Analyse temporelle de la diversité en régime autogame approches théorique et empirique

Margaux Jullien

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE MONTPELLIER SUPAGRO

En Génétique et Génomique

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau
Portée par

Unité de recherche AGAP

Analyse temporelle de la diversité en régime autogame : Approches théorique et empirique

Présentée par Margaux JULLIEN
Le 14 mai 2019

Sous la direction de Joëlle RONFORT, Laurène GAY
et Miguel NAVASCUÉS

Devant le jury composé de

Mme Agnès MIGNOT, Professeur, Université de Montpellier	Présidente
Mme Frédérique VIARD, Directrice de recherche, CNRS	Rapporteur
M. Denis ROZE, Chargé de recherche, CNRS	Rapporteur
M. Jérôme ENJALBERT, Directeur de recherche, INRA	Examinateur
Mme Joëlle RONFORT, Directrice de recherche, INRA	Directrice de thèse
Mme Laurène GAY, Chargée de recherche, INRA	Co-encadrante de thèse
M. Miguel NAVASCUÉS, Chargé de recherche, INRA	Co-encadrant de thèse

UNIVERSITÉ DE MONTPELLIER

Montpellier SupAgro

“And now, let us step out into the night and pursue that flighty temptress, adventure.”

Remerciements

Après trois ans et demi de thèse, il y a beaucoup de personnes que je souhaite remercier. La thèse n'est pas toujours une période facile, il y a parfois du découragement et des moments difficiles, mais je pense que j'ai eu la chance d'être entourée des bonnes personnes au bon moment. Je vais probablement oublier des gens, du coup si je vous ai croisé à un moment donné pendant ces années, il y a de bonnes chances pour que je vous remercie ☺

Tout d'abord, je veux remercier mes encadrants : Laurène (élue meilleure encadrante par un panel de thésards), Joëlle et Miguel. Merci de m'avoir donné l'opportunité de réaliser cette thèse avec vous et de m'avoir conseillée et accompagnée pendant ces trois ans et demi. Merci pour votre soutien, votre disponibilité et votre gentillesse tout au long de ce parcours.

Merci à Frédérique Viard, Denis Roze, Agnès Mignot et Jérôme Enjalbert d'avoir accepté d'évaluer mon travail en participant à mon jury. Merci également à Ophélie Ronce, Sylvain Glémin, Patrice David, Renaud Vitalis, et Guillaume Achaz pour les longues discussions qui m'ont grandement aidée à faire avancer mon projet de thèse lors de mes comités. Merci également à Mathieu Siol et John Pannell pour leurs remarques constructives sur mon article.

Je voudrais remercier toute l'équipe GE^{2}POP qui m'a accueillie pendant ces trois ans et demi. Je n'aurais pas pu espérer être mieux entourée pendant ma thèse. Cette équipe est composée des personnes les plus gentilles que l'on puisse imaginer, je pense que c'est même le critère de sélection principal pour en faire partie. Nathalie, Hélène, Jean-Marie, Brigitte, Pierre, Martin, Jacques, Marie-Hélène, Vincent, Sylvain, Muriel, Christine, Véro, Audrey, Morgane, Pascal, merci à tous de m'avoir entourée et accompagnée pendant cette thèse. Je tiens à remercier en particulier Karine pour son aide précieuse et sa patience infinie pendant le génotypage, et une mention spéciale à Cédric pour m'avoir sauvée en me trouvant des clusters de calcul de secours (encore désolée d'avoir utilisé toutes tes heures de calcul…). Et bien sûr, merci à l'équipe de postdocs, Diala, Nicolas, Ingrid, Yann qui m'ont également donné des conseils au cours de leur passage à GE^{2}POP.

J'ai également eu la chance d'avoir de super compagnons de bureau : Germain, Josselin et Yacine, mes co-thésards qui m'ont rejoint alors que j'étais encore la seule thésarde de l'équipe. Plusieurs stagiaires sont aussi passés par ce bureau : Arnaud (j'ai encore peur d'avoir un fond d'écran avec la

poule ou le chien le plus moche du monde), Audrey, Pierre-Etienne, Alban, Marie-Charlotte, Iris. D'autres personnes de passage, Elsa et Mathilde, merci pour les pauses thé et discussions méthodes d'organisation de travail, j'ai toute une liste de choses à tester maintenant que j'aurai un peu plus de temps devant moi.

Je voudrais aussi dire un grand merci à la bande du LEPSE : Maéva, Romain, Diane, Sandy, merci de m'avoir accueillie parmi vous (vive les projections Game of Thrones dans la salle de réunion), et pour tous ces moments de rigolade avec vous.

Pendant ma thèse, j'ai aussi eu la chance de passer quelques mois à Uppsala et je tiens à remercier Martin Lascoux et Sylvain Glémin pour m'avoir accueillie à EBC. Et également merci à Pascal et Marion, mon séjour n'aurait pas été aussi agréable sans vous, Tom et Barbie. Merci de m'avoir laissée squatter chez vous et pour ce super voyage en Laponie au milieu des aurores boréales et des rennes.

Sydney, je suis vraiment très heureuse de t'avoir rencontrée (et que tu aies été la collègue de ma maman). Et un grand merci pour t'être occupée de moi pendant la rédaction de mon manuscrit, prépare-toi pour l'oral je serai probablement pire ;)

Je ne peux pas ne pas mentionner mes copines de prépa : Lola, Marylou et bien sûr Mrs Robinson/Amandine, ou de SupAgro : Delphine, merci pour ce super voyage en Écosse, un jour on verra des puffins et des loutres pour de vrai, moi j'y crois ; Camille, Barbara, et Ségo.

Merci à mes compagnons d'APIMET : Benjamin, Alizée, Coralie, Adama, Jeannot, il va vraiment falloir se faire ce selfie avec les lions au Sénégal.

Alexia, toujours mon exemple et depuis si longtemps. On ne se voit pas aussi souvent que ce que je voudrais mais malgré la distance je sais que je peux compter sur toi à tout moment.

Sophie, la meilleure partenaire d'urban trail (team basilic forever) et de concerts (on n'en a plus de prévu d'ailleurs, ça ne va pas du tout…). Mes autres copines de la danse Sarah et Gabrielle. Et bien sûr, Christine, merci pour tes super cours de danse depuis toutes ces années. Je t'assure, tu arrives très bien à me changer les idées, que ce soit pendant le cours ou même plusieurs jours après avec les courbatures.

En parlant de courbatures, merci Mélodie de m'avoir fait découvrir mon transverse et plein d'autres muscles dont j'ignorais l'existence, un jour je décontracterai ma mâchoire.

Agathe, merci pour tous les cours de sport où tu m'as suivie (ou collée, j'hésite encore) depuis quatre ans et plus particulièrement de t'être occupée de moi pendant ces derniers mois difficiles. Ne t'inquiète pas, je n'oublie pas tes cours de yoga en illimité, prépare ta visio à Fontainebleau ;)

Et pour finir, je voudrais remercier ma famille qui est à mes côtés depuis 28 ans maintenant. Vous êtes avec moi dans les moments difficiles comme dans les bons moments. Je ne le montre pas forcément, mais vous m'aidez à aller de l'avant tous les jours. Merci à mes parents d'être mes plus grands fans et de me soutenir quelles que soient les circonstances. Clément, merci de m'avoir donné le meilleur filleul du monde, j'attends ma nièce avec impatience ☺ Théo, je sais que tu me demandes quand je ne suis pas là, merci d'être mon soutien silencieux.

J'espère qu'avec ma soutenance vous comprendrez un peu plus ce que j'ai fait pendant ces trois ans (je vous épargne la lecture du manuscrit si vous ne vous sentez pas de le lire), et sachez que j'ai tout à fait confiance en l'avenir parce que s'il y a quelque chose dont je suis sure, c'est que vous ferez la route avec moi ☺

Sommaire

INTRODUCTION

I. Démographie et diversité génétique des populations

Les processus démographiques régulent le nombre d'individus présents dans une population (taille démographique) et leurs mouvements d'une population à l'autre (migration). Ils influencent donc la manière dont la variabilité génétique est organisée dans et entre les populations. A l'inverse, la connaissance de la diversité génétique présente dans une population à un moment donné (la population a-t-elle beaucoup de diversité ?), son organisation (comment se distribue la diversité génétique entre les individus ?) ou sa variation sur un intervalle de temps permettent de faire des inférences sur le fonctionnement de la population et nous renseignent sur les processus démographiques en jeu. Parce qu'il détermine comment les allèles d'un individu sont transmis d'une génération à l'autre, le système de reproduction affecte lui aussi la diversité et la répartition des allèles dans et entre les individus. Il est donc nécessaire de le considérer lorsqu'on étudie une population. La génétique des populations fournit des attendus théoriques basés sur des modèles (généralement simplifiés) permettant de relier la diversité aux processus démographiques et sélectifs à l'œuvre dans les populations naturelles. La diversité génétique peut être considérée à plusieurs échelles : à l'échelle de l'espèce, l'échelle de temps évolutif, qui régit les relations phylogénétiques, ou bien à l'échelle de la population, en focalisant sur l'histoire ancestrale ou contemporaine. Les travaux présentés dans le cadre de ma thèse visent à décrire et estimer les processus contemporains (environ 20 générations) dans des populations locales (quelques centaines de mètres). Je présente ci-dessous les effets de la dérive génétique et de la migration sur la diversité génétique d'une population. Puis je considèrerai comment le régime de reproduction par autofécondation peut accentuer ces effets.

1. La dérive génétique

Le terme « dérive génétique » correspond aux fluctuations des fréquences alléliques attendues au sein d'une population de taille finie du fait de l'échantillonnage aléatoire des gamètes lors de la reproduction pour produire la génération suivante. Ces fluctuations sont d'autant plus importantes que l'échantillon (et donc la taille de la population) est petit. Elles peuvent conduire à la fixation ou à

la disparition d'allèles dans la population, ce qui augmente l'homozygotie et réduit la diversité génétique. Ce résultat est illustré dans la Figure 1 à l'aide de simulations de l'évolution de fréquences alléliques pour deux tailles démographiques contrastées. Crow et Kimura (1970) ont décrit analytiquement cette diminution de diversité au cours du temps : la diversité génétique présente au temps t dans une population, mesurée par l'indice de diversité de Nei (He_t, Nei 1973), diminue de $1/2N$ à chaque génération selon l'équation : $He_t = He_0(1 - \frac{1}{2N})^t$, où N est la taille démographique et He_0 est l'hétérozygotie initiale. En accord avec cet attendu, Frankham (1996) a montré, sur des données empiriques récoltées par une revue de la littérature, que la diversité génétique est corrélée à la taille des populations. Des niveaux de diversité réduits ont également été observés chez des espèces dont l'aire de répartition est limitée, reflétant de faibles effectifs (Hamrick et Godt 1990). Les fluctuations de taille démographique de type goulots d'étranglement, ou les effets de fondation, accentuent aussi la dérive et donc la diminution de diversité génétique. L'impact des évènements de fondation est notamment visible chez les populations insulaires, qui présentent des niveaux de diversité en moyenne 29% plus faibles que les populations des mêmes espèces localisées sur le continent (Frankham 1997). La force de la dérive génétique est estimée à travers le concept de taille efficace N_e (Encadré 1).

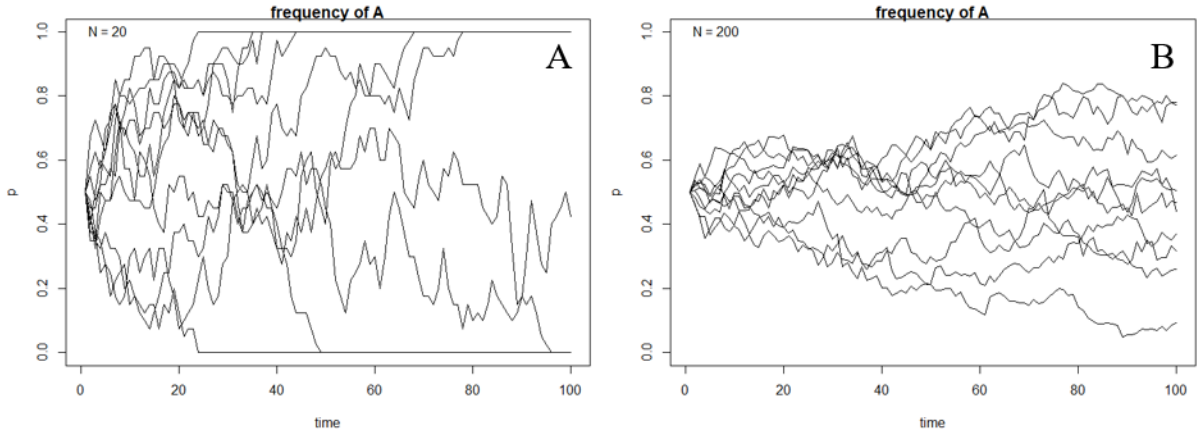


Figure 1 : Simulations de la dérive génétique à un locus bi-allélique neutre pour différentes tailles de population.
Chaque courbe correspond à l'évolution au cours du temps (en générations) de la fréquence p de l'allèle A à un locus bi-allélique dans une population panmictique de N individus sous le seul effet de la dérive. Pour chacune des simulations indépendantes, la fréquence initiale de l'allèle est 0.5. (A) Simulations avec N = 20 individus diploïdes. L'allèle A est soit fixé, soit perdu en 100 générations dans la majorité des simulations. (B) simulations avec N = 200 individus diploïdes. Les deux allèles A et a sont conservés après 100 générations. (Figures générées à l'aide du package R learnPopGen, Revell 2018).

Encadré 1 : Qu'est-ce que la taille efficace?

Le concept de la taille efficace a été introduit par Wright (1931, 1933) et est défini comme la taille qu'aurait une population si elle fonctionnait comme une population idéale de Wright-Fisher. Une population de Wright-Fisher est une population composée de N individus diploïdes, panmictique et avec des générations non chevauchantes. La taille de la population est constante au cours du temps et le nombre de descendants par individus est tiré dans une distribution de Poisson d'espérance 1. Ce modèle de population idéale constitue la base de nombreuses hypothèses en génétique des populations.

La taille efficace d'une population réelle (N_e), a été définie comme la taille d'une population idéale de Wright-Fisher qui subirait la même intensité de dérive génétique que la population considérée (Crow et Kimura 1970). Selon l'aspect de la dérive génétique considéré, plusieurs tailles efficaces ont été définies et les plus couramment utilisées sont :

● « Inbreeding effective size, N_{eI} » : la taille d'une population idéale avec la même probabilité que des allèles échantillonnés au hasard dans le pool de gamètes soient identiques-par-descendance que dans la population considérée. N_{eI} permet de prédire la diminution d'hétérozygotie (ou l'augmentation d'homozygotie).

● « Variance effective size, N_{eV} » : la taille d'une population idéale avec la même variance des fréquences alléliques entre génération que la population réelle considérée.

Ces deux tailles efficaces sont identiques dans le cas d'une population à l'équilibre de Wright-Fisher, mais elles peuvent être différentes dans certaines situations démographiques plus complexes (par exemple population structurée, fluctuations de taille démographique au cours du temps). L'estimation de la taille efficace permet de quantifier l'importance de plusieurs processus évolutifs : $N_e\mu$, où μ est le taux de mutation, traduit le niveau de diversité génétique neutre attendu à l'équilibre mutation-dérive (Kimura 1983) ; le produit N_es, où s est le coefficient de sélection, détermine l'efficacité de la sélection (Kimura 1983) ; N_em, avec m le taux de migration, traduit l'importance de la migration (Wright 1931). De nombreux facteurs, comme par exemple le sexe- ratio, le système de reproduction, la sélection ou les fluctuations de taille démographique, peuvent affecter la taille efficace d'une population (décrits dans Charlesworth 2009), et ils doivent donc être pris en compte lors de l'estimation de N_e.

Echelles de temps

La taille efficace peut être considérée à différentes échelles de temps. La théorie de la coalescence utilise le concept de taille efficace pour estimer la probabilité de coalescence au cours du temps jusqu'à l'ancêtre commun le plus récent. Dans ce contexte, on peut adresser des questions à l'échelle de l'espèce et du temps évolutif (variations historiques de la taille efficace, évènements d'admixture, etc.). A l'autre extrême, la taille efficace reflétant la variance des fréquences alléliques au cours du temps, elle permet aussi de décrire la force de la dérive dans des populations et ses éventuelles variations contemporaines (si on dispose de plusieurs intervalles de temps).

2. La migration

Le concept de population est central en génétique des populations, mais parfois difficile à définir clairement (voir Waples et Gaggiotti 2006 pour une revue). La plupart des espèces sont distribuées de manière non continue, soit du fait de leur comportement ou parce que les habitats favorables sont fragmentés. Cette subdivision en populations et la différenciation entre elles sont accentuées par la reproduction non aléatoire des individus, qui se reproduisent préférentiellement au sein d'une population. Au contraire, la différenciation est atténuée par la migration, qui tend à homogénéiser les populations (Wright 1931). Chez les plantes, la migration peut se produire via le pollen (migration haploïde), ou via les graines (migration diploïde). Plusieurs modèles théoriques de migration ont été définis, parmi lesquels le modèle « stepping stone » et le modèle de « migration en îles » sont les plus utilisés (Wright 1943). Dans le modèle en îles, on considère une population divisée en n sous-populations (les îles) de taille constante N. Dans chaque sous-populations, les individus se reproduisent en panmixie et une proportion m de migrants est échangée avec les n îles. Ainsi, les taux de migration entre les îles sont symétriques (Figure 2). En supposant n grand, à l'équilibre migration-dérive, on attend une différenciation entre îles inversement proportionnelle à la taille des îles et au taux de migration, approximativement égale à $F_{ST} = \frac{1}{1+4Nm}$ (Wright 1943).

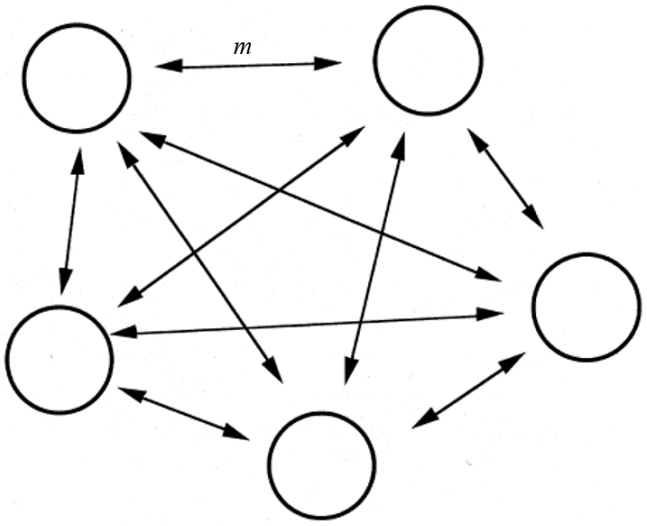


Figure 2 : Modèle de migration en îles avec *n* = 5 dèmes de taille *N* et un taux de migration *m*.

La migration interagit avec la dérive génétique pour façonner la diversité d'une population. Elle augmente la diversité génétique de chaque dème en apportant de nouveaux individus différenciés, et contre ainsi les effets de la dérive locale. A une échelle plus large, les effets de la migration sur la diversité peuvent être complexifiés par des évènements démographiques. En particulier, le modèle en îles peut être étendu au concept de métapopulation (Hanski 1998), dans lequel on considère des extinctions récurrentes des dèmes et leur recolonisation par migration depuis un ou plusieurs autres dèmes. Ce renouvellement récurrent des dèmes réduit la diversité de la métapopulation totale ainsi que la diversité intra-dème car les extinction-recolonisations réduisent la taille efficace des dèmes (Whitlock et Barton 1997 ; Wang et Caballero 1999).

II. Les systèmes de reproduction chez les plantes

Les modèles classiques de génétique des populations pour étudier la dérive et la migration supposent la panmixie. Pourtant, les populations sont rarement panmictiques, en particulier quand elles ont un système de reproduction qui inclut de l'autofécondation.

