60" at 70C)] and sequenced using PCR primers defined for studies in *A. lyrata* [Wright et al., 2006]. Contaminating salts, unincorporated dNTPs and primers were removed using Millipore-Multiscreen purification kits (Company, City). PCR fragments were sequenced using BigDye Terminator Kit 3.1 (Applied Biosystems, City) and run on an ABI-3130 capillary sequencer (Applied Biosystems). All sequences were checked manually using SeqScape V2.5. All polymorphic sites were confirmed by sequencing on both strands. In total we sequenced 26.54 Kb per diploid individual in *A. halleri*. Sequence data have been submitted to GenBank (accessions XXXXXX- XXXXXX).

Gene	Number of Sites (bp)			GO Terms
	Total	**Synonymous**	**Non Synonymous**	
At1g01040	437	96	341	DCL1
At1g03560	484	110	374	unknown protein
At1g04650	459	108	351	unknown protein
At1g06520	447	101	346	acyltransferase activity expressed in flower buds and siliques
At1g06530	447	86	361	unknown protein
At1g10900	481	110	371	1phosphatidylinositol 4phosphate 5kinase
At1g10980	495	113	382	unknown protein
At1g11050	468	115	353	kinase activity
At1g15240	429	91	338	unknown protein
At1g59720	483	109	374	unknown protein
At1g62310	456	94	362	transcription factor jumonji (jmjC)
At1g62390	483	99	384	unknown protein
At1g62520	424	99	325	unknown protein
At1g64170	423	110	313	unknown protein
At1g72390	371	78	293	unknown protein
At1g74600	513	120	393	unknown protein
At2g16870	543	121	422	disease resistance protein (TIR NBS LRR class)
At2g23170	448	108	340	IAA amido synthase
At2g26140	435	100	335	ftsH protease that is localized to the mitochondrion
At2g26730	349	86	263	leucine-rich repeat transmembrane protein kinase
At2g43680	506	120	386	calmodulin-binding family protein, similar to SF16
At2g44900	447	108	339	ubiquitin-protein ligase activity
At2g46550	430	99	331	unknown protein
At3g20820	489	124	365	unknown protein
At3g23590	522	132	390	unknown protein
At3g48690	444	91	353	similar to PrMC3
At3g50740	444	107	337	DP glucoronosyl UDP glucosyl transferase family
At3g55060	465	96	369	unknown protein
At3g62890	454	104	350	unknown protein
Total	**13,276**	**3,035**	**10,241**	

TAB. 2.1 – Description of the loci surveyed using identification, chromosomal location, and annotation based on *A. thaliana* genome.

2.3.3 Data Analysis.

We used the PHASE algorithm implemented in DNAsp v.4.50.3 [Librado and Rozas, 2009] to reconstruct haplotypes in the *A. halleri* and *A. lyrata* data sets. The algorithm was run with 100,000 iterations, a thinning interval value equal to 1 and a burn-in period of 10,000. For loci with more than 5 polymorphic sites in *A. halleri*, we estimated the intragenic population recombination rate $\rho = 4.N_e.r$ using the LDHAT program [McVean et al., 2002]. The ABC analysis assumes that all sequences have evolved neutrally. To test this, Tajimas's D [Tajima, 1989], Fu's FS [Fu, 1996] and Ramos-Onsins and Rozas $R2$ [Ramos-Onsins and Rozas, 2002] statistics were computed on synonymous positions to test the neutral evolution hypothesis in both populations. Observed values were compared to simulated distributions obtained with 3000 coalescent simulations under the neutral null hypothesis using MANVa (http ://www.ub.edu/softevol/manva). A maximum-likelihood multilocus test of the standard neutral model based on the Hudson-Kreitman-Aguade test was applied to polymorphism data for 29 genes from *A. halleri* and *A. lyrata* by using divergence data from *A. thaliana* [Wright and Charlesworth, 2004]. We used a routine written in C (AnalMS, available upon request to xavier.vekemans@univ-lille1.fr) to compute classical diversity estimators at biallelic synonymous sites (nucleotide diversity π_{syn} ; Watterson's θ_{syn}, population differentiation statistic F_{ST}, computed as $1 - \frac{\pi_s}{\pi_{tot}}$ where π_s is the average pairwise nucleotide diversity within population and π_{tot} is the total pairwise nucleotide diversity of the pooled sample across populations). By using sequences from the *A. thaliana* reference genome as an outgroup, we also partitioned the biallelic synonymous polymorphic sites into 7 different classes defined by Ramos-Onsins [Ramos-Onsins et al., 2004] with the routine AnalMS : (i) exclusive polymorphisms noted $S_{x.hal}$ (or $S_{x.lyr}$), i.e. polymorphic sites for which the frequency of the derived allele $f(d)$ is equal to 0 in *A. lyrata* (or in A. halleri) but $0 < f(d) < 1$ in *A. halleri* (or A. lyrata) ; (ii) fixed differences between species, noted $S_{f.hal}$ (or $S_{x.lyr}$), where $f(d) = 1$ in *A. halleri* and $f(d) = 0$ in *A. lyrata* (or the reverse) ; (iii) shared polymorphic sites (noted S_s), i.e. sites where $0 < f(d) < 1$ in both species ; and (iv) exclusive polymorphisms that are fixed for the derived allele in the other species, noted $S_{x.hal\ f.lyr} = 25$ (or $S_{x.lyr\ f.hal}$), i.e. $f(d) = 1$ in *A. lyrata* (or in *A. halleri*) but $0 < f(d) < 1$ in *A. halleri* (or in *A. lyrata*).

2.3.4 Approximate Bayesian Computation (ABC) analysis.

We performed the ABC analysis based on a set of 34 summary statistics capturing various aspects of polymorphism and divergence between *A. lyrata* and *A. halleri* (Table 2.2, and used a neutral network approach to identify the most relevant combination of statistics for the rejection step. In order to check the robustness of the analyses to the sampling strategy in A.lyrata and to the choice of loci, we performed the ABC analyses on four distinct datasets combining "Plech" and "Pool" population samples with two sets of loci (see text in annexe, "28" referring to the full set of loci excluding a single non-polymorphic locus ; "19" refers to a subsample of loci excluding additionally nine loci showing departure from neutral expectations at three tests of neutrality).

2.3.5 Coalescent simulations

We generated distributions of summary statistics under different demographic scenarios by extensive simulations of the neutral coalescent with recombination using a modified version of Hudson's ms [Hudson, 2002] implemented in the software msnsam [Ross-Ibarra et al., 2008]. For each locus i, coalescent simulations were performed based on corresponding sample sizes for *A. halleri* and *A. lyrata*, and based on observed sequence length L_i, where L_i is the corresponding

number of synonymous sites. The mutation parameters for both populations ($\theta_i(hal)$ and $\theta_i(lyr)$) and for the ancestral population ($\theta_i(A)$) were specified as a function of a reference population of effective size arbitrarily set at 100,000 individuals. Then, $\theta_i(ref)$ for locus i is computed as $\theta_i(ref) = 4 * 100,000 * \mu_i * L_i$, where μ_i is the mutation rate of locus i. Locus specific values of μ_i were obtained by dividing the average net nucleotide divergence at synonymous sites between [*A. halleri* − *A. lyrata*] and *A. thaliana* at locus i, by the divergence time (5 million years ; Koch & Matschinger, 2007 (12)), assuming a generation time of two years. For recombination parameters, we assumed $\rho_i(ref)$=$\theta_i(ref)$, as this corresponds to observations in *A. lyrata* [Hansson et al., 2006, Kawabe et al., 2006], as well as our own observations in *A. halleri*. Some scenarios assumed the occurrence of gene flow affecting equally all loci from population *A. lyrata* to *A. halleri* at rate M_1, and from population *A. halleri* to *A. lyrata* at rate M_2. For each of the 5,000,000 replicated multilocus datasets we computed an array of 34 summary statistics (S_{sim}) (Table 2.2).

		Model choice	Estimation of parameters	Goodness-of-fit
Biallelic Positions	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{f.hal}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{f.lyr}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{x.hal}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{x.lyr}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{x.half.lyr}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
$S_{x.lyrf.hal}$	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
S_s	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used
π_{hal}	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
π_{lyr}	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
θ_{hal}	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
θ_{lyr}	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
Tajima's D_{hal}	Average	Used	Not Used	Used
	Standard Deviation	Used	Not Used	Used
Tajima's D_{lyr}	Average	Used	Not Used	Used
	Standard Deviation	Used	Not Used	Used
Gross Divergence	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
Net Divergence	Average	Not Used	Not Used	Used
	Standard Deviation	Not Used	Not Used	Used
F_{ST}	Average	Used	Used	Used
	Standard Deviation	Used	Used	Used

TAB. 2.2 – Statistics used in the three following steps of the ABC procedure : model selection, parameter estimates and goodness of fit.

2.3.6 Demographic scenarios

We simulated four main demographic scenarios as described in [Ross-Ibarra et al., 2009] (Figure 2.1) classified according the chronological pattern of gene exchanges. For each scenario, two alternative models were simulated assuming either constant population size or exponential population growth. The 8 different models were simulated according to sample size and number of loci of each of the four datasets (Plech19, Plech28, Pool19, and Pool28), making a total of 160 millions of multilocus dataset used in our demographic inference. We used large uniform prior distributions for all parameters, and used identical prior distributions for parameters common to all models. Prior distributions for the ratios $N_e(hal)/N_e(ref)$ and $N_e(lyr)/N_e(ref)$ were uniform on the interval 0-3, prior distribution for $N_e(A)/N_e(ref)$ was uniform on the interval 0-10. Prior distributions for migration rates in both directions were uniform on the interval 0-20. We sampled $T_{split}/(4.N_e(ref))$ from the interval 0-4, translating T_{split} to 0-3,200,000 years. The parameters T_{iso} and T_{SC} were drawn from a uniform distribution on the interval 0-T_{split}. For each of the 8 models we performed 5,000,000 multilocus simulations.

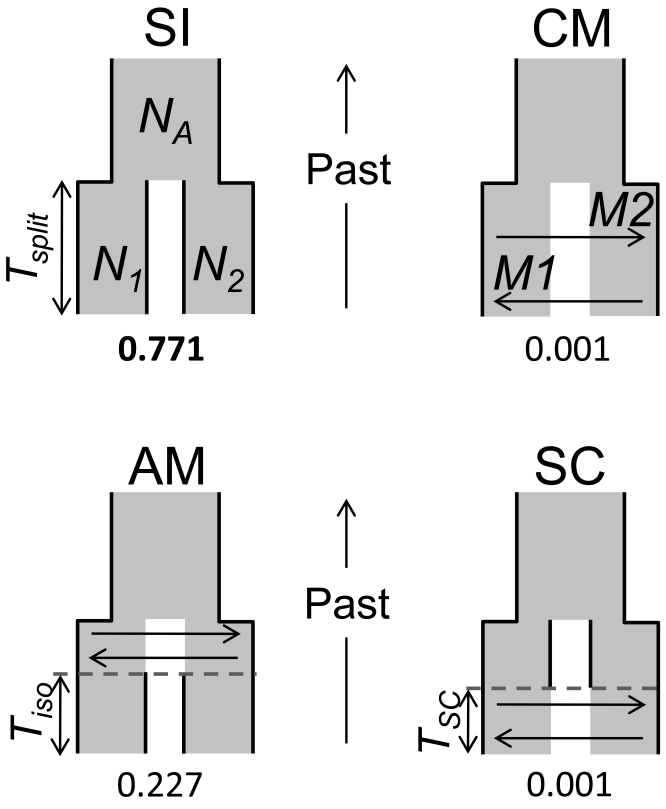

FIG. 2.1 – Alternative scenarios of speciation in *Arabidopsis* genus. Strict Isolation (SI), Constant Migration (CM), Ancient Migration (AM) and Secondary Contact (SC). The posterior probabilities of the different scenarios is given under each scenario. N is the effective population size in number of individuals. The migration rate M is in $4Nm$ units, where m is the proportion of a population made up each generation by migrants from the other population.

2.3.7 Procedure for model testing

For model testing we followed a two step hierarchical procedure [Fagundes et al., 2007]. First, we evaluated posterior probabilities of the two alternative models for each of the four main scenarios separately. Second, we then compared the main scenarios using the best model for each. Posterior probabilities for each candidate model were estimated using a feed forward neural-network implementing non-linear multivariate regression, by considering the model itself as an additional parameter to be inferred under the ABC framework using the R functions "model selection abc nnet" [Blum and François, 2010, R Development Core Team, 2008, Venables and Ripley, 2002]. The 0.1% replicate simulations nearest to the observed values for 22 summary-statistics (S_{obs}) (Table 2.2) were retained, and these were weighted by an Epanechnikov kernel that has a maximum when $S_{obs} = S_{sim}$. Computations were made by using 50 trained neural networks and 10 hidden networks in the regression.

2.3.8 Procedure for parameters estimation

We estimated the posterior distribution of the parameters for the best model of each of the four main scenarios using a regression procedure. Before performing the regression analysis, the parameters were transformed according to a log-tangent transformation [Hamilton et al., 2005]. We considered only the 2,000 replicate simulations with the smallest associated Euclidean distance $\delta = ||S_{obs} - S_{sim}||$. The joint posterior distribution of parameters describing the best model was obtained by means of weighted non-linear multivariate regressions of the parameters on 18 summary-statistics (mean and standard variation of the number of biallelic sites, $S_{f.hal}$, $S_{f.lyr}$, $S_{x.hal}$, $S_{x.lyr}$, $S_{x.hal\ f.lyr}$, $S_{x.lyr\ f.hal}$, S_s, and F_{ST}). One hundred of feed forward neural-networks were trained for each regression, using the R function "abc nnet multivar" [Blum and François, 2010] and results were averaged over the replicate networks in order to reduce the variance of the estimated posterior distribution. We used 50 trained neural networks and 15 hidden networks to perform this analysis.

2.3.9 Estimation of HMA4 duplication times.

We made a multiple alignment of paralogous and orthologous sequences of *HMA4* (Heavy Metal ATPase 4, *At2g19110* in the TAIR database), a gene encoding a plasma membrane protein of the 1B family of transition metal pumps in the P-type ATPase superfamily [Hanikenne et al., 2008]. The complete coding sequences of the three copies of *HMA4* found in *A. halleri* were obtained from the GenBank BACs sequence BAC7C17 (*AhHMA4-1*) and BAC17L07 (*AhHMA4-2* and *AhHMA4-3* [Hanikenne et al., 2008, Willems et al., 2007]. The orthologous copy of *AlHMA4* found on linkage group 3 in *A. lyrata* was obtained from (*AlHMA4* ; JGI database ; http ://www.jgi.doe.gov/). The single copy found on chromosome 2 in *A. thaliana* was obtained from the TAIR database (http ://www.arabidopsis.org/). Maximum likelihood phylogenetic analysis were conducted in PhyML [Dereeper et al., 2008, Guindon and Gascuel, 2003, Anisimova and Gascuel, 2006] using the best substitution model determined according to the software MODELTEST [Posada and Crandall, 1998]. BEAST (v.1.5.3) [Drummond et al., 2005] was used to date duplication events. The molecular clock model used was the strict molecular clock. The analysis performed on third codon positions was calibrated by using a normal prior on the age of the *A. thaliana*-[*A. halleri*/*A. lyrata*] divergence (median 5 MY, with 95% of the distribution lying between 4.5 and 5.5 MY ; based on [Koch and Matschinger, 2007]. Although the estimate of divergence between *A. thaliana* and the other species has been challenged re-

cently [Beilstein et al., 2010], we used the same calibration in both analyses of speciation time and *HMA4* duplication, so that this should not modify our conclusions. A Yule process assuming a constant speciation rate per lineage was used for the speciation model. Posterior distributions were obtained by Markov chain Monte Carlo (MCMC) sampling, with 24,000 samples drawn from a total of 60,000,000 steps, and a 30,000,000 steps long burn-in. Quality of mixing and convergence to the stationary distribution were assessed from three independent runs by using Tracer v1.5 [Drummond and Rambaut, 2007], Tracer v1.4, http ://beast.bio.ed.ac.uk/Tracer).

2.4 Results.

2.4.1 Patterns of polymorphism and divergence.

To evaluate the demographical and historical context of speciation, we estimated levels of nucleotide diversity in the genomic background of *A. halleri* and *A. lyrata*. In *A. halleri*, we resequenced 29 unlinked nuclear genes (totaling 26kb of coding sequence per individual Table 2.1) on a geographically broad sample of 31 individuals from five European populations. In *A. lyrata*, we used published sequence data [Ross-Ibarra et al., 2008] for the orthologs in 48 individuals from four European populations. Over both species, we observed a total of 850 biallelic polymorphic sites (Tables 2.3-2.4). Levels of synonymous polymorphism estimated at these loci with Tajima's π [Tajima, 1983] ($\pi_{syn} = 0.0206 \pm 0.0207$ vs. 0.0240 ± 0.0208 ; Figure 2.2, Table 2.7) and Watterson's θ_W [Watterson, 1975] ($\theta_{Wsyn} = 0.0174 \pm 0.0140$ vs. 0.0190 ± 0.0137, for *A. halleri* and *A. lyrata* respectively ; Figure 2.2, Table 2.7 in indexes) were very similar in both species, and the differences were not significant (Wilcoxon signed-rank test, W=383, P=0.5650 for π_{syn} ; and W=368, P=0.4187 for θ_{syn}). The site frequency spectrum (as measured by Tajimas's D, [Tajima, 1989]) showed no departure from neutrality in any species, supporting the absence of recent variation in population size (Figure 2.2).

Locus	*Bp.syn*	$S_{x.hal}$ **syn**	$S_{x.lyr}$ **syn**	$S_{f.hal}$ **syn**	$S_{f.lyr}$ **syn**	$S_{x.half.lyr}$ **syn**	$S_{x.lyrf.hal}$ **syn**	S_s **syn**
At1g01040	96	6	8	0	0	0	2	0
At1g03560	110	6	7	2	0	0	8	0
At1g04650	109	5	2	1	0	1	0	0
At1g06520	101	4	4	2	3	0	0	0
At1g06530	86	1	1	0	1	1	0	0
At1g10900	110	4	3	0	0	0	1	2
At1g10980	112	4	9	0	1	0	0	2
At1g11050	116	13	5	0	0	3	1	4
At1g15240	90	0	5	4	5	0	0	0
At1g59720	109	7	9	0	0	2	0	24
At1g62310	94	5	3	0	0	0	1	4
At1g62390	99	11	3	0	0	1	0	3
At1g62520	99	9	7	0	0	3	0	3
At1g64170	110	6	6	0	1	3	1	3
At1g72390	78	0	0	0	0	0	0	0
At1g74600	120	1	14	0	0	0	9	0
At2g16870	121	6	10	0	0	0	2	4
At2g23170	108	5	16	1	1	1	3	0
At2g26140	100	3	2	1	0	0	0	0
At2g26730	86	5	2	0	0	1	1	3
At2g43680	120	7	3	0	0	0	3	3
At2g44900	108	4	1	1	2	1	0	0
At2g46550	99	5	8	2	2	1	0	0
At3g20820	124	8	9	1	1	1	0	1
At3g23590	132	7	3	0	0	1	1	0
At3g48690	92	9	7	0	0	2	1	7
At3g50740	107	4	7	0	0	1	6	0
At3g55060	96	7	9	0	0	0	1	0
At3g62890	103	5	8	3	1	2	6	1
Total	**3035**	**157**	**171**	**18**	**18**	**25**	**47**	**64**

TAB. 2.3 – Partition of synonymous polymorphic sites into different categories of mutations when all *A. lyrata* populations are pooled.

Locus	*Bp.asyn*	$S_{x.hal}$ **asyn**	$S_{x.lyr}$ **asyn**	$S_{f.hal}$ **asyn**	$S_{f.lyr}$ **asyn**	$S_{x.halflyr}$ **asyn**	$S_{x.lyrf.hal}$ **asyn**	S_s **asyn**
At1g01040	339	4	3	0	0	1	0	0
At1g03560	373	4	7	0	0	0	2	1
At1g04650	347	8	6	0	1	2	0	0
At1g06520	340	4	4	0	1	0	0	0
At1g06530	361	5	2	0	1	1	0	0
At1g10900	370	4	7	0	0	0	1	1
At1g10980	383	8	14	2	0	0	1	0
At1g11050	352	2	5	1	0	3	1	0
At1g15240	336	3	5	5	2	0	0	0
At1g59720	374	13	10	0	0	1	0	11
At1g62310	362	6	10	0	0	1	0	1
At1g62390	384	2	4	0	0	1	1	0
At1g62520	324	4	1	0	1	0	0	0
At1g64170	313	4	2	0	0	1	0	0
At1g72390	285	3	1	0	1	0	0	0
At1g74600	393	2	5	0	0	0	1	3
At2g16870	422	10	13	0	1	2	0	3
At2g23170	339	0	3	1	1	0	0	0
At2g26140	335	0	0	0	0	0	0	0
At2g26730	262	0	0	0	0	0	0	0
At2g43680	393	7	3	0	0	0	0	1
At2g44900	339	7	2	1	0	0	0	0
At2g46550	336	8	3	2	0	0	0	0
At3g20820	365	1	5	0	0	0	0	0
At3g23590	390	5	6	0	1	1	0	0
At3g48690	352	6	1	0	3	1	0	2
At3g50740	337	5	5	0	0	0	0	0
At3g55060	369	8	3	0	1	0	0	0
At3g62890	350	2	7	3	0	1	3	0
Total	**10225**	**135**	**137**	**15**	**14**	**16**	**10**	**23**

TAB. 2.4 – Partition of non-synonymous polymorphic sites into different categories of mutations when all *A. lyrata* populations are pooled.

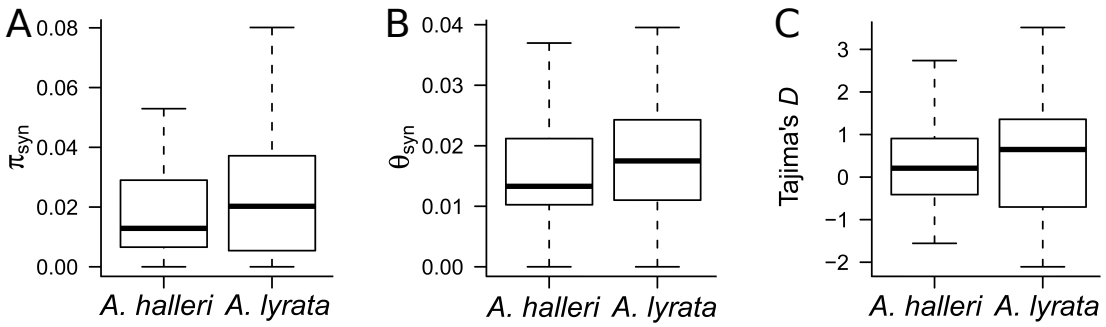

FIG. 2.2 – Comparison of average synonymous nucleotide diversity measured by (A) π_{syn} and (B) θ_{syn}. (C) Empirical distribution of Tajima's D in the species *Arabidopsis halleri* and *A. lyrata*, when all *A. lyrata* populations are pooled.

The joint frequency spectra of derived synonymous sites in *A. halleri* and *A. lyrata* as compared to the outgroup *A. thaliana* (Figure 2.3, Table 2.3-2.4) clearly rejected complete homogeneity between the two species, since most observed polymorphisms were private to a single species (31.4% and 34.2% of all polymorphic sites in *A. halleri* and *A. lyrata* respectively, Figure 2.6 in annexes).