1. Diversité des systèmes de reproduction chez les plantes

L'étude des systèmes de reproduction et de leur évolution constitue un axe de recherche majeur en biologie évolutive. En effet, les plantes à fleurs présentent une diversité de systèmes de reproduction remarquable, visible à travers la grande diversité des morphologies florales. D'un côté, les espèces portant des fleurs mâles et des fleurs femelles sur des individus différents sont qualifiées de dioïques (4% des plantes à fleurs, Richards 1997). Les espèces monoïques portent des fleurs mâles et des fleurs femelles séparées mais sur le même individu (5% des plantes à fleurs, Richards 1997). Enfin, à l'autre extrême, les espèces hermaphrodites portent des fleurs qui regroupent à la fois les organes reproducteurs mâles et femelles sur les mêmes individus. Ces espèces hermaphrodites sont les plus fréquentes (72% des plantes à fleurs, Richards 1997). Il existe également des régimes intermédiaires plus rares comme la gynodioécie (coexistence d'individus à fleurs hermaphrodites et d'individus à fleurs femelles), la gynomonoécie (coexistence de fleurs hermaphrodites et de fleurs femelles sur un même individu), ou la trioécie (coexistence d'individus hermaphrodites, mâles et femelles dans la population). La particularité principale de l'hermaphrodisme est que ce système de reproduction permet en théorie l'autofécondation, au contraire de la dioécie qui impose l'allogamie.

L'autofécondation n'est cependant pas automatique chez les hermaphrodites. En effet, de nombreux mécanismes permettent d'éviter l'autofécondation. Ils impliquent notamment des systèmes d'auto-incompatibilité moléculaires (Allen et Hiscock 2008), de séparation spatiale des organes mâles et femelles au sein de la fleur (herkogamie, Webb et Lloyd 1986), ou de décalage temporel de la maturation des organes reproducteurs (dichogamie, Lloyd et Webb 1986). Les taux d'autofécondation (σ) sont variables entre les espèces et des études ont montré une distribution continue et légèrement bimodale avec un pic d'allogamie stricte ($\sigma = 0$) et un pic d'autogamie quasi-complète ($\sigma \sim 1$; Figure 3 ; Schemske et Lande 1985 ; Goodwillie et al. 2005).

2. Evolution des systèmes de reproduction

De nombreux modèles théoriques ont été proposés afin d'expliquer l'évolution des taux d'autofécondation et considèrent des facteurs génétiques (par exemple Lloyd 1979 ; Lande et Schemske 1985), et/ou des facteurs écologiques (par exemple Porcher et Lande 2005 ; Johnston *et al.* 2009). Les modèles génétiques originaux se basent sur l'opposition entre l'avantage de transmission de l'autofécondation (Fisher 1941) qui postule qu'une plante autogame transmettra son patrimoine génétique 50% plus souvent qu'une allogame, et la dépression de consanguinité qui est la réduction relative de la valeur sélective des individus issus d'autofécondation par rapport à ceux issus d'allofécondation. Ces modèles prédisent que le taux d'autofécondation devrait évoluer vers 0 lorsque la dépression de consanguinité est supérieure à 0.5, et vers 1 si la dépression de consanguinité est inférieure à 0.5 (Lande et Schemske 1985). Cependant, ils n'expliquent pas la forte proportion de taux d'autofécondation intermédiaires observés dans les populations. Des modèles génétiques plus complexes ont été développés, et prédisent la stabilité des régimes mixtes dans certaines conditions (Uyenoyama 1986). D'autres modèles prennent en compte l'écologie de la pollinisation et son effet sur l'évolution du taux d'autofécondation. En particulier, ces modèles considèrent l'effet du « *pollen discounting* » (Holsinger *et al.* 1984), la réduction de pollen exporté due à l'utilisation de pollen pour l'autofécondation, ainsi que l'effet du « *seed discounting* » (Lloyd 1992), la réduction du nombre d'ovules allofécondés due à l'autofécondation. Ces modèles prédisent que le *pollen discounting* peut réduire l'avantage de transmission de l'autofécondation lorsque les quantités de pollen produit par les plantes autogames sont d'autant plus limitées que le taux d'autofécondation est fort (Lloyd 1992 ; Johnston 1998).

En lien avec les prédictions de ces modèles, on définit en général trois grandes catégories pour le régime de reproduction : le régime allogame ($0 < \sigma < 20\%$), le régime mixte ($20\% < \sigma < 80\%$), et le régime autogame ($80\% < \sigma < 100\%$, Figure 3). Au cours de mon travail de thèse, je me suis intéressée plus particulièrement aux espèces ayant un régime dit autogame et qui représentent 10 à 15% des plantes à fleurs.

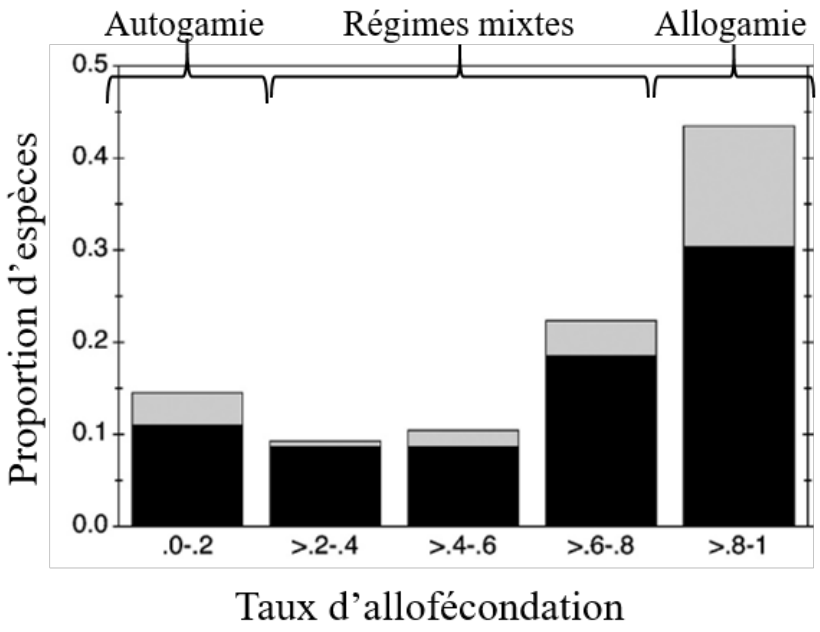


Figure 3 : Distribution des taux d'allofécondation estimés chez 345 espèces de plantes à fleurs. Les espèces pollinisées par des vecteurs abiotiques (eau, vent) sont représentées en gris, les espèces pollinisées par des vecteurs biotiques (insectes, oiseaux, chauves-souris) sont représentées en noir. Figure adaptée de Goodwillie *et al.* (2005).

3. Variabilité du taux d'autofécondation et ses déterminants

Afin de confronter les données biologiques à leur modèle théorique, Schemske et Lande (1985) ont fait une revue des taux d'autofécondation chez différentes espèces et ont reporté des taux variables d'une espèce à l'autre, mais aussi entre populations d'une même espèce (Figure 4). Plus récemment, Whitehead *et al.* (2018) ont effectué une revue des taux d'autofécondation estimés pour 741 populations provenant de 105 espèces et ont également montré d'importantes variations de taux d'autofécondation entre populations. Cette variabilité est particulièrement importante chez les espèces à régime mixte, mais de la variabilité est également observée chez les espèces majoritairement autogames ou allogames. Ils recommandent donc d'étudier plusieurs populations, préférentiellement avec plusieurs mesures dans le temps, afin de caractériser le régime de reproduction d'une espèce.

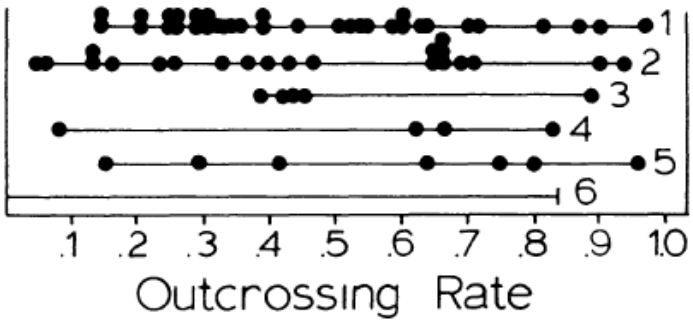


Figure 4 : Variation du taux d'allofécondation entre populations de six espèces.
Les points correspondent aux taux d'autofécondation d'une population. 1 = *Lupinus succulentus*, 2 = *L. nanus*, 3 = *Clarkia exilis*, 4 = *C. tembloriensis*, 5 = *Gillia achilleifolia*, 6 = *Collinsia sparsiflora* var. *arvensis*. Figure issue de Schemske et Lande (1985).

Cette variabilité peut être due à des facteurs environnementaux ou génétiques (détaillés par Barrett et Eckert 1990). Les facteurs environnementaux affectant les proportions d'auto et d'allo-pollen se déposant sur les stigmates sont notamment la densité de la population (Karron *et al.* 1995), ou le comportement des pollinisateurs (voir la revue de Devaux *et al.* 2014). Des traits floraux qui peuvent être sous contrôle génétique peuvent également entrainer des variations du taux d'autofécondation, comme par exemple l'herkogamie (la séparation spatiale du stigmate et des étamines) qui est associée à une diminution de l'autofécondation (Herlihy and Eckert 2007) ou la protandrie (décalage temporel entre la réceptivité des stigmates et la libération du pollen).

III. Conséquences génétiques de l'autofécondation

1. Autofécondation et diversité génétique neutre

L'autogamie a deux effets directs : elle diminue le nombre de gamètes indépendants échantillonnés au moment de la reproduction et augmente l'homozygotie sur l'ensemble du génome (Figure 5, Caballero et Hill 1992). En effet, dans une population se reproduisant exclusivement par autofécondation, la proportion d'hétérozygotes H_{obs_t} à un locus donné est divisée par deux à chaque génération : $H_{obs_t} = H_{obs_0} (1/2)^t$ (Crow and Kimura 1970). Ces deux effets conduisent à une réduction de la taille efficace N_e selon l'équation $N_e = 1/(1 + F_{IS})$ (Pollak 1987), où F_{IS} est l'écart à la panmixie (c'est-à-dire le déficit en hétérozygotes, $F_{IS} = 1 - \frac{H_e}{H_{obs}}$). Ainsi, dans le cas extrême d'une population complètement autogame (σ = 1 et donc F_{IS} = 1), la taille efficace est divisée par deux par rapport à une population allogame. On s'attend donc à une augmentation de la dérive génétique dans les populations autogames et à une diversité génétique réduite.

Une autre conséquence de l'augmentation d'homozygotie en régime autogame est une réduction de la recombinaison efficace (Figure 5, Golding et Strobeck 1980 ; Nordborg 2000). En effet, si les locus sont homozygotes, la recombinaison ne change pas les haplotypes. Des associations entre locus (autrement dit du déséquilibre de liaison) se créent alors et ne peuvent pas être cassées par l'autofécondation. Cet effet se voit très clairement lorsqu'on compare l'étendue du déséquilibre de liaison chez *Arabidopsis thaliana*, une plante fortement autogame, et chez le maïs qui est allogame. On détecte du déséquilibre de liaison sur plus de 100 kb chez *A. thaliana* (Nordborg *et al.* 2002), alors qu'il est limité à quelques centaines de nucléotides chez le maïs (Tenaillon *et al.* 2001). De manière générale, on observe de plus importants déséquilibres de liaison dans les populations de plantes autogames que dans les populations allogames (Brown 1979 ; Glémin *et al.* 2006). Si la recombinaison est moins efficace, on s'attend à des phénomènes d'auto-stop génétique (Maynard Smith et Haigh 1974) lorsque la sélection sur un allèle favorable (balayage sélectif) ou la contre-sélection d'une mutation délétère (sélection d'arrière-plan) vont entraîner les allèles neutres liés. Ces deux effets de la sélection appauvrissent encore la diversité génétique neutre (Barton 2000), en particulier chez les populations autogames.

Contrairement au cas de l'allogamie, un individu autogame seul peut fonder une nouvelle population. Les évènements de fondation peuvent donc avoir des effets très importants chez les populations autogames (voir la loi de Baker, Baker 1967). Comme détaillé dans le paragraphe

introductif sur la dérive génétique, les effets de fondation ainsi que les goulots d'étranglement, ou encore les extinctions/recolonisations qui sont supposées plus fréquentes en régime autogame (Glémin 2007), résultent en une dérive génétique accrue (Whitlock et Barton 1997). Ingvarsson (2002) s'est appuyé sur ces attendus et a proposé que les dynamiques de métapopulation seraient une cause supplémentaire de réduction de la diversité génétique chez les autogames.

L'ensemble des effets de l'autofécondation sur la taille efficace peut se résumer avec l'équation suivante : $N_e = \frac{\alpha N}{1+F}$ (Glémin 2007), où N est la taille démographique (*census size*), $F = \frac{\sigma}{2-\sigma}$ est l'indice de fixation de Wright (1931), et α représente la diminution de la taille efficace due à la démographie et à l'auto-stop génétique ($\alpha \in [0; 1]$). On s'attend à ce que cette réduction de taille efficace entraîne une importante perte de diversité génétique chez les espèces autogames par rapport aux allogames. Ceci a effectivement été vérifié sur des données empiriques (Schoen et Brown 1991 ; Hamrick et Godt 1997 ; Glémin *et al.* 2006).

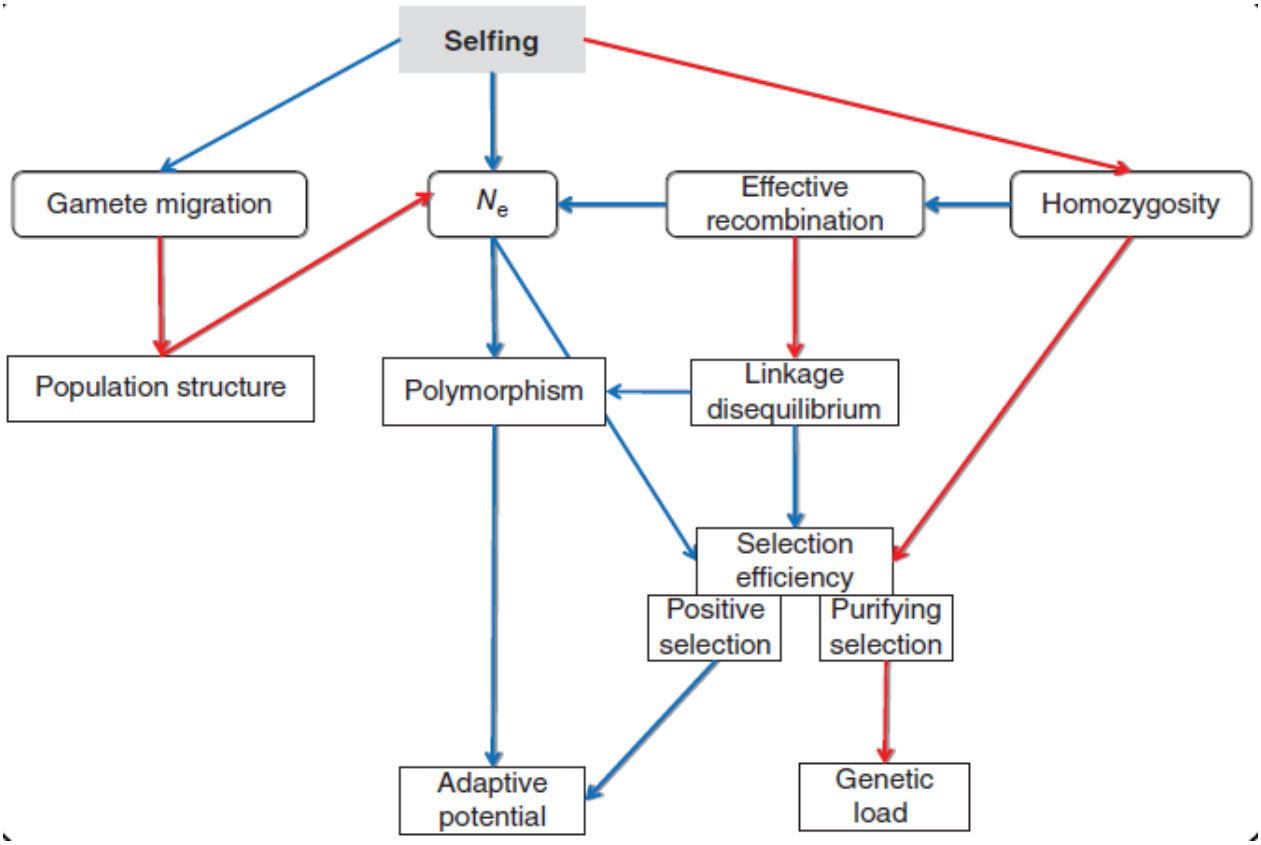


Figure 5 : Conséquences de l'autofécondation.
Les flèches bleues représentent un effet de diminution, les rouges d'augmentation. Figure issue de Burgarella et Glémin (2017).

2. L'autofécondation, un cul-de-sac évolutif ?

En s'appuyant sur cette prédiction d'une diversité génétique réduite chez les espèces autogames, Stebbins (1957) a proposé que ces espèces auraient un faible potentiel adaptatif, limitant leur capacité à répondre à des changements environnementaux et conduisant à des taux d'extinction plus forts chez les lignées autofécondantes que chez les lignées allogames. Il conclue que l'autogamie est un « cul-de-sac évolutif ». L'hypothèse du « cul-de-sac évolutif » permettrait d'expliquer que les espèces autogames sont rares, alors que les transitions de régime de reproduction semblent se faire principalement de l'allogamie vers l'autogamie (Barrett 2002). Goldberg *et al.* (2010) ont montré que chez les Solanacées, le taux de diversification net est plus élevé dans les lignées auto-incompatibles et que les lignées auto-compatibles (qui sont susceptibles de faire de l'autofécondation) ont effectivement plus tendance à s'éteindre. Toutefois, les démonstrations expérimentales des effets négatifs de l'autofécondation sur le potentiel adaptatif à long terme sont rares (voir Noël *et al.* 2017 pour un exemple chez un escargot autogame). Quantifier la diversité chez les populations autogames et sa dynamique reste donc une question centrale en Évolution et nécessite à la fois le développement d'attendus théoriques (pour la diversité multilocus) et des comparaisons avec des données empiriques.

IV. Structure de la diversité des populations autogames en milieu naturel

Les différents effets de l'autofécondation présentés ci-dessus résultent en des structures de population particulières que je vais maintenant décrire.

1. Variabilité de la diversité entre populations locales

Schoen et Brown (1991) ont comparé la diversité génétique de populations autogames et de populations allogames à partir de données sur des allozymes. Ils ont mis en évidence une diversité réduite chez les autogames, mais également des variations entre populations plus importantes que chez les plantes allogames. Le tableau 1 est issu d'une revue des données disponibles dans la littérature sur la diversité génétique mesurée à l'aide de marqueurs moléculaires dans des populations naturelles d'espèces très autogames. On observe que la diversité génétique, mesurée par l'hétérozygotie selon Nei est effectivement très variable entre populations et au cours du temps (études temporelles), surtout lorsque les marqueurs utilisés sont des microsatellites, qui peuvent être hautement polymorphes (Tableau 1).

Tableau 1 : Hétérozygotie selon Nei estimée dans des populations naturelles d'espèces autogames.

Dans le cas d'études temporelles, la gamme de valeurs fournie inclue les valeurs estimées dans chaque population à chaque échantillon temporel.