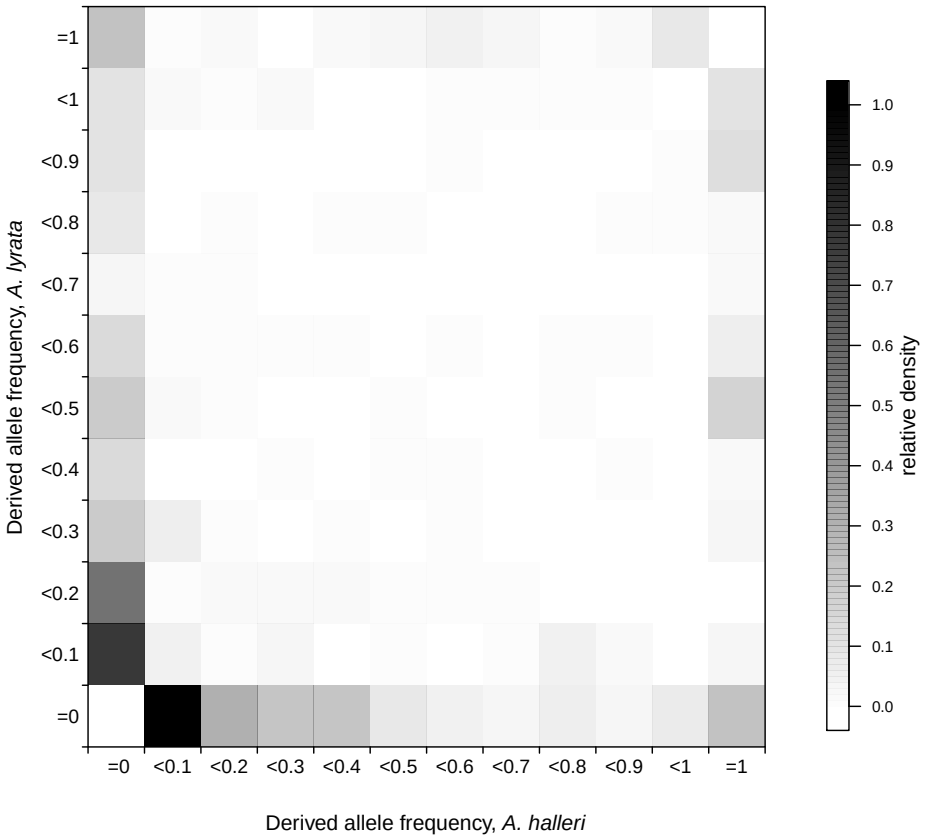

FIG. 2.3 – Distributions of derived synonymous SNP frequencies in *A. halleri* and *A. lyrata* calculated by using *A. thaliana* as an outgroup. Exclusive polymorphic sites (lower row and first column) are defined as positions where the derived allele frequency is between >0 and <1 in one species, but a frequency of zero in the second. Fixed differences (upper left corner and lower right corner) are positions where the derived allele frequency is zero in one species and one in the second. Shared polymorphic sites (all cells except the peripheral ones) are positions where the derived frequencies are >0 and <1 in both species. Putatively ancestral polymorphic sites (upper row and last column) are positions where the derived allele frequency is one in one species and between zero and one in the second species.

The predominance of low frequency derived alleles exclusive to each species suggests recent divergence between *A. halleri* and *A. lyrata*. Yet, in the same time 12.8% of all polymorphic sites were shared between the two species and only 7.2% were fixed for a derived allele in either species, suggesting that the time since divergence has not been sufficient to complete the process of lineage sorting. These shared and fixed mutations were distributed across 13 and 14 different loci respectively, suggesting that the overall pattern is not due to locus-specific effects of genomic divergence. Five loci presented both fixed and shared sites, which may arise only in the presence of intragenic recombination [McVean et al., 2002]. Indeed estimates of population recombination rates were high, of the order of magnitude of the synonymous population mutation rate (Table 2.7 in annexes), as already noted for *A. lyrata* [Hansson et al., 2006, Kawabe et al., 2006]. In addition, 72 sites showed polymorphisms in one species for a derived allele that was fixed in the other species ($S_{xhalflyr}$= 25 and $S_{xlyrfhal}$=47, using the notation of [Ramos-Onsins et al., 2004]. As these sites can be considered as segregating for ancestral polymorphisms [Charlesworth et al., 2005], the overall estimate of the number of sites showing shared polymorphisms is 136 (27.2%), which is much larger than the total number of sites fixed for a derived allele ($S_f = 36$). Such pattern could in principle be due to either a relatively ancient divergence with ongoing gene flow, or to a very recent speciation, and we now aim to disentangle these two scenarios.

2.4.2 Inferring the historical and demographic context of speciation.

We used an Approximate Bayesian Computation (ABC) approach based on 34 summary statistics capturing various aspects of polymorphism and divergence for the two species, to compare four alternative demographic scenarios (Figure 2.1, Table 2.10). For each scenario we simulated five millions replicates. The "strict isolation" scenario assumed a panmictic ancestral population of constant effective population size $N_e(A)$ suddenly split T_{split} generations ago into two panmictic populations of effective population sizes $N_e(hal)$ and $N_e(lyr)$ with a complete arrest of gene flow. The "constant migration" scenario consisted of an ancestral population splitting at time T_{split} in two populations which subsequently diverged but remained connected by constant gene flow at rates $M_1 = M_{lyr-hal}$ and $M_2 = M_{hal-lyr}$. The "ancient migration" scenario was similar to the "constant migration" scenario with the exception that gene flow between both populations stopped T_{iso} generations ago, after which the two populations diverged under strict isolation. The "secondary contact" scenario was similar to the "strict islation" scenario but at time T_{SC} in the past, the populations experienced a secondary contact and started exchanging migrants at constant rates $M_{lyr-hal}$ and $M_{hal-lyr}$. The polymorphism data clearly were not consistent with ongoing migration, since both models allowing for ongoing migration (the "constant migration" and the "secondary contact" models) had very low posterior probability ($P = 0.0009$ and 0.0012, respectively, Table 2.10 in annexes). In contrast, the "strict isolation" and "ancient migration" models had high posterior probability, the former being substantially better supported ($P = 0.7707$ and 0.2273, respectively). To further confirm the reliability of these two models, we performed goodness of fit tests by generating 2,000 simulations from the posterior distribution of parameter values 95% percentiles of all 34 summary statistics. While this confirmed that both models provided excellent fit to the data, the "ancient migration" model was slighly better supported. Indeed, for this model, the entire 34 observed summary statistics were within their 95% percentiles, while one summary statistic was outside the 95 percentile for the "strict isolation" model (Table 2.11). These results are qualitatively robust to changes in the sampling scheme (considering a single *A. lyrata* population, as suggested by [Ross-Ibarra et al., 2008], to the exclusion of nine loci showing departures from neutral expectations and to the possibility of

exponential expansion associated with speciation (see Text in annexe). Hence, our data strongly suggest that these two species are not currently exchanging migrants and that the pattern of polymorphism-sharing may indeed be ascribed to a recent split.

Parameter estimation under the two well-supported models pointed to very similar demographical histories for the divergence process. Under the "strict isolation" scenario, the ancestral population would have split 337,434 years ago (95% HPD interval : [272,799-438,240], Table 2.5, Figure 2.7 in annexes). Under the "ancient migration" scenario, the split would have started more anciently (492,958 years ago, 95% HPD interval : [395,267-931,018]) but migration would have totally ceased at approximately the same time (333,267 years ago, 95% HPD interval : [195,976-449,580]), suggesting that complete reproductive isolation occurred about 160,000 years after the split (Table 2.5, Figure 2.7 in annexes).

Model	$N_{e(hal)}$	$N_{e(lyr)}$	$N_{e(A)}$	T_{split}	$M_{lyr-hal}$	$M_{hal-lyr}$	T_{iso}
1. Strict Isolation	82	79.2	532.9	337.4	-	-	-
	(65.2-98.9)	(65.2-103.9)	(440.2-657.7)	(272.8-438.2)			
2. Ancient Migration	79.8	91.4	543.2	492.9	9.291	8.084	333.3
	(60.5-119.5)	(76.2-126.9)	(484.4-761.1)	(395.27-931.0)	(3.496-14.381)	(2.496-17.722)	(196.0-449.6)

TAB. 2.5 – Modes of parameter estimates under a range of ABC analyses for the Isolation and the Ancient Migration models, with 95% HPD intervals in parentheses.
Effective population sizes $N_{e(hal)}$, $N_{e(lyr)}$ and $N_{e(A)}$ are expressed in x10^3 individuals.
Times T_{split} and T_{iso} are expressed in x10^3 years.

Interestingly, in both cases, ancestral population sizes were predicted to be substantially larger than current sizes ($N_e(A)$=532,985 vs. $N_e(hal)$=81,979 and $N_e(lyr)$=79,245 for the "strict isolation" model and $N_e(A)$=543,223 vs. $N_e(hal)$=79,836 and $N_e(lyr)$=91,428 for the "ancient migration" scenario).

2.4.3 The first HMA4 duplication coincides with speciation time.

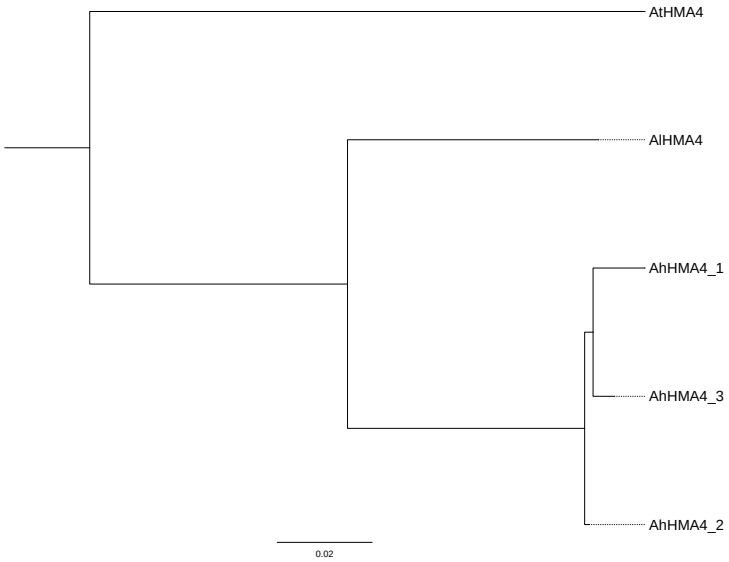

FIG. 2.4 – Phylogram representing preferred tree of HM4 locus in *Arabidopsis* using PhyML.

We compared the inferred speciation times with the timing of the first duplication of *AhHMA4* (Figure 2.4), a mutation conferring a major ecological adaptation in *A. halleri*, i.e. tolerance and hyperaccumulation of Zinc and Cadmium [Hanikenne et al., 2008]. To obtain this time estimate, we compared paralogous nucleotide sequences of *AhHMA4* in *A. halleri* with orthologous sequences in *A. thaliana* and *A. lyrata*.

Nodes	**mode**	**95% HPD**		
Fisrt $AhHMA4$ duplication	387,338	247,929	-	609,928
Second $AhHMA4$ duplication	191,851	104,970	-	372,366
Coalescent time between $AhHM4$ and $AlHMA4$	2,671,806	2,125,305	-	3,320,684

TAB. 2.6 – Estimation of node ages in years, in the *HMA4* genealogy.

We found clear evidence that the two duplications occurred specifically along the *A. halleri* lineage (Table 2.6, Figure 2.5). We estimated the time of the first *AhHMA4* duplication events in *A. halleri* by applying a molecular clock calibrated assuming 5 MY divergence between *A. thaliana* and the pair *A. halleri/A. lyrata*. Our estimates indicated that the first duplication occurred $\approx 387,338$ years ago (95% CI : 247,929-609,928), and thus was contemporary with the speciation between *A. halleri* and *A. lyrata*.

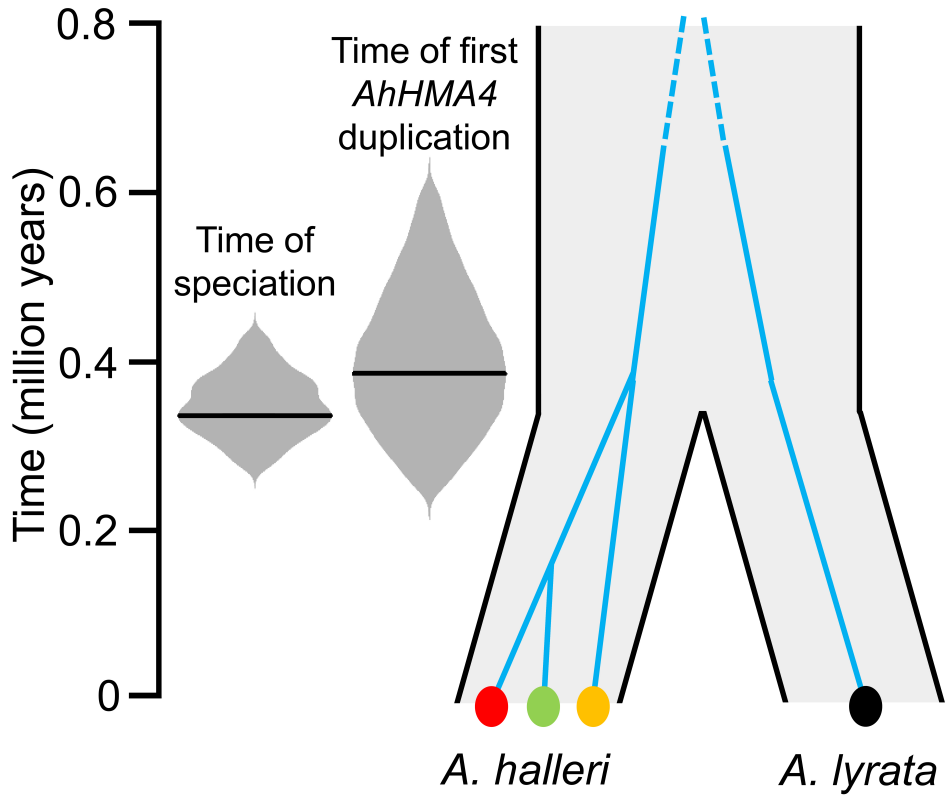

FIG. 2.5 – Coincidence of the speciation time of *Arabidopsis halleri* and *A. lyrata* and the first duplication of the *AhMA4* gene. The genealogy of *HMA4* locus is embedded into a demographic projection of *A. halleri – A. lyrata* speciation. The coalescence between *A. halleri and A. lyrata*'s lineages occurred $\approx$ 2.67 million years ago in the common ancestor (Table 2), and is not represented here. Distributions show the 95% *HPD* of the time of speciation between *A. halleri* and *A. lyrata* under the best model obtained by our ABC approach, and the time of first *AhHMA4* duplication estimated by BEAST software. Thick lines represent the mode of each distribution.

2.5 Discussion.

Research on the genetics of speciation has mainly focused on the description of secondary Dobzhansky-Muller genetic incompatibilities reducing the probability of gene exchange between extent species [Ting et al., 2000, Masly and Presgraves, 2007, Phadnis and Orr, 2009, Mihola et al., 2009]. While equally important, the initial causes of divergence remain much more poorly documented at the genetic level [Schluter and Conte, 2009]. By combining adaptation molecular genetics with population genomic approaches, we find that a major adaptive change specific to *A. halleri* was contemporary with the divergence from the *A. lyrata* lineage. This strongly suggests that ecological differentiation occurred at the onset of speciation in this species pair. This association between speciation and molecular changes at a locus strongly impacting plant fitness in relation to the environment (the self-incompatibility locus enforcing outcrossing in hermaphrodites) was also reported in the *Capsella* genus, where the breakdown of self-incompatibility, the establishment of a "selfing syndrome" and the geographic

expansion in *C. rubella* were inferred to have occurred in a very narrow timescale associated with the divergence between *C. rubella* and *C. grandiflora* [Foxe et al., 2009, Guo et al., 2009]. Interestingly, recent investigations on *A. thaliana* showed that similar features evolved very recently and simultaneously [Charlesworth and Vekemans, 2005, Bechsgaard et al., 2006, Sherman-Broyles et al., 2007, Tsuchimatsu et al.,], but the time of divergence between *A. thaliana* and its closely related species is orders of magnitude higher [Koch et al., 2001, Beilstein et al., 2010]. Hence, these contrasted patterns suggest that the shift in mating system may have been a key element of the speciation process in *C. rubella* but not in *A. thaliana*. These examples highlight the power of combining inference on species trees based on data from the genomic background, with estimates of the time of molecular changes in target genes of adaptation.

The mechanisms by which divergent natural selection on phenotypic traits associated to ecological differentiation may promote the onset of reproductive isolation between populations are still largely unknown [Rundle and Nosil, 2005]. A key issue is to determine whether reproductive isolation occurs mostly by direct or indirect effects of the adaptive molecular changes at target genes. In *A. halleri*, while *AhHMA4* is necessary for the accumulation and tolerance of heavy metals, this particular phenotypic trait is the result of a complex genetic architecture involving other genes of smaller effects [Willems et al., 2007]. Indeed, expression of *AhHMA4* in *A. thaliana* leads to elevated sensitivity to metals as a result of enhanced transfer from roots to shoots [Hanikenne et al., 2008]. This negative effect of *AhHMA4* introgression into the *A. thaliana* genomic background suggests that increased expression of *HMA4* has necessitated the prior establishment of an adapted genetic network involving metal chelators, antioxidants or transporters controlling metal transport into the vacuole [Shahzad et al., 2010]. Understanding the chronology of this genetic architecture by studying other candidate genes will be a very exciting challenge. Also, it suggests that genetic exchanges between tolerant and non tolerant populations may have been hampered as a direct consequence of the adaptation process, as was reported previously for copper tolerance in Monkey flowers [MacNair and Christie, 1983].

The presence of *A. halleri* on numerous anthropogenically contaminated soils was initially proposed to result from multiple and independent recent colonization events and as many parallel adaptive walks towards tolerance. Phylogeographic analyses were consistent with this interpretation, by suggesting that contaminated environments had been colonized independently [Pauwels et al., 2005]. Our ABC approach strongly rejected this hypothesis, as well as the observation of high retention of ancestral polymorphism in both *A. halleri* and *A. lyrata* species.

Since *AhHMA4* duplication was a key step toward metal homeostasis evolution to hypertolerance and hyperaccumulation, its occurrence in the early ages of *A. halleri* is coherent with the species-wide pattern of zinc and cadmium tolerance and hyperaccumulation observed in *A. halleri* from Central Europe, Eastern Europe, Taiwan and Japan as compared to its sister species *A. lyrata* [Pauwels et al., 2005, Kubota and Takenaka, 2003, Kashem et al., 2010]. In addition, the sudden transition to species-wide metal tolerance long before the expansion of anthropogenic environments raises the issue of ecological conditions having selected for this genetic architecture. An emerging hypothesis is the important role of metal hyperaccumulation in plant leaves as a defense mechanism against pathogens or herbivores [Boyd, 2007, Freeman et al., 2006, Rascio and Navari-Izzo,]. Alternatively, the presence of naturally high concentrations of heavy metals in soils have been reported [Alloway, 1995], but their restricted geographic distribution make it difficult to understand how they may have played a major role, considering that the level of polymorphism observed in *A. halleri* allows to reject scenarios with a strong bottleneck at speciation.

In summary, our results support the hypothesis that a major mutation leading to heavy metal tolerance and hyperaccumulation is associated with the speciation of *A. halleri*. In this scenario, there is a single ancient origin from a tolerant and hyper-accumulator *A. halleri* population to all extent natural populations, explaining why all origins studied to date exhibit metal tolerance and hyperaccumulation. Because the origin of *A. halleri* is well anterior to the emergence of modern humans, we believe that metal hyper-accumulation has evolved as a response to selection pressures other than those exerted by elevated metal concentrations in polluted soils, possibly in relation to herbivorism or pathogen defense.

2.6 Annexes.

2.6.1 Neutrality tests3

We report values of statistics and results of tests of neutrality (Tajima's *D*, Fu's *Fs* and Ramos-Onsins and Rozas's *R2*) for each locus, as well as results from a multilocus HKA test performed for each locus in comparison to all other loci (Table 2.9). For *A. halleri* and *A. lyrata*, two and one loci had significantly negative values of *D*, and 4 and 6 loci had significantly positive values of *D*, respectively. We removed three loci (*At1g01040*, *At3g50740* and *At3g62890*) from further analyses because they produced significant neutrality tests for all three statistics *D*, *Fs* and *R2* in at least one species. In addition, HKA tests showed significant departure from neutrality for 7 of the 29 loci, because of an excess of divergence in *A. halleri* (*At1g15240*, *At1g72390*, *At1g74600*) and in *A. lyrata* (*At1g72390*, *At2g26140*), or because of an excess of polymorphism in *A. halleri* (*At1g59720*, *At1g62390*, *At3g48690*) and in *A. lyrata* (*At1g59720*). These loci were also removed from Plech19 and Pool19 analyses. Plech28 and Pool28 analyses were conducted with all sequenced loci except *At1g72390* which presents absolutely no polymorphism in either species.

2.6.2 ABC results are robust to demographic changes associated to speciation.

For each scenario, we tested for demographic changes asociated with speciation by assuming that population sizes either remained constant or growed exponentially after the speciation event. Among the four scenarios SI, IM, AM and SC, the alternative models of constant population sizes are favored with posterior probabilities of 0.585, 0.600, 0.749 and 0.635 respectively for the Plech19 analysis, (posterior probabilities of 0.525, 0.561, 0.755 and 0.503 for the Plech28 analysis) suggesting that both species haven't experimented any particular population expansions since the speciation event.

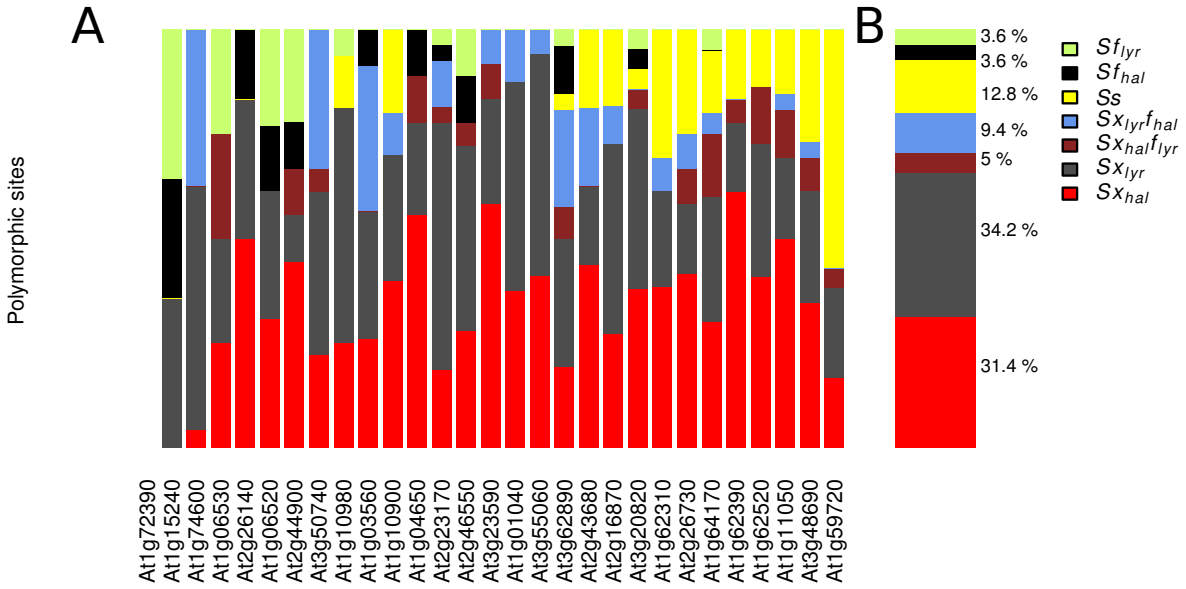

FIG. 2.6 – Nucleotide variation at synonymous sites. (A) Relative proportions within each locus of the 7 classes of polymorphic sites and (B) total proportions of the 7 classes of polymorphic sites.