Référence	Espèce	Type de marqueurs	Nombre de populations	H_e
Abbott et Gomes (1989)	*Arabidopsis thaliana*	Allozymes	1	0.1483
Kuittinen *et al.* (1997)	*Arabidopsis thaliana*	Microsatellites	6	0 - 0.51
Stenøien *et al.* (2005)	*Arabidopsis thaliana*	Microsatellites	10	0.01 - 0.21
Bakker *et al.* (2006)	*Arabidopsis thaliana*	Microsatellites		0.042 - 0.946
Picó *et al.* (2008)	*Arabidopsis thaliana*	microsatellites & SNP	7	0.13 - 0.28
Lundemo *et al.* (2009) (étude temporelle)	*Arabidopsis thaliana*	SNP	27	0 - 0.17
Montesinos *et al.* (2009)	*Arabidopsis thaliana*	SNP	10	0 - 0.26
Bomblies *et al.* (2010)	*Arabidopsis thaliana*	SNP	77	0 - 0.318
Gomaa *et al.* (2011) (étude temporelle)	*Arabidopsis thaliana*	SNP	9	0 - 0.28
Falahati-Anbaran *et al.* (2014) (étude temporelle)	*Arabidopsis thaliana*	SNP	10	0 - 0.12
Bonnin *et al.* (1996)	*Medicago truncatula*	RAPD	4	0.096 - 0.242
Bonnin *et al.* (2001)	*Medicago truncatula*	Microsatellites	1 (divisée en 3 sous-populations)	0.348 - 0.476
Siol *et al.* (2007) (étude temporelle)	*Medicago truncatula*	Microsatellites	2	0.18 - 0.47
Siol *et al.* (2008)	*Medicago truncatula*	Microsatellites	1	0.259
Chauvet *et al.* (2004)	*Mycelis muralis*	Microsatellites	17	0.24 - 0.68

2. Une diversité structurée en génotypes multilocus

Dans les populations autogames, la diversité est organisée de façon très particulière, avec une structure en génotypes multilocus (MLG) : on observe généralement quelques MLGs fortement homozygotes et en forte fréquence et plusieurs MLG uniques (par exemple Gomaa *et al.* 2011). Certaines populations peuvent être complètement monomorphes et ne contenir qu'un seul MLG (par exemple voir Todokoro *et al.* 1995 ; Kuittinen *et al.* 1997 ; Stenøien *et al.* 2005). De plus, les génotypes multilocus au sein d'une population sont soit très similaires entre eux, soit très différents (Bakker *et al.* 2006 ; Bomblies *et al.* 2010, Figure 6). Cette distribution bimodale des distances génétiques entre individus traduit l'accumulation de différences entre groupes de MLGs qui s'autoféconden t et peut s'avérer informative sur le fonctionnement des populations, par exemple sur les effets de la recombinaison et de la migration (qui introduirait des génotypes fortement différenciés).

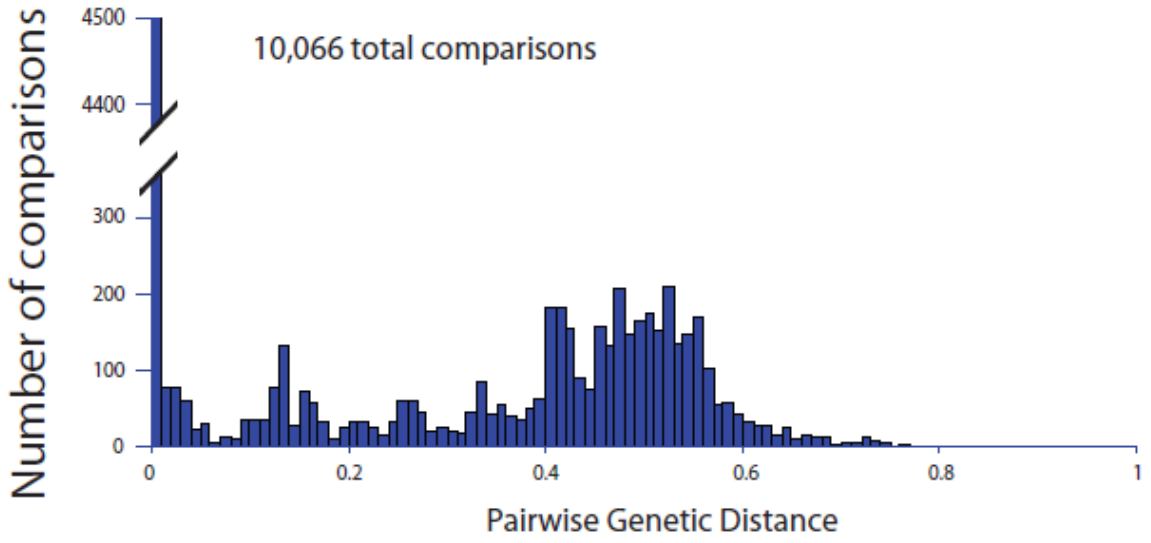


Figure 6 : Distribution des distances génétiques entre paires d'individus au sein de 77 populations d'*Arabidopsis thaliana*.
Figure issue de Bomblies *et al.* (2010).

L'analyse de cette composition en MLG au cours du temps a montré des patrons variables : certaines populations conservent des MLGs (parfois sur des longues périodes, voir le chapitre 2), et au contraire d'autres populations se renouvellent entièrement (Bomblies *et al.* 2010 ; Gomaa *et al.* 2011). Certains auteurs ont invoqué de forts taux de migration dans des métapopulations pour expliquer les niveaux de diversité génétique importants parfois observés (Chauvet *et al.* 2004). Dans d'autres cas, des évènements récurrents d'extinction-recolonisation pourraient expliquer des diversités génétiques faibles (Bergelson *et al.* 1998 ; Innan and Stephan 2000). Globalement, ces populations fortement autogames semblent souvent fonctionner en métapopulation (Ingvarsson

2002). Elles sont d'ailleurs également fortement structurées au niveau spatial, avec les génotypes multilocus répétés agrégés à une échelle très locale (Bonnin *et al.* 1996 ; Bomblies *et al.* 2010).

Les populations autogames ont donc une structure génétique multilocus très caractéristique où la diversité génétique est organisée au sein de « lignées » autogames. Du fait de cette non-indépendance entre les locus, les indices de diversité monolocus sont peu adaptés à la description de ces populations. De fait, des méthodes basées sur la diversité multilocus sont disponibles mais, contrairement à la diversité monolocus, peu d'attendus analytiques sont disponibles.

3. Effets de la recombinaison

Contrairement à l'allogamie, l'autogamie stricte n'est jamais observée en populations naturelles. En effet, un résidu d'allofécondation est régulièrement observé (Bonnin *et al.* 2001 ; Bomblies *et al.* 2010). Une des conséquences de l'allofécondation résiduelle est que les populations autogames sont composées de deux « compartiments » : un compartiment majoritairement homozygote, issu de nombreuses générations d'autofécondation, et un compartiment issu d'évènements d'allofécondation résiduelle récents, composé d'individus d'hétérozygotie variable (Bomblies *et al.* 2010). Cette organisation particulière peut être utilisée pour déterminer des classes d'autofécondation récente à partir de l'hétérozygotie multilocus individuelle (Enjalbert et David 2000). En effet, le nombre de locus hétérozygotes d'un individu dépend du nombre de générations d'autofécondation dont il est issu (hétérozygotie divisée par deux à chaque génération d'autofécondation) et de la diversité génétique de la population considérée (Figure 7). Ainsi, après plusieurs générations d'autofécondation, l'homozygotie est restaurée et il se forme des lignées recombinantes entre les deux MLGs impliqués dans l'évènement d'allofécondation. La méthode développée par Enjalbert et David (2000) permet donc d'estimer un taux d'autofécondation moyen sur quelques générations et également de détecter des variations au cours du temps.

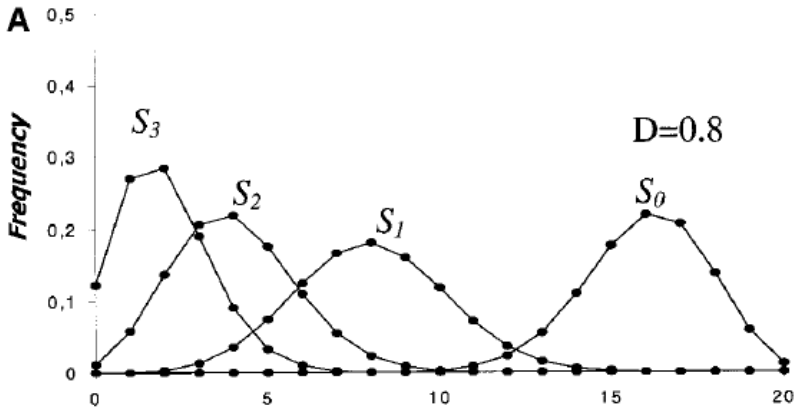


Figure 7 : Nombre de locus hétérozygotes attendus chez des individus de classe d'autofécondation S_0, S_1, S_2 et S_3.
D est la diversité selon Nei, S_i est la classe d'individus qui ont *i* générations d'autofécondation depuis le dernier évènement d'allofécondation dans leur généalogie. Figure issue de Enjalbert et David (2000).

Par ailleurs, les familles de lignées recombinantes potentielles peuvent être identifiées à partir des distances génétiques entre génotypes parents et recombinants. Si on note $d_{p_1 - p_2}$ la distance génétique entre les deux MLGs parentaux p_1 et p_2, $d_{p_1 - hybride}$ et $d_{p_2 - hybride}$ les distances génétiques entre chaque parent p_1 ou p_2 et le recombinant potentiel « $hybride$ », alors « $hybride$ » peut être issu de l'allofécondation entre p_1 et p_2 si $\frac{d_{p_1 - hybride} + d_{p_2 - hybride}}{2} = \frac{d_{p_1 - p_2}}{2}$. Cette méthode est implémentée dans le programme genetHaplo (développé par V. Ranwez) et permet d'identifier des lignées recombinantes au sein des populations. Des études en populations naturelles comparent quant à elles les génotypes sur des fragments de chromosomes pour identifier des recombinants potentiels (Bomblies *et al.* 2010).

V. Méthodes d'estimation de paramètres démographiques

Au cours de ma thèse, j'ai cherché notamment à caractériser finement des populations naturelles d'une espèce très autogame afin de comprendre leur histoire démographique récente et de déterminer comment les processus démographiques façonnent la diversité des populations lorsqu'elles se reproduisent préférentiellement par autogamie. Pour cela, j'ai eu à estimer différents paramètres tels que le taux d'autofécondation et la taille efficace des populations. Je présente ici une revue des méthodes existantes et des hypothèses sous-jacentes à l'estimation de ces deux paramètres.

1. Estimation du taux d'autofécondation σ

Une étape indispensable pour caractériser le système de reproduction d'un organisme repose sur l'estimation de son taux d'autofécondation. Le développement de marqueurs moléculaires a permis de mettre au point différentes techniques d'estimation du taux d'autofécondation à partir de données génétiques. Je présente ici trois grandes catégories de méthodes d'estimation.

a) Approche monolocus : à partir du déficit en hétérozygotes

Comme décrit précédemment, l'autogamie a pour effet d'augmenter l'homozygotie ce qui génère un déficit en hétérozygotes (F_{IS}) comparé à une population qui se reproduit en panmixie. Le taux d'autofécondation σ peut ainsi être estimé de manière indirecte à l'aide du déficit en hétérozygotes selon l'équation : $F_{IS} = \frac{\sigma}{(2-\sigma)}$ (Hartl et Clark 1998). Cette relation suppose une population diploïde de taille infinie et à l'équilibre, et dans laquelle la seule source de consanguinité est l'autofécondation. De plus, cela suppose que l'allofécondation est distribuée de manière uniforme entre les individus et qu'il n'y a pas de sélection. Le taux d'autofécondation ainsi estimé reflète la consanguinité accumulée sous l'effet de l'autofécondation sur plusieurs générations. Cependant, les croisements entre individus apparentés (*biparental inbreeding*) réduisent aussi l'hétérozygotie ; si on néglige ces croisements, la méthode surestime le taux d'autofécondation par rapport à sa valeur réelle. Une autre source d'erreur de cette méthode provient des erreurs de génotypage affectant le taux d'hétérozygotie : les allèles nuls (dus à une mutation dans la séquence de l'amorce, qui empêche l'amplification PCR) ou les erreurs de lecture peuvent par exemple augmenter artificiellement l'hétérozygotie (Taberlet *et al.* 1999). Jarne et David (2008) ont montré que la précision de l'estimation de σ est améliorée en augmentant le nombre d'individus échantillonnés et recommandent d'utiliser des locus non liés pour limiter les biais.

Malgré les erreurs d'estimation possibles, il reste très facile d'estimer un taux d'autofécondation à l'aide du F_{IS}, et cette méthode est très régulièrement utilisée en génétique des populations (Jarne et Auld 2006).

b) Approche multilocus : à partir du déséquilibre d'identité

Il est également possible d'exploiter les associations entre locus pour estimer σ. David *et al.* (2007) ont proposé une méthode basée sur la distribution multilocus de l'hétérozygotie. En effet, la consanguinité crée du déséquilibre d'identité, c'est-à-dire des corrélations d'hétérozygotie par paires de locus (Weir et Cockerham 1973). Lorsqu'il y a du déséquilibre d'identité dans une population, si un individu est hétérozygote à un locus, il y a une plus forte probabilité qu'il soit hétérozygote à d'autres locus. Un évènement d'allofécondation récent favorise le déséquilibre d'identité (par exemple chez les individus F1 où tous les locus différenciés entre les parents seront à l'état hétérozygote). Le déséquilibre d'identité g_2 traduit l'excès de génotypes hétérozygotes à deux locus. En supposant l'équilibre de consanguinité et l'équilibre de liaison, on a l'équation : $g_2 = \frac{\sigma}{(4-\sigma)(1-\sigma)}$ (David *et al.* 2007).

L'avantage de cette méthode d'estimation par rapport à celle du F_{IS} est qu'elle n'est pas sensible aux erreurs de génotypage car celles-ci se produisent de manière indépendante d'un locus à l'autre (David *et al.* 2007). Cette méthode ne résout cependant pas les problèmes de surestimation du taux d'autofécondation dans le cas de croisements entre individus apparentés.

De manière similaire, Enjalbert et David (2000) proposent une méthode d'estimation des taux d'autofécondation récents basée sur l'hétérozygotie multilocus individuelle (détaillé en IV.3).

c) Estimation directe à partir de descendances maternelles

Le taux d'autofécondation peut également être estimé en comparant les génotypes d'une mère à ceux de ses descendants, ce qui constitue l'approche dite « sur descendances maternelles ». Cette approche a été initialement proposée par Jones (1916) à l'aide de marqueurs morphologiques. Il a intercalé des plants de tomates naines (trait récessif) avec des plants de tomates de taille normale (trait dominant) et a compté le nombre d'individus de taille standard dans la descendance des plantes naines, ce qui lui a donné le nombre de descendants issus d'allofécondation. Cette méthode a ensuite été améliorée avec l'utilisation des marqueurs moléculaires en considérant le fait qu'un descendant A_1A_2 dont la mère est A_1A_1 est forcément issu d'allofécondation (en négligeant la mutation, Tableau 2). La fréquence de ce type de descendants donne une estimation du taux d'allofécondation (1-σ). Contrairement aux deux méthodes décrites précédemment, l'utilisation de descendances maternelles fournit une estimation directe du taux d'autofécondation à une génération donnée (« *here and now*

estimate »). Shaw *et al.* (1981) ont introduit l'utilisation des informations multilocus pour améliorer la précision des estimations. Ainsi, en incluant un plus grand nombre de locus, on augmente la probabilité de détecter des évènements d'allofécondation, même entre individus fortement apparentés (Tableau 2). Par ailleurs, la comparaison des estimations monolocus et multilocus ($t_s - t_m$) permet d'estimer la contribution des croisements entre individus apparentés. En effet, les estimations monolocus ne permettent pas de différencier une autofécondation d'un croisement entre individus apparentés (et donc similaires), ce qui résulte en une surestimation du taux d'autofécondation. L'estimation des croisements entre individus apparentés peut ainsi donner des informations sur la structure de la population considérée.

Tableau 2 : Exemple de jeu de données de descendances maternelles.
Les génotypes en gras permettent de discriminer entre auto- et allofécondation, la dernière colonne indique le statut inféré du descendant. Le descendant 1 serait considéré comme étant issu d'autofécondation si l'on ne considérait que les locus A à D. Le locus E permet de rejeter l'hypothèse d'autofécondation en négligeant la mutation. (D'après Shaw *et al.* 1981)

Locus	A	B	C	D	E	Statut
Mère	A_1A_1	B_2B_2	C_1C_2	D_1D_3	E_2E_3	
Descendant 1	A_1A_1	B_2B_2	C_1C_2	D_1D_3	$\mathbf{E_1E_3}$	Allofécondé
Descendant 2	A_1A_1	B_2B_2	C_2C_2	D_1D_1	E_3E_3	Autofécondé
Descendant 3	A_1A_1	$\mathbf{B_1B_2}$	C_1C_2	$\mathbf{D_2D_3}$	$\mathbf{E_1E_3}$	Allofécondé

L'utilisation de descendances maternelles donne également des informations sur la variance du taux d'autofécondation intra-famille (entre tous les descendants d'une même mère, Ritland 2002). Cela permet donc de caractériser finement les régimes de reproduction des populations et en particulier d'analyser les taux d'autofécondation à différentes échelles : échelle du génotype, de la plante ou de l'inflorescence. Une telle analyse caractérisant le système de reproduction d'une plante majoritairement autogame a été réalisée au cours de cette thèse et est présentée au Chapitre 1.

Cette méthode d'estimation directe est régulièrement utilisée chez les plantes (Goodwillie *et al.* 2005 ; Whitehead *et al.* 2018) mais moins souvent chez les animaux. En effet, elle requiert un effort d'échantillonnage important car il faut échantillonner des familles entières, ce qui peut se révéler difficile chez les animaux car ces derniers se dispersent.

2. Estimation de la taille efficace N_e

De nombreuses méthodes d'estimation de la taille efficace existent et sont implémentées par un nombre encore plus important de programmes ou logiciels. Je me limite ici aux méthodes d'estimation de la taille efficace contemporaine et les plus couramment utilisées, mais des méthodes existent également pour estimer la taille efficace ancestrale d'une population (voir Wang 2005 pour une revue).

a) Méthode temporelle (plusieurs échantillons)

En l'absence de sélection, mutation et migration, les changements de fréquences alléliques au cours du temps dans une population de taille finie sont dus uniquement à la dérive génétique (comme illustré dans la Figure 1). On peut donc utiliser les changements de fréquences alléliques entre générations pour estimer la force de la dérive, autrement dit la taille efficace de la population considérée, N_e (encadré 1). Cela constitue le principe de base de la méthode temporelle pour estimer N_e, initialement proposée par Krimbas et Tsakas (1971). Cette méthode permet d'estimer un N_e moyen sur une période de temps t à partir de la variance des fréquences alléliques, $\hat{F} = \frac{1}{k}\sum_{i=1}^{k}\frac{(x_i-y_i)^2}{\frac{x_i+y_i}{2}-x_iy_i}$ (Nei et Tajima 1981), où x_i et y_i sont les fréquences de l'allèle i à un locus avec k allèles dans le premier et le deuxième échantillon temporel respectivement. En supposant que la taille démographique N est largement supérieure à N_e, une estimation de la taille efficace est alors : $\widehat{N_e} = \frac{t}{2(\hat{F}-\frac{1}{2S_0}-\frac{1}{2S_t})}$ (Waples 1989), où t est l'intervalle de temps entre les échantillons temporels, S_0 et S_t sont les tailles du premier et deuxième échantillon respectivement. L'introduction de S_0 et S_t dans le calcul de N_e permet de prendre en compte l'effet d'échantillonnage dans les fréquences alléliques de la population. Waples (1989) a montré à l'aide de simulations que la précision de l'estimation dépendait notamment de la taille démographique et de l'intervalle de temps entre les échantillons. En effet, il existe un biais important lorsque $N < 2N_e$ ou lorsque $t < 3$ générations. Plus récemment, Frachon *et al.* (2017) ont proposé un estimateur de la taille efficace basé sur la différenciation temporelle entre deux échantillons (F_{ST}) : $\widehat{N_e} = \frac{t(1-F_{ST})}{4F_{ST}}$. L'utilisation du F_{ST} multilocus (Weir et Cockerham 1984) permet d'avoir un estimateur qui présente un biais et une variance moindre qu'en utilisant le $\hat{F}$ de Waples (Waples 1989 ; Frachon *et al.* 2017).