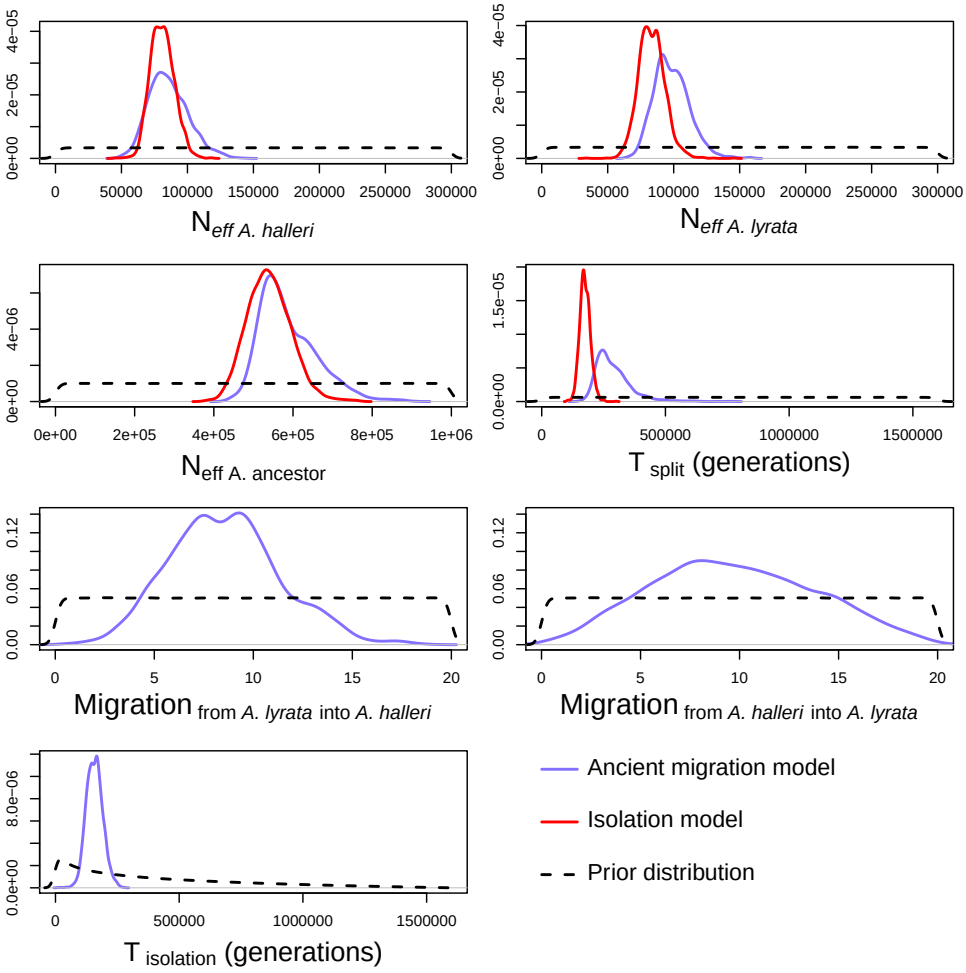

FIG. 2.7 – Posterior distributions for the parameters of the SI and AM speciation models. Dashed curves represent the Bayesian prior for each parameter. Point estimates of the parameters for each population and for the Plech28 analysis are shown in Table 2.5.

Locus	**Species**	**n samp**	π_{syn}	π_{asyn}	π_{asyn}/π_{syn}	θ_{syn}	θ_{asyn}	$\theta_{asyn}/\theta_{syn}$	ρ **(pb)**
At1g01040	*A.halleri*	56	0.020	0.001	0.051	0.013	0.003	0.237	0.022
	A.lyrata (Plech)	18	0.033	0.002	0.047	0.018	0.001	0.047	-
	A.lyrata (Pool)	72	0.027	0.003	0.109	0.021	0.002	0.085	0.000
At1g03560	*A.halleri*	56	0.013	0.002	0.152	0.012	0.003	0.247	0.006
	A.lyrata (Plech)	22	0.045	0.006	0.139	0.037	0.007	0.178	-
	A.lyrata (Pool)	84	0.061	0.008	0.129	0.027	0.005	0.197	0.000
At1g04650	*A.halleri*	58	0.005	0.003	0.544	0.012	0.006	0.516	0.003
	A.lyrata (Plech)	18	0.002	0.002	1.193	0.005	0.003	0.619	-
	A.lyrata (Pool)	84	0.000	0.002	5.296	0.004	0.003	0.942	0.004
At1g06520	*A.halleri*	62	0.007	0.001	0.108	0.008	0.002	0.294	0.016
	A.lyrata (Plech)	12	0.007	0.001	0.198	0.010	0.003	0.294	-
	A.lyrata (Pool)	68	0.005	0.002	0.311	0.008	0.002	0.297	0.031
At1g06530	*A.halleri*	60	0.011	0.002	0.193	0.005	0.004	0.713	0.098
	A.lyrata (Plech)	22	0.002	0.002	0.790	0.003	0.002	0.475	-
	A.lyrata (Pool)	44	0.004	0.001	0.344	0.003	0.001	0.476	0.007
At1g10900	*A.halleri*	56	0.010	0.002	0.223	0.012	0.003	0.246	0.012
	A.lyrata (Plech)	24	0.021	0.004	0.177	0.015	0.005	0.345	-
	A.lyrata (Pool)	80	0.020	0.003	0.154	0.011	0.005	0.446	0.021
At1g10980	*A.halleri*	52	0.007	0.003	0.440	0.012	0.005	0.393	0.083
	A.lyrata (Plech)	16	0.024	0.005	0.215	0.021	0.007	0.332	-
	A.lyrata (Pool)	76	0.013	0.004	0.299	0.020	0.008	0.399	0.002
At1g11050	*A.halleri*	60	0.041	0.003	0.080	0.037	0.003	0.082	0.013
	A.lyrata (Plech)	22	0.002	0.001	0.445	0.005	0.001	0.164	-
	A.lyrata (Pool)	78	0.004	0.002	0.472	0.017	0.003	0.198	0.000
At1g15240	*A.halleri*	62	0	0	NA	0	0.002	NA	NA
	A.lyrata (Plech)	20	0.013	0.001	0.045	0.012	0.002	0.136	-
	A.lyrata (Pool)	40	0.021	0.006	0.301	0.013	0.003	0.268	NA
At1g59720	*A.halleri*	52	0.087	0.015	0.167	0.067	0.015	0.220	0.040
	A.lyrata (Plech)	14	0.087	0.012	0.133	0.081	0.010	0.125	-
	A.lyrata (Pool)	54	0.080	0.012	0.148	0.066	0.012	0.185	0.037
At1g62310	*A.halleri*	52	0.029	0.002	0.073	0.021	0.005	0.233	0.007
	A.lyrata (Plech)	16	0.020	0.005	0.239	0.025	0.007	0.262	-
	A.lyrata (Pool)	76	0.005	0.003	0.509	0.017	0.006	0.357	0.013
At1g62390	*A.halleri*	58	0.064	0.002	0.038	0.032	0.002	0.052	0.003
	A.lyrata (Plech)	18	0.025	0.003	0.116	0.017	0.003	0.174	-
	A.lyrata (Pool)	70	0.020	0.002	0.117	0.013	0.003	0.215	0.000

Suite page suivante ...

Locus	Species	n samp	π_{syn}	π_{asyn}	π_{asyn}/π_{syn}	θ_{syn}	θ_{asyn}	$\theta_{asyn}/\theta_{syn}$	ρ (pb)
At1g62520	*A.halleri*	60	0.052	0.002	0.040	0.037	0.003	0.072	0.017
	A.lyrata (Plech)	14	0.024	0	0.000	0.019	0	0	-
	A.lyrata (Pool)	70	0.026	0.000	0.007	0.021	0.001	0.031	0.000
At1g64170	*A.halleri*	56	0.028	0.002	0.054	0.024	0.003	0.148	0.017
	A.lyrata (Plech)	20	0.012	0.001	0.087	0.018	0.001	0.051	-
	A.lyrata (Pool)	76	0.027	0.002	0.075	0.019	0.001	0.070	0.014
At1g72390	*A.halleri*	52	0.000	0.004	NA	0	0.002	NA	0.000
	A.lyrata (Plech)	20	0	0	NA	0	0	NA	-
	A.lyrata (Pool)	82	0	0.000	NA	0	0.001	NA	NA
At1g74600	*A.halleri*	56	0	0.003	8.800	0.002	0.003	1.523	0.013
	A.lyrata (Plech)	22	0.088	0.011	0.124	0.048	0.006	0.131	-
	A.lyrata (Pool)	72	0.041	0.006	0.142	0.040	0.005	0.119	0.002
At2g16870	*A.halleri*	56	0.021	0.009	0.442	0.018	0.008	0.463	0.006
	A.lyrata (Plech)	18	0.033	0.011	0.317	0.026	0.008	0.315	-
	A.lyrata (Pool)	52	0.037	0.009	0.253	0.029	0.008	0.287	0.031
At2g23170	*A.halleri*	54	0.005	0	0.000	0.012	0	0.000	0.000
	A.lyrata (Plech)	22	0.047	0.001	0.017	0.033	0.002	0.049	-
	A.lyrata (Pool)	84	0.054	0.002	0.029	0.035	0.002	0.050	0.004
At2g26140	*A.halleri*	62	0.006	0	0.000	0.006	0	0	0
	A.lyrata (Plech)	22	0	0	NA	0	0	NA	-
	A.lyrata (Pool)	74	0.003	NA	NA	0.004	NA	NA	0.000
At2g26730	*A.halleri*	62	0.032	0	0	0.022	0	0	0
	A.lyrata (Plech)	18	0.005	0	0	0.013	0	0	-
	A.lyrata (Pool)	32	0.013	NA	NA	0.017	NA	NA	0.006
At2g43680	*A.halleri*	62	0.009	0.003	0.365	0.018	0.004	0.250	0.047
	A.lyrata (Plech)	10	0.020	0.002	0.099	0.020	0.003	0.130	-
	A.lyrata (Pool)	50	0.032	0.004	0.124	0.017	0.002	0.136	0.016
At2g44900	*A.halleri*	60	0.008	0.003	0.370	0.010	0.004	0.450	0.002
	A.lyrata (Plech)	12	0.002	0.002	1.257	0.003	0.002	0.643	-
	A.lyrata (Pool)	72	0.000	0.001	5.685	0.002	0.001	0.637	0.004
At2g46550	*A.halleri*	60	0.014	0.006	0.463	0.013	0.005	0.403	0.015
	A.lyrata (Plech)	20	0.030	0.003	0.107	0.020	0.002	0.086	-
	A.lyrata (Pool)	82	0.020	0.004	0.180	0.016	0.002	0.110	0.014
At3g20820	*A.halleri*	52	0.022	0.000	0.009	0.018	0.001	0.034	0.043
	A.lyrata (Plech)	18	0.030	0.003	0.091	0.021	0.002	0.113	-
	A.lyrata (Pool)	48	0.022	0.003	0.126	0.018	0.003	0.170	0.071

Suite page suivante ...

Locus	**Species**	**n samp**	π_{syn}	π_{asyn}	π_{asyn}/π_{syn}	θ_{syn}	θ_{asyn}	$\theta_{asyn}/\theta_{syn}$	ρ **(pb)**
At3g23590	*A.halleri*	54	0.016	0.003	0.211	0.013	0.003	0.256	0.003
	A.lyrata (Plech)	24	0.004	0.003	0.740	0.004	0.003	0.682	-
	A.lyrata (Pool)	90	0.007	0.006	0.825	0.006	0.003	0.508	0.006
At3g48690	*A.halleri*	56	0.053	0.005	0.103	0.045	0.006	0.123	0.014
	A.lyrata (Plech)	24	0.055	0	0.008	0.035	0.001	0.022	-
	A.lyrata (Pool)	80	0.045	0.001	0.029	0.033	0.002	0.052	0.029
At3g50740	*A.halleri*	54	0.005	0.003	0.552	0.010	0.003	0.320	0.025
	A.lyrata (Plech)	22	0.011	0.003	0.241	0.010	0.002	0.240	-
	A.lyrata (Pool)	84	0.049	0.003	0.056	0.024	0.003	0.122	0.016
At3g55060	*A.halleri*	50	0.009	0.003	0.269	0.016	0.005	0.297	0.011
	A.lyrata (Plech)	22	0.008	0	0.000	0.011	0	0	-
	A.lyrata (Pool)	90	0.011	0.002	0.143	0.021	0.002	0.078	0.002
At3g62890	*A.halleri*	52	0.034	0.003	0.097	0.017	0.002	0.111	0.009
	A.lyrata (Plech)	20	0.010	0.005	0.440	0.011	0.003	0.297	-
	A.lyrata (Pool)	74	0.051	0.011	0.210	0.030	0.006	0.196	0.007
Average	***A.halleri***		**0.021**	**0.003**	**0.513**	**0.018**	**0.004**	**0.285**	**0.019**
	A.lyrata (Plech)		**0.023**	**0.003**	**0.269**	**0.019**	**0.003**	**0.219**	**-**
	A.lyrata (Pool)		**0.024**	**0.004**	**0.618**	**0.019**	**0.004**	**0.255**	**0.012**

TAB. 2.7 – Estimates of nucleotide variation and recombination rates

Locus	Pair	K_{syn}	K_{asyn}	K_{asyn}/K_{syn}	F_{ST}	μ_{syn}
At1g01040	*Arabidopsis halleri versus Arabidopsis lyrata*	0.049	0.002	0.041	0.16	2.25E-08
	A. halleri versus *A. thaliana*	0.133	0.006	0.047		
	A. lyrata versus *A. thaliana*	0.145	0.009	0.059		
At1g03560	*A. halleri* versus *A. lyrata*	0.103	0.011	0.107	0.44	2.85E-08
	A. halleri versus *A. thaliana*	0.192	0.025	0.129		
	A. lyrata versus *A. thaliana*	0.152	0.022	0.146		
At1g04650	*A. halleri* versus *A. lyrata*	0.029	0.012	0.413	0.74	1.83E-08
	A. halleri versus *A. thaliana*	0.111	0.017	0.151		
	A. lyrata versus *A. thaliana*	0.079	0.019	0.241		
At1g06520	*A. halleri* versus *A. lyrata*	0.065	0.004	0.061	0.69	3.12E-08
	A. halleri versus *A. thaliana*	0.151	0.032	0.211		
	A. lyrata versus *A. thaliana*	0.175	0.036	0.205		
At1g06530	*A. halleri* versus *A. lyrata*	0.025	0.008	0.313	0.59	2.25E-08
	A. halleri versus *A. thaliana*	0.109	0.035	0.324		
	A. lyrata versus *A. thaliana*	0.129	0.041	0.318		
At1g10900	*A. halleri* versus *A. lyrata*	0.029	0.005	0.159	0.19	1.77E-08
	A. halleri versus *A. thaliana*	0.099	0.015	0.15		
	A. lyrata versus *A. thaliana*	0.109	0.012	0.113		
At1g10980	*A. halleri* versus *A. lyrata*	0.049	0.016	0.326	0.33	1.58E-08
	A. halleri versus *A. thaliana*	0.077	0.039	0.502		
	A. lyrata versus *A. thaliana*	0.111	0.032	0.286		
At1g11050	*A. halleri* versus *A. lyrata*	0.063	0.017	0.265	0.55	3.74E-08
	A. halleri versus *A. thaliana*	0.22	0.016	0.073		
	A. lyrata versus *A. thaliana*	0.196	0.018	0.093		
At1g15240	*A. halleri* versus *A. lyrata*	0.131	0.024	0.183	0.87	2.82E-08
	A. halleri versus *A. thaliana*	0.119	0.033	0.279		
	A. lyrata versus *A. thaliana*	0.177	0.033	0.187		
At1g59720	*A. halleri* versus *A. lyrata*	0.097	0.016	0.162	0.04	1.34E-08
	A. halleri versus *A. thaliana*	0.157	0.055	0.351		
	A. lyrata versus *A. thaliana*	0.151	0.058	0.387		
At1g62310	*A. halleri* versus *A. lyrata*	0.051	0.006	0.124	0.33	2.22E-08
	A. halleri versus *A. thaliana*	0.144	0.022	0.154		
	A. lyrata versus *A. thaliana*	0.127	0.024	0.187		

Suite page suivante ...

Locus	Pair	K_{syn}	K_{asyn}	K_{asyn}/K_{syn}	F_{ST}	μ_{syn}
At1g62390	*A. halleri* versus *A. lyrata*	0.065	0.006	0.087	0.28	2.11E-08
	A. halleri versus *A. thaliana*	0.156	0.014	0.09		
	A. lyrata versus *A. thaliana*	0.144	0.016	0.112		
At1g62520	*A. halleri* versus *A. lyrata*	0.069	0.004	0.062	0.32	4.67E-08
	A. halleri versus *A. thaliana*	0.259	0.022	0.085		
	A. lyrata versus *A. thaliana*	0.285	0.022	0.078		
At1g64170	*A. halleri* versus *A. lyrata*	0.064	0.006	0.093	0.51	2.45E-08
	A. halleri versus *A. thaliana*	0.116	0.016	0.136		
	A. lyrata versus *A. thaliana*	0.17	0.019	0.11		
At1g72390	*A. halleri* versus *A. lyrata*	0	0.006	NA	NA	2.50E-08
	A. halleri versus *A. thaliana*	0.125	0.028	0.224		
	A. lyrata versus *A. thaliana*	0.125	0.029	0.23		
At1g74600	*A. halleri* versus *A. lyrata*	0.071	0.01	0.141	NA	2.71E-08
	A. halleri versus *A. thaliana*	0.175	0.027	0.157		
	A. lyrata versus *A. thaliana*	0.185	0.028	0.153		
At2g16870	*A. halleri* versus *A. lyrata*	0.073	0.017	0.236	0.34	2.34E-08
	A. halleri versus *A. thaliana*	0.148	0.05	0.335		
	A. lyrata versus *A. thaliana*	0.141	0.045	0.318		
At2g23170	*A. halleri* versus *A. lyrata*	0.091	0.006	0.069	0.42	2.71E-08
	A. halleri versus *A. thaliana*	0.124	0.015	0.124		
	A. lyrata versus *A. thaliana*	0.199	0.014	0.072		
At2g26140	*A. halleri* versus *A. lyrata*	0.015	0	0	0.66	2.49E-08
	A. halleri versus *A. thaliana*	0.136	0	0		
	A. lyrata versus *A. thaliana*	0.119	0	0		
At2g26730	*A. halleri* versus *A. lyrata*	0.037	0	0	0.43	4.52E-08
	A. halleri versus *A. thaliana*	0.246	0.004	0.016		
	A. lyrata versus *A. thaliana*	0.243	0.004	0.016		
At2g43680	*A. halleri* versus *A. lyrata*	0.048	0.006	0.122	0.24	2.48E-08
	A. halleri versus *A. thaliana*	0.154	0.01	0.063		
	A. lyrata versus *A. thaliana*	0.123	0.012	0.095		
At2g44900	*A. halleri* versus *A. lyrata*	0.034	0.007	0.214	0.68	2.05E-08
	A. halleri versus *A. thaliana*	0.104	0.021	0.203		
	A. lyrata versus *A. thaliana*	0.111	0.017	0.152		

Suite page suivante ...

Locus	Pair	K_{syn}	K_{asyn}	K_{asyn}/K_{syn}	F_{ST}	μ_{syn}
At2g46550	*A. halleri* versus *A. lyrata*	0.083	0.013	0.157	0.47	3.16E-08
	A. halleri versus *A. thaliana*	0.161	0.038	0.238		
	A. lyrata versus *A. thaliana*	0.198	0.032	0.16		
At3g20820	*A. halleri* versus *A. lyrata*	0.08	0.003	0.032	0.42	2.31E-08
	A. halleri versus *A. thaliana*	0.137	0.004	0.027		
	A. lyrata versus *A. thaliana*	0.146	0.002	0.015		
At3g23590	*A. halleri* versus *A. lyrata*	0.014	0.009	0.616	0.3	2.20E-08
	A. halleri versus *A. thaliana*	0.12	0.023	0.195		
	A. lyrata versus *A. thaliana*	0.12	0.032	0.266		
At3g48690	*A. halleri* versus *A. lyrata*	0.115	0.015	0.129	0.32	2.18E-08
	A. halleri versus *A. thaliana*	0.153	0.015	0.101		
	A. lyrata versus *A. thaliana*	0.174	0.024	0.138		
At3g50740	*A. halleri* versus *A. lyrata*	0.018	0.005	0.251	0.27	4.60E-08
	A. halleri versus *A. thaliana*	0.228	0.017	0.074		
	A. lyrata versus *A. thaliana*	0.248	0.017	0.069		
At3g55060	*A. halleri* versus *A. lyrata*	0.038	0.006	0.148	0.58	2.27E-08
	A. halleri versus *A. thaliana*	0.139	0.008	0.06		
	A. lyrata versus *A. thaliana*	0.106	0.009	0.086		
At3g62890	*A. halleri* versus *A. lyrata*	0.128	0.029	0.228	0.68	4.32E-08
	A. halleri versus *A. thaliana*	0.27	0.037	0.137		
	A. lyrata versus *A. thaliana*	0.207	0.032	0.153		
Average	***A. halleri* versus *A. lyrata***	**0.06**	**0.009**	**0.154**	**0.438**	**2.68E-08**
	A. halleri* versus *A. thaliana	**0.154**	**0.022**	**0.145**		
	A. lyrata* versus *A. thaliana	**0.158**	**0.023**	**0.143**		

TAB. 2.8 – Levels of synonymous and non-synonymous divergence (K_{syn} and K_{nonsyn}, genetic differentiation (F_{ST}) and synonymous mutation rate per pb (μ_{syn})

Locus	Species	Tajima's D	Fu's Fs	*R2*	K(MLHKA)
At1g01040	*A. halleri*	1.672*	1.963*	0.185*	1.037
	A. lyrata	2.598***	6.399***	0.264***	1.171
At1g03560	*A. halleri*	0.866	1.991*	0.156	0.48
	A. lyrata	0.781	5.595	0.165	2.171
At1g04650	*A. halleri*	-1.407	-1.57	0.048*	0.975
	A. lyrata	-1.508	-1.744*	0.157	0.525
At1g06520	*A. halleri*	-0.411	-0.938	0.086	0.451
	A. lyrata	-0.829	-1.256	0.142	0.366
At1g06530	*A. halleri*	2.017*	0.847	0.233*	0.329
	A. lyrata	-0.641	-0.176	0.086*	0.23
At1g10900	*A. halleri*	-0.377	-2.413	0.092	1.091
	A. lyrata	1.391	0.375	0.202	1.274
At1g10980	*A. halleri*	-0.667	-1.363	0.078	1.576
	A. lyrata	0.446	-2.493	0.159	1.76
At1g11050	*A. halleri*	0.103	0.717	0.109	1.614
	A. lyrata	-1.515	-1.974	0.144	0.207
At1g15240	*A. halleri*	NA	NA	NA	0†
	A. lyrata	0.82	0.225	0.184	0.548
At1g59720	*A. halleri*	0.953	-6.896	0.142	9.261†
	A. lyrata	0.252	2.18**	0.159	13.747†
At1g62310	*A. halleri*	0.831	1.327	0.141	1.708
	A. lyrata	-0.294	-0.781	0.129	2.108
At1g62390	*A. halleri*	2.923***	8.831**	0.213**	2.643†
	A. lyrata	1.494	2.349	0.211	2.117
At1g62520	*A. halleri*	1.483	-1.161	0.158	1.411
	A. lyrata	0.947	0.517	0.199	0.679
At1g64170	*A. halleri*	0.586	-1.879	0.131	2.113
	A. lyrata	-1.168	-0.124	0.11	1.06
At1g72390	*A. halleri*	NA	NA	NA	0†
	A. lyrata	NA	NA	NA	0†
At1g74600	*A. halleri*	-1.091	-1.731	0.132	0.069†
	A. lyrata	3.111***	14.062***	0.252***	2.157
At2g16870	*A. halleri*	0.596	-2.486	0.133	1.006
	A. lyrata	0.827	0.279	0.176	1.738
At2g23170	*A. halleri*	-1.557*	-1.941	0.047*	0.837
	A. lyrata	1.376*	-0.827	0.191	1.39
At2g26140	*A. halleri*	-0.029	-0.143	0.108	0.354
	A. lyrata	NA	NA	NA	0†
At2g26730	*A. halleri*	1.252	4.06*	0.155	0.725
	A. lyrata	-1.853*	1.116	0.229*	0.497
At2g43680	*A. halleri*	-1.212	-7.262	0.058	0.853
	A. lyrata	-0.382	1.176**	0.187	1.435
At2g44900	*A. halleri*	-0.163	-0.586	0.1	0.9
	A. lyrata	-1.141	-0.476	0.276*	0.215
At2g46550	*A. halleri*	-0.034	-0.154	0.105	0.627
	A. lyrata	1.65*	0.866	0.211*	0.88
At3g20820	*A. halleri*	0.702	-4.506	0.137	1.129

Suite page suivante ...

Locus	Species	Tajima's D	Fu's Fs	*R2*	K(MLHKA)
	A. lyrata	1.507*	-2.995	0.206*	1.35
At3g23590	*A. halleri*	0.575	0.216	0.133	1.144
	A. lyrata	0.062	0.102	0.138	0.324
At3g48690	*A. halleri*	0.473	-1.846	0.124	3.974†
	A. lyrata	2.012*	1.384*	0.208*	2.624
At3g50740	*A. halleri*	-1.587*	-3.959*	0.052*	0.32
	A. lyrata	0.261	-1.56	0.148	0.26
At3g55060	*A. halleri*	-1.149	-2.718	0.064	0.978
	A. lyrata	-0.736	-1.542	0.106	0.915
At3g62890	*A. halleri*	2.017*	3.029*	0.191*	0.581
	A. lyrata	-0.317	1.806	0.126	0.53

TAB. 2.9 – Neutrality tests measured by Tajima's *D*, Fu's *Fs*, *R2* and HKA tests at synonymous positions. Loci in bold were excluded from Plech19 and Pool19 ABC analysis.

Analysis	Scenario	Posterior probabilities of different models within each scenario		Posterior probabilities of different scenarios
		Constant population size model	Exponential population growth model	
Plech19	SI	**0.585**	0.415	0.3801
	CM	**0.6**	0.4	0.0546
	AM	**0.749**	0.251	**0.4386**
	SC	**0.632**	0.368	0.1267
Plech28	SI	**0.525**	0.475	**0.7187**
	CM	**0.561**	0.439	0.0007
	AM	**0.755**	0.245	0.2792
	SC	**0.503**	0.497	0.0014
Pool19	SI	**0.827**	0.173	**0.8238**
	CM	**0.684**	0.316	0.0115
	AM	**0.8**	0.2	0.1416
	SC	**0.63**	0.37	0.0231
Pool28	SI	**0.623**	0.377	**0.7707**
	CM	**0.699**	0.301	0.0009
	AM	**0.714**	0.286	0.2273
	SC	**0.652**	0.348	0.0012

TAB. 2.10 – Posterior probabilities of alternative speciation models in four different analyses according sampling scheme and the number of studied loci.

		Plech19		Plech28		Pool19		Pool28	
		SI	AM	SI	AM	SI	SI	SI	AM
Bialsites	avg	0.38	0.422	0.446	0.346	0.207	0.415	0.31	0.322
	std	0.17	0.155	0.073	0.289	0.175	0.306	0.264	0.38
S_{fhal}	avg	0.359	0.463	0.4	0.332	0.488	0.281	0.473	0.36
	std	0.162	0.344	0.244	0.457	0.214	0.482	0.313	0.391
S_{flyr}	avg	0.498	0.317	0.432	0.252	0.471	0.281	0.474	0.257
	std	0.24	0.337	0.27	0.425	0.306	0.464	0.337	0.404
S_{xhal}	avg	0.346	0.226	0.318	0.186	0.191	0.474	0.24	0.311
	std	0.203	0.17	0.31	0.445	0.015	0.104	0.068	0.26
S_{xlyr}	avg	0.439	0.275	0.478	0.45	0.486	0.387	0.311	0.325
	std	0.481	0.399	0.369	0.455	0.397	0.353	0.295	0.365
$S_{xhalflyr}$	avg	0.373	0.374	0.46	0.33	0.279	0.432	0.414	0.334
	std	0.351	0.232	0.237	0.431	0.066	0.279	0.042	0.17
$S_{xlyrfhal}$	avg	0.179	0.16	0.126	0.186	0.142	0.224	0.088	0.199
	std	0.211	0.146	0.1	0.154	0.307	0.26	0.105	0.308
S_s	avg	0.447	0.326	0.396	0.363	0.366	0.448	0.341	0.475
	std	0.267	0.208	0.135	0.141	0.08	0.128	0.114	0.212
π_{hal}	avg	0.261	0.265	0.222	0.415	0.033	0.268	0.202	0.455
	std	0.211	0.202	0.234	0.441	0.015	0.129	0.433	0.464
π_{lyr}	avg	0.452	0.467	0.325	0.392	0.258	0.38	0.478	0.319
	std	0.345	0.459	0.48	0.42	0.245	0.435	0.29	0.287
θ_hal	avg	0.432	0.405	0.338	0.393	0.178	0.43	0.45	0.411
	std	0.198	0.095	0.26	0.386	0.02	0.078	0.371	0.28
θ_lyr	vg	0.426	0.327	0.463	0.4	0.442	0.46	0.456	0.44
	std	0.123	0.311	0.145	0.304	0.1	0.285	0.472	0.499
Tajima's D_{hal}	avg	0.079	0.104	0.098	0.262	0.023	0.094	0.086	0.19
	std	0.145	0.186	0.26	0.435	0.061	0.123	0.179	0.22
Tajima's D_{lyr}	avg	0.161	0.128	0.429	0.431	0.122	0.185	0.271	0.444
	std	0.348	0.358	0.188	0.14	0.168	0.037	0.329	0.137
Gross divergence	avg	0.395	0.433	0.406	0.228	0.191	0.341	0.182	0.374
	std	0.05	0.097	0.193	0.341	0.01	0.086	0.059	0.115
Net divergence	avg	0.489	0.416	0.35	0.247	0.326	0.302	0.214	0.442
	std	0.067	0.119	0.232	0.436	0.011	0.099	0.066	0.203
F_{ST}	avg	0.252	0.187	0.416	0.198	0.054	0.07	0.3	0.159
	std	0.005	0.05	0.123	0.348	0.247	0.332	0.359	0.354

TAB. 2.11 – 1-tailed probabilities of the observed data measured at 34 different summary statistics from simulations under the posterior parameter distributions for 8 ABC analysis.

Chapitre 3

Extend of linkage to a locus under balancing selection

3.1 Résumé.

Les systèmes d'auto-incompatibilité chez les plantes hermaphrodites sont à l'origine de mécanismes physiologiques qui empêchent l'auto-fécondation, à savoir la fécondation d'une oosphère au sein de l'ovule par les noyaux spermatiques produits par du pollen du même individu. Plus généralement la réaction d'auto-incompatibilité est un mécanisme actif faisant intervenir la reconnaissance puis le rejet de tous les grains de pollen ayant le même génotype au niveau du locus d'auto-incompatibilité (locus-S) que celui du pistil. Plusieurs systèmes de reconnaissances ont évolué indépendamment dans différentes lignées d'angiospermes. Chez le couple d'espèces allogames *Arabidopsis halleri* et *A. lyrata*, le locus-S est un complexe de deux gènes physiquement liés : *SCR*, produit par les anthères, et *SRK*, produit par les pistils. Le gène *SCR* code pour une petite protéine présentée à l'extérieur du grain de pollen qui sera reconnue par le domaine extra-cellulaire de la protéine *SRK* s'ils partagent une même spécificité. Dans ce cas, le domaine intra-cellulaire de *SRK* est activé et ceci conduira au déclenchement d'un processus biochimique empêchant l'hydratation du grain de pollen nécessaire pour la germination de son tube pollinique. Sur le plan évolutif, le mécanisme d'auto-incompatibilité génère une sélection balancée de type fréquence-dépendante agissant au locus-S. En effet, les individus dont les allèles au locus-S (allèles-S) sont les plus fréquents auront un succès reproducteur plus faible que les individus possédant un allèle rare. Cette forme de sélection balancée réduit la probabilité de perdre les allèles-S rares par dérive génétique et conduit au maintien d'un grand nombre d'allèles sur de longues durées évolutives. De par sa structure génétique, la recombinaison entre deux allèles-S différents entre les gènes *SCR* et *SRK* peut former des haplotypes auto-compatibles à cause de la non-reconnaissance de l'allèle *SCR* par *SRK* au sein du nouvel haplotype. Le maintien à long terme de lignées auto-incompatibles suggère que la recombinaison est fortement réduite, ou que les recombinants sont fortement sélectionnés. Les deux effets combinés, persistance dans le temps des allèles et suppression locale de la recombinaison intra-génique conduisent à une grande divergence nucléotidique entre paires d'allèles-S. Ils suggèrent également la formation d'un contexte génomique susceptible d'accumuler dans chaque haplotype des mutations délétères récessives. La présence de ce fardeau génétique lié, confirmée chez *A. halleri*, peut également intervenir dans le maintien de ce système de reproduction par la contre sélection des individus recombinants, autogames et homozygotes pour le fardeau lié. Selon des travaux théoriques, la région génomique

entrainée par un locus soumis à sélection balancée sur le long terme serait néanmoins restreinte. Cette prédiction est restée peu mais elle a été empiriquement confirmé par le peu d'études publiées montrant une décroissance rapide du pic de polymorphisme autour d'un locus impliqué dans la résistance aux pathogènes, également soumis à sélection balancée.

Dans ce chapitre nous étudions chez les deux espèces *A. halleri* et *A. lyrata* les patrons de polymorphisme de 12 locus proches du locus-S afin d'avoir une estimation de la taille de la région génomique où le fardeau lié peut s'accumuler. Dans le genre Arabidopsis, les gènes *SCR* et SRK sont situés dans une large région présentant une absence de synthénie entre différents allèles-S (appelée ici, région non-homologue). Ceci suggère à nouveau une suppression presque complète de la recombinaison dans une région d'environ 40 kb. Par contre, à l'extérieur de cette région il semblerait que la reprise de la recombinaison vers des taux génomiques moyens se fasse de façon abrupte. Les taux de recombinaison intragénique mesurés autour de la région non-homologue semblent suffisants pour ne plus observer la signature laissée par la sélection balancée agissant au niveau du locus-S au-delà d'une région flanquante d'environ 15 à 28 kb autour de la région non-homologue, indiquant un faible contenu en gènes dans la région où le fardeau lié a évolué.

3.2 Introduction.

The extent to which natural selection shapes genome-wide patterns of nucleotide diversity remains a controversial issue [Galtier and Duret, 2007]. In a seminal paper, Maynard-Smith and Haigh [Maynard Smith and Haigh, 1974] proposed the "genetic hitchhiking" concept, an important conceptual advance towards understanding how natural selection shapes genome diversity. In this simple model, an advantageous mutation arising in a sexual population increases in frequency, eventually reaching fixation. Because of linkage disequilibrium, i.e. the statistical association between alleles at different loci, neutral alleles linked to this advantageous mutation are also expected to increase in frequency, hence reducing polymorphism in neutral flanking regions even though these regions are not the direct targets of selection. The size of the chromosomal region exhibiting the signature of genetic hitchhiking has been found to depend on the strength of selection on the target site, on the time since the occurrence of the selective sweep, and the recombination rate in this region [Kim and Stephan, 2002]. Accordingly, hitchhiking is believed to be the predominant cause of the lower nucleotide diversity observed in regions of low as compared to high recombination in Drosophila and humans [Begun and Aquadro, 1993, Payseur and Nachman, 2002].

However, natural selection can act differently than sweeping diversity around selected loci. Indeed, balancing selection is maintaining multiple allelic lineages for long evolutionary time [Takahata and Nei, 1990, Vekemans and Slatkin, 1994], thus increasing, rather than decreasing, levels of polymorphism at the selected locus [Maruyama and Nei, 1981], and at closely linked neutral sites [Strobeck, 1983, Hudson and Kaplan, 1988, Meagher and Potts, 1997, H. Schierup et al., 2000]. Long-term balancing selection also causes patterns of trans-specific polymorphism, where several pairs of sequences from two species show stronger identity than some pairs of sequences within the same species [Ioerger et al., 1990, Wu et al., 1998, **?**]. It has been suggested that such trans-specific polymorphisms can extend to neighboring regions due to hitchhiking [Charlesworth et al., 2006]. Yet, the extent of hitchhiking to a locus under balancing selection is still little studied empirically [Charlesworth, 2006]. On the one hand, locally deeper genealogies around the direct target of balancing selection may a priori suggest increased diversity across long chromosomal tracts. Yet, on the other hand, Schierup et al. [Schierup et al., 2001b] showed that longer genealogies also meant longer time for the accumulation of recombination

events, potentially uncoupling the target locus from its local genomic background. How these two processes interact remains poorly known.

Among the promises of the ongoing sequencing revolution stands the exhaustive catalogue of patterns of molecular variation in natural populations, revealing the signature of natural selection at a genome-wide scale. However, the relative importance of the different forms of natural selection remains largely undocumented. A major step to achieve this goal will be to precisely document the footprint of natural selection in contexts where its selective processes and genomic contexts are well documented. Self-incompatibility in the Brassicaceae is one such system. It is a widespread genetic system preventing pollen grains to fertilize ovules from plants with the same mating phenotype. In Brassicaceae, this system is controlled by a single genomic region (the S-locus) composed of two linked genes, one encoding a surface-pollen protein (*SCR*) and a receptor kinase expressed at the surface of papilla cells (*SRK*). Haplotype-specific recognition between the two encoded proteins triggers a biochemical pathway, which ultimately leads to the rejection of pollen. The main evolutionary force driving allele frequency changes at the S-locus is strong negative frequency-dependent selection [Wright, 1939] favoring rare alleles, a form of balancing selection. *Arabidopsis halleri* and *A. lyrata* are closely related outcrossing plant species with functional SI. They both show high allelic diversity at the S-locus [Schierup et al., 2001a, Castric and Vekemans, 2007], and a high proportion of shared allelic lineages [Castric et al., 2008]. Kamau &al [Kamau and Charlesworth, 2005] found that polymorphism at two genes flanking the S-locus (*B80* and *B120*) in A. lyrata were significantly elevated according to the Hudson-Kreitman-Aguade (HKA) test, and interpreted these results as a consequence of an hitchhiking effect. Similar results were found by Ruggiero et al. [Ruggiero et al., 2008] in *A. halleri*, where four genes flanking the S-locus (*B80*, *B120*, *ARK3*, and *B160*) showed significantly higher diversity as compared to control loci. To determine the extent of the genomic region influenced by selection on the S-locus, Kamau &al [Kamau et al., 2007] investigated levels of polymorphism at four additional flanking genes located between 189 kb and 554 kb from *SRK*. No evidence for elevated diversity was observed so great a distance from the S-locus. Although these results suggest that the genomic impact of the S-locus is rather limited, the precise size of the genomic region influenced by selection on the S-locus is still unknown.

Here, we obtained nucleotide sequence data for population samples in *A. halleri* and *A. lyrata* at 8 additional loci flanking the S-locus. We compared patterns of nucleotide polymorphism within and between both species in a total dataset of 12 loci in order to determine the extent of the genomic influence of selection on the S-locus. Because demographic history and locus-specific mutation rate [Wright and Gaut, 2005] may affect the levels and patterns of variation, we contrasted the levels of polymorphism detected in our sample against the null expectation obtained by coalescent simulations under the best fitting demographic model of divergence between *A. halleri* and *A. lyrata* species (Roux &al, in prep). Our study confirms that the signature of hitchhiking, both in terms of excess diversity and shared polymorphism, is restricted to a very narrow genomic region (of the order of 10 kb from each side of the S-locus region). We also found that the elevated number of shared polymorphic sites and high exclusive polymorphism both contribute to the excess of polymorphism observed in regions flanking the S-locus.

3.3 Material and Methods.

3.3.1 Plant material.

We compared species-wide levels of polymorphism and divergence at 12 genes located in a 260kb wide region centered on the S-locus. For *Arabidopsis halleri*, we sampled 31 individuals from 6 populations scattered throughout the European distribution of the species : Auby, France [N=6], St Leonhard in Passeier, Italy [N=5], Harz, Germany [N=5], Stojnci, Slovenia [N=5], Katowice, Poland [N=5] and Zaton, Czech-Republic [N=5]. The same samples were used in a preliminary study on four genes flanking the S-locus in *A. halleri* [Ruggiero et al., 2008] and in a study of the demographical history of *A. halleri* and *A. lyrata* (Roux et al. in prep). For *A. lyrata*, we obtained from O. Savolainen samples from five individuals from each of four populations : Stubbsand (Iceland), Spiterstulen (Norway), Karhumäki (Russia) and Plech (Germany). For some genes, we used previous data on several *A. lyrata* populations from Iceland [Kamau et al., 2007]. Leaves were collected in the field, dried and DNA was extracted as described in Pauwels et al. [Pauwels et al., 2006].

3.3.2 DNA sequencing.

Nucleotide sequences for three genes directly flanking the S-locus [*B80* (*At4G21350*, using *A. thaliana*'s genome annotation), *ARK3* (*At4G21380*), and *B120* (*At4G21390*) ; see Figure 3.1], and a third gene very close to the S-locus [*B160* (*At4g21430*)], were already available from the literature for both *A. halleri* [Ruggiero et al., 2008] and *A. lyrata* [Hagenblad et al., 2006, Kamau and Charlesworth, 2005].

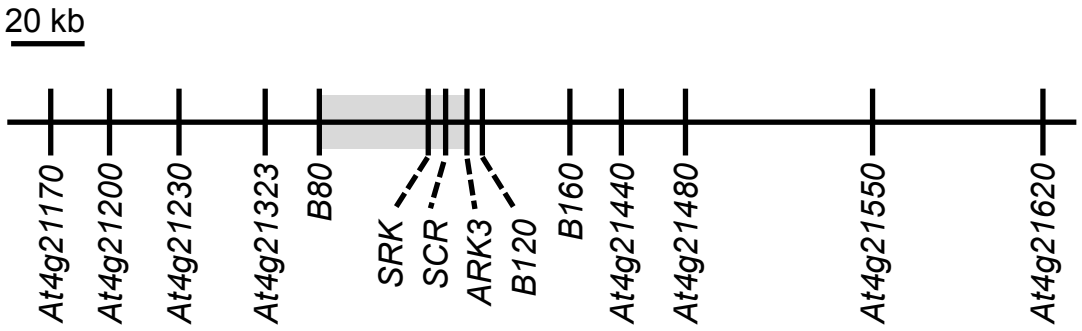

FIG. 3.1 – Diagram of the surveyed region surrounding the S-locus. The shaded box indicates the non-recombining region. Vertical segments indicate the position of studied fragments. The stigma receptor kinase *SRK*, and *SCR*, the pollen ligand recognized by the S-domain of *SRK* compose the S-locus. Gray shading indicates the non-recombining area.

We sequenced additional flanking genes further away from the S-locus (up to 150 kb) so as to provide a regular overview of genetic diversity on a more expanded genomic region. We chose to focus on genes comprising large exons to allow the design of a more efficient direct sequencing strategy, avoiding intronic sequences presenting frequent insertion/deletion variants [Ross-Ibarra et al., 2008]. Physical distances among genes were estimated using the annotation of the *A. lyrata* genome assembly (http ://www.phytozome.net/alyrata.php). Morover, since the S-locus region is determined by a sudden and complete lack of homology between different S-alleles spanning from 300bp upstream the *B80*'s start codon to the immediately 3' non-coding region of *ARK3*'s (V. Castric, personal observation), we assumed that the high level of divergence between alleles in this region is a barrier to recombination. Hence, we used the two flanking genes *B80* and *ARK3* as starting points to calculate physical distances from the S-locus to each of the flanking genes (Figure 3.1 ; Table 3.1). PCR amplifications were carried out in $50\mu L$ and conditions were the following : 30 cycles of 30 seconds at 95C, 45 seconds at 55C and 60 seconds at 70C. Contaminating salts, unincorporated dNTPs and primers were removed from PCR products by using the Millipore-Multiscreen purification kit. PCR fragments were sequenced using the BigDye Terminator Kit 3.1 (Applied Biosystems) and run on an ABI-3130 capillary sequencer (Applied Biosystems). All sequences were checked manually using SeqScape V2.5 and only data that could be confirmed on both strands were included in the analysis. The GenBank accession numbers for the new sequences obtained in this study are as follows : *At4g21170*, XXX-XXX ; *At4g21200*, XXX-XXX ; *At4g21230*, XXX-XXX ; *At4g21323*, XXX-XXX ; *At4g21440*, XXX-XXX ; *At4g21480*, XXX-XXX ; *At4g21550*, XXX-XXX ; *At4g21620*, XXX-XXX.

Loci	Distance from *B80* in kb	Distance from *ARK3* in kb	$n_{A.\ halleri}$	$n_{A.\ lyrata}$	bp_{all}	bp_{syn}	Gene product
At4g21170	73.24	-	60	34	539	136.99	pentatricopeptide (PPR) repeat protein
At4g21200	57.3	-	62	40	354	76.65	ATGA2OX8 (GIBBERELLIN 2-OXIDASE 8)
At4g21230	38.48	-	60	40	473	112.55	protein kinase family protein
At4g21323	15.09	-	54	37	484	120.38	subtilase family protein
B80	-	-	58	45	666	173.68	binding / ubiquitin-protein ligase
ARK3	-	-	43	42	292	67.13	Arabidopsis Receptor Kinase 3
B120	-	4.24	57	44	520	115.46	Protein kinase/ sugar binding
B160	-	28.04	47	12	761	155.52	Transcription factor
At4g21440	-	42.03	46	26	479	115.55	ATM4/ATMYB102; DNA binding / transcription factor
At4g21480	-	59.43	60	38	548	132.28	glucose transporter
At4g21550	-	110.11	56	36	430	97.91	transcriptional factor B3 family protein
At4g21620	-	156.29	48	38	260	68.48	glycine-rich protein

TAB. 3.1 – Sample size per species, length and distance from the non-homologous region of the 12 surveyed loci.