Des méthodes temporelles basées sur la vraisemblance ont également été développées, d'abord pour des marqueurs bialléliques (Williamson et Slatkin 1999), et ensuite pour des marqueurs multi-alléliques (Anderson *et al.* 2000). Malgré une meilleure précision de ces méthodes par rapport à la méthode des moments présentée ci-dessus, elles requièrent un temps de calcul très long et présentent

des problèmes de convergence lorsque des marqueurs hautement polymorphes sont utilisés. Wang (2001) a proposé une méthode de « pseudo-vraisemblance » pour estimer N_e sur des marqueurs multi-alléliques avec un temps de calcul moindre. Il montre que cette méthode est plus précise que la méthode des moments, notamment en présence d'allèles rares.

Malgré la difficulté de devoir disposer de plusieurs échantillons temporels, l'estimation de la taille efficace à partir de données temporelles est largement utilisée dans la littérature (Palstra et Ruzzante 2008). Cependant, les différentes méthodes temporelles présentées ci-dessus reposent sur l'hypothèse que les locus considérés sont indépendants, ce qui n'est pas le cas en régime autogame. On s'attend donc à ce que l'autofécondation réduise la précision des estimations temporelles en diminuant le nombre de locus indépendants (ou loci efficaces). La méthode suppose aussi que la population considérée est isolée, et qu'il n'y a donc pas de changements de fréquences alléliques au cours du temps dus à de la migration. Cette hypothèse a cependant de fortes chances de ne pas être respectée en populations naturelles et les estimations qui en résultent seront donc biaisées. Le programme MLNe (Wang et Whitlock 2003) permet d'estimer conjointement N_e et le taux de migration m à partir d'échantillons temporels de la population d'intérêt et d'une population source de migrants. Dans une revue qui compare plusieurs estimateurs de la taille efficace, Gilbert et Whitlock (2015) déconseillent cependant l'utilisation de MLNe pour obtenir des estimations précises du taux de migration malgré le fait qu'il se comporte mieux que les autres logiciels testés pour estimer N_e dans les scénarios avec de forts taux de migration et/ou de structure de population. De plus, la question se pose de savoir si l'on estime la taille efficace locale (de la sous-population échantillonnée), ou la taille efficace de l'ensemble de la métapopulation.

b) Méthodes d'estimation à partir d'un seul échantillon

Obtenir des données génétiques à plusieurs pas de temps (et potentiellement pour plusieurs populations) peut se révéler difficile selon l'espèce étudiée. Dans certains cas, on ne dispose que d'un seul échantillon et des méthodes d'estimation de N_e adaptées à ces cas de figure ont été développées.

La méthode d'excès d'hétérozygotes permet d'estimer le nombre efficace de reproducteurs (N_b) dans une population parentale à partir d'un échantillon de la descendance. Dans une population de taille finie, on s'attend à une déviation aux proportions de Hardy-Weinberg avec notamment un excès en hétérozygotes $\alpha_0 = -\frac{1}{2N-1}$ (Kimura et Crow 1963), où N est la taille d'une population de Wright-Fisher. Dans le cas d'une population avec des sexes séparés, un écart aux proportions de Hardy-Weinberg est également attendu du fait de l'échantillonnage d'un nombre fini de parents mâles et femelles à chaque génération. L'excès d'hétérozygotes dans la descendance est alors $\alpha_p = -\frac{1}{8N_m} -$

$\frac{1}{8N_f} = -\frac{1}{2N_e}$ (Robertson 1965), où N_m et N_f sont le nombre de parents mâles et femelles respectivement et $N_e = \frac{4N_mN_f}{(N_m+N_f)}$ (Caballero 1994). Un estimateur de la taille efficace pour des locus bialléliques a été dérivé par Pudovkin *et al.* (1996) : $\widehat{N_b} = \frac{1}{2D} + \frac{1}{2(D+1)}$, où $D = \frac{H_{exp}}{H_{exp}-H_{obs}}$, $H_{exp} = 2p(1-p)$ est l'hétérozygotie attendue calculée à partir de la fréquence génétique *p* observée dans un échantillon de *N* descendants, et H_{obs} est l'hétérozygotie observée dans l'échantillon. Dans le cas de locus multi-alléliques, *D* est la moyenne sur tous les allèles par locus, et sur tous les locus (Luikart and Cornuet 1999). Cette méthode est néanmoins peu précise (elle résulte souvent en des estimations de N_e infinies pour des populations de petite taille), et est très sensible aux écarts à la panmixie (Beaumont 2004). La méthode d'excès d'hétérozygotes est notamment implémentée dans le logiciel Colony2 (Wang 2009), Nb_HetEx (Zhdanova and Pudovkin 2008) et NeEstimator v2 (Do *et al.* 2013).

Des méthodes d'estimation de N_e basées sur le déséquilibre de liaison (l'association non aléatoire d'allèles à différents locus) dans un échantillon ont également été développées. Elles sont basées sur l'hypothèse que dans une population panmictique isolée, la dérive génétique crée aléatoirement des associations entre allèles à différents locus neutres avec un taux inversement proportionnel à N_e (Hill 1981). Ainsi, un estimateur du carré de la corrélation des fréquences alléliques par paire de locus (r^2) est $E(\hat{r}^2) \simeq \frac{(1-c)^2+c^2}{2N_ec(2-c)} + \frac{1}{S}$ (Weir et Hill 1980), où *c* est le taux de recombinaison entre locus et *S* est le nombre d'individus échantillonnés. En supposant les locus indépendants (*c* = 0.5), on a $E(\hat{r}^2) \approx \frac{1}{3N_e} + \frac{1}{S}$. Des exemples de logiciels permettant d'estimer N_e à partir du déséquilibre de liaison sont : LDNe (Waples et Do 2008) ou Estim (Vitalis et Couvet 2001). Waples et Do (2010) ont évalué la précision de la méthode et ont montré qu'elle était particulièrement précise pour des populations de petite taille (N_e < 200). En revanche, elle a des difficultés à distinguer les populations de grande taille et les populations de taille infinie. De plus, ils ont montré que doubler le nombre de marqueurs résultait en une augmentation de la précision des estimations plus grande avec la méthode du déséquilibre de liaison qu'avec la méthode temporelle. Toutefois, comme pour les méthodes présentées précédemment, l'estimation de N_e à partir du déséquilibre de liaison repose sur des hypothèses qui peuvent se révéler limitantes en populations naturelles. En effet, la relation entre r^2 et N_e suppose l'absence d'autofécondation car celle-ci crée du déséquilibre de liaison, et peut donc biaiser les estimations. De plus, on suppose ici aussi que la population est isolée. Or, la migration peut avoir des effets contraires sur le déséquilibre de liaison : d'un côté, la présence de migrants augmente la taille du pool de parents, ce qui réduit le déséquilibre de liaison et on s'attend donc à des estimations de N_e local biaisées vers le haut. D'un autre côté, les migrants sont génétiquement différenciés des individus

locaux, ce qui crée du déséquilibre de liaison (Nei et Li 1973) et on s'attend à ce que les estimations de N_e soient biaisées vers le bas. Waples et England (2011) ont évalué les performances des estimations de taille efficace à partir du déséquilibre de liaison lorsque l'hypothèse de population isolée n'est pas respectée et ont montré qu'un modèle de migration à l'équilibre n'a que peu d'effet sur l'estimation de N_e, sauf si m > 5-10%. Par ailleurs, ils ont montré que des apports soudains de migrants biaisaient $\widehat{N_e}$ vers le bas.

De nombreuses autres méthodes existent et utilisent des méthodologies d'estimation variées. On peut par exemple citer le programme ONeSAMP (Tallmon *et al.* 2008) qui estime N_e par calcul Bayésien approché (voir encadré 2 dans le chapitre 3 pour une description de l'ABC) à partir de données microsatellites. Huit statistiques résumées sont utilisées : le nombre d'allèles divisé par la gamme de longueurs des allèles, la différence des logarithmes de la variance de la taille allélique et de l'hétérozygotie, l'hétérozygotie attendue, le nombre d'allèles par locus, le F_{IS}, la moyenne et la variance de l'homozygotie multi-locus, et le carré de la corrélation entre allèles à différents locus. Cette méthode est cependant déconseillée par Gilbert et Whitlock (2015) car elle surestime N_e pour des valeurs faibles et les temps de calcul lorsque N_e est grand sont prohibitifs (plus de 30 jours). Un autre exemple de méthode est celle de Wang (2009), la méthode de « sibship assignment » qui estime N_e à partir de la probabilité qu'une paire de descendants pris au hasard dans la population soient demi- ou plein frères : $\frac{1}{N_e} = \frac{1+3\alpha}{4}(Q_{HS} + 2Q_{FS}) - \frac{\alpha}{2}(\frac{1}{N_1} + \frac{1}{N_2})$, où α est une mesure de l'écart aux proportions de Hardy-Weinberg (équivalent du F_{IS}) ; Q_{HS} et Q_{FS} sont les probabilités que les individus de la paire soient demi- ou plein frères respectivement ; et N_1 et N_2 sont le nombre d'individus mâles et femelles dans la population. Wang (2009) reporte une meilleure précision que les méthodes des moments, du déséquilibre de liaison, ou de l'excès d'hétérozygotes sur des simulations et des données empiriques.

3. Conclusion sur l'estimation des paramètres démographiques en populations autogames : peut-on séparer les effets de la dérive génétique, de la migration et du taux d'autofécondation ?

La grande diversité des méthodes d'estimation présentée ici souligne l'intérêt d'estimer les paramètres démographiques des populations. Cette rapide revue des méthodes fait apparaitre le fait que ce sont souvent les mêmes statistiques qui sont informatives pour l'estimation de paramètres différents : le déficit (ou l'excès) d'hétérozygotes et le déséquilibre de liaison par exemple sont utilisés à la fois dans des méthodes d'estimation de taux d'autofécondation et de taille efficace. En régime autogame, ceci peut être problématique pour différencier les effets des différentes forces évolutives

et démographiques. En particulier, estimer séparément le taux d'autofécondation, la migration et la dérive génétique contemporaine n'est pas satisfaisant car chaque estimation individuelle néglige les autres facteurs. Il existe des méthodes d'inférence conjointe de N_e et m, mais elles ne prennent pas en compte l'autofécondation. De plus, les méthodes d'estimation reposent très souvent sur des modèles dont les hypothèses ne sont pas respectées en régime autogame (à commencer par la panmixie). En effet, comme décrit plus haut, les populations autogames semblent souvent suivre des dynamiques de métapopulations et ne respectent donc pas les hypothèses de population isolée de taille constante. Ainsi, l'étude des populations autogames est rendue difficile par le manque de méthodes adaptées. Il paraît donc nécessaire de développer et décrire des attendus sur des indicateurs informatifs en autofécondation, comme par exemple des indices multilocus (voir chapitre 2). Ces statistiques multilocus pourraient ensuite permettre de développer des méthodes d'inférence adaptées (voir chapitre 3).

VI. Exemple d'une espèce majoritairement autogame : *Medicago truncatula*

1. Présentation de l'espèce

Medicago truncatula est une plante annuelle appartenant à la famille des légumineuses (Fabaceae). Elle est originaire du pourtour méditerranéen et est considérée comme une plante colonisatrice mais peu compétitive. En Australie, elle est cultivée en tant que plante fourragère et également comme engrais vert (Pearson *et al.* 1997). Du fait de sa facilité de culture, de son génome diploïde de petite taille (~ $5x10^{-8}$ pb) et de sa capacité à s'associer avec les bactéries fixatrices d'azote, *M. truncatula* est devenue une plante modèle pour les études de génétique moléculaire sur les légumineuses (Cook 1999 ; Young *et al.* 2003 ; Choi *et al.* 2004). Plusieurs études en populations naturelles ont montré le caractère majoritairement autogame de *M. truncatula* avec des taux d'autofécondation estimés qui varient la plupart du temps entre 0.8 et 1 (Bonnin *et al.* 2001 ; Bataillon et Ronfort 2006 ; Siol *et al.* 2007, 2008). Au cours de ce travail de thèse, j'ai pu tirer profit de larges jeux de données disponibles sur *M. truncatula*, avec notamment des données temporelles sur neuf populations naturelles sur des pas de temps de 20 ans et des données récoltées sur des descendances maternelles.

a) Organisation de la diversité génétique à l'échelle de l'espèce

Ronfort *et al.* (2006) ont analysé 346 lignées de *Medicago truncatula* provenant de l'ensemble de l'aire de répartition de l'espèce à l'aide de treize marqueurs microsatellites. Ils ont montré une

importante diversité génétique avec une hétérozygotie selon Nei (1973) moyenne de 0.75. Une analyse de stratification génétique à l'aide du logiciel Structure (Pritchard *et al.* 2000) a permis d'identifier quatre groupes génétiques distincts. Lorsque l'on considère uniquement les accessions présentant une forte probabilité d'appartenance à l'un des groupes (> 0.60), ces groupes correspondent aux origines géographiques des accessions : nord-est du bassin méditerranéen (triangles rouges, Figure 8), Espagne et Maroc (triangles bleus, Figure 8), sud-est du bassin méditerranéen (triangles jaunes, Figure 8), et sud de France (triangles verts, Figure 8). La correspondance avec l'origine géographique est moins claire lorsque l'on considère les accessions avec une probabilité d'appartenance moindre (environ 30% des accessions étudiées). Cela semble indiquer soit des flux de gènes récents entre les groupes, soit de la migration à longue distance combinée à des évènements de recombinaison (Ronfort *et al.* 2006). La structure en quatre groupes génétiques, bien que significative, n'explique qu'une petite partie de la variabilité totale (F_{ST} global = 0.08). Cela suggère que la variance de niveau de diversité génétique est distribuée à des échelles géographiques plus faibles (échelle de la région ou de la population).

Figure 8 : Structure génétique de *Medicago truncatula* inférée à partir de marqueurs microsatellites grâce au logiciel STRUCTURE.
Seules les accessions avec une probabilité d'appartenance à un groupe génétique de plus de 60% sont représentées. Les couleurs correspondent aux quatre groupes identifiés sur la seule base des données génétiques (treize marqueurs microsatellites). D'après Ronfort *et al.* (2006).

b) Organisation de la diversité génétique à l'échelle régionale

Bonnin *et al.* (1996) ont comparé quatre populations naturelles séparées au maximum par 200 km. Ils ont montré que la variance intra-population représentait 55% de la variance totale mesurée à l'aide de marqueurs RAPD, alors que 45% de la variabilité génétique totale était distribuée entre populations. De plus, les différentes populations ne partageaient aucun génotype multilocus (MLG).

Ainsi, *M. truncatula* présente une forte structure spatiale de sa diversité génétique entre populations, mais des niveaux de diversité non négligeables semblent être conservés au sein des populations.

c) Organisation de la diversité génétique à l'échelle des populations

La structure génétique intra-populations de *Medicago truncatula* à très fine échelle a été décrite par plusieurs études à l'aide de marqueurs RAPD (Bonnin *et al.* 1996), et de marqueurs microsatellites (Bonnin *et al.* 2001 ; Siol *et al.* 2007, 2008). Ces études montrent que les populations sont fortement structurées, même à des échelles spatiales très réduites (~ 50 mètres). Les populations (ou sous-populations), sont constituées de quelques génotypes multilocus (MLG) répétés un grand nombre de fois, et de génotypes uniques plus ou moins nombreux (Figure 9). De plus, les MLG sont peu ou pas partagés entre sous-populations (Bonnin *et al.* 1996; Siol *et al.* 2007).

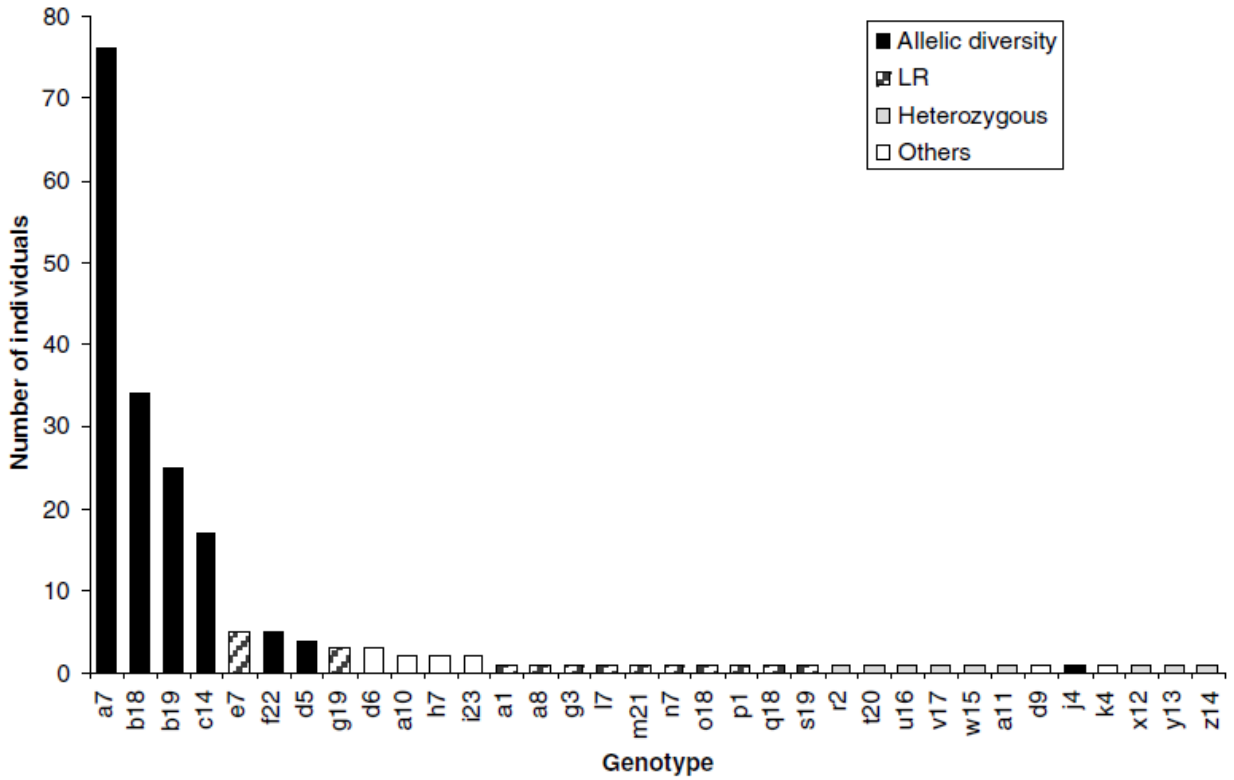


Figure 9 : Exemple de distribution des génotypes multilocus (MLG) au sein d'une population naturelle de *M. truncatula*.
34 génotypes multilocus distincts ont été détectés parmi 200 individus. Les MLG représentés en noirs expliquent l'ensemble de la variation allélique observée dans la population. Les barres hachurées correspondent à des MLG potentiellement issus de recombinaison entre les MLG majoritaires. D'après Siol *et al.* (2008).

Bonnin *et al.* (2001) montrent que les MLG sont distribués en patch au sein des sous-populations. De même, Siol *et al.* (2008) montrent que l'apparentement entre individus diminue très rapidement avec la distance (Figure 10), synonyme d'isolement par la distance. Enfin, une étude temporelle a été conduite sur une population. Malgré des changements de fréquence (parfois importants) des MLG au cours du temps, cette étude montre que l'organisation spatiale des MLG reste relativement stable au cours du temps (Siol *et al.* 2007).

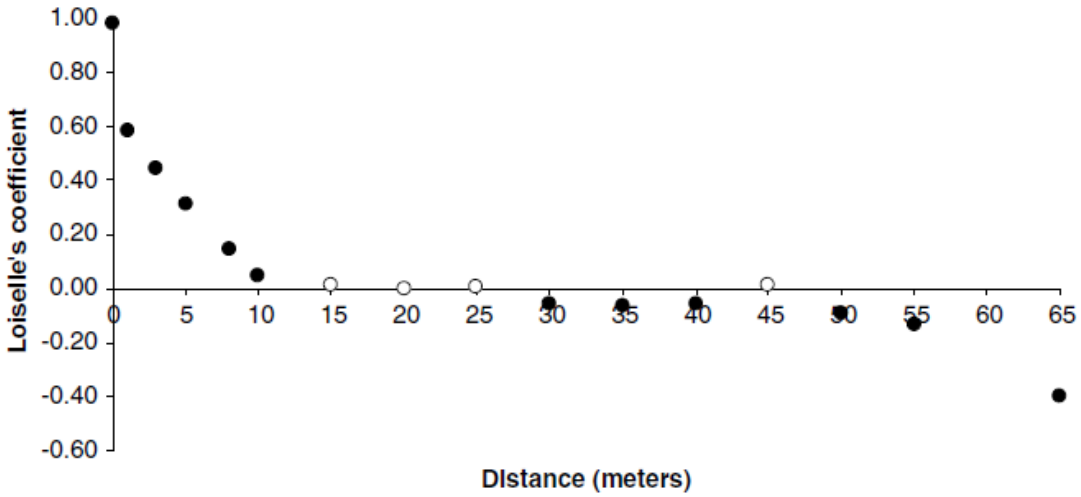


Figure 10 : Analyse d'autocorrélation spatiale au sein d'une population naturelle de *Medicago truncatula*.
Le graphique représente l'apparentement moyen (calculé selon Loiselle *et al.* 1995) entre deux individus situés à *x* mètres de distance l'un de l'autre au sein d'une population. D'après Siol *et al.* (2008).

2. Système de reproduction

Comme décrit précédemment, *M. truncatula* est une plante majoritairement autogame. L'autofécondation est autonome et se déroule à l'intérieur du bouton floral, avant que la fleur ne s'ouvre (autofécondation prioritaire, Lloyd 1992). Elle est facilitée par la structure de la fleur, où les étamines entourent le stigmate et sont directement à son niveau (Figure 11).

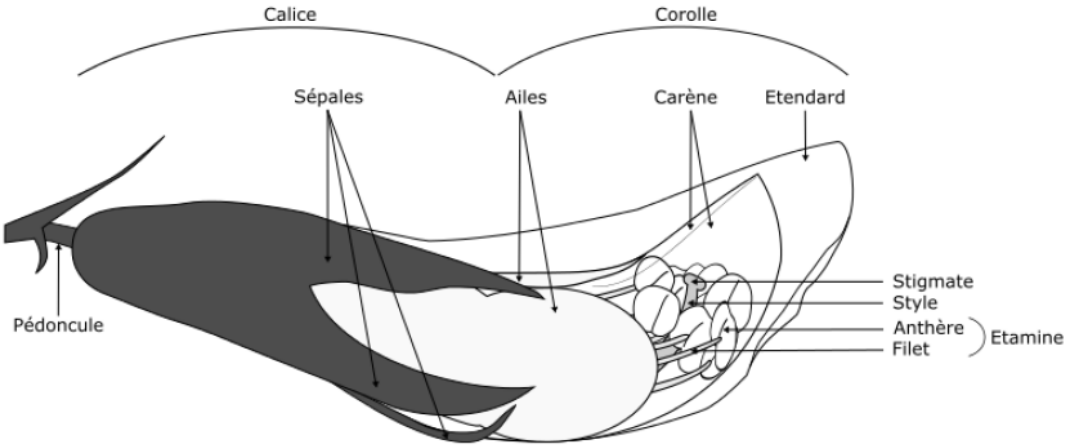


Figure 11 : Structure de la fleur de *Medicago truncatula*.
D'après Bosseno *et al.* (2017)

La fleur dispose cependant d'un mécanisme dit de déclenchement (*tripping mechanism*) qui projette les étamines sur le corps d'un pollinisateur qui viendrait se poser à la base de la fleur sur le pétale étendard (Small 2010). Ce déclenchement semble donc être un mécanisme facilitant l'allofécondation et est conservé chez l'ensemble du genre *Medicago*. De fait, les populations de *M. truncatula* ne sont pas complètement autogames et des niveaux plus ou moins variables d'allofécondation résiduelle sont mesurés selon les populations (voir Chapitre 2). Les estimations disponibles sont cependant des estimations moyennes sur l'ensemble des individus analysés et ne permettent pas de savoir à quelle échelle se distribue l'allofécondation résiduelle : sur l'ensemble des individus, chez seulement quelques-uns d'entre eux, chez certains génotypes en particulier, sur quelques fleurs seulement ?

Présentation de la thèse

L'objectif de cette thèse était de comprendre de manière fine comment la diversité génétique neutre se structure dans les populations autogames et comment elle évolue au cours du temps. A travers l'exploration de patrons de diversité contrastés, je cherche à développer des indices et méthodes adaptés aux spécificités des populations autogames. Pour cela, je me suis appuyée sur des approches théoriques que j'ai ensuite confrontées à des données empiriques avant de développer une méthode d'inférence. Ma démarche se découpe en trois axes :

Chapitre 1 : Allofécondation résiduelle chez *Medicago truncatula*

Je me suis tout d'abord intéressée aux variations du système de reproduction que l'on peut observer dans une population naturelle de *Medicago truncatula*. L'objectif de ce chapitre était de décrire la variabilité de l'allofécondation résiduelle à différents niveaux : au cours de la saison de floraison, entre génotypes, entre plantes mères, et entre fleurs. J'ai donc utilisé une approche par descendances maternelles pour estimer précisément le taux d'autofécondation et tenter de caractériser les déterminismes (génétique, environnemental) de ses variations. En particulier, j'ai tiré parti de l'organisation en génotypes multilocus répétés caractéristique des populations autogames pour évaluer un possible déterminisme génétique de l'allofécondation résiduelle.

Une autre étude du déterminisme de l'allofécondation résiduelle, cette fois-ci basée sur des conditions expérimentales plus contrôlées se trouve également en Annexe.

Chapitre 2 : Structure de la diversité génétique multilocus en régime autogame

Dans ce chapitre, je cherche à caractériser comment l'autofécondation affecte la diversité mono- et surtout multilocus. En effet, cette dernière nous semble particulièrement informative pour comprendre les processus démographiques à l'œuvre dans les populations autogames. Cependant, nous ne disposons pas d'attendus théoriques neutres et il n'est pas clair si la structure en génotypes multilocus répétés caractéristique des populations autogames peut être générée seulement avec des scénarios démographiques neutres. J'utilise donc un cadre de simulations dans lequel je développe des scénarios démographiques contrastés, me permettant d'évaluer les effets de l'autofécondation, de la taille de population, de la migration, de l'admixture et des évènements d'extinction-recolonisation sur la diversité génétique au cours du temps. En particulier, je regarde comment l'analyse conjointe de la diversité mono- et multilocus peut permettre de différencier les effets de plusieurs paramètres démographiques. Je compare ensuite les patrons obtenus avec des données

temporelles sur des populations de naturelles de *Medicago truncatula* afin de valider la pertinence de mes simulations.

Chapitre 3 : Inférence conjointe du taux d'autofécondation, de la taille efficace et de la migration chez des populations naturelles de *Medicago truncatula* à partir de données temporelles

L'objectif de ce dernier chapitre est de développer une méthode d'inférence de l'histoire démographique contemporaine adaptée aux populations autogames à partir de données temporelles. J'utilise donc le cadre de simulations mis en place précédemment avec une méthode hautement flexible, le calcul Bayésien approché avec forêts aléatoires, pour tester trois scénarios démographiques sur cinq populations naturelles de *Medicago truncatula* : population isolée, population avec un taux de migration constant, et population subissant des évènements d'admixture. Une fois le scénario démographique inféré, je tente d'inférer les paramètres démographiques qui le caractérisent et j'analyse les limites et les développements futurs de la méthode d'inférence.

Bibliographie

Abbott RJ and Gomes MF. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity* **62**: 411–8.

Allen AM and Hiscock SJ. 2008. Evolution and phylogeny of self-incompatibility systems in Angiosperms. In: Franklin-Tong VE (Ed). Self-Incompatibility in Flowering Plants: Evolution, Diversity, and Mechanisms. Berlin, Heidelberg: Springer Berlin Heidelberg.

Anderson EC, Williamson EG, and Thompson EA. 2000. Monte Carlo evaluation of the likelihood for Ne from temporally spaced samples. *Genetics* **156**: 2109–18.

Baker HG. 1967. Support for Baker's law as a rule. *Evolution* **21**: 853–6.

Bakker EG, Stahl EA, Toomajian C, *et al.* 2006. Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* **15**: 1405–18.

Barrett SCH. 2002. The evolution of plant sexual diversity. *Nature Reviews Genetics* **3**: 274–84.

Barrett SCH and Eckert CG. 1990. Variation and evolution of mating systems in seed plants. *Biological Approaches and Evolutionary Trends in Plants*: 229–54.

Barton NH. 2000. Genetic hitchhiking. *PhilTrans R Soc Lond B* **355**: 1553–62.

Bataillon T and Ronfort J. 2006. Evolutionary and ecological genetics of *Medicago truncatula*

Beaumont MA. 2004. Conservation genetics. In: Handbook of Statistical Genetics. American Cancer Society.

Bergelson J, Stahl E, Dudek S, and Kreitman M. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* **148**: 1311–23.

Bomblies K, Yant L, Laitinen RA, *et al.* 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet* **6**: e1000890.

Bonnin I, Huguet T, Gherardi M, *et al.* 1996. High level of polymorphism and spatial structure in a selfing plant species, *Medicago truncatula* (Leguminosae), shown using RAPD markers. *American Journal of Botany* **83**: 843–55.

Bonnin I, Ronfort J, Wozniak F, and Olivieri I. 2001. Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). *Molecular Ecology* **10**: 1371–83.

Bosseno M, Lambert A, Beucher D, *et al.* 2017. Protocole simple de rétrocroisement chez *Medicago truncatula*. INRA.

Brown AHD. 1979. Enzyme polymorphism in plant populations. *Theoretical Population Biology* **15**: 1–42.

Burgarella C and Glémin S. 2017. Population genetics and genome evolution of selfing species (John Wiley & Sons Ltd, Ed). *eLS*: 1–8.

Caballero A. 1994. Developments in the prediction of effective population size. *Heredity* **73 (Pt 6)**: 657–79.

Caballero A and Hill WG. 1992. Effects of partial inbreeding on fixation rates and variation of mutant genes. *Genetics* **131**: 493–507.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**: 195–205.

Chauvet S, Velde MVD, Imbert E, *et al.* 2004. Past and current gene flow in the selfing, wind-dispersed species *Mycelis muralis* in Western Europe. *Molecular Ecology* **13**: 1391–407.

Choi H-K, Kim D, Uhm T, *et al.* 2004. A sequence-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. *Genetics* **166**: 1463–502.

Cook DR. 1999. *Medicago truncatula* - a model in the making! *Current Opinion in Plant Biology* **2**: 301–4.

Crow JF and Kimura M. 1970. An introduction to population genetics theory. New York: Harper & Row.

David P, Pujol B, Viard F, *et al.* 2007. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* **16**: 2474–87.

Devaux C, Lepers C, and Porcher E. 2014. Constraints imposed by pollinator behaviour on the ecology and evolution of plant mating systems. *Journal of Evolutionary Biology* **27**: 1413–30.

Do C, Waples RS, Peel D, *et al.* 2013. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* **14**: 209–14.

Enjalbert J and David JL. 2000. Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics* **156**: 1973–82.

Falahati-Anbaran M, Lundemo S, and Stenøien HK. 2014. Seed dispersal in time can counteract the effect of gene flow between natural populations of *Arabidopsis thaliana*. *New Phytologist* **202**: 1043–54.

Fisher RA. 1941. Average excess and average effect of a gene substitution. *Annals of Eugenics* **11**: 53–63.

Frachon L, Libourel C, Villoutreix R, *et al.* 2017. Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nature Ecology & Evolution* **1**: 1551.

Frankham R. 1996. Relationship of genetic variation to population size in wildlife. *Conservation Biology* **10**: 1500–8.

Frankham R. 1997. Do island populations have less genetic variation than mainland populations? *Heredity* **78**: 311–27.

Gilbert KJ and Whitlock MC. 2015. Evaluating methods for estimating local effective population size with and without migration. *Evolution* **69**: 2154–66.

Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* **177**: 905–16.

Glémin S, Bazin E, and Charlesworth D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc R Soc B* **273**: 3011–9.

Goldberg EE, Kohn JR, Lande R, *et al.* 2010. Species selection maintains self-incompatibility. *Science* **330**: 493–5.

Golding GB and Strobeck C. 1980. Linkage disequilibrium in a finite population that is partially selfing. *Genetics* **94**: 777–89.

Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, and Picó FX. 2011. Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. *Molecular Ecology* **20**: 3540–54.

Goodwillie C, Kalisz S, and Eckert CG. 2005. The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics* **36**: 47–79.

Hamrick JL and Godt MJW. 1990. Allozyme diversity in plant species. *Plant population genetics, breeding, and genetic resources*: 43–63.

Hamrick JL and Godt MJW. 1997. Allozyme diversity in cultivated crops. *Crop Science* **37**: 26–30.

Hanski I. 1998. Metapopulation dynamics. *Nature* **396**: 41–9.

Hartl D and Clark AG. 1998. Principles of population genetics. Sinauer Associates.

Herlihy CR and Eckert CG. 2007. Evolutionary analysis of a key floral trait in *Aquilegia Canadensis* (ranunculaceae): genetic variation in herkogamy and its effect on the mating system. *Evolution* **61**: 1661–74.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet Res* **38**: 209–16.

Holsinger KE, Feldman MW, and Christiansen FB. 1984. The evolution of self-fertilization in plants: a population genetic model. *The American Naturalist* **124**: 446–53.

Ingvarsson PK. 2002. A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* **56**: 2368–73.

Innan H and Stephan W. 2000. The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. *Genetics* **155**: 2015–9.

Jarne P and Auld JR. 2006. Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. *Evolution* **60**: 1816–25.

Jarne P and David P. 2008. Quantifying inbreeding in natural populations of hermaphroditic organisms. *Heredity* **100**: 431–9.

Johnston MO. 1998. Evolution of intermediate selfing rates in plants: pollination ecology versus deleterious mutations. *Genetica* **102**: 267.

Johnston MO, Porcher E, Cheptou P, *et al.* 2009. Correlations among fertility components can maintain mixed mating in plants. *The American Naturalist* **173**: 1–11.

Jones DF. 1916. Natural cross-pollination in the tomato. *Science* **43**: 509–10.

Karron JD, Thumser NN, Tucker R, and Hessenauer AJ. 1995. The influence of population density on outcrossing rates in *Mimulus ringens*. *Heredity* **75**: 175–80.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kimura M and Crow JF. 1963. The measurement of effective population number. *Evolution* **17**: 279–88.

Krimbas CB and Tsakas S. 1971. The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control: selection or drift? *Evolution* **25**: 454–60.

Kuittinen H, Mattila A, and Savolainen O. 1997. Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis Thaliana*. *Heredity* **79**: 144–52.

Lande R and Schemske DW. 1985a. The evolution of self-fertilization and inbreeding depression in plants: I. Genetic models. *Evolution* **39**: 24–40.

Lande R and Schemske DW. 1985b. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**: 24–40.

Lloyd DG. 1979. Some reproductive factors affecting the selection of self-fertilization in plants. *The American Naturalist* **113**: 67–79.

Lloyd DG. 1992a. Self- and cross-fertilization in plants: II. The selection of self-fertilization. *International Journal of Plant Sciences* **153**: 370–80.

Lloyd DG. 1992b. Self- and cross-fertilization in plants. II. The selection of self- fertilization. *Int J Plant Sci* **153**: 370–80.

Lloyd DG and Webb CJ. 1986. The avoidance of interference between the presentation of pollen and stigmas in angiosperms I. Dichogamy. *New Zealand Journal of Botany* **24**: 135–62.

Loiselle BA, Sork VL, Nason J, and Graham C. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **82**: 1420–5.

Luikart G and Cornuet J-M. 1999. Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* **151**: 1211–6.

Lundemo S, Falahati-Anbaran M, and Stenøien HK. 2009. Seed banks cause elevated generation times and effective population sizes of *Arabidopsis thaliana* in Northern Europe. *Molecular Ecology* **18**: 2798–811.

Maynard Smith J and Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23–35.

Montesinos A, Tonsor SJ, Alonso-Blanco C, and Picó FX. 2009. Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments. *PLoS ONE* **4**: e7213.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* **70**: 3321–3.

Nei M and Li W-H. 1973. Linkage disequilibrium in subdivided populations. *Genetics* **75**: 213–9.

Nei M and Tajima F. 1981. Genetic drift and estimation of effective population size. *Genetics* **98**: 625–40.

Noël E, Jarne P, Glémin S, *et al.* 2017. Experimental evidence for the negative effects of self-fertilization on the adaptive potential of populations. *Current Biology* **27**: 237–42.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–9.

Nordborg M, Borevitz JO, Bergelson J, *et al.* 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **30**: 190–3.

Palstra FP and Ruzzante DE. 2008. Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology* **17**: 3428–47.

Pearson CJ, Brown R, Collins WJ, *et al.* 1997. An Australian temperate pastures database. *Aust J Agric Res* **48**: 453–66.

Picó FX, Méndez-Vigo B, Martínez-Zapater JM, and Alonso-Blanco C. 2008. Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian Peninsula. *Genetics* **180**: 1009–21.

Pollak E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**: 353–60.

Porcher E and Lande R. 2005. The evolution of self-fertilization and inbreeding depression under pollen discounting and pollen limitation. *J Evol Biol* **18**: 497–508.

Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–59.

Pudovkin AI, Zaykin DV, and Hedgecock D. 1996. On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144**: 383–7.

Revell LJ. 2018. learnPopGen: Population genetic simulations & numerical analysis.

Richards AJ. 1997. Plant Breeding Systems. London: Chapman & Hall.

Ritland K. 2002. Extensions of models for the estimation of mating systems using *n* independent loci. *Heredity* **88**: 221–8.

Robertson A. 1965. The interpretation of genotypic ratios in domestic animal populations. *Animal Science* **7**: 319–24.

Ronfort J, Bataillon T, Santoni S, *et al.* 2006. Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*. *BMC Plant Biology* **6**: 28.

Schemske DW and Lande R. 1985. The evolution of self-fertilization and inbreeding depression in plants: II. Empirical observations. *Evolution* **39**: 41–52.

Schoen DJ and Brown AH. 1991. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc Natl Acad Sci USA* **88**: 4494–7.

Sharbel TF, Haubold B, and Mitchell-Olds T. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology* **9**: 2109–18.

Shaw DV, Kahler AL, and Allard RW. 1981. A multilocus estimator of mating system parameters in plant populations. *Proc Natl Acad Sci USA* **78**: 1298–302.

Siol M, Bonnin I, Olivieri I, *et al.* 2007. Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual *Medicago truncatula*. *J Evol Biol* **20**: 2349–60.

Siol M, Prosperi J-M, Bonnin I, and Ronfort J. 2008. How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. *Heredity* **100**: 517–25.

Small E. 2010. Alfalfa and relatives: Evolution and classification of Medicago. Canada: NRC Research Press.

Stebbins GL. 1957. Self-fertilization and population variability in the higher plants. *The American Naturalist* **91**: 337–54.

Stenøien HK, Fenster CB, Tonteri A, and Savolainen O. 2005. Genetic variability in natural populations of *Arabidopsis thaliana* in Northern Europe. *Molecular Ecology* **14**: 137–48.

Taberlet P, Waits LP, and Luikart G. 1999. Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution* **14**: 323–7.

Tallmon DA, Koyuk A, Luikart G, and Beaumont MA. 2008. ONeSAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources* **8**: 299–301.

Tenaillon MI, Sawkins MC, Long AD, *et al.* 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *PNAS* **98**: 9161–6.

Todokoro S, Terauchi R, and Kawano S. 1995. Microsatellite polymorphisms in natural populations of Arabidopsis thaliana in Japan. *Jpn J Genet* **70**: 543–54.

Uyenoyama MK. 1986. Inbreeding and the cost of meiosis: The evolution of selfing in populations practicing biparental inbreeding. *Evolution* **40**: 388–404.

Vitalis R and Couvet D. 2001. Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**: 911–25.

Wang J. 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genet Res* **78**: 243–57.

Wang J. 2005. Estimation of effective population sizes from data on genetic markers. *Phil Trans R Soc B* **360**: 1395–409.

Wang J. 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology* **18**: 2148–64.