3.3.3 Data Analysis.

Sequences were aligned using the ClustalW program , and slightly modified manually with the MEGA version 4 program [Tamura et al., 2007]. Reading frames were determined by comparison with A. thaliana orthologs. Haplotypic data were inferred using the PHASE algorithm implemented in the DNAsp (v.4.50.3) software [Librado and Rozas, 2009]. The algorithm was run with 100,000 replicates, a thinning interval value equal to 10 and a burn-in period of 10,000. DNAsp (v.4.50.3) was also used to calculate the Watterson's θ_w and Tajima's π estimators of diversity on synonymous positions.

3.3.4 Intralocus recombination analysis.

The population intragenic recombination rate statistics ρ ($\rho = 4Nr$, where N is the effective population size and r the recombination rate) and R_{min} (the minimum number of recombination events, [Hudson and Kaplan, 1985]) were estimated using the PAIRWISE program in the LDhat package. This program implements Hudson's composite-likelihood approach to estimate the population recombination rate conditioned on the mutation rate per site θ_W from an approximate finite-sites version of the Watterson estimate [Hudson, 2001, McVean et al., 2002]. To test the null hypothesis of no recombination ($\rho = 0$), we used the likelihood permutation test. Computer simulations have shown that balancing selection does not affect the accuracy of recombination rate estimates by LDhat [Richman et al., 2003].

To jointly describe the pattern of polymorphism in the two species, we then computed the number of polymorphic sites belonging to each of seven different classes (according to [Ramos-Onsins et al., 2004]). We used sequences from the A. thaliana reference genome (Col-0) as an outgroup to determine ancestral and derived states. These seven classes included : (i) fixed differences between species ($S_{f.hal}$ and $S_{f.lyr}$), *i.e.* polymorphic sites whose derived allele frequency $f(i)$ was equal to 1 in *A. halleri* or *A. lyrata* respectively and 0 in the other species ; (ii) shared polymorphic sites (S_s), where $0 < f(i) < 1$ in both species ; (iii) exclusive polymorphisms, ($S_{x.hal}$ and $S_{x.lyr}$), where $f(i) = 0$ in *A. lyrata* or *A. halleri* respectively but $0 < f(i) < 1$ in the other species ; and (iv) two last categories ($S_{x.hal\ f.lyr}$ and $S_{x.lyr\ f.hal}$) corresponding to polymorphic sites where $f(i) = 1$ in *A. lyrata* or *A. halleri* respectively but $0 < f(i) < 1$ in the other species. A software to perform these computations is available upon request to X. Vekemans).

3.3.5 Coalescent simulations for the neutral model.

The observation of high levels of polymorphism at a locus in the S-locus region relative to the rest of the genomic background may be the effect of hitchhiking to the S-locus subject to balancing selection, according to our working hypothesis, but may also result from a local increase in mutation rate [Hudson et al., 1987], or from direct independent selection on the locus. To disentangle the relative effects of neutral processes, i.e., mutation rate and genetic drift, from the effects of hitchhiking to the S-locus, we generated distributions of summary statistics under the neutral hypothesis by performing coalescent simulations according to a demographic model of divergence between *A. halleri* and *A. lyrata* inferred from a previous study using data from 29 unlinked control loci (roux &al, in prep). Because these simulations were performed using the actual posterior distribution of parameter values obtained in an Approximated Bayesian Computation framework (roux &al, in prep) rather than the modes of the posterior distribution, we obtained broade ranges of the summary statistics investigated, providing conservative tests. The

MSnsam program [Hudson, 2002, Ross-Ibarra et al., 2008] was used to perform 10,000 coalescent simulations under the most plausible divergence scenario between *A. halleri* and *A. lyrata* based on Approximate Bayesian Computation (Roux &al, in prep). This model entails strict isolation since 168,700 generations (95%HPD : 136,400-219,100) and comprises four parameters : the current effective population sizes of *A. halleri* and *A. lyrata* (N_{hal} and N_{lyr}), the effective population size of the common ancestor (N_{anc}) and the time of the split between *A. halleri* and *A. lyrata* (T_{split}) (Figure 3.2). Values of the parameters were drawn from the posterior distribution obtained from the previous analysis. We calibrated the mutation rate for each locus specifically based on divergence from *A. thaliana* ($\mu = K/2T$ where K is the measured synonymous divergence with *A. thaliana*), assuming divergence T=5 million years [Koch and Matschinger, 2007] and population recombination ρ estimated for each locus individually. Null distributions of the seven summary statistics were obtained for each of the 12 studied loci and compared to the observed values to estimate P values.

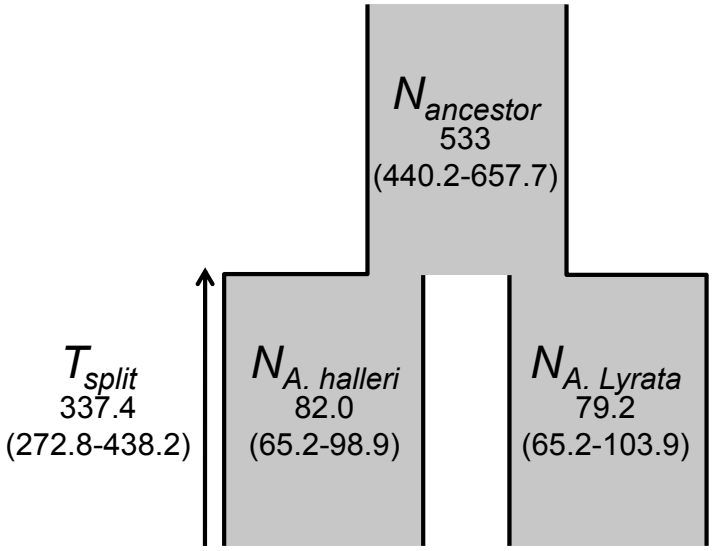

FIG. 3.2 – Schematic representation of the demographic model used to obtain the neutral distribution through coalescent simulations with four parameters : the time of population splitting in x10^3 years (T_{split}), the population sizes $N_{A.halleri}$, $N_{A.lyrata}$, and $N_{ancestor}$ are in x10^3 effective individuals.

3.3.6 Tests of diversifying selection.

To distinguish the effect of hitchhiking to the S-locus from the independent effect of direct positive diversifying selection on neighbouring genes, we tested for the presence of sites showing signs of positive selection by comparing alternative models of sequence evolution implemented in CodeML from the PAML (v.4.4) package [Yang, 2007]. The program compares the likelihood of models accounting for or not accounting for the heterogeneity among sites of $\omega = d_n/d_s$ (where d_n is the number of nonsynonymous substitutions per nonsynonymous site and d_s the number of synonymous substitutions per synonymous site). A phylogeny using neighbor joining [Tamura et al., 2007] was reconstructed using the best substitution model estimated by Modeltest [Posada and Crandall, 1998, Leitner et al., 1997], and its topology was used as the input tree

for the PAML analysis. Six models were implemented in our study : (i) M_0 assumes identical values of ω for all codon sites ; (ii) M_{1a} assumes a proportion p_0 of sites under purifying selection ($0 < \omega < 1$) and a proportion $1 - p_0$ of selectively neutral sites at which $\omega = 1$; (iii) M_{2a} assumes a supplemental class of sites with ω as a free parameter ; (iv) M_3 assumes three classes of sites in proportions p_0, p_1 and p_2 with ω values ω_0, ω_1 and ω_2 estimated from the data ; (v) M_7 is based on the β distributions and allows the ω ratio to take values between zero and one ; (vi) M_8 assumes an extra class of sites to M_7 with the proportion and the ω ratio estimated from the data, allowing this latter ratio to be greater than one. Hence, comparing the likelihood of $M_{1a} vs. M_{2a}$ and of $M_7 vs. M_8$ is a test for the presence of sites showing signs of positive selection.

3.3.7 Sliding window analysis within flanking genes.

Because the peak of increased polymorphism was found to be very narrow (see results), we used a sliding window analysis of θ_W and π to analyse the distribution of polymorphism at a very narrow scale, i.e. within the two genes directly flanking the S-locus, *B80* and *ARK3*. Briefly, we used windows of 30 bp moved by steps of 15 sites. We then calculated Spearman's rank correlation between the observed level of polymorphism in a window and the position of this window in the sequenced fragment. To assess the statistical significance of the correlation, we obtained its null distribution by 10,000 random permutations between positions.

3.4 Results.

3.4.1 Contrasting the observed data with neutral expectation.

sThe effect of linkage to the S-locus was empirically investigated by the comparison of the observed level of synonymous polymorphism at 12 loci at different distances from the S-locus with their neutral expectation in the absence of linkage, obtained by simulations under a relevant demographic scenario (Figure 3.1, Table 3.1).Overall, the fit of the observed polymorphism data against neutral simulations in the absence of linkage to the S-locus was excellent for most loci, suggesting that the signature of hitchhiking was very narrow. However, it is remarkable that the three genes directly flanking the S-locus (*ARK3*, *B80* and *B120*) revealed individually a significant departure from neutral expectations for both estimators of diversity, Watterson's θ_w and nucleotide diversity π, in both *A. halleri* and *A. lyrata* (Figure 3.3, Table 3.2, Table 3.5 in annexes).

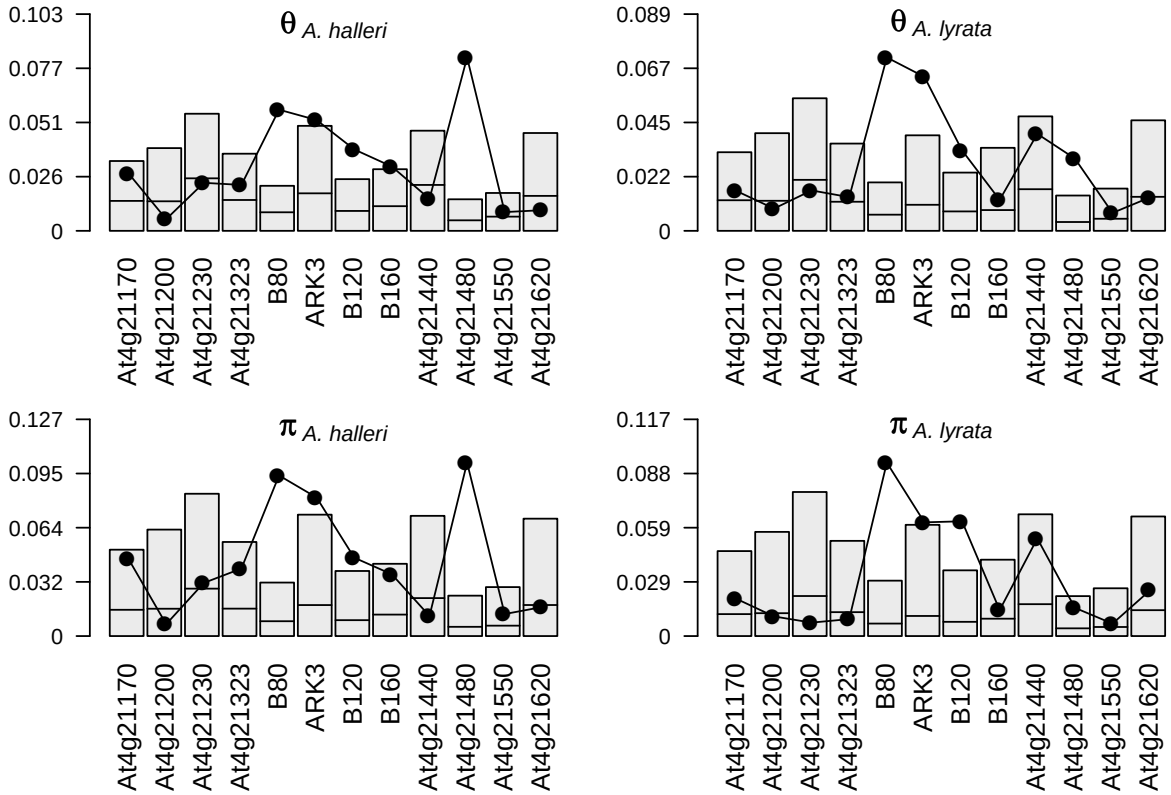

FIG. 3.3 – Distribution of the estimators of diversity θ_W and π per nucleotide, obtained from coalescent simulations at synonymous positions. Shaded boxes indicate the one tail 95% confidence intervals for each locus. Solid dots indicate the observed values. A thick horizontal line indicates the median for each distribution.

locus	π_{hal}	π_{lyr}	θ_{hal}	θ_{lyr}	$S_{f.hal}$	$S_{f.lyr}$	$S_{x.hal}$	$S_{x.lyr}$	$S_{x.hal\ f.lyr}$	$S_{x.lyr\ f.hal}$	S_s
At4g21170	0.0451	0.0201	0.0272	0.0164	1	1	11	5	2	0	4
At4g21200	0.007	0.0106	0.0055	0.0092	1	0	2	3	0	0	0
At4g21230	0.0312	0.0073	0.0228	0.0166	0	8	10	7	2	1	0
At4g21323	0.0396	0.0092	0.0218	0.0139	2	2	10	7	2	0	0
B80	0.0941***	0.0937***	0.0574***	0.0714***	0	0	13*	23***	2	0	31***
ARK3	0.0812*	0.0614*	0.0525*	0.0634**	0	0	9	10*	1	3*	5*
B120	0.0459*	0.0620**	0.0386**	0.0331*	0	0	12*	7	0	1	9**
B160	0.0362	0.014	0.0305*	0.0127	0	0	19**	6	2	0	0
At4g21440	0.0118	0.0527	0.0152	0.0399	1	1	6	12	0	3	1
At4g21480	0.1016***	0.0153	0.0822***	0.0298***	0	0	33***	3	1	0	12***
At4g21550	0.0129	0.0067	0.0089	0.0074	0	0	4	3	0	0	0
At4g21620	0.017	0.0248	0.0098	0.0138	1	0	3	3	0	1	0

TAB. 3.2 – Estimates of nucleotide variation by the Watterson θ and the Tajima's estimator π for synonymous positions.
$*P < 0.05$ $**P < 0.01$ $***P < 0.001$
$S_{f.hal}$ and $S_{f.lyr}$ are the numbers of the differences fixed in the *A. halleri* and *A. lyrata* branches.
$S_{x.hal}$ and $S_{x.lyr}$ are the numbers of segregating sites exclusive to *A. halleri* and *A. lyrata*.
$S_{x.hal\ f.lyr}$ and $S_{x.lyr\ f.hal}$ are mutations that either segregate in one species (x) and are fixed in the other species (f) relative to the outgroup.
S_s is the number of mutations that segregate in both species.

For *A. halleri*, values of θ_w for *ARK3*, *B80* and *B120* were 0.0525 ($P = 0.0364$), 0.0574 ($P < 0.0001$) and 0.0386 ($P = 0.0028$) respectively, and associated values of π were 0.0812 ($P = 0.0302$), 0.0941 ($P < 0.0001$) and 0.0459 ($P = 0.023$). The *A. lyrata*'s values of θ_w for *ARK3*, *B80* and *B120* were 0.0634 ($P = 0.0045$), 0.0714 ($P < 0.0001$) and 0.0331 ($P = 0.0105$) respectively, and values of π for the same genes were 0.0614 ($P = 0.0473$), 0.0937 ($P < 0.0001$) and 0.0620 ($P = 0.0043$). Among the 12 surveyed loci, *At4g21480* showed a significant excess of polymorphism for both diversity estimators in *A. halleri* ($\theta_w = 0.0822$,$P = 0$ and $\pi = 0.1016$, $P = 0$) while only Watterson's estimator θ_w was significant in *A. lyrata* ($\theta_w = 0.0298$, $P = 0.0007$ and $\pi = 0.0153$, $P = 0.1143$). While *B160* and *At4g21440* are closer to the S-locus than *At4g21480*, no particular departure from neutrality was strongly supported in these two loci in either species. Thus, the elevated diversity observed at *At4g21480* might be due to some form of balancing selection acting directly on this locus, rather than to hitchhiking to the S-locus. However, likelihood-ratio tests did not support this hypothesis, since models of nucleotide site evolution allowing for positive selection applied to the dataset, i.e. M_{2a}, M_3 and M_8 did not fit the data significantly better than the simpler models M_{1a}, M_0 and M_7 neither in *A. halleri* nor in *A. lyrata* (In annexes : tables 3.6, 3.7, 3.8 and 3.9). Nevertheless, we note that this locus contained polymorphic stop codons and an insertion/deletion polymorphism of a 13 amino-acids long motif. This premature stop codon was found in *A. halleri* from Slovenia, Poland and Czech-Republic but not in *A. lyrata*. Finally, our results show that the genomic region expected to be linked to the S-locus extends from *At4g21323* to *B160*. We count 9 loci embedded in the linked region according the annotation of *A. lyrata*'s genome (Table 3.3).

Gene	Distance from *B80* (kb)	Distance from *ARK3* (kb)	Gene product
At4g21320	13.78	-	HSA32 (Heat-Stress-Associated 32)
At4g21330	8.53	-	DYT1 (Dysfunctionnal Tapetum 1) ; Transcription factor
At4g21340	4.93	-	B70 ; Transcription factor activity
B80	-	-	Binding / ubiquitin-protein ligase
ARK3	-	-	Arabidopsis Receptor Kinase 3
B120	-	4.24	Protein kinase/ Sugar binding
At4g21400	-	14.03	Protein kinase family protein
At4g21410	-	22.02	Cysteine-Rich Receptor-Like Protein Kinase 29 (Cysteine-rich RLK29)
At4g21420	-	25.6	Transposable element gene

TAB. 3.3 – Loci embedded in the region linked to S-locus, comprised between *At4g21323* and *B160* according the complete sequence of *A. lyrata*'s genome. *SRK* and *SCR* are excluded from this list.

3.4.2 Origin of the elevated variation in the region linked to the S-locus.

The elevated levels of nucleotide polymorphism observed at *ARK3*, *B80* and *B120* were associated with a significantly elevated number of shared polymorphic sites in these three genes as compared to neutral expectations, suggesting the retention of an excess of ancestral polymorphism (Figure 3.4, Table 3.2). Indeed, 5 ($P = 0.0258$), 31 ($P < 0.0001$), and 9 ($P = 0.0046$)

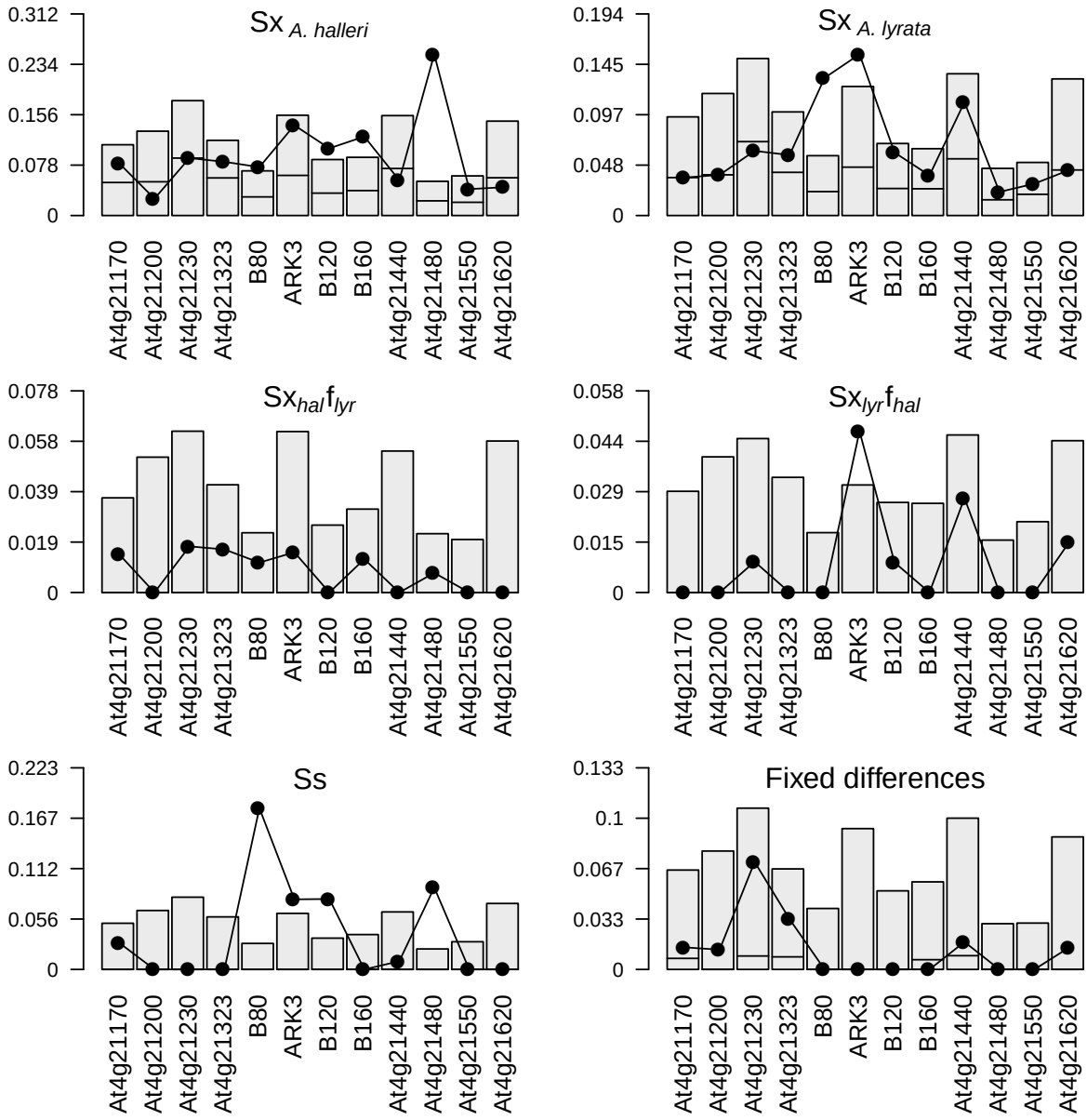

FIG. 3.4 – Distribution of the number of each synonymous polymorphic sites category divided by the number of synonymous positions. Shaded boxes indicate the one tail 95% confidence intervals for each locus. Solid dots indicate the observed values. A thick horizontal line indicates the median for each distribution.

synonymous shared polymorphic sites were observed at the *ARK3*, *B80* and *B120* genes res-

pectively. Another type of putative ancestral polymorphism, sites with a derived allele polymorphic in one species but fixed in the other species, was found in excess at *ARK3* in *A. lyrata* ($S_{x.hal\ f.lyr} = 3$, $P = 0.0256$), but not in *A. halleri* ($S_{x.hal\ f.lyr} = 1$, $P = 0.1824$). Yet, shared polymorphic sites were not the only cause of elevated polymorphism at these three genes. Indeed, an excess of derived exclusive polymorphism was also found at the *B80* gene in both *A. halleri* ($S_{x.hal} = 13$, $P = 0.0255$) and *A. lyrata* ($S_{x.lyr} = 23$, $P < 0.0001$), at *ARK3* in *A. lyrata* ($S_{x.lyr} = 10$, $P = 0.013$) but not in *A. halleri* ($S_{x.hal} = 9$, $P = 0.0598$), and at *B120* in *A. halleri* ($S_{x.hal} = 12$, $P = 0.0113$) but not in *A. lyrata* ($S_{x.lyr} = 12$, $P = 0.077$).

3.4.3 Sliding window analysis of polymorphism.

A sliding window analysis of synonymous diversity along the sequenced fragment of the gene directly adjacent to the S-locus (*B80*) revealed a negative correlation between the distance from the S-locus and θ_w (Spearman's $r = -0.428$; $P = 0.0025$) or π (Spearman's $r = -0.300$; $P = 0.0248$) (Figure 3.5) in *A. halleri*, suggesting a very sharp decline of polymorphism even at very short distances from the S-locus. This correlation was not found in *A. lyrata* for θ_w (Spearman's $r = -0.428$; $P = 0.4415$) or π (Spearman's $r = -0.300$; $P = 0.6122$) , nor at other genes in either species.

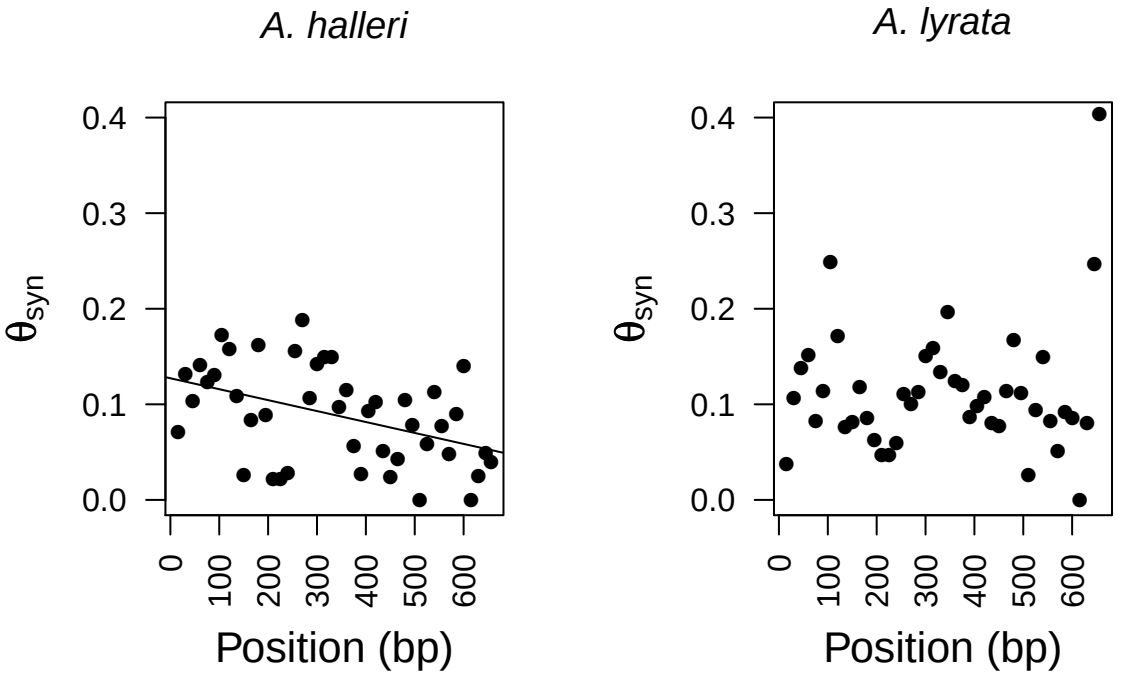

FIG. 3.5 – Sliding window analysis of density in synonymous segregating sites measured by Watterson's θ along the *B80*'s coding sequence in (A) *A. halleri* and (B) *A. lyrata*. The x-axis is base position relative to the closest nucleotide from the S-locus in the sequenced fragment.

3.4.4 Levels of intragenic recombination in the region flanking the S-locus.

Strong hitchhiking signature in the genes flanking the S-locus suggests the absence or very low levels of recombination in this region. However, we found evidence for intralocus recombination in *ARK3*, *B80* and *B120* in both *A. halleri* and *A. lyrata* species (Table 3.4). The population recombination rates per nucleotide (ρ) in *A. halleri* were 0.068, 0.1397 and 0.0346 and the relative amounts of recombination compared to mutation (ρ/θ) in the genealogies were 1.2952, 2.4338 and 0.8964 for *ARK3*, *B80* and *B120* respectively. The ρ/θ ratios measured at 29 anonymous loci in A. halleri range from 0 to 19.6 with a median of 0.529. Recombination was somewhat less important in A. lyrata than in A. halleri, with $\rho = 0.0136$, 0.0837 and 0.0074 and ρ/θ=0.2145, 1.1723 and 0.2236 respectively. The ρ/θ ratios measured at 29 anonymous loci in *A. lyrata* range from 0 to 3.91 with a median of 0.658 (Roux & al, in prep). Our estimates of ρ/θ ratio in *B80* are very close to reported estimates of the recombination rate in *A. lyrata*'s genomic background [Hansson et al., 2006, Kawabe et al., 2006]. Hence, our results suggest a sudden switch in recombination rates outside the non-homologous region rather than a slow progressive increase from *SRK*.

Locus	Species	R_{min}	$\rho = 4Nr$	P	ρ/θ_w
At4g21170	*A. halleri*	5	0.0323	0	1.1875
	A. lyrata	4	0.0148	0.004	0.9024
At4g21200	*A. halleri*	0	0.0791	0.333	14.3818
	A. lyrata	0	0.048	0.7	5.2174
At4g21230	*A. halleri*	3	0.0296	0.013	1.2982
	A. lyrata	2	0.0374	0.276	2.253
At4g21323	*A. halleri*	6	0.0103	0.126	0.4725
	A. lyrata	0	0	0.558	0
B80	*A. halleri*	27	0.1397	0	2.4338
	A. lyrata	18	0.0837	0	1.1723
ARK3	*A. halleri*	4	0.068	0	1.2952
	A. lyrata	4	0.0136	0.02	0.2145
B120	*A. halleri*	4	0.0346	0	0.8964
	A. lyrata	2	0.0074	0	0.2236
B160	*A. halleri*	5	0.0094	0	0.3082
	A. lyrata	0	0	0.05	0
At4g21440	*A. halleri*	5	0.0563	0.103	3.7039
	A. lyrata	8	0.0689	0	1.7268
At4g21480	*A. halleri*	7	0	0.975	0
	A. lyrata	1	0	0.136	0
At4g21550	*A. halleri*	3	0.0463	0	5.2022
	A. lyrata	0	0	0.622	0
At4g21620	*A. halleri*	0	0.0346	0	3.5306
	A. lyrata	0	0.0192	0.105	1.3913

TAB. 3.4 – Estimates of Recombination rate.
R_{min} : Hudson and Kaplan's lower bound on the number of recombination.
ρ : recombination rate per base pair.
P : significance of the evidence for recombination tested by nonparametric permutation-based tests.
ρ/θ_w : equivalent to the r/μ ratio.

3.5 Discussion.

Previous theoretical studies on hitchhiking have predicted local effects of long-term balancing selection on neutral linked regions. Whether the physical size of the genomic region whose evolutionary dynamics is affected by a locus under balancing selection depends on the local recombination rate and population size, reasonable values of recombination rates are sufficient to restrict the elevation of neutral polymorphism to a small region of just a few hundred base pairs [Hudson and Kaplan, 1988]. In outcrossing species, the signal of balancing selection is restricted to a narrow genomic region because of the accumulation of recombination events in the genealogies of flanking loci [Charlesworth et al., 1997]. Our results confirm that in spite of the strong balancing selection operating at the S-locus, the expected elevation of nucleotide diversity due to genetic hitchhiking was only observed in the genes immediately flanking the S-locus on each side. In the downstream region of the S-locus, these signatures of hitchhiking are observed within 4.24 kb after the end of the non-homologous region, but balancing selection has no measurable influence in the next sequenced locus located at 28.04 kb. In the upstream region of the S-locus, an exceptionally high level of synonymous polymorphism was estimated at *B80* in both *A. halleri* and *A. lyrata* species, but the levels measured at 15kb upstream from *B80* did not differ from neutral expectations. Moreover, this rapid erosion of the signature of hitchhiking to the S-locus was observed at a fine scale in *B80* through a robust negative correlation between local synonymous SNPs density and distance from the non-homologous region in *A. halleri*. This observation indicates that intragenic recombination does occur in *B80*, although this gene is also subject to a strong hitchhiking effect to the S-locus. This apparent paradox had been predicted theoretically by Schierup &al. [Schierup et al., 2001b] as they showed that the increase in the timescale of gene genealogies in regions partially linked to a site under balancing selection would cause an increase in effective recombination rate. Hence, even if the recombination rate in the region flanking the S-locus is lower than the genomic average, as suggested [Hansson et al., 2006, Kawabe et al., 2006] a signature of recombination could be produced through this increase in effective recombination. Indeed, our estimates of intragenic recombination in *B80* in *A. halleri* indicate rather high values of effective recombination. Our estimates in *A. lyrata* are lower than in *A. halleri*, which may reflect differences in actual recombination rates, but could also be due to the more restricted geographical sample used for sequencing the genes *B80*, *ARK3* and *B120* (only populations from Iceland were used).

Although few empirical studies have focused on the signature of balancing selection in different species, they all conclude on the rapid erosion of the signal even in species with low recombination rate [Nordborg, 1997] as *A. thaliana* [Tian et al., 2002]. In *A. thaliana*, the elevation of polymorphism in neutral linked regions due to long term balancing selection was not measurable at 10 kb from *RPS5*, a gene involved in pathogen resistance. The small sizes of such linked regions were also reported in the human genome, complicating the identification of new targets of balancing selection [Akey et al., 2002, Akey et al., 2004, Bubb et al., 2006, Andrés et al., 2009].

Interestingly, our results suggest that the observed excess of polymorphism derived from both an elevated number of shared polymorphic sites and an excess of polymorphisms exclusive to only one species. Hence, this reveals that the excess of polymorphism has two distinct origins. First, an ancestral origin with the retention of segregating alleles for longer evolutionary times than in neutral regions. Second, a contemporary origin, whereby flanking genes exhibit longer external branches in coalescent trees, causing recent neutral mutations to be more likely retained at these loci than elsewhere in the genome. Hence, this excess of polymorphism seems to also reflect an ongoing process.

A recent empirical study in *A. halleri* suggested the existence of a sheltered genetic load linked to the S-locus [Llaurens et al., 2009], i.e., the accumulation of slightly recessive but deleterious mutations in the genomic neighborhood around the S-locus [Uyenoyama, 1997]. According to the annotation of *A. lyrata*'s genome, the non-recombining region only embedded *SRK* and *SCR* as coding sequences. Outside this region, a maximum of nine coding sequences are likely under the influence of linkage to the S-locus. Hence, these nine genes are particularly relevant candidates to investigate the genetic basis of the sheltered load. Since the sheltered load is believed to be specific to each allele, it will now be exciting to investigate haplotype structure of allelic variants of these nine genes and their association with specific *SRK*/*SCR* combinations.

3.6 Annexes.

Locus	$P(\theta_{hal})$	$P(\theta_{lyr})$	$P(\pi_{hal})$	$P(\pi_{lyr})$	$P(S_{x.hal})$	$P(S_{x.lyr})$	$P(S_{x.hal\ f.lyr})$	$P(S_{x.lyr\ f.hal})$	$P(S_s)$
At4g21170	0.1143	0.3135	0.0793	0.3076	0.1606	0.456	0.1679	0.291	0.0996
At4g21200	0.9183	0.688	0.7665	0.5474	0.7729	0.461	0.3095	0.2228	0.2624
At4g21230	0.5803	0.6384	0.4467	0.8473	0.4819	0.5353	0.2366	0.2371	0.3994
At4g21323	0.3077	0.4834	0.1394	0.616	0.1773	0.2541	0.1499	0.2899	0.3291
B80	**0**	**0**	**0**	**0**	**0.0255**	**0**	0.1159	0.2542	**0**
ARK3	**0.0364**	**0.0045**	**0.0302**	**0.0473**	0.0598	0.013	0.1824	0.0256	**0.0258**
B120	**0.0028**	**0.0105**	**0.023**	**0.0043**	**0.0113**	0.077	0.2952	0.1155	**0.0046**
B160	**0.0397**	0.4221	0.0872	0.3633	**0.0052**	0.2115	0.1667	0.2565	0.2931
At4g21440	0.6805	**0.0999**	0.7356	0.1045	0.6621	0.1112	0.4345	0.0902	0.2625
At4g21480	**0**	**7.00E-04**	**0**	0.1143	**0**	0.2133	0.1075	0.1607	**2.00E-04**
At4g21550	0.3822	0.4112	0.2511	0.387	0.1429	0.1555	0.2118	0.1436	0.1925
At4g21620	0.8058	0.5566	0.5219	0.3059	0.612	0.4597	0.2999	0.1211	0.2577

TAB. 3.5 – P is the probability of having a greater value under the best-fit demographic model than that observed.

Models compared	*df*	Test statistic	Significance
M1a v M2a	2	0	1
M0 v M3	4	6.041368	0.196080113
M8 v M7	2	1.40E-05	0.999993

TAB. 3.6 – Summary of test for the likelihood-ratio tests of *At4g21480* gene in *A. halleri*. Test was computed as $2(L_b - L_a)$, where L_a and L_b are log-likelihood values for each of the nested models being compared.
df = degrees of freedom

Model code	*P*	Log-likelihood	Parameter estimates
M_0 (one ratio)	1	-1235.938	ω=0.167
M_{1a} (NearlyNeutral)	1	-1233.549	p_0=0.871, p_1=0.129, ω_0=0.069
M_{2a} (PositiveSelection)	3	-1233.549	p_0=0.871 , p_1=0.109 , p_2=0.02 ,ω_0=0.07 , ω_1=1 , ω_2=1
M_3 (discrete)	5	-1232.918	p_0=0.652, p_1=0.28, p_2=0.067, ω_0=0, ω_1=0.5, ω_2=0.5
M_7 (β)	2	-1233.124	p=0.201, q=0.931
M_8 (β and ω)	4	-1233.124	p_0=0.999 , p_1=0.001 , p=0.201 , q=0.931 , ω=1

TAB. 3.7 – Results of maximum likelihood models of *At4g21480* gene in *A. halleri*

Models compared	*df*	Test statistic	Significance
$M_{1a}vM_{2a}$	2	0.0005	0.9997
M_0vM_3	4	6.00E-05	1
M_8vM_7	2	1.00E-04	0.9999

TAB. 3.8 – Summary of test for the likelihood-ratio tests of *At4g21480* gene in *A. lyrata*. Test was computed as $2(L_b - L_a)$, where L_a and L_b are log-likelihood values for each of the nested models being compared.
df = degrees of freedom

Model code	*P*	Log-likelihood	Parameter estimates
M_0 (one ratio)	1	-880.837	ω=0.078
M_{1a} (NearlyNeutral)	1	-880.837	p_0=0.999, p_1=0.000, ω_0=0.078
M_{2a} (PositiveSelection)	3	-880.837	p_0=1, p_1=0, p_2=0, ω_0=0.078, ω_1=1 , ω_2=1
M_3 (discrete)	5	-880.837	p_0=0.217, p_1=0.385, p_2=0.397, ω_0=0.078, ω_1=0.078, ω_2=0.078
M_7 (β)	2	-880.842	p=8.41, q=99
M_8 (β and ω)	4	-880.842	p_0=0.999, p_1=0.000 , p=8.409 , q=99 , ω=1

TAB. 3.9 – Results of maximum likelihood models of *At4g21480* gene in *A. lyrata*.

Discussion et perspectives.

A l'échelle d'une espèce, la variation observée d'un trait phénotypique est la résultante des processus stochastiques et du potentiel adaptatif de ce trait vis à vis d'un ou de plusieurs facteurs environnementaux. Pour qu'un trait phénotypique donné soit considéré comme étant un trait adaptatif en relation avec un facteur environnemental donné, sa variation doit être corrélée avec la divergence écologique mesurée à ce facteur environnemental. Dans ce cadre, un objectif important des études portant sur l'adaptation est de caractériser quels phénotypes peuvent être en interaction avec une pression environnementale particulière. En fonction de l'étendue de la période évolutive où cette interaction a lieu, il est possible de distinguer une adaptation partagée par un couple d'espèces proches par rapport à une adaptation spécifique à l'une des deux lignées évolutives, puis d'étudier le rôle potentiel de la sélection naturelle sur le processus de divergence des populations. L'essor de la génomique constitue une révolution en biologie évolutive en permettant d'associer la relation phénotype/environnement à une ou plusieurs entités physiques constituées par des segments génomiques. Un des exemples les plus frappants de la puissance de cette approche est la caractérisation du rôle du gène *Frigida* chez Arabidopsis thaliana, un gène majeur impliqué dans la date de floraison. Il existe un polymorphisme parmi les accessions naturelles d'*A. thaliana* pour la présence/absence de copies fonctionnelles du gène *Frigida*, les accessions porteuses de la mutation inactivante fleurissant précocement sans avoir intégré le signal de la vernalisation. La copie fonctionnelle de *Frigida* permet d'entrer en floraison lorsque les conditions environnementales paraissent favorables. Ainsi, il existe une corrélation entre la latitude d'origine des accessions naturelles d'*A. thaliana* et le temps nécessaire pour entrer en floraison pour les accessions ayant l'allèle fonctionnel de *Frigida*, corrélation non-observée chez les mutants naturels présentant une perte de fonction de *Frigida* [Stinchcombe et al., 2004]. Au delà de ces progrès dans l'identification des bases moléculaires de traits phénotypiques importants, la génomique est également une révolution en biologie par l'émergence de méthodes analytiques pour détecter la signature de la sélection naturelle à grande échelle, en particulier grâce à la disponibilité de données de séquençage complet de génomes. Identifier ainsi des régions génomiques sous sélection est essentiel pour relier génotypes et traits phénotypiques adaptatifs. Leur détection peut se faire indirectement en tirant profit du déséquilibre de liaison entre la cible directe de la sélection et les régions sélectivement neutres dans son voisinage. L'entrainement moléculaire des variants neutres dans les régions flanquantes par le locus sélectionné modifie la forme de la généalogie de ces régions par rapport à l'attendu pour des généalogies de régions neutres non-liées à un locus soumis à sélection. Ainsi, une catégorie de tests de neutralités estime le biais par rapport à l'attendu nul des patrons de polymorphisme observés au niveau de régions génomiques (D de Tajima,test HKA, F de Fu and Li). Cet attendu nul est formé de l'ensemble des généalogies possibles des individus composant une population de taille invariable dans le temps, et en complète neutralité sélective (modèle standard neutre). Une autre approche pour détecter la signature moléculaire de la sélection naturelle repose sur la quantification des niveaux

de différentiation (F_{ST}) entre deux populations de génomes. Comme pour les tests précédents, son principe repose sur la détection d'un biais par rapport à un attendu nul.

Cependant, même en absence de sélection, des écarts au modèle standard neutre peuvent conduire à des biais d'interprétations des tests pour détecter l'adaptation dans les génomes. Des études par simulation ont montré que la migration d'allèles entre populations entraine un fort risque de biais des patrons de polymorphismes par rapport à l'attendu nul [Stadler et al., 2008]. De même, les fluctuations dans les tailles efficaces des populations au cours du temps ont des conséquences importantes sur les patrons de diversité, mimant parfois les effets de la sélection naturelle dans le modèle standard neutre. Un nombre excessif de faux positifs lors de tests de sélection peuvent être la conséquence d'un goulot d'étranglement qui augmente la variance entre les gènes pour les niveaux de diversité [Wright and Gaut, 2005]. Les fluctuations des effectifs efficaces produisent en effet des patrons de polymorphisme différents de ceux du modèle standard neutre à l'échelle du génome. Ainsi, les premières mesures de polymorphisme à l'échelle d'un génome entier chez une espèce végétale connue pour avoir expérimenté une expansion géographique récente a révélé un décalage de la distribution empirique du D de Tajima vers des valeurs plus négatives que la distribution théorique selon le modèle standard neutre [Nordborg et al., 2005]. Ce besoin d'étudier à l'échelle génomique l'écart entre l'histoire démographique d'un modèle biologique et le modèle standard neutre a contribué à l'essor d'une discipline récente : la génomique des populations en divergence. Elle permet d'étudier les biais démographiques par rapport au modèle standard en estimant les valeurs des paramètres ancestraux à partir de populations échantillonnées. Le développement des méthodes informatiques IM, IMa et dérivés permettent d'estimer des tailles de populations ancestrales différentes de celles des populations filles, puis d'estimer également les flux géniques entre deux populations. Les analyses publiées à ce jour montrent que lors des évènements de spéciation il est fréquent que la taille ancestrale soit supérieure à celle des populations filles [Hey, 2006]. Ne pas tenir compte de cette transition de tailles de populations au moment de la spéciation par la considération d'un simple modèle de Wright et Fisher risque d'entrainer une surestimation des réels effectifs efficaces actuels, et donc une sous-estimation de l'importance de la dérive génétique contemporaine. En effet, la relation $\pi/4\mu$ pour estimer un effectif efficace actuel suppose que la population considérée est à l'équilibre, cette relation ne tient en effet pas compte de l'élongation des branches internes du coalescent lorsque la taille de la population ancestrale est plus grande que la taille actuelle. Les comparaisons de scénarios démographiques dans un cadre ABC à partir de données réelles (chapitre 1) montrent qu'en plus de l'importance de ne pas contraindre la taille de la population ancestrale, il peut également être nécessaire d'intégrer des changements de tailles efficaces au cours du temps dans les lignées filles. Par extension, afin d'obtenir des attendus démographiques nuls plus réalistes que le modèle standard ou le modèle IM, il faudrait idéalement complexifier les modèles. Une des limites méthodologiques actuelles, associée à l'inférence des histoires démographiques, est l'hétérogénéité des précisions dans les estimations parmi les paramètres décrivant les modèles. Des efforts importants sont ainsi à fournir pour utiliser plus efficacement l'information contenue dans les données de polymorphisme. Egalement, des efforts moins techniques que méthodologiques consisteraient à définir le niveau de complexité pour lequel un modèle démographique puisse être considéré comme "utile", puis d'estimer la puissance statistique avec laquelle un tel scénario pourrait être étudié.

Il serait dommage de penser que la détermination d'un modèle démographique plus adéquate que le modèle standard permettrait d'établir un attendu nul uniquement pour corriger par la démographie des tests de neutralités dans la détection de régions fortement soumises à sélection naturelle. Connaitre l'histoire démographique d'un couple d'espèces forme un cadre de travail qui

permet également de tester des hypothèses sur le rôle de la sélection naturelle dans le processus de spéciation. Etudier ainsi dans un cadre de spéciation la chronologie de la mise en place d'une architecture génomique impliquée dans un trait adaptatif permet en effet d'exclure ou non l'implication de ce trait dans la séparation de deux lignées évolutives. Ainsi chez les Brassicaceae, deux évènements de transition dans les systèmes de reproductions sont très étudiées, une dans le genre *Capsella*, l'autre dans le genre Arabidopsis. Elles consistent à la fixation du caractère autogame à l'échelle de l'espèce chez *A. thaliana* et *C. rubella* à partir d'un ancêtre auto-incompatible [Bechsgaard et al., 2006, Foxe et al., 2009, Guo et al., 2009]. Avoir une connaissance de l'histoire démographique de ces espèces permet de tester l'association de la transition dans leur système de reproduction avec de possibles goulots d'étranglements ou avec des changements dans les taux et les orientations des flux géniques. S'il est actuellement admis que la rupture de l'auto-incompatibilité est responsable de la spéciation de *C. rubella*, la séparation d'*A. thaliana* avec l'ancêtre commun de *A. halleri* et *A. lyrata* semble être indépendante d'un tel changement (mais voir les résultats du chapitre 1 qui remettent en cause ces datations). De même, la compréhension de l'évolution du caractère hyperaccumulateur et hypertolérant aux métaux d'*A. halleri* a beaucoup profité du transfert d'un cadre de travail intraspécifique à interspécifique. Dans un tel cadre il a été possible de quantifier les proportions relatives des différents types de sites polymorphes, et de constater que très peu de mutations dérivées dans la lignée *A. halleri* sont actuellement fixées par dérive génétique dans l'espèce. Ainsi, l'hyperaccumulation constitutive à l'ensemble de l'espèce *A. halleri* suggère un rôle important de la sélection naturelle dans la fixation de ce caractère. La comparaison intraspécifique permet également de comparer les fluctuations des effectifs efficaces et de rejeter des hypothèses en cours dans un cadre intraspécifique. Les effectifs efficaces très similaires entre *A. halleri* et *A. lyrata*, la rétention importante du polymorphisme ancestral chez les deux espèces ainsi que l'absence d'expansion ou de goulot d'étranglement récent chez l'une des deux espèces exclut l'hypothèse d'un effet fondateur ou d'un balayage sélectif à l'échelle du génome accompagnant l'évolution de l'hyperaccumulation en conséquence de l'adaptation à des sites métallifères.