Wang J and Caballero A. 1999. Developments in predicting the effective size of subdivided populations. *Heredity* **82**: 212–26.

Wang J and Whitlock MC. 2003. Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163**: 429–46.

Waples RS. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**: 379–91.

Waples RS and Do C. 2008. LDNe: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**: 753–6.

Waples RS and Do C. 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* **3**: 244–62.

Waples RS and England PR. 2011. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics* **189**: 633–44.

Waples RS and Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**: 1419–39.

Webb CJ and Lloyd DG. 1986. The avoidance of interference between the presentation of pollen and stigmas in angiosperms II. Herkogamy. *New Zealand Journal of Botany* **24**: 163–78.

Weir BS and Cockerham CC. 1973. Mixed self and random mating at two loci. *Genet Res* **21**: 247–62.

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–70.

Weir BS and Hill WG. 1980. Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**: 477–88.

Whitehead MR, Lanfear R, Mitchell RJ, and Karron JD. 2018. Plant mating systems often vary widely among populations. *Front Ecol Evol* **6**.

Whitlock MC and Barton NH. 1997. The effective size of a subdivided population. *Genetics* **146**: 427–41.

Williamson EG and Slatkin M. 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**: 755–61.

Wright S. 1931a. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright S. 1931b. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright S. 1933. Inbreeding and homozygosis. *PNAS* **19**: 411–20.

Wright S. 1943. Isolation by distance. *Genetics* **28**: 114–38.

Young ND, Mudge J, and Ellis TN. 2003. Legume genomes: more than peas in a pod. *Current Opinion in Plant Biology* **6**: 199–204.

Zhdanova OL and Pudovkin AI. 2008. Nb_HetEx: a program to estimate the effective number of breeders. *Journal of Heredity* **99**: 694–5.

Chapitre 1

Allofécondation résiduelle chez *Medicago truncatula*

Présentation générale

Nous avons vu en Introduction qu'il existe une grande variabilité de systèmes de reproduction chez les Angiospermes et que les taux d'autofécondation ont une distribution continue et légèrement bimodale avec un pic de fréquence pour les allogames stricts (0) et un autre pic pour les majoritairement autogames (1). Cependant, des taux d'autofécondation de 1 ne sont presque jamais observés en populations naturelles. Des modèles théoriques prédisent la maintenance d'un taux d'allofécondation résiduelle à travers l'avantage de la recombinaison (Charlesworth *et al.* 1991 ; David *et al.* 1993 ; Kamran-Disfani et Agrawal 2014), un avantage écologique (Porcher et Lande 2005), ou à travers un état transitoire jusqu'à l'autofécondation complète. Le déterminisme de l'autofécondation pourrait impliquer des facteurs génétiques (déterminant des traits floraux par exemple) comme environnementaux (densité, ressources, etc.).

Dans ce chapitre, je cherche à identifier la variabilité et le déterminisme de l'allofécondation résiduelle chez *Medicago truncatula* en évaluant les variations du taux d'autofécondation à différents niveaux : au cours de la saison de floraison, entre génotypes, entre plantes mères et entre fleurs. Pour cela, j'utilise des descendances maternelles échantillonnées dans une population située dans le sud de la France. Pour chaque plante mère, la première et la dernière gousse de la saison de floraison ont été récoltées. Un descendant par gousse a été sélectionné au hasard afin de réaliser une analyse de diversité de la population. Les taux d'autofécondation ont été estimés à partir des génotypes des descendants et un test de ratio de vraisemblance appliqué pour tester une variation de taux d'autofécondation entre le début et la fin de la saison. Par ailleurs, les génotypes maternels ont été inférés et une probabilité d'être issu d'autofécondation (P_{self}) calculée pour chaque descendant. Les effets du génotype, de la plante mère, du patch d'origine et de la fleur sur P_{self} ont été testés à l'aide d'un modèle généralisé mixte.

L'analyse de diversité montre une forte structure spatiale et que la population est composée à plus de 50% de génotypes multilocus répétés et agrégés dans l'espace. Le test de ratio de vraisemblance met en évidence une augmentation du taux d'allofécondation résiduelle au cours de la saison de floraison. Cet effet n'est cependant plus significatif lorsqu'on considère les effets génétiques, maternel et environnementaux. Nous avons également détecté des variations de P_{self} entre génotypes maternels, résultant en une héritabilité de 9%. Cette analyse suggère qu'il pourrait y avoir une part génétique à l'allofécondation résiduelle, mais que la variance génétique disponible dans cette population est faible. Ainsi, même si l'allofécondation présente un avantage en termes de recombinaison ou de pollinisation, il est peu probable que ce trait évolue dans cette population.

Les résultats obtenus sont présentés et discutés de manière détaillée dans le manuscrit ci-dessous.

Par ailleurs, une étude du déterminisme de l'allofécondation résiduelle basée sur un dispositif expérimental permettant de mieux randomiser les génotypes maternels dans l'environnement est détaillée en Annexe.

How and when does outcrossing occur in the predominantly selfing species *Medicago truncatula?*

Margaux Jullien[1], Joëlle Ronfort[1], Laurène Gay[1]

[1] AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro – Montpellier, France

Corresponding author: Margaux Jullien; INRA, 2 place Pierre Viala, 34060 Montpellier Cedex 1, FRANCE, Telephone: +33(0)4.99.61.39.89; Fax: +33.(0)4.99.61.20.64; E-mail: margauxjullien3@gmail.com

Running title: Residual outcrossing in *Medicago truncatula*

Word count (excluding references, tables and figures): 4961

Abstract (294 words)

Background and aims: Empirical studies on natural populations of *Medicago truncatula* revealed selfing rates higher than 80% but never up to 100%. Those studies show variability in the level of residual outcrossing between populations and also between temporal samples of the same population. However, these studies measure global selfing rates at the scale of the population and we do not know whether there is intra-population variation and how it is distributed (variance between genotypes, plants, flowers or seeds). Here, we focus on one natural population of *M. truncatula* in order to describe precisely its mating system. In particular, we investigated the determinants of the selfing rate by testing for seasonal variation (environmental or developmental determinism), and variation between genotypes (genetic determinism).

Methods: We measured selfing rates in maternal progenies from plants collected widely across a natural population. For each plant, we collected pods from flowers produced at the beginning and at the end of the flowering season to test for the occurrence of outcrossing rate seasonal variation. Using the individual probability of being issued from a self-fertilization event, we tested the distribution of variation at different scales: between genotypes (available because of the multilocus structure of the population), between mother plants, between flowers and within flower.

Key results: There was a significant, albeit small, increase in multilocus outcrossing rate in progenies collected at the end of the flowering season (t_m = 0.137 (0.025)) compared to those collected at the beginning (t_m = 0.083 (0.016)) of the flowering period. Between genotype variation in selfing rate was also detected, resulting in a heritability of 9%.

Conclusions: Despite a predominantly selfing mating system, *M. truncatula* displays variations in residual outcrossing rate. We show that these variations are genetically determined, but that environmental and developmental effects are prevalent.

Key words: *Medicago truncatula*, residual outcrossing, predominantly selfing

Introduction

The diversity of plant mating systems

Plant mating systems present a remarkable diversity, which results in a continuous distribution of selfing rates over angiosperm species (Igic and Kohn 2006). Despite widespread hermaphroditism (around 70% of the angiosperms, Richards 1997), only about 15% of flowering plants are "predominantly selfing" and present outcrossing rates lower than 10% (Igic and Kohn 2006). Residual outcrossing has been observed in experimental and natural populations (e.g. Allard and Workman 1963; Kahler *et al.* 1975; Bonnin *et al.* 2001; Bomblies *et al.* 2010). Actually, complete selfing is seldom encountered in the wild, as shown by a meta-analysis where only 1% of the species reached an estimated outcrossing rate of zero (Winn et al. 2011). Little is known about the maintenance of this residual outcrossing and it remains a question of interest.

What favours the maintenance of residual outcrossing?

The evolution of plant mating systems is mainly driven by the balance between two opposing forces: the twofold automatic transmission advantage of selfing (Fisher 1941), and inbreeding depression (reviewed in Charlesworth 2006). However, continued selfing eliminates most of the inbreeding depression by purging deleterious mutations (Lande and Schemske 1985). As a result, once self-fertilization has evolved, reversion to outcrossing is very unlikely, which should theoretically allow the fixation of a selfing rate of one (Charlesworth 1980; Lande and Schemske 1985; Charlesworth *et al.* 1990). Yet, several factors could explain the maintenance of residual outcrossing. We will detail here three hypotheses: the advantage of recombination, residual outcrossing as a transient phase, and the ecological advantage.

1. Residual outcrossing and the advantage of recombination

In a genetic model considering synergistic epistasis between deleterious alleles (i.e. when the effects of mutations alone are smaller than when combined with others), Charlesworth *et al.* (1991) showed that an evolutionary stable selfing rate slightly below one could be reached and that complete selfing could be selected against, even if inbreeding depression was low. In addition, due to the reduced effective recombination rate in predominantly selfing species (Nordborg 2000), deleterious mutations are more likely to accumulate through Muller's ratchet (Muller 1964). As a consequence, negative disequilibrium, i.e. associations between deleterious and beneficial alleles at different loci, can arise (Hartfield 2016). Kamran-Disfani and Agrawal (2014) used simulations to show that negative disequilibrium is rapidly reduced when the selfing rate is below one and that, when the

selfing rate is allowed to evolve, low levels of outcrossing can be maintained by selection. Furthermore, in completely selfing populations we expect low genetic diversity (Charlesworth and Charlesworth 1995; Clo *et al.* 2019), structured into highly differentiated homozygous multilocus genotypes (MLGs). Residual outcrossing creates recombinant genotypes with new allele combinations, thereby potentially allowing the association of favourable alleles that were otherwise confined into different genetic backgrounds (lines). Some outcrossing is therefore beneficial to break apart selection interference between mutations (Hartfield and Glémin 2016). Using simulations, David *et al.* (1993) showed that under strong directional selection for fitness, low levels of residual outcrossing can be conserved in selfing populations and this allows the population to rapidly adapt to a changing environment.

2. "Almost-complete" selfing could just be a transient phase

The above-mentioned genetic models of mating system evolution all assume a simple genetic determinism of selfing (a single modifier locus) and rarely consider the time to fixation of the selfing allele. Yet, once a sufficiently large selfing rate has been reached, the fitness advantages of increasing the selfing rate (twofold automatic transmission advantage and reproductive assurance) might become negligible, which would result in a weaker selection gradient. Moreover, the effective population size is expected to be extremely small at high selfing rates (Glémin 2007), which could further decrease the efficiency of selection (Burgarella and Glémin 2017). In light of these arguments, residual outcrossing could just be a prolonged transient state towards complete selfing.

3. Residual outcrossing has an ecological advantage in the context of pollen discounting and pollen limitation

Finally, besides its advantage in terms of increased recombination, residual outcrossing could also be maintained by ecological factors related to pollination biology. Pollen discounting is a reduction of the male reproductive success through outcrossing due to a reduction in the amount of exported pollen in predominantly selfing plants. It is hypothesized to allow the evolution of stable mixed mating systems (Holsinger 1991; Johnston 1998). Pollen limitation can reinforce the reproductive assurance advantage of selfing because selfing avoids the occurrence of unfertilized seeds when outcross pollen is limited. Indeed, Larson and Barrett (2000) found that self-compatibility and autogamy were associated with reduced pollen limitation. In a genetic model combining pollen discounting and pollen limitation, Porcher and Lande (2005) showed that selfing rates close to but less than one can be maintained.

Genetic and ecological determinism of the selfing rate

These three hypotheses for the maintenance of residual outcrossing outline some of the complexity of the determinism of selfing, which most likely combines both genetic and environmental factors. Autonomous selfing is defined as fertilization between male and female gametes produced by the same flower and can happen (i) in an unopened bud (prior selfing), (ii) at the same time as outcrossing (competing selfing), or (iii) after outcrossing has occurred (delayed selfing) (Lloyd 1992). Floral morphology and flowering phenology impose constraints on pollination and thereby likely affect the selfing rate. High heritability values have been found for floral morphology traits (e.g. Campbell 1996; Ashman and Majetic 2006) and studies of the genetic bases of floral evolution have identified several QTLs associated with floral traits involved in the selfing syndrome (reviewed in Sicard and Lenhard 2011). It is therefore likely that there is some genetic determinism of the selfing rate. Indeed, genetic variation in herkogamy, the spatial separation of style and stigma, was correlated with variation in outcrossing rate in several mixed mating species (Karron *et al.* 1997; Herlihy and Eckert 2007). Demographic factors such as plant density can also have an impact on the selfing rate by increasing the opportunities for outcrossing in high density populations (Karron *et al.* 1995). Finally, environmental variations could also influence residual outcrossing rates through their effects on floral traits such as cleistogamy, which is the production of closed flowers (cleistogamous) as well as open flowers (chasmogamous) within a single plant (Richards 1997). Indeed, variation in the proportion of chasmogamous flowers, depending for example on light intensity and resources availability (Paoletti and Holsinger 1999), was shown to result in variation of outcrossing rate in *Lamium amplexicaule* (Stojanova *et al.* 2014). Another environmental factor, pollinators foraging behavior, was also shown to generate variations in outcrossing rate at the plant level (Barrett *et al.* 1994; Karron and Mitchell 2012). The selfing rate is thus determined by complex interactions of factors, both deterministic and stochastic, and we can expect within population variations, even in the case of predominant selfing.

Aims of the study

Here, we aim to describe thoroughly the variability in mating regime by focussing on the level of residual outcrossing in a natural population of the predominantly selfing species *Medicago truncatula*. *M. truncatula* is an annual plant found in open areas around the Mediterranean Sea. Although highly selfing, it displays a floral characteristic allowing an explosive tripping mechanism (Small 2010), shared among the whole *Medicago* genus. When a pollinator visits the flower, the stigma and anthers are projected on the insect's body, thus coating it with pollen and allowing the stigma to gather pollen

from other flowers. Yet, self-fertilization occurs within the flower bud, before flower opening (prior selfing, Lloyd and Schoen 1992). Nevertheless, complete selfing is rarely observed within population. Selfing rates estimated from allozyme data have been reported to range from 0.65 to 1, with a mean selfing rate of 0.96 (Bataillon and Ronfort 2006). Estimates based on microsatellite data are also strongly skewed towards high rates of selfing and vary between 0.95 and 0.98 (Bonnin *et al.* 2001; Siol *et al.* 2008). The determinism of residual outcrossing in *M. truncatula* is not well understood and it is unclear whether only a few individuals outcross or whether residual outcrossing is homogenous between individuals. Flowers produced early in the season remain in a closed bud longer than later flowers (L. Gay field observation). We thus hypothesize that later flowers open more quickly and are therefore more likely to be outcrossed. Moreover, flowering is continuous in *M. truncatula*, so that flowers are still produced while pods are maturing, resulting in competition for resources between flowers and pods. This could result in a reduced production of self-pollen (or lower quality) later in the flowering season and thus an increase in outcrossing. In this study, we aim to identify the determinisms of residual outcrossing in *M. truncatula* through a precise characterization of the variations in selfing rate at different levels in a natural population: during the flowering season, between genotypes, mother plants, and flowers. As *M. truncatula* natural populations are composed of repeated multilocus genotypes, we are able to test for a genetic determinism of residual outcrossing by examining the variance in selfing rates between maternal multilocus genotypes.

Material and methods

Study population and sampling

The FR3 population is a large and stable population located in southern France (Aude). It typically comprises several thousands of individuals. Previous studies have shown that FR3 harbours substantial genetic diversity between 2004 and 2014 (Jullien *et al.* 2019) and is strongly spatially structured (Siol 2007). The population was monitored during spring 2010. 221 mother plants were randomly chosen in 22 quadrats randomly located in eight different areas (hereafter denominated patches) of the population. For each mother plant, two pods were collected: one early in the season (often the pod formed by the first flower in May, E) and one towards the end of the season (among the last flowers in late June, L). Between one and eight seeds were gathered from these pods, and the effect of early and late pods on the production of seeds was analysed. The collected seeds were germinated the following autumn in the lab in Petri dishes. Germination success was recorded.

Seedlings were grown in a greenhouse until we could collect enough leaf material for DNA extraction for genetic analyses using 20 microsatellite loci (Baquerizot-Audiot et al. 2001).

Microsatellite genotyping

As a first step, equal proportions of leaf material from each of the seedlings collected in a single pod (between 1 and 5, whenever the number of seedlings per pod exceeded five, the seedlings were split in two pools) were pooled together for the extraction and microsatellite genotyping on a subset of 10 loci. For all the bulks in which two alleles, or more, were detected in at least one locus, we reiterated the leaf sampling on each individual of the pool and extracted DNA again from each individual. In a second step, the homozygous bulks and the individuals from heterozygous bulks were genotyped on the full set of 20 microsatellites markers.

DNA was extracted from 200 mg of frozen leaves. Amplification reactions were performed in a final volume of 20 µL and 50 ng of template DNA, 4 pmol of the reverse primer and 1 pmol of the forward primer, 0.2 mM of each deoxynucleotide, 2 mM of $MgCl_2$ and 0.5 unit Taq polymerase. After 5 min at 94°C, 35 cycles were performed of 30 s at 94°C, 1 min at 55 °C and 1 min at 72°C, followed by the final extension step of 7 min at 72 °C. Diluted amplification products were analysed on an ABI prism 3100 Genetic Analyser and results were read using Genemapper 2.5 (Applied Biosystems, Foster City, USA). Individuals with more than 10% missing data were filtered out, resulting in a dataset of 1729 progenies among which 826 came from early pods and 903 from late pods.

Population genetic structure

In order to characterize the spatial genetic structure of the population, one seed per maternal plant was randomly selected. It resulted in a dataset of 209 individuals genotyped on 16 loci after applying a filter to remove individuals and loci with more than 10% missing data. To describe genetic diversity, we computed Nei's gene diversity (H_E), the observed heterozygosity (H_O), the number of alleles per locus (n_A) and the inbreeding coefficient (F_{IS}) with the R package hierfstat (Goudet 2005). In order to assess the spatial structure within the population, we computed pairwise F_{ST} according to Weir and Cockerham (1984) between patches with the package hierfstat. We also used the R package poppr (Kamvar *et al.* 2014) to perform an analysis of molecular variance (AMOVA) with the following hierarchical levels: between patches, between individuals within patches, and within individuals. Significance testing was performed using a randomization test with 1000 permutations. Individuals

were grouped in multilocus genotypes (MLGs) based on their allelic composition using the R package poppr. A threshold error rate of 10% (corresponding to one genotyping error on one locus) was tolerated while assigning individuals to a MLG. Minimum spanning networks based on the genetic distance between MLGs (computed as the proportion of different alleles) were also computed using poppr.

Progeny array analyses

For the following analyses, maternal sibships were defined by specifying the identity of the mother plant for each seed, as well as the pod from which each seed originated.

Selfing rate variation over the flowering season: We used the MLTR software, version 3.2 (Ritland 2002) to compute maximum likelihood estimates of single (t_s) and multilocus (t_m) outcrossing rates. The difference between the two parameters (t_m – t_s) provides an estimation of the contribution of biparental inbreeding, or mating between relatives. Indeed, multilocus estimates exclude apparent selfing due to biparental inbreeding. However, outcrossing between individuals of identical genotypes cannot be detected. The correlation of selfing within progeny arrays (r_s), which measures the normalized variance in outcrossing rates among families, was also estimated. The estimations were conducted using the Newton-Raphton likelihood optimization algorithm and 1000 bootstraps with whole family resampling were performed. Allele frequencies were assumed equal between the pollen cloud and ovule pool. We reiterated the analyses to check for convergence of the model. In order to test for an effect of the flowering time, we first estimated a global outcrossing rate by constraining it to be equal for early and late progenies (pooling together progenies from early and late pods). We then reiterated the estimation separately for the progenies from early and late pods. A likelihood ratio test was performed to compare these two models and test whether outcrossing rates were significantly different between early and late progenies. The LRT was performed by testing whether $\Delta_{dev} = 2(\sum_{i=1}^{N_{group}} lnLik_i - lnLik_{constrained})$ follows a χ^2 distribution with N_{group}-1 degrees of freedom, where N_{group} is the number of subdivisions of the data (Early/Late), $lnLik_i$ is the log-likelihood of the model with independent outcrossing rate for each group and $lnLik_{constrained}$ is the log-likelihood of the model constrained to estimate a single outcrossing rate. The same analysis was also performed within each patch.