L'association d'un cadre interspécifique à une étude sur l'histoire démographique permet également de comprendre la dynamique de la sélection naturelle sur les patrons de polymorphismes. En simulant les attendus nuls pour différentes catégories de sites polymorphes à des locus autour du locus-S chez *A. halleri* et *A. lyrata*, il est possible de voir que l'excès de polymorphisme du à l'entrainement moléculaire par le locus-S soumis à sélection balancée est expliquée par à la fois un excès de polymorphisme ancestral partagé et par un excès de mutations exclusives. Cette distinction entre les catégories de sites polymorphes possible uniquement dans un contexte interspécifique renseigne sur le mode d'action de la sélection au niveau du locus-S. Ainsi, la sélection balancée agissant sur le locus-S est à la fois un processus ancien qui entretient une quantité importante de polymorphisme ancestral chez les deux espèces, et un processus toujours en cours qui maintient un nombre élevé de mutations dérivées en ségrégations chez les deux espèces. Cette observation faite pour un couple d'espèces connues être allogames, il sera intéressant de tester par simulations l'effet sur les patrons de polymorphismes de la rupture de l'auto-incompatibilité chez une espèce au moment de sa spéciation puis d'appliquer cette démarche au couple d'espèces *C. rubella* et *C. grandiflora*. Enfin, il sera intéressant d'explorer par simulations l'effet sur les mêmes patrons de polymorphisme de la rupture de l'auto-incompatibilité dans une lignée évolutive bien après que la spéciation eue lieu.

Finalement, l'intégration des aspects démographiques dans un contexte interspécifique forme à ce jour un cadre de travail particulièrement excitant. Si des efforts importants sont en cours pour caractériser la diversité génétique des principales espèces modèles en biologie, le travail

présenté dans cette thèse permet d'anticiper que l'analyse conjointe de leurs espèces apparentées fournira une plus-value importante. En effet, au-delà des questions directement liées au processus de spéciation, la caractérisation conjointe des patrons de polymorphisme dans plusieurs espèces permet l'intégration des histoires démographiques et des histoires adaptatives ce qui ouvre des perspectives liées à la dynamique de l'action de la sélection naturelle sur les polymorphismes nucléaires.

Bibliographie

[Akey et al., 2004] Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., and Kruglyak, L. (2004). Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. *PLoS Biol*, 2(10) :e286. Available from : `http://dx.doi.org/10.1371%2Fjournal.pbio.0020286`.

[Akey et al., 2002] Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (2002). Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12(12) :1805–1814. Available from : `http://genome.cshlp.org/content/12/12/1805.abstract`.

[Al-Shehbaz and O'Kane, 2002] Al-Shehbaz, I. A. and O'Kane, S. L. (2002). Taxonomy and phylogeny of Arabidopsis (Brassicaceae). *The Arabidopsis Book*.

[Alloway, 1995] Alloway, B. (1995). *Heavy Metals in Soils*. Blackie Academic and Professional, London.

[Andolfatto, 2005] Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062) :1149–1152. 10.1038/nature04107. Available from : `http://dx.doi.org/10.1038/nature04107http://www.nature.com/nature/journal/v437/n7062/suppinfo/nature04107_S1.html`.

[Andolfatto, 2008] Andolfatto, P. (2008). Controlling Type-I Error of the McDonald-Kreitman Test in Genomewide Scans for Selection on Noncoding DNA. *Genetics*, 180(3) :1767–1771. Available from : `http://www.genetics.org/cgi/content/abstract/180/3/1767`.

[Andrés et al., 2009] Andrés, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., and Nielsen, R. (2009). Targets of Balancing Selection in the Human Genome. *Molecular Biology and Evolution*, 26(12) :2755–2764. Available from : `http://mbe.oxfordjournals.org/content/26/12/2755.abstract`.

[Anisimova and Gascuel, 2006] Anisimova, M. and Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches : A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55(4) :539–552. Available from : `http://sysbio.oxfordjournals.org/content/55/4/539.abstract`.

[Arndt et al., 2003] Arndt, P., Burge, C., and Hwa, T. (2003). DNA Sequence Evolution with Neighbor-Dependent Mutation. *Journal of Computational Biology*, 10(3-4) :313–322.

[Beaumont et al., 2010] Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S. A., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J. M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D., and Excoffier, L. (2010). In defence of model-based inference in phylogeography. *Molecular Ecology*, 19(3) :436–446. Available from : `http://dx.doi.org/10.1111/j.1365-294X.2009.04515.x`.

[Beaumont et al., 2002] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4) :2025–2035. Available from : `http://www.genetics.org/cgi/content/abstract/162/4/2025`.

[Becher et al., 2004] Becher, M., Talke, I. N., Krall, L., and Kramer, U. (2004). Cross-species microarray transcript profiling reveals high constitutive expression of metal homeostasis genes in shoots of the zinc hyperaccumulator Arabidopsis halleri. *Plant J.*, 37 :251.

[Bechsgaard et al., 2006] Bechsgaard, J. S., Castric, V., Charlesworth, D., Vekemans, X., and Schierup, M. H. (2006). The transition to self-compatibility in Arabidopsis thaliana and evolution within S-haplotypes over 10 Myr. *Mol Biol Evol*, 23 :1741 – 1750.

[Becquet and Przeworski, 2007] Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17(10) :1505–1519. Available from : `http://genome.cshlp.org/content/17/10/1505.abstract`, `doi:10.1101/gr.6409707`.

[Becquet and Przeworski, 2009] Becquet, C. and Przeworski, M. (2009). Learning about modes of speciation by computational approaches. *Evolution*, 63(10) :2547–2562. Available from : `http://dx.doi.org/10.1111/j.1558-5646.2009.00662.x`.

[Begun and Aquadro, 1993] Begun, D. J. and Aquadro, C. F. (1993). African and North American populations of Drosophila melanogaster are very different at the DNA level. *Nature*, 365(6446) :548–550. 10.1038/365548a0. Available from : `http://dx.doi.org/10.1038/365548a0`.

[Beilstein et al., 2010] Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*. Available from : `http://www.pnas.org/content/early/2010/09/28/0909766107.abstract`.

[Blum and François, 2010] Blum, M. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1) :63–73. 10.1007/s11222-009-9116-0. Available from : `http://dx.doi.org/10.1007/s11222-009-9116-0`.

[Box and Draper, 1987] Box, G. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, Oxford.

[Boyd, 2007] Boyd, R. (2007). The defense hypothesis of elemental hyperaccumulation : status, challenges and new directions. *Plant and Soil*, 293(1) :153–176. Available from : `http://dx.doi.org/10.1007/s11104-007-9240-6`.

[Bradshaw and Schemske, 2003] Bradshaw, H. D. and Schemske, D. W. (2003). Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, 426(6963) :176–178. 10.1038/nature02106. Available from : `http://dx.doi.org/10.1038/nature02106`.

[Bubb et al., 2006] Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., and Olson, M. V. (2006). Scan of Human Genome Reveals No New Loci Under Ancient Balancing Selection. *Genetics*, 173(4) :2165–2177. Available from : `http://www.genetics.org/cgi/content/abstract/173/4/2165`.

[Castric et al., 2008] Castric, V., Bechsgaard, J., Schierup, M. H., and Vekemans, X. (2008). Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection. *PLoS Genet*, 4(8) :e1000168. Available from : `http://dx.doi.org/10.1371%2Fjournal.pgen.1000168`.

[Castric and Vekemans, 2007] Castric, V. and Vekemans, X. (2007). Evolution under strong balancing selection : how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae ? *BMC Evolutionary Biology*, 7(1) :132. Available from : `http://www.biomedcentral.com/1471-2148/7/132`.

[Cavalli-Sforza and Feldman, 2003] Cavalli-Sforza, L. L. and Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet*.

[Charlesworth et al., 2005] Charlesworth, B., Bartolom, Eacute, Carolina, No, Euml, L, V., and Ronique (2005). The detection of shared and ancestral polymorphisms. *Genetics Research*, 86(02) :149–157. Available from : `http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=360623&fulltextType=RA&fileId=S0016672305007743`.

[Charlesworth et al., 1997] Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). *The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations*, volume 70.

[Charlesworth, 2006] Charlesworth, D. (2006). Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genet*, 2(4) :e64. Available from : `http://dx.plos.org/10.1371%2Fjournal.pgen.0020064`.

[Charlesworth et al., 2006] Charlesworth, D., Kamau, E., Hagenblad, J., and Tang, C. (2006). Trans-specificity at loci near the self-incompatibility loci in Arabidopsis. *Genetics*, 172 :2699 – 2704.

[Charlesworth and Vekemans, 2005] Charlesworth, D. and Vekemans, X. (2005). How and when did Arabidopsis thaliana become highly self-fertilising. *BioEssays*, 27(5) :472–476. Available from : `http://dx.doi.org/10.1002/bies.20231`, `doi:10.1002/bies.20231`.

[Clark, 1997] Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15) :7730–7734. Available from : `http://www.pnas.org/content/94/15/7730.abstract`.

[Clauss and Koch, 2006] Clauss, M. J. and Koch, M. A. (2006). Poorly known relatives of Arabidopsis thaliana. *Trends Plant Sci.*, 11 :449.

[Clauss and Mitchell-Olds, 2006] Clauss, M. J. and Mitchell-Olds (2006). Population genetic structure of Arabidopsis lyrata in Europe. *Molecular Ecology*, 15 :2753–2766. doi : 10.1111/j.1365-294X.2006.02973.x.

[Clemens, 2001] Clemens, S. (2001). Molecular mechanisms of plant metal tolerance and homeostasis. *Planta*, 212 :475.

[Consortium, 2001] Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. 10.1038/35057062. Available from : `http://dx.doi.org/10.1038/35057062http://www.nature.com/nature/journal/v409/n6822/suppinfo/409860a0_S1.html`.

[Coyne and Orr, 2004] Coyne, J. and Orr, H. (2004). Sinauer Associates, Sunderland.

[Csillery et al., 2010] Csillery, K., Francois, O., and Blum, M. G. B. (2010). *abc : estimation and model selection with Approximate Bayesian Computation (ABC)*. 1.

[Csilléry et al., 2010] Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7) :410–418. doi : DOI : 10.1016/j.tree.2010.04.001.

Available from : http://www.sciencedirect.com/science/article/B6VJ1-503WMFS-1/2/64bcd19f311b55ad28dc6bc889caad6d.

[Darwin, 1859] Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London.

[Dereeper et al., 2008] Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J. F., Guindon, S., Lefort, V., Lescot, M., Claverie, J. M., and Gascuel, O. (2008). Phylogeny.fr : robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36(suppl 2) :W465–W469. Available from : http://nar.oxfordjournals.org/content/36/suppl_2/W465.abstract, doi:10.1093/nar/gkn180.

[Dobzhansky, 1935] Dobzhansky, T. (1935). *Phil. Sci.*, 2 :344–355. 10.1086/286379. Available from : http://dx.doi.org/10.1086/286379.

[Dobzhansky, 1937] Dobzhansky, T. (1937). Genetic Nature of Species Differences. *The American Naturalist*, 71(735) :404. Available from : http://www.journals.uchicago.edu/doi/abs/10.1086/280726.

[Doherty and Zinkernagel, 1975] Doherty, P. C. and Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 256 :50–52. 10.1038/256050a0. Available from : http://dx.doi.org/10.1038/256050a0.

[Drummond and Rambaut, 2007] Drummond, A. and Rambaut, A. (2007). BEAST : Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1) :214. Available from : http://www.biomedcentral.com/1471-2148/7/214.

[Drummond et al., 2005] Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22 :1185 – 1192.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5) :1792–1797. Available from : http://nar.oxfordjournals.org/content/32/5/1792.abstract.

[Fagundes et al., 2007] Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45) :17614–17619. Available from : http://www.pnas.org/content/104/45/17614.abstract, doi:10.1073/pnas.0708280104.

[Force et al., 1999] Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, 151(4) :1531–1545. Available from : http://www.genetics.org/cgi/content/abstract/151/4/1531.

[Foxe et al., 2009] Foxe, J. P., Slotte, T., Stahl, E. A., Neuffer, B., Hurka, H., and Wright, S. I. (2009). Recent speciation associated with the evolution of selfing in Capsella. *Proceedings of the National Academy of Sciences*, 106(13) :5241–5245. Available from : http://www.pnas.org/content/106/13/5241.abstract.

[Freeman et al., 2006] Freeman, J. L., Quinn, C. F., Marcus, M. A., Fakra, S., and Pilon-Smits, E. A. H. (2006). Selenium-Tolerant Diamondback Moth Disarms Hyperaccumulator Plant Defense. *Current biology : CB*, 16(22) :2181–2192. Available from : http://linkinghub.elsevier.com/retrieve/pii/S0960982206022081.

[Fu, 1996] Fu, Y. X. (1996). New Statistical Tests of Neutrality for DNA Samples From a Population. *Genetics*, 143(1) :557–570. Available from : http://www.genetics.org/cgi/content/abstract/143/1/557.

[Gaffney and Keightley, 2005] Gaffney, D. J. and Keightley, P. D. (2005). The scale of mutational variation in the murid genome. *Genome Research*, 15(8) :1086–1094. Available from : http://genome.cshlp.org/content/15/8/1086.abstract, doi:10.1101/gr.3895005.

[Galtier and Duret, 2007] Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion ? extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6) :273 – 277. Available from : http://www.sciencedirect.com/science/article/B6TCY-4NDVGR0-3/2/31b3783138b1b4c8ecce404c18bbb1ef.

[Gavrilets, 2006] Gavrilets, S. (2006). The Maynard Smith model of sympatric speciation. *Journal of Theoretical Biology*, 239(2) :172–182. doi : DOI : 10.1016/j.jtbi.2005.08.041. Available from : http://www.sciencedirect.com/science/article/B6WMD-4HCDK4W-3/2/8321361affb9816e09087521f39a2297.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5) :696–704. Available from : http://sysbio.oxfordjournals.org/content/52/5/696.abstract, doi:10.1080/10635150390235520.

[Guo et al., 2009] Guo, Y.-L., Bechsgaard, J. S., Slotte, T., Neuffer, B., Lascoux, M., Weigel, D., and Schierup, M. H. (2009). Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. *Proceedings of the National Academy of Sciences*, 106(13) :5246–5251. Available from : http://www.pnas.org/content/106/13/5246.abstract.

[H. Schierup et al., 2000] H. Schierup, M., Vekemans, X., and Charlesworth, D. (2000). The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetical Research. *Genetical Research*, 76 :51–62.

[Hagenblad et al., 2006] Hagenblad, J., Bechsgaard, J., and Charlesworth, D. (2006). Linkage Disequilibrium Between Incompatibility Locus Region Genes in the Plant Arabidopsis lyrata. *Genetics*, 173(2) :1057–1073. Available from : http://www.genetics.org/cgi/content/abstract/173/2/1057, doi:10.1534/genetics.106.055780.

[Hamilton et al., 2005] Hamilton, G., Stoneking, M., and Excoffier, L. (2005). Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21) :7476–7480. Available from : http://www.pnas.org/content/102/21/7476.abstract.

[Hanikenne et al., 2008] Hanikenne, M., Talke, I. N., Haydon, M. J., Lanz, C., Nolte, A., Motte, P., Kroymann, J., Weigel, D., and Kramer, U. (2008). Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature*, 453(7193) :391–395.

[Hansson et al., 2006] Hansson, B., Kawabe, A., Preuss, S., Kuittinen, H., and Charlesworth, D. (2006). Comparative gene mapping in Arabidopsis lyrata chromosomes 1 and 2 and the corresponding A. thaliana chromosome 1 : recombination rates, rearrangements and centromere location. *Genetics Research*, 87(02) :75–85. Available from : http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=439699&fulltextType=RA&fileId=S0016672306008020.

[Hey, 2006] Hey, J. (2006). Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*, 16(6) :592–596. doi : DOI : 10.1016/j.gde.2006.10.005. Available from : http://www.sciencedirect.com/science/article/B6VS0-4M51F97-5/2/b6ab42c640c81c6291cd880f4efd9849.

[Hey and Nielsen, 2004] Hey, J. and 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(2) :747–760. Available from : http://www.genetics.org/cgi/content/abstract/167/2/747, doi:10.1534/genetics.103.024182.

[Hey and Nielsen, 2007] Hey, J. and 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(8) :2785–2790. Available from : http://www.pnas.org/content/104/8/2785.abstract, doi:10.1073/pnas.0611164104.

[Hobolth et al., 2007] Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. (2007). Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genet*, 3(2) :e7. Available from : http://dx.plos.org/10.1371%2Fjournal.pgen.0030007.

[Hodgkinson et al., 2009] Hodgkinson, A., Ladoukakis, E., and Eyre-Walker, A. (2009). Cryptic Variation in the Human Mutation Rate. *PLoS Biol*, 7(2) :e1000027. Available from : http://dx.doi.org/10.1371%2Fjournal.pbio.1000027.

[Hudson, 2001] Hudson, R. R. (2001). Two-Locus Sampling Distributions and Their Application. *Genetics*, 159(4) :1805–1817. Available from : http://www.genetics.org/cgi/content/abstract/159/4/1805.

[Hudson, 2002] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2) :337–338. Available from : http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/2/337.

[Hudson and Coyne, 2002] Hudson, R. R. and Coyne, J. A. (2002). MATHEMATICAL CONSEQUENCES OF THE GENEALOGICAL SPECIES CONCEPT. *Evolution*, 56(8) :1557–1565. Available from : http://dx.doi.org/10.1111/j.0014-3820.2002.tb01467.x, doi:10.1111/j.0014-3820.2002.tb01467.x.

[Hudson and Kaplan, 1985] Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1) :147–164. Available from : http://www.genetics.org/cgi/content/abstract/111/1/147.

[Hudson and Kaplan, 1988] Hudson, R. R. and Kaplan, N. L. (1988). The Coalescent Process in Models With Selection and Recombination. *Genetics*, 120(3) :831–840. Available from : http://www.genetics.org/cgi/content/abstract/120/3/831.

[Hudson et al., 1987] Hudson, R. R., Kreitman, M., and Aquade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116 :153–159.

[I. Wu, 2001] I. Wu, C. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6) :851–865. 10.1046/j.1420-9101.2001.00335.x. Available from : http://dx.doi.org/10.1046/j.1420-9101.2001.00335.x.

[Ioerger et al., 1990] Ioerger, T. R., Clark, A. G., and Kao, T. H. (1990). Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24) :9732–9735.

[Kamau et al., 2007] Kamau, E., Charlesworth, B., and Charlesworth, D. (2007). Linkage Disequilibrium and Recombination Rate Estimates in the Self-Incompatibility Region of Arabidopsis lyrata. *Genetics*, 176(4) :2357–2369. Available from : http://www.genetics.org/cgi/content/abstract/176/4/2357.

[Kamau and Charlesworth, 2005] Kamau, E. and Charlesworth, D. (2005). Balancing Selection and Low Recombination Affect Diversity near the Self-Incompatibility Loci of the Plant Arabidopsis lyrata. *Current biology : CB*, 15(19) :1773–1778. Available from : http://linkinghub.elsevier.com/retrieve/pii/S0960982205010213.

[Kashem et al., 2010] Kashem, M., Singh, B., Kubota, H., Sugawara, R., Kitajima, N., Kondo, T., and Kawai, S. (2010). Zinc tolerance and uptake by &lt ;i&gt ;Arabidopsis halleri ssp. gemmifera&lt ;/i&gt ; grown in nutrient solution. *Environmental Science and Pollution Research*, 17(5) :1174–1176. Available from : http://dx.doi.org/10.1007/s11356-009-0193-6, doi:10.1007/s11356-009-0193-6.

[Kawabe et al., 2006] Kawabe, A., Hansson, B., Forrest, A., Hagenblad, J., and Charlesworth, D. (2006). Comparative gene mapping in Arabidopsis lyrata chromosomes 6 and 7 and A. thaliana chromosome IV : evolutionary history, rearrangements and local recombination rates. *Genetics Research*, 88(01) :45–56. Available from : http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=504888&fulltextType=RA&fileId=S0016672306008287.

[Kekalainen et al., 2009] Kekalainen, J., Vallunen, J. A., Primmer, C. R., Rattya, J., and Taskinen, J. (2009). Signals of major histocompatibility complex overdominance in a wild salmonid population. *Proceedings of the Royal Society B : Biological Sciences*, 276(1670) :3133–3140. Available from : http://rspb.royalsocietypublishing.org/content/276/1670/3133.abstract.

[Kim and Stephan, 2002] Kim, Y. and Stephan, W. (2002). Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. *Genetics*, 160(2) :765–777. Available from : http://www.genetics.org/cgi/content/abstract/160/2/765.

[Kimura, 1955] Kimura, M. (1955). SOLUTION OF A PROCESS OF RANDOM GENETIC DRIFT WITH A CONTINUOUS MODEL. *Proceedings of the National Academy of Sciences of the United States of America*, 41(3) :144–150. Available from : http://www.pnas.org/content/41/3/144.short.

[Kliman et al., 2000] Kliman, R. M., Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A. J., McCarter, J., Wakeley, J., and Hey, J. (2000). The Population Genetics of the Origin and Divergence of the Drosophila simulans Complex Species. *Genetics*, 156(4) :1913–1931. Available from : http://www.genetics.org/cgi/content/abstract/156/4/1913.

[Koch et al., 2001] Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2001). Molecular systematics of the Brassicaceae : evidence from coding plastidic MATK and nuclear CHS sequences. *Am J Bot*, 88 :534 – 544.

[Koch and Matschinger, 2007] Koch, M. A. and Matschinger, M. (2007). Evolution and genetic differentiation among relatives of Arabidopsis thaliana. *Proc Natl Acad Sci USA*, 104 :6272 – 6277.

[Krings, 1997] Krings, M. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90 :19–30. 10.1016/S0092-8674(00)80310-4. Available from : http://dx.doi.org/10.1016/S0092-8674(00)80310-4.

[Kubota and Takenaka, 2003] Kubota, H. and Takenaka, C. (2003). Field Note : ¡i¿Arabis gemmifera¡/i¿ is a Hyperaccumulator of Cd and Zn. *International Journal of Phytoremediation*, 5(3) :197 – 201. Available from : http://www.informaworld.com/10.1080/713779219.