Genetic variation of residual outcrossing: We used the software COLONY version 2.0.6.4 (Jones and Wang 2010) to perform further parentage analyses. If the maternal genotypes are unknown, COLONY reconstructs them based on the genotypes of each progeny array. After reconstructing maternal genotypes, the program returns for each seed its probability of being self-fertilized (P_{self}). The full likelihood (FL) method was used as it is the most accurate method available in COLONY (Wang 2012) and we chose to estimate the likelihood with a high precision. Allele frequencies were calculated from the data without updating by accounting for the inferred relationships.

We tested for an effect of the maternal genotype on P_{self} using a generalized linear mixed model (GLMM), assuming a binomial distribution function and a *logit*-link function, with the R package MCMCglmm (Hadfield 2010). Pod type (early or late) was considered as fixed effect, whereas maternal genotype was a random effect. We controlled for the non-genetic maternal and pod effects, as well as the spatial structure using maternal ID, pod ID and patch as random effects. We used the default weakly informative priors in MCMCglmm (Hadfield 2010). The model was run using 50 million iterations, with a burn-in period of 10000 and a thinning interval of 5000 to minimize the autocorrelation between samples and ensure a reasonable effective sample size for each effect. Model convergence and mixing were verified by visual examination of posterior traces and autocorrelation values. Heritability was estimated based on the method described by Villemereuil *et al.* (2016) using the R package QGglmm. Best linear unbiased predictors (BLUPs) for the maternal genotypes were extracted from the posterior distributions and compared with the mean number of seeds per pod per maternal genotype and the mean germination rate per pod per maternal genotype.

Results

Genetic diversity analyses

All the microsatellite loci were polymorphic and the number of alleles per locus ranged from 2 to 17, with a mean number n_A = 8.05 per locus. The mean genetic diversity as measured by H_E was 0.49 and ranged from 0 to 0.85 among loci. The mean F_{IS} value was 0.90 and ranged from 0.83 to 1 among patches (for details per patch, see Table S1), which corresponds to a mean selfing rate of 0.82 for the maternal population. Pairwise F_{ST} between patches were variable, their value ranging from -0.03 to 0.40, Fig 1). The AMOVA showed that 19.4% of the total genetic variation was located between patches (pval = 0.001), 70.4% between individuals within patches (pval = 0.001), and 10.2% within individuals (pval = 0.001).

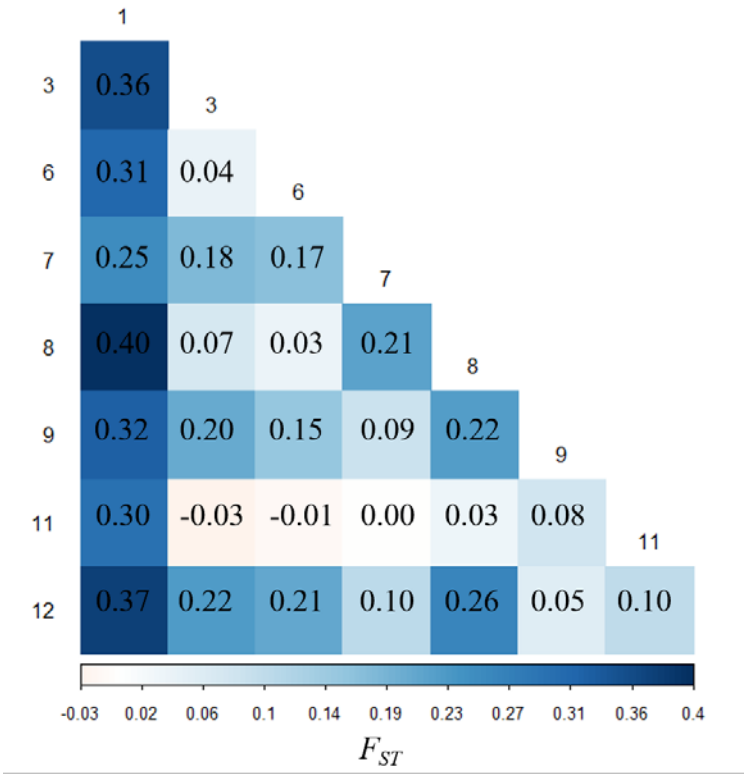


Figure 1: Pairwise F_{ST} between patches

The 209 individuals could be grouped in 119 MLGs, among which 47% were represented by only one individual. The most frequent MLG represented 14% of all the studied individuals. The minimum spanning network on figure 2 shows how those MLGs are distributed among the different patches, as well as how similar the MLGs are with each other. In accordance with results from the population structure analyses (pairwise F_{ST} and AMOVA), the repetitions of a given MLG are clustered together in space with only one MLG (the most frequent one) observed in three different patches. The other MLGs are either found in a single patch or in only two different patches. Figure 2 also shows that some patches are composed of several low frequency MLGs differing by a few loci.

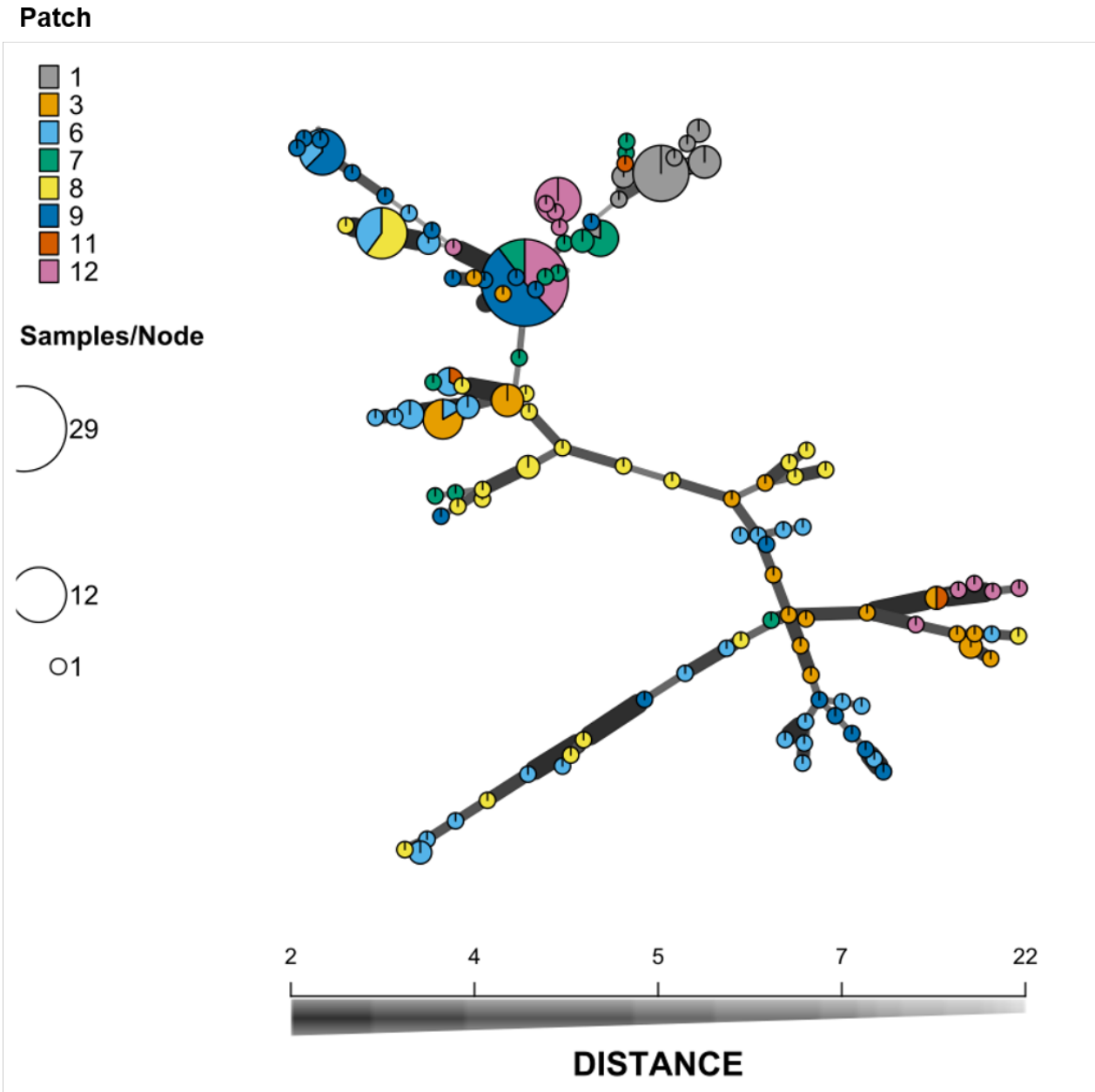


Figure 2: Minimum spanning network of the offspring mulilocus genotypes.
Each circle represents a MLG and the diameter of the circle represents the frequency of the MLG in the overall sample. Multi-coloured circles represent MLGs found in several patches. The length and thickness of the lines linking the MLGs represents their genetic distance computed as the number of different alleles.

Outcrossing rate variation over the flowering season

No difference in seed production between early and late pods was detected (Figure S1). The average germination success rate was about 95.6%, with no difference between seeds from early or

late pod (Figure S2). The outcrossing rate estimated on the whole progeny data was around 10% (see multilocus estimates t_m in Table 1). All the reiterations of the analyses gave similar results, which confirmed that the model had reached convergence. The likelihood ratio test (LRT) showed that the outcrossing rate was significantly higher in late flowers than in early flowers (Δ_{dev} = 62.766; pval = $2.3x10^{-15}$, Table 1). The single locus estimates of the outcrossing rate were slightly lower than the multilocus estimates, which means that moderate biparental inbreeding occurs, i.e. that there are outcrossing events between related individuals. The proportion of these crosses between relatives tended to increase over the flowering season. Moreover, we estimated a non-null correlation of outcrossing within families (r_s), suggesting that outcrossing events were not randomly distributed between mother plants.

Table 1: MLTR estimates of outcrossing rates and results of the likelihood ratio test.
t_m is the multilocus outcrossing rate; t_s is the single locus outcrossing rate; r_s is the correlation of outcrossing within families.

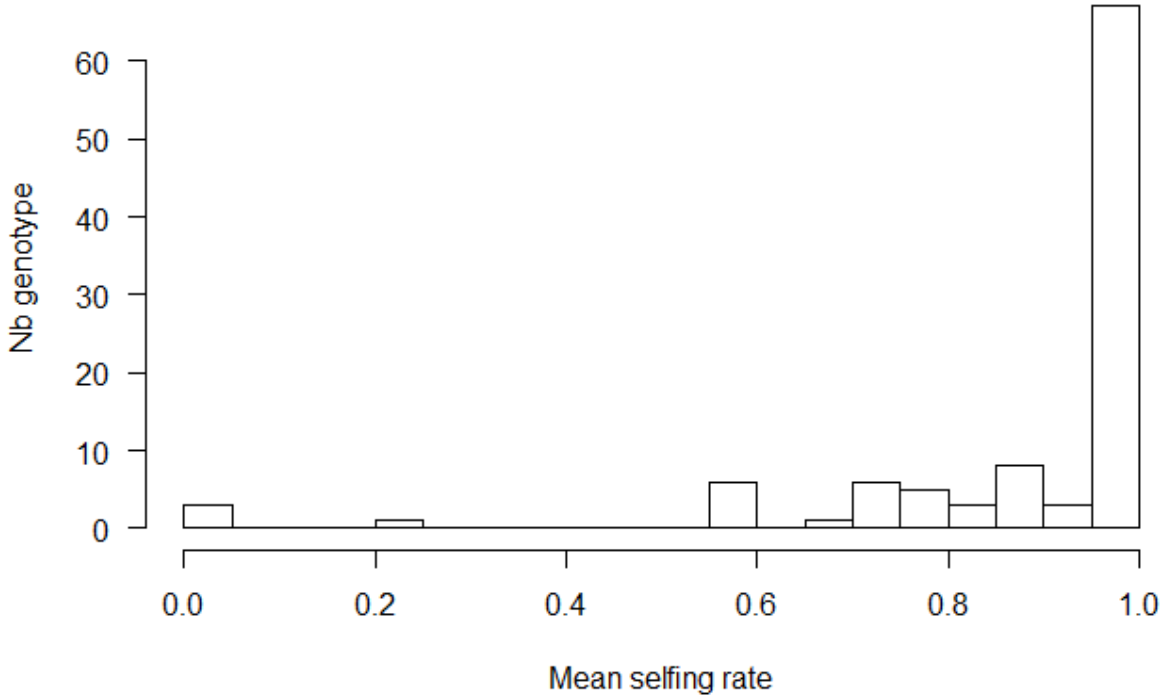
Data	logLik(tm)	t_m (SD)	t_s (SD)	$t_m - t_s$	r_s
Total	-6077,986	0.093 (0.021)	0.025 (0.005)	0.068 (0.018)	0.302 (0.044)
Early	-2666,568	0.083 (0.016)	0.019 (0.004)	0.064 (0.014)	0.368 (0.075)
Late	-3380,035	0.137 (0.025)	0.031 (0.006)	0.106 (0.020)	0.485 (0.065)

Genetic determinism of residual outcrossing

COLONY estimated 103 maternal genotypes although sampling was performed on 222 distinct mother plants, confirming that there are repeated MLGs in the population. A description of the genetic diversity of the maternal population based on the reconstructed maternal genotypes can be found in table S2 and figure S3. In accordance with MLTR, COLONY estimated a global selfing rate of 0.92 (95% CI: 0.91 – 0.94). For each progeny, the probability of being selfed (P_{self}) was either 0 or 1, except for two progenies for which P_{self} = 0.93 and which we converted to 1 in order to obtain binary data to perform the GLMM analysis. The distribution of the average P_{self} per maternal genotype shows that, in accordance with the high selfing rates estimated, most of the offspring are selfing progenies (1605

selfed versus 124 outcrossed seeds). In three instances, mother plants carrying unique genotypes produced completely outcrossed progeny. On the other hand, all the offspring were self-fertilized for 62 genotypes (Figure 3). Most of the reconstructed maternal genotypes exhibited selfing rates higher than 0.6.

Figure 3: Distribution of selfing rates per maternal genotype (n=103), computed as the mean of P_{self} from the progeny of each genotype.



To determine whether there is an effect of the genotype of the mother on the probability of being selfed, P_{self}, we performed a GLMM analysis. No autocorrelation between samplings of the MCMC chains was found and the effective sample sizes for each effect were satisfyingly high (> 9000, Table 2). We found a non-significant decrease in the probability of being selfed for seeds produced later in the flowering season (Table 2). No significant effects of the patch of origin or of the maternal plant were detected. Significant pod and genotype effects were detected (Table 2), indicating that outcrossing events are not randomly distributed among flowers, nor among maternal genotypes. Using the R package QGglmm, we estimated a mean observed probability of being selfed of 0.906. The observed variance was 0.085 among which 0.008 was genetic variance. This resulted in a heritability of 9% for the probability of being selfed. The BLUPs of the probability of being selfed for each maternal genotypes were not correlated with the number of seeds produced nor with the germination rate (Figures S4 and S5).

Table 2: Results of the generalized linear mixed model performed with MCMCglmm on P_{self}. Effects significantly different from zero are in bold. Posterior means are given on the logit scale.

Effect	Type	Effective sample size	Posterior mean of the fixed effect [95% CI]	Posterior mean of the variance of random effects
Intercept	Fixed	9060	6.371 [4.926 – 7.881]	-
Late	Fixed	9998	-0.610 [-1.713 – 0.492]	-
Patch	Random	9998	-	0.706 [$9x10^{-9}$ – 2.4]
Pod	Random	9348	**-**	**7.678 [3.303 – 12.42]**
MotherID	Random	9599	-	2.053 [$5x10^{-8}$ – 5.947]
MotherGT	Random	9221	**-**	**6.003 [1.581 – 10.85]**

Discussion

In this study, we performed a detailed analysis of the mating system of a predominantly selfing species. We confirm that, despite an autogamous mating strategy (sensus Richards 1997, i.e. self-pollination occurs within flower), *M. truncatula* flowers are not fully cleistogamous and occasional outcrossing occurs. Interestingly, we detected a higher rate of outcrossing in later flowers in accordance with our hypothesis that the end of the flowering season offers more outcrossing opportunities, although the signal was less clear when environmental effects were considered. Finally, we also detected between genotype variations of the selfing rate, suggesting that residual outcrossing could be an adaptive trait.

Residual outcrossing and population genetic diversity

The natural population of *M. truncatula* we investigated here presents substantial levels of residual outcrossing (~10%), as well as genetic diversity. Such level of outcrossing is higher than usually found in *M. truncatula* populations (Bonnin *et al.* 2001; Siol *et al.* 2007, 2008). As already reported in other studies on *M. truncatula* (Bonnin *et al.* 1996, 2001; Siol *et al.* 2008), the population is strongly spatially structured over a small spatial scale. The most frequent genotype represented 14% of the studied plants and occurred in three different patches. Higher frequencies of genotype repetitions were found in other *M. truncatula* natural populations (Jullien *et al.* 2019), which highlights the level of multilocus diversity encountered in the FR3 population. Contrary to what was

found by Siol *et al.* (2008) in another *M. truncatula* population, the comparison of single and multilocus estimates of selfing rates revealed biparental inbreeding. Due to the aggregation of MLGs in space, our estimates of residual outcrossing are thus probably underestimated. Because outcrossing events between two plants bearing exactly the same genotype cannot be detected through progeny array analysis, this biparental inbreeding may occur between MLGs differing at a low number of loci, or within families of recombinant MLGs.

Environmental variations of the residual outcrossing rate

Interestingly, we detected a significant variation of the population outcrossing rate along the flowering season. Indeed, when we compared offspring from early and late pods collected on the same mother plant, we found that offspring from flowers produced early in the season had higher selfing rates than offspring from late flowers. When we consider flower, maternal, spatial and genetic effects in the statistical model, the effect of flowering time on the probability of being selfed is no longer significant. Yet, this may be a consequence of high estimation variance for individual probabilities of being selfed (as estimated by Colony). In addition, the flower effect was significant, which implies that outcrossing events were not randomly distributed between pods. Similarly, an increase in outcrossing rate between the first and second flowers was also observed in inflorescences of *Aquilegia buergeriana* var. *oxysepala* (Itagaki *et al.* 2016). The authors also reported variations in the number of pollen grains produced between the two types of flowers which explained the higher selfing rate in the first flowers. In *Lychnis flos-cuculi*, the outcrossing rate was also higher in late-opening flowers because protandry resulted in a lack of outcross pollen for early-opening flowers (Dulya and Mikryukov 2016).

This increase in the outcrossing rates for later flowering times suggests delayed outcrossing in the predominantly selfing *M. truncatula* and in these two mixed-mating species. It contrasts with the more commonly observed delayed selfing (Goodwillie and Weber 2018). Delayed selfing is described as a "best-of-both-worlds" mating system because it combines the advantages of allogamy and selfing through reproductive assurance when outcrossing is not possible. In the case of *M. truncatula*, the situation is different as the species is predominantly selfing, but residual outcrossing always occurs and is slightly more frequent towards the end of the flowering season. Several hypotheses related to flower traits could explain the residual outcrossing variations we observed. First, faster flower development late in the season could speed up the opening of the flowers and provide more opportunities for outcrossing. This contrasts with observations in *Incarvillea sinensis* for which

senescence is associated with a mechanism of "corolla dragging" enabling delayed selfing (Qu *et al.* 2007). As a consequence, higher selfing rates were measured across the flowering season in a natural population of this species, although they remained high during the whole season (Yin *et al.* 2016). Besides the effects on flower corolla, the quantity of pollen or ovules could also vary between early and late flowers. Yet, studies measuring the evolution of the pollen-ovule (P:O) ratios over the flowering season in controlled greenhouse experiments have shown that, contrary to their outcrossing sister species counterpart, selfing species present stable P:O ratios over time (Mazer *et al.* 2009). However, this result was less clear in natural conditions (Delesalle *et al.* 2008). Finally, flower senescence could also reduce pollen viability. Phenotypic plasticity due to the perception of environmental characteristics varying over the flowering season, such as temperature or humidity, may also result in a variation of outcrossing rate.