[Leitner et al., 1997] Leitner, T., Kumar, S., and Albert, J. (1997). Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.*, 71(6) :4761–4770. Available from : http://jvi.asm.org/cgi/content/abstract/71/6/4761.

[Lewontin and Krakauer, 1973] Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74 :175–195.

[Li et al., 2010] Li, Y., Stocks, M., Hemmilä, S., Källman, T., Zhu, H., Zhou, Y., Chen, J., Liu, J., and Lascoux, M. (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(5) :1001–1014. Available from : http://mbe.oxfordjournals.org/content/27/5/1001.abstract.

[Librado and Rozas, 2009] Librado, P. and Rozas, J. (2009). DnaSP v5 : a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11) :1451–1452. Available from : http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/11/1451.

[Linnen et al., 2009] Linnen, C. R., Kingsley, E. P., Jensen, J. D., and Hoekstra, H. E. (2009). On the Origin and Spread of an Adaptive Allele in Deer Mice. *Science*, 325(5944) :1095–1098. Available from : http://www.sciencemag.org/cgi/content/abstract/325/5944/1095.

[Liu et al., 2006] Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *American journal of human genetics*, 79(2) :230–237. Available from : http://linkinghub.elsevier.com/retrieve/pii/S0002929707631310.

[Llaurens et al., 2009] Llaurens, V., Gonthier, L., and Billiard, S. (2009). The Sheltered Genetic Load Linked to the S Locus in Plants : New Insights From Theoretical and Empirical Approaches in Sporophytic Self-Incompatibility. *Genetics*, 183(3) :1105–1118. Available from : http://www.genetics.org/cgi/content/abstract/183/3/1105.

[Lynch and Conery, 2003] Lynch, M. and Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, 302(5649) :1401–1404. Available from : http://www.sciencemag.org/cgi/content/abstract/302/5649/1401, doi:10.1126/science.1089370.

[MacNair and Christie, 1983] MacNair, M. R. and Christie, P. (1983). Reproductive isolation as a pleiotropic effect of copper tolerance in Mimulus guttatus ? *Heredity*, 50(3) :295–302. Available from : http://dx.doi.org/10.1038/hdy.1983.31.

[Makova and Li, 2002] Makova, K. D. and Li, W.-H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. *Nature*, 416(6881) :624–626. 10.1038/416624a. Available from : http://dx.doi.org/10.1038/416624ahttp://www.nature.com/nature/journal/v416/n6881/suppinfo/416624a_S1.html.

[Maruyama and Nei, 1981] Maruyama, T. and Nei, M. (1981). Genetic variability maintained by mutation and overdominant selection in finite population. *Genetics*, 98(2) :441–459. Available from : http://www.genetics.org/cgi/content/abstract/98/2/441.

[Masly and Presgraves, 2007] Masly, J. P. and Presgraves, D. C. (2007). High-Resolution Genome-Wide Dissection of the Two Rules of Speciation in ¡italic¿Drosophila¡/italic¿. *PLoS Biol*, 5(9) :e243. Available from : http://dx.doi.org/10.1371%2Fjournal.pbio.0050243.

[Maynard Smith and Haigh, 1974] Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research.*, 23 :23–35.

[Mayr, 1942] Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press, New-York.

[McVean et al., 2002] McVean, G., Awadalla, P., and Fearnhead, P. (2002). A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics*, 160(3) :1231–1241. Available from : http://www.genetics.org/cgi/content/abstract/160/3/1231.

[Meagher and Potts, 1997] Meagher, S. and Potts, W. K. (1997). A Microsatellite-Based MHC Genotyping System for House Mice (Mus domesticus). *Hereditas*, 127 :75–82.

[Merlo et al., 1995] Merlo, A., Herman, J. G., Mao, L., Lee, D. J., Gabrielson, E., Burger, P. C., Baylin, S. B., and Sidransky, D. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med*, 1(7) :686–692. 10.1038/nm0795-686. Available from : http://dx.doi.org/10.1038/nm0795-686.

[Mihola et al., 2009] Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., and Forejt, J. (2009). A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. *Science*, 323(5912) :373–375. Available from : http://www.sciencemag.org/cgi/content/abstract/323/5912/373.

[Mitchell-Olds, 2001] Mitchell-Olds, T. (2001). Arabidopsis thaliana and its wild relatives : a model system for ecology and evolution. *Trends Ecol Evol*, 16 :693 – 700.

[Muirhead et al., 2002] Muirhead, C. A., Glass, N. L., and Slatkin, M. (2002). Multilocus Self-Recognition Systems in Fungi as a Cause of Trans-Species Polymorphism. *Genetics*, 161(2) :633–641. Available from : http://www.genetics.org/cgi/content/abstract/161/2/633.

[Muller, 1942] Muller, H. J. (1942). Isolation mechanisms, evolution and temperature. *Biol Symp*, 6 :71–125.

[Nielsen and Wakeley, 2001] Nielsen, R. and Wakeley, J. (2001). Distinguishing Migration From Isolation : A Markov Chain Monte Carlo Approach. *Genetics*, 158(2) :885–896. Available from : http://www.genetics.org/cgi/content/abstract/158/2/885.

[Nordborg, 1997] Nordborg, M. (1997). Structured Coalescent Processes on Different Time Scales. *Genetics*, 146(4) :1501–1514. Available from : http://www.genetics.org/cgi/content/abstract/146/4/1501.

[Nordborg et al., 2005] Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005). The pattern of polymorphism in ¡italic¿arabidopsis thaliana¡/italic¿. *PLoS Biol*, 3 :e196. Available from : http://dx.doi.org/10.1371%2Fjournal.pbio.0030196.

[Oliver et al., 2009] Oliver, M., Telfer, S., and Piertney, S. (2009). Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole

(Arvicola terrestris). *Proceedings of the Royal Society B : Biological Sciences*, 276(1659) :1119–1128. Available from : http://rspb.royalsocietypublishing.org/content/276/1659/1119.abstract.

[Orr, 1995] Orr, H. A. (1995). The population genetics of speciation : the evolution of hybrid incompatibilities. *Genetics*, 139 :1805–1813.

[Patterson et al., 2006] Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097) :1103–1108. 10.1038/nature04789. Available from : http://dx.doi.org/10.1038/nature04789http://www.nature.com/nature/journal/v441/n7097/suppinfo/nature04789_S1.html.

[Pauwels et al., 2006] Pauwels, M., Frérot, H., Bonnin, I., and Saumitou-Laprade, P. (2006). A broad-scale analysis of population differentiation for Zn tolerance in an emerging model species for tolerance study : *Arabidopsis halleri* (Brassicaceae). *Journal of Evolutionary Biology*, 19(6) :1838–1850. Available from : http://dx.doi.org/10.1111/j.1420-9101.2006.01178.x.

[Pauwels et al., 2005] Pauwels, M., Saumitou-Laprade, P., Holl, A. C., Petit, D., and Bonnin, I. (2005). Multiple origin of metallicolous populations of the pseudometallophyte ¡i¿Arabidopsis halleri¡/i¿ (Brassicaceae) in central Europe : the cpDNA testimony. *Molecular Ecology*, 14(14) :4403–4414. 10.1111/j.1365-294X.2005.02739.x. Available from : http://dx.doi.org/10.1111/j.1365-294X.2005.02739.x.

[Pauwels et al., 2008] Pauwels, M., Willems, G., Roosens, N., Frérot, H., and Saumitou-Laprade, P. (2008). Merging methods in molecular and ecological genetics to study the adaptation of plants to anthropogenic metal-polluted sites : implications for phytoremediation. *Molecular Ecology*, 17 :108–119.

[Payseur and Nachman, 2002] Payseur, B. A. and Nachman, M. W. (2002). Natural selection at linked sites in humans. *Gene*, 300(1-2) :31–42. Available from : http://www.sciencedirect.com/science/article/B6T39-474GKTP-1/2/49e165de994b106eec5fc7c0c9cfb847.

[Penn et al., 2002] Penn, D. J., Damjanovich, K., and Potts, W. K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17) :11260–11264. Available from : http://www.pnas.org/content/99/17/11260.abstract.

[Phadnis and Orr, 2009] Phadnis, N. and Orr, H. A. (2009). A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids. *Science*, 323(5912) :376–379. Available from : http://www.sciencemag.org/cgi/content/abstract/323/5912/376, doi:10.1126/science.1163934.

[Pinho and Hey, 2010] Pinho, C. and Hey, J. (2010). Divergence with Gene Flow : Models and Data. *Annual Review of Ecology, Evolution, and Systematics*, 41(1). Available from : http://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-102209-144644, doi:doi:10.1146/annurev-ecolsys-102209-144644.

[Posada and Crandall, 1998] Posada, D. and Crandall, K. A. (1998). MODELTEST : testing the model of DNA substitution. *Bioinformatics*, 14(9) :817–818. Available from : http://bioinformatics.oxfordjournals.org/cgi/content/abstract/14/9/817.

[Prasad, 2009] Prasad, M. (2009). Springer Berlin Heidelberg, Berlin.

[Pritchard et al., 1999] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12) :1791–1798. Available from : http://mbe.oxfordjournals.org/content/16/12/1791.abstract.

[R Development Core Team, 2008] R Development Core Team (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from : http://www.R-project.org.

[Ramos-Onsins and Rozas, 2002] Ramos-Onsins, S. E. and Rozas, J. (2002). Statistical Properties of New Neutrality Tests Against Population Growth. *Mol Biol Evol*, 19(12) :2092–2100. Available from : http://mbe.oxfordjournals.org/cgi/content/abstract/19/12/2092.

[Ramos-Onsins et al., 2004] Ramos-Onsins, S. E., Stranger, B. E., Mitchell-Olds, T., and Aguade, M. (2004). Multilocus Analysis of Variation and Speciation in the Closely Related Species Arabidopsis halleri and A. lyrata. *Genetics*, 166(1) :373–388. Available from : http://www.genetics.org/cgi/content/abstract/166/1/373.

[Rascio and Navari-Izzo,] Rascio, N. and Navari-Izzo, F. Heavy metal hyperaccumulating plants : How and why do they do it ? And what makes them so interesting ? *Plant Science*, In Press, Uncorrected Proof. doi : DOI : 10.1016/j.plantsci.2010.08.016. Available from : http://www.sciencedirect.com/science/article/B6TBH-511BYSY-1/2/6b50d73711c67e4a1565d828c6974774.

[Richman et al., 2003] Richman, A., Herrera, L. G., Nash, D., and Schierup, M. H. (2003). Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in Peromyscus mainculatus. *Genet Res*, 82 :89–99. 10.1017/S0016672303006347. Available from : http://dx.doi.org/10.1017/S0016672303006347.

[Ronce and Kirkpatrick, 2001] Ronce, O. and Kirkpatrick, M. (2001). When sources become sinks : migrational meltdown in heterogeneous habitats. *Evolution*, 55 :1520–1531.

[Ross-Ibarra et al., 2009] Ross-Ibarra, J., Tenaillon, M., and Gaut, B. S. (2009). Historical Divergence and Gene Flow in the Genus Zea. *Genetics*, page genetics.108.097238. Available from : http://www.genetics.org/cgi/content/abstract/genetics.108.097238v1.

[Ross-Ibarra et al., 2008] Ross-Ibarra, J., Wright, S. I., Foxe, J. P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., and Gaut, B. S. (2008). Patterns of Polymorphism and Demographic History in Natural Populations of ¡italic¿Arabidopsis lyrata¡/italic¿. *PLoS ONE*, 3(6) :e2411. Available from : http://dx.plos.org/10.1371%2Fjournal.pone.0002411.

[Ruggiero et al., 2008] Ruggiero, M., Jacquemin, B., Castric, V., and Vekemans, X. (2008). Hitch-hiking to a locus under balancing selection : high sequence diversity and low population subdivision at the S-locus genomic region in Arabidopsis halleri. *Genetical Research.*, 90(1) :37–46.

[Rundle and Nosil, 2005] Rundle, H. D. and Nosil, P. (2005). Ecological speciation. *Ecology Letters*, 8(3) :336–352. 10.1111/j.1461-0248.2004.00715.x. Available from : http://dx.doi.org/10.1111/j.1461-0248.2004.00715.x.

[Schierup et al., 2001a] Schierup, M. H., Mable, B. K., Awadalla, P., and Charlesworth, D. (2001a). Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of Arabidopsis lyrata. *Genetics*, 158 :387 – 399.

[Schierup et al., 2001b] Schierup, M. H., Mikkelsen, A. M., and Hein, J. (2001b). Recombination, Balancing Selection and Phylogenies in MHC and Self-Incompatibility Genes. *Genetics*, 159(4) :1833–1844.

[Schluter, 2001] Schluter, D. (2001). Ecology and the origin of species. *Trends in Ecology & Evolution*, 16(7) :372–380. doi : DOI : 10.1016/S0169-5347(01)02198-X. Available from : http://www.sciencedirect.com/science/article/B6VJ1-436W013-8/2/17c8bbbd67d1a60c09a0dbc59ad7cc57.

[Schluter and Conte, 2009] Schluter, D. and Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1) :9955–9962. Available from : http://www.pnas.org/content/106/suppl.1/9955.abstract.

[Sequencing, 2005] Sequencing, A. C. C. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437 :69–87.

[Shahzad et al., 2010] Shahzad, Z., Gosti, F., Frérot, H., Lacombe, E., Roosens, N., Saumitou-Laprade, P., and Berthomieu, P. (2010). The Five ¡italic¿AhMTP1¡/italic¿ Zinc Transporters Undergo Different Evolutionary Fates towards Adaptive Evolution to Zinc Tolerance in ¡italic¿Arabidopsis halleri¡/italic¿. *PLoS Genet*, 6(4) :e1000911. Available from : http://dx.doi.org/10.1371%2Fjournal.pgen.1000911.

[Sherman-Broyles et al., 2007] Sherman-Broyles, S., Boggs, N., Farkas, A., Liu, P., Vrebalov, J., Nasrallah, M. E., and Nasrallah, J. B. (2007). S Locus Genes and the Evolution of Self-Fertility in Arabidopsis thaliana. *Plant Cell*, 19(1) :94–106. Available from : http://www.plantcell.org/cgi/content/abstract/19/1/94.

[Slatkin, 1985] Slatkin, M. (1985). Gene flow in natural populations. *Ann Rev Ecol Syst*, 16 :393–430. 10.1146/annurev.ecolsys.16.1.393. Available from : http://dx.doi.org/10.1146/annurev.ecolsys.16.1.393.

[Stadler et al., 2008] Stadler, T., Arunyawat, U., and Stephan, W. (2008). Population Genetics of Speciation in Two Closely Related Wild Tomatoes (Solanum Section Lycopersicon). *Genetics*, 178(1) :339–350. Available from : http://www.genetics.org/cgi/content/abstract/178/1/339.

[Stewart et al., 2003] Stewart, C. N., Halfhill, Matthew, D., and Warwick, S. I. (2003). Transgene introgression from genetically modified crops to their wild relatives. *Nat Rev Genet*, 4 :806–817.

[Stinchcombe et al., 2004] Stinchcombe, J. R., Weinig, C., Ungerer, M., Olsen, K. M., Mays, C., Halldorsdottir, S. S., Purugganan, M. D., and Schmitt, J. (2004). A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13) :4712–4717. Available from : http://www.pnas.org/content/101/13/4712.abstract.

[Strobeck, 1983] Strobeck, C. (1983). Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics*, 103(3) :545–555. Available from : http://www.genetics.org/cgi/content/abstract/103/3/545.

[Tajima, 1983] Tajima, F. (1983). EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics*, 105(2) :437–460. Available from : http://www.genetics.org/cgi/content/abstract/105/2/437.

[Tajima, 1989] Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3) :585–595. Available from : http://www.genetics.org/cgi/content/abstract/123/3/585.

[Takahata, 1990] Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proceedings of the National Academy of Sciences of the United States of America*, 87(7) :2419–2423. Available from : http://www.pnas.org/content/87/7/2419.abstract.

[Takahata and Nei, 1990] Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, 124 :967–978.

[Tamura et al., 2007] Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4 : Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24 :1596 – 1599.

[Tavare et al., 1997] Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2) :505–518. Available from : http://www.genetics.org/cgi/content/abstract/145/2/505.

[Templeton, 2002] Templeton, A. (2002). Out of Africa again and again. *Nature*, 416(6876) :45–51. 10.1038/416045a. Available from : http://dx.doi.org/10.1038/416045ahttp://www.nature.com/nature/journal/v416/n6876/suppinfo/416045a_S1.html.

[Tenaillon and Tiffin, 2008] Tenaillon, M. I. and Tiffin, P. L. (2008). The quest for adaptive evolution : a theoretical challenge in a maze of data. *Current Opinion in Plant Biology*, 11(2) :110–115. doi : DOI : 10.1016/j.pbi.2007.12.003. Available from : http://www.sciencedirect.com/science/article/B6VS4-4RS3TPN-1/2/ff1cac0e77451eacb70f7a86ec7ebfbc.

[Thalmann et al., 2007] Thalmann, O., Fischer, A., Lankester, F., Paabo, S., and Vigilant, L. (2007). The Complex Evolutionary History of Gorillas : Insights from Genomic Data. *Molecular Biology and Evolution*, 24(1) :146–158. Available from : http://mbe.oxfordjournals.org/content/24/1/146.abstract, doi:10.1093/molbev/msl160.

[Tian et al., 2002] Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M. (2002). Signature of balancing selection in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17) :11525–11530. Available from : http://www.pnas.org/content/99/17/11525.abstract.

[Ting et al., 2000] Ting, C.-T., Tsaur, S.-C., and Wu, C.-I. (2000). The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10) :5313–5316. Available from : http://www.pnas.org/content/97/10/5313.abstract.

[Tsuchimatsu et al.,] Tsuchimatsu, T., Suwabe, K., Shimizu-Inatsugi, R., Isokawa, S., Pavlidis, P., Stadler, T., Suzuki, G., Takayama, S., Watanabe, M., and Shimizu, K. K. Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. *Nature*, 464(7293) :1342–1346. 10.1038/nature08927. Available from : http://dx.doi.org/10.1038/nature08927http://www.nature.com/nature/journal/v464/n7293/suppinfo/nature08927_S1.html.

[Uyenoyama, 1997] Uyenoyama, M. K. (1997). Genealogical Structure Among Alleles Regulating Self-Incompatibility in Natural Populations of Flowering Plants. *Genetics*, 147(3) :1389–1400. Available from : http://www.genetics.org/cgi/content/abstract/147/3/1389.

[Vekemans and Slatkin, 1994] Vekemans, X. and Slatkin, M. (1994). Gene and Allelic Genealogies at a Gametophytic Self-Incompatibility Locus. *Genetics*, 137(4) :1157–1165. Available from : http://www.genetics.org/cgi/content/abstract/137/4/1157.

[Venables and Ripley, 2002] Venables, W. and Ripley, B. (2002). Modern Applied Statistics with S. Springer. ISBN 0-387-95457-0. Available from : http://www.stats.ox.ac.uk/pub/MASS4.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides,

P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., et al. (2001). The Sequence of the Human Genome. *Science*, 291(5507) :1304–1351. Available from : `http://www.sciencemag.org/cgi/content/abstract/291/5507/1304`, `doi:10.1126/science.1058040`.

[Via, 2001] Via, S. (2001). Sympatric speciation in animals : the ugly duckling grows up. *Trends in Ecology & Evolution*, 16(7) :381–390. doi : DOI : 10.1016/S0169-5347(01)02188-7. Available from : `http://www.sciencedirect.com/science/article/B6VJ1-436W013-9/2/5169c399cc345a8cfcd8df7f2a428d5d`.

[Via, 2009] Via, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, 106(Supplement 1) :9939–9946. Available from : `http://www.pnas.org/content/106/suppl.1/9939.abstract`, `doi:10.1073/pnas.0901397106`.

[Via and West, 2008] Via, S. and West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, 17(19) :4334–4345. Available from : `http://dx.doi.org/10.1111/j.1365-294X.2008.03921.x`, `doi:10.1111/j.1365-294X.2008.03921.x`.

[Wagner, 1873] Wagner, M. (1873). *The Darwinian theory and the law of the migration of organisms*. Edward Stanford, London.

[Wakeley and Hey, 1997] Wakeley, J. and Hey, J. (1997). Estimating Ancestral Population Parameters. *Genetics*, 145(3) :847–855. Available from : `http://www.genetics.org/cgi/content/abstract/145/3/847`.

[Watterson, 1975] Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2) :256–276. doi : DOI : 10.1016/0040-5809(75)90020-9. Available from : `http://www.sciencedirect.com/science/article/B6WXD-4F1HNYK-C/2/5ea232159a58446ac9a72553b6659b46`.

[Willems et al., 2007] Willems, G., Drager, D. B., Courbot, M., Gode, C., Verbruggen, N., and Saumitou-Laprade, P. (2007). The Genetic Basis of Zinc Tolerance in the Metallophyte Arabidopsis halleri ssp. halleri (Brassicaceae) : An Analysis of Quantitative Trait Loci. *Genetics*, 176(1) :659–674. Available from : `http://www.genetics.org/cgi/content/abstract/176/1/659`, `doi:10.1534/genetics.106.064485`.

[Won and Hey, 2005] Won, Y.-J. and Hey, J. (2005). Divergence Population Genetics of Chimpanzees. *Molecular Biology and Evolution*, 22(2) :297–307. Available from : `http://mbe.oxfordjournals.org/content/22/2/297.abstract`.

[Wright, 1939] Wright, S. (1939). The distribution of self-sterility alleles in populations. *Genetics*, 24 :538–552.

[Wright and Charlesworth, 2004] Wright, S. I. and Charlesworth, B. (2004). The HKA Test Revisited : A Maximum-Likelihood-Ratio Test of the Standard Neutral Model. *Genetics*, 168(2) :1071–1076. Available from : `http://www.genetics.org/cgi/content/abstract/168/2/1071`.

[Wright et al., 2006] Wright, S. I., Foxe, J. P., DeRose-Wilson, L., Kawabe, A., Looseley, M., Gaut, B. S., and Charlesworth, D. (2006). Testing for Effects of Recombination Rate on Nucleotide Diversity in Natural Populations of Arabidopsis lyrata. *Genetics*, 174(3) :1421–1430. Available from : `http://www.genetics.org/cgi/content/abstract/174/3/1421`.

[Wright and Gaut, 2005] Wright, S. I. and Gaut, B. S. (2005). Molecular Population Genetics and the Search for Adaptive Evolution in Plants. *Molecular Biology and Evolution*, 22(3) :506–519. Available from : `http://mbe.oxfordjournals.org/content/22/3/506.abstract`.

[Wu et al., 1998] Wu, J., Saupe, S. J., and Glass, N. L. (1998). Evidence for balancing selection operating at the het-c heterokaryon incompatibility locus in a group of filamentous fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21) :12398–12403. Available from : `http://www.pnas.org/content/95/21/12398.abstract`.

[Yang and Seto, 2007] Yang, X. J. and Seto, E. (2007). HATs and HDACs : from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene*, 26(37) :5310–5318. Available from : `http://dx.doi.org/10.1038/sj.onc.1210599`.

[Yang, 2007] Yang, Z. (2007). PAML 4 : Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8) :1586–1591. Available from : `http://mbe.oxfordjournals.org/content/24/8/1586.abstract`.

[Zhao et al., 2000] Zhao, F. J., Lombi, E., Breedon, T., and M, S. P. (2000). Zinc hyperaccumulation and cellular distribution in Arabidopsis halleri. *Plant, Cell & Environment*, 23(5) :507–514. Available from : `http://dx.doi.org/10.1046/j.1365-3040.2000.00569.x`.