Besides the effect of flowering time, the local environment could also influence the selfing rate. We did not detect significant spatial effects on the seed's individual probability of being selfed (no significant maternal or patch effects). Nonetheless, studies have shown that environmental conditions could be responsible for variation in selfing rate through for example metal pollution (Dulya and Mikryukov 2016; Mousset *et al.* 2016), pollinator behaviour (Barrett *et al.* 1994; Karron *et al.* 2009; Karron and Mitchell 2012), or resources availability (Waller 1980). Although highly selfing, *M. truncatula* mating system may thus remain plastic to cope with temporally and spatially changing environments.

Genetic determinism of residual outcrossing

The correlation of selfing within families (r_s) we found with the MLTR analysis turned out to be due not to maternal effects, but to a significant genotype effect. Our study thus provides evidence for genetic variation of the selfing rate and consequently of the residual outcrossing rate in this population of *M. truncatula*. However, our diversity analysis highlights shortcomings of the experimental design used to answer this question. Repeated genotypes were found in the population but were strongly aggregated over space, with most of the repeated MLGs located in the same patch. Such strong spatial structure is typical for predominantly selfing species, and has already been described in natural populations of *M. truncatula* (Bonnin *et al.* 1996). Without a randomization step of the genotypes, such a strong spatial structure can create confusion between genetic variance and local environmental effects. Moreover, even though the presence of repeated MLGs allows testing for a genetic determinism of the trait, we nevertheless observed a large proportion of unique maternal

genotypes (47%). This limits our power to partition the variability in selfing rate between genotypic and environmental or maternal effects. Only 9% of the variation of the individual probability of being selfed can be explained by the maternal genotype. This is a stark difference with the genetic contributions to floral traits related to the selfing syndrome where large effect QTLs could be detected (Sicard and Lenhard 2011), and it suggests that environmental effects are predominant. In order to test more thoroughly for a genetic determinism of the residual outcrossing rate, a further study with another experimental design, such as a randomized assay in a common garden, would be required. Such a study in a common garden with randomized transplants was carried out in spring 2015 using eleven *M. truncatula* genotypes from Spain, Corsica and the south of France (detailed in Annexe 1). The whole progenies were sampled in order to estimate outcrossing rates per genotype. However, no outcrossing was detected, probably due to further environmental effects such as plant density. Indeed, germination rates were low and as a consequence, plant density in the experimental design was lower than what is usually observed in natural populations, which has been shown to negatively affect outcrossing rates (Karron *et al.* 1995).

In conclusion, residual outcrossing rate in *M. truncatula* appears to be variable during the flowering season, between flowers and between genotypes, although the selfing rate always remains very high. Our study shows that residual outcrossing has a complex determinism and its variations seem to be largely driven by environmental effects. Studies comparing the development of flowers and their pollen production during the flowering season could be helpful to better understand the temporal variations of outcrossing rate. Overall, this analysis suggests that there could be a genetic component underlying residual outcrossing, but the genetic variance available for this trait in this population is low. Consequently, even if residual outcrossing is beneficial in terms of recombination or pollination, it is unlikely to evolve in this population.

Bibliography

Allard RW and Workman PL. 1963. Population studies in predominantly self-pollinated species. IV. Seasonal fluctuations in estimated values of genetic parameters in Lima bean populations. *Evolution* **17**: 470–80.

Ashman T-L and Majetic CJ. 2006. Genetic constraints on floral evolution: a review and evaluation of patterns. *Heredity* **96**: 343–52.

Baquerizo-Audiot E, Desplanque B, Prosperi JM, and Santoni S. 2001. Characterization of microsatellite loci in the diploid legume *Medicago truncatula* (barrel medic). *Mol Ecol Notes* **1**: 1–3.

Barrett SCH, Harder LD, and Cole WW. 1994. Effects of flower number and position on self-fertilization in experimental populations of *Eichhornia paniculata* (Pontederiaceae). *Funct Ecol* **8**: 526–35.

Bataillon T and Ronfort J. 2006. Evolutionary and ecological genetics of *Medicago truncatula*

Bomblies K, Yant L, Laitinen RA, *et al.* 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLOS Genet* **6**: e1000890.

Bonnin I, Huguet T, Gherardi M, *et al.* 1996. High level of polymorphism and spatial structure in a selfing plant species, *Medicago truncatula* (Leguminosae), shown using RAPD markers. *Am J Bot* **83**: 843–55.

Bonnin I, Ronfort J, Wozniak F, and Olivieri I. 2001. Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). *Mol Ecol* **10**: 1371–83.

Burgarella C and Glémin S. 2017. Population genetics and genome evolution of selfing species (John Wiley & Sons Ltd, Ed). *eLS*: 1–8.

Campbell DR. 1996. Evolution of floral traits in a hermaphroditic plant: field measurements of heritabilities and genetic correlations. *Evolution* **50**: 1442–53.

Charlesworth B. 1980. The cost of sex in relation to mating system. *J Theor Biol* **84**: 655–71.

Charlesworth D. 2006. Evolution of plant breeding systems. *Curr Biol* **16**: R726–35.

Charlesworth D and Charlesworth B. 1995. Quantitative genetics in plants: the effect of breeding system on genetic variability. *Evol Int J Org Evol* **49**: 911–20.

Charlesworth D, Morgan MT, and Charlesworth B. 1990. Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. *Evolution* **44**: 1469–89.

Charlesworth B, Morgan MT, and Charlesworth D. 1991. Multilocus models of inbreeding depression with synergistic selection and partial self-fertilization. *Genet Res* **57**: 177–94.

Clo J, Gay L, and Ronfort J. 2019. How does selfing affect the genetic variance of quantitative traits? An updated meta-analysis on empirical results in angiosperm species. *Evol Accept*.

David JL, Savy Y, and Brabant P. 1993. Outcrossing and selfing evolution in populations under directional selection. *Heredity* **71**: 642–51.

Delesalle VA, Mazer SJ, and Paz H. 2008. Temporal variation in the pollen:ovule ratios of Clarkia (Onagraceae) taxa with contrasting mating systems: field populations. *J Evol Biol* **21**: 310–23.

Dulya OV and Mikryukov VS. 2016. Genetic variation and selfing rate in *Lychnis flos-cuculi* along an industrial pollution gradient. *New Phytol* **209**: 1083–95.

Fisher RA. 1941. Average excess and average effect of a gene substitution. *Ann Eugen* **11**: 53–63.

Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* **177**: 905–16.

Goodwillie C and Weber JJ. 2018. The best of both worlds? A review of delayed selfing in flowering plants. *Am J Bot* **105**: 641–55.

Goudet J. 2005. hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes* **5**: 184–6.

Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *J Stat Softw* **33**: 1–22.

Hartfield M. 2016. Evolutionary genetic consequences of facultative sex and outcrossing. *J Evol Biol* **29**: 5–22.

Hartfield M and Glémin S. 2016. Limits to adaptation in partially selfing species. *Genetics* **203**: 959–74.

Herlihy CR and Eckert CG. 2007. Evolutionary analysis of a key floral trait in *Aquilegia Canadensis* (ranunculaceae): genetic variation in herkogamy and its effect on the mating system. *Evolution* **61**: 1661–74.

Holsinger KE. 1991. Mass-action models of plant mating systems: the evolutionary stability of mixed mating systems. *Am Nat* **138**: 606–22.

Igic B and Kohn JR. 2006. The distribution of plant mating systems: study bias against obligately outcrossing species. *Evol Int J Org Evol* **60**: 1098–103.

Itagaki T, Kimura MK, Maki M, and Sakai S. 2016. Differential self-fertilization rates in response to variation in floral traits within inflorescences of *Aquilegia buergeriana* var. oxysepala (Ranunculaceae). *Bot J Linn Soc* **181**: 294–304.

Johnston MO. 1998. Evolution of intermediate selfing rates in plants: pollination ecology versus deleterious mutations. *Genetica* **102**: 267.

Jones OR and Wang J. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Mol Ecol Resour* **10**: 551–5.

Jullien M, Navascués M, Ronfort J, *et al.* 2019. Structure of multilocus genetic diversity in predominantly selfing populations. *Heredity*: 1.

Kahler AL, Clegg MT, and Allard RW. 1975. Evolutionary changes in the mating system of an experimental population of barley (Hordeum vulgare L.). *Proc Natl Acad Sci U S A* **72**: 943–6.

Kamran-Disfani A and Agrawal AF. 2014. Selfing, adaptation and background selection in finite populations. *J Evol Biol* **27**: 1360–71.

Kamvar ZN, Tabima JF, and Grünwald NJ. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**.

Karron JD, Holmquist KG, Flanagan RJ, and Mitchell RJ. 2009. Pollinator visitation patterns strongly influence among-flower variation in selfing rate. *Ann Bot* **103**: 1379–83.

Karron JD, Jackson RT, Thumser NN, and Schlicht SL. 1997. Outcrossing rates of individual *Mimulus ringens* genets are correlated with anther–stigma separation. *Heredity* **79**: 365–70.

Karron JD and Mitchell RJ. 2012. Effects of floral display size on male and female reproductive success in *Mimulus ringens*. *Ann Bot* **109**: 563–70.

Karron JD, Thumser NN, Tucker R, and Hessenauer AJ. 1995. The influence of population density on outcrossing rates in *Mimulus ringens*. *Heredity* **75**: 175–80.

Lande R and Schemske DW. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**: 24–40.

Larson BMH and Barrett SCH. 2000. A comparative analysis of pollen limitation in flowering plants. *Biol J Linn Soc* **69**: 503–20.

Lloyd DG. 1992. Self- and cross-fertilization in plants. II. The selection of self- fertilization. *Int J Plant Sci* **153**: 370–80.

Lloyd DG and Schoen DJ. 1992. Self- and cross-fertilization in plants. I. Functional dimensions. *Int J Plant Sci* **153**: 358–69.

Mazer SJ, Dudley LS, Delesalle VA, *et al.* 2009. Stability of pollen–ovule ratios in pollinator-dependent versus autogamous Clarkia sister taxa: testing evolutionary predictions. *New Phytol* **183**: 630–48.

Mousset M, David P, Petit C, *et al.* 2016. Lower selfing rates in metallicolous populations than in non-metallicolous populations of the pseudometallophyte *Noccaea caerulescens* (Brassicaceae) in Southern France. *Ann Bot* **117**: 507–19.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res Mol Mech Mutagen* **1**: 2–9.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–9.

Paoletti and Holsinger. 1999. Spatial patterns of polygenic variation in *Impatiens capensis*, a species with an environmentally controlled mixed mating system. *J Evol Biol* **12**: 689–96.

Porcher E and Lande R. 2005. The evolution of self-fertilization and inbreeding depression under pollen discounting and pollen limitation. *J Evol Biol* **18**: 497–508.

Qu R, Li X, Luo Y, *et al.* 2007. Wind-dragged corolla enhances self-pollination: a new mechanism of delayed self-pollination. *Ann Bot* **100**: 1155–64.

Richards AJ. 1997. Plant Breeding Systems. London: Chapman & Hall.

Ritland K. 2002. Extensions of models for the estimation of mating systems using *n* independent loci. *Heredity* **88**: 221–8.

Sicard A and Lenhard M. 2011. The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann Bot* **107**: 1433–43.

Siol M. 2007. Organisation de la diversité dans les populations autogames : études empiriques chez *Medicago truncatula* et utilisation pour l'analyse des caractères quantitatifs *in natura*.

Siol M, Bonnin I, Olivieri I, *et al.* 2007. Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual *Medicago truncatula*. *J Evol Biol* **20**: 2349–60.

Siol M, Prosperi J-M, Bonnin I, and Ronfort J. 2008. How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. *Heredity* **100**: 517–25.

Small E. 2010. Alfalfa and relatives: Evolution and classification of Medicago. Canada: NRC Research Press.

Stojanova B, Cheptou P-O, and Maurice S. 2014. Does cleistogamy variation translate into outcrossing variation in the annual species *Lamium amplexicaule* (Lamiaceae)? *Plant Syst Evol* **300**: 2105–14.

Villemereuil P de, Schielzeth H, Nakagawa S, and Morrissey M. 2016. General methods for evolutionary quantitative genetic inference from generalized mixed models. *Genetics* **204**: 1281–94.

Waller DM. 1980. Environmental determinants of outcrossing in *Impatiens Capensis* (balsaminaceae). *Evolution* **34**: 747–61.

Wang J. 2012. Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics* **191**: 183–94.

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–70.

Yin G, Barrett SCH, Luo Y-B, and Bai W-N. 2016. Seasonal variation in the mating system of a selfing annual with large floral displays. *Ann Bot* **117**: 391–400.

Supporting Information

The following Supporting Information is available for this article:

Supporting Figure S1 Distribution of the number of seeds per pod

Supporting Figure S2 Distribution of the germination rate for early or late pods

Supporting Table S1 Genetic diversity within the experimental patches computed from the progeny genotypes

Supporting Table S2 Single and multilocus genetic diversity within the experimental patches of the maternal genotypes as inferred by COLONY

Supporting Figure S3 Minimum spanning network based on the maternal genotypes inferred by COLONY

Supporting Figure S4 Best linear unbiased predictors (BLUP) of the maternal genotypes for the number of seeds per pod

Supporting Figure S5 Best linear unbiased predictions (BLUP) of the maternal genotypes for the mean germination rate per pod

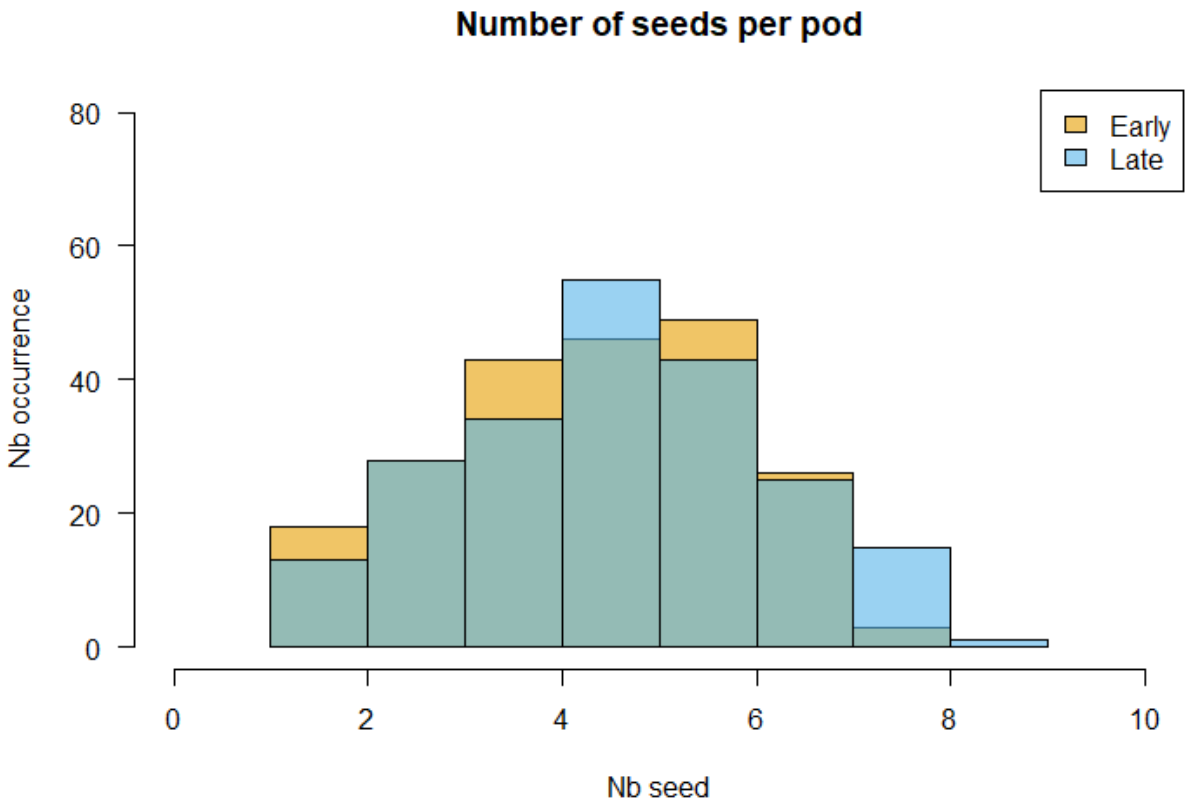


Figure S1: Distribution of the number of seeds per pod.
Early pods are represented in orange and late pods in blue. A Pearson's Chi-squared test revealed no significant difference between the number of seeds in pods produced early in the flowering season and pods produced at the end of the flowering season.

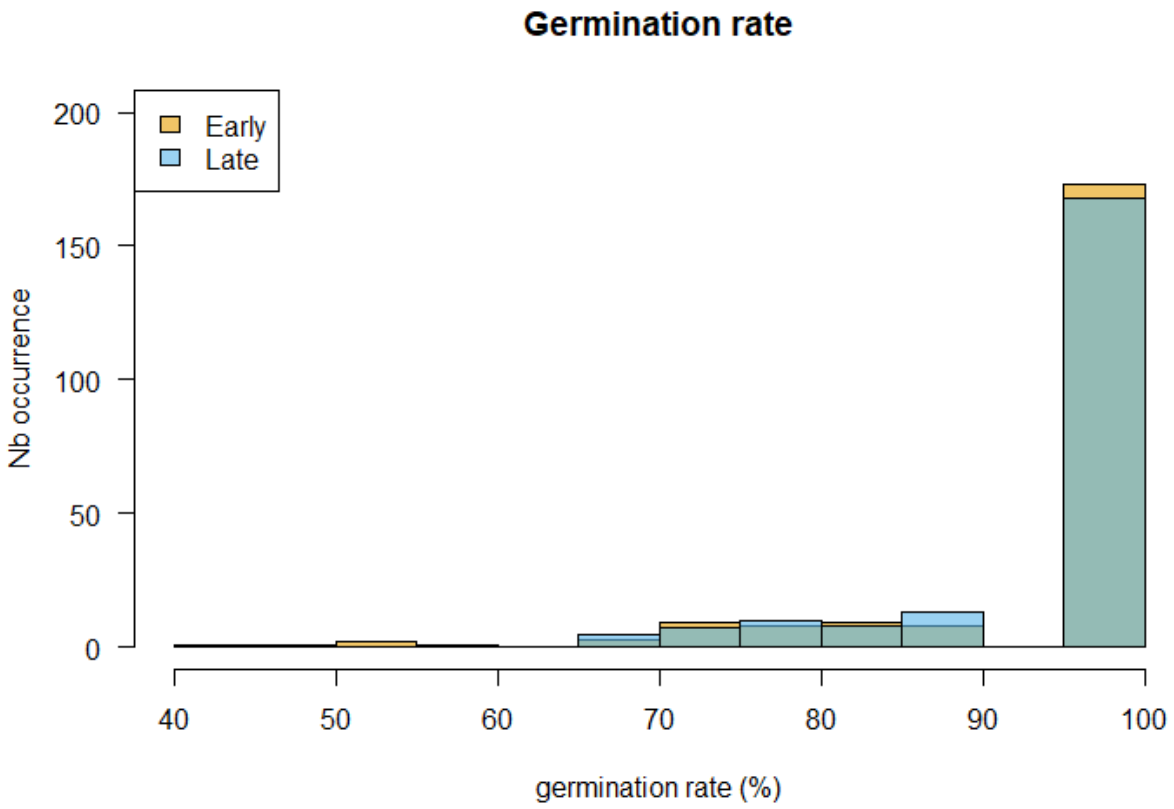


Figure S2: Distribution of the germination rates for early (orange) or late (blue) pods.
A Pearson's Chi-squared test revealed no significant difference between the germination rates of seeds produced early or late in the flowering season.

Table S1: Genetic diversity within the experimental patches computed from the progeny genotypes.

N_{plant} is the number of maternal plants sampled in the patch, N is the number of offspring in each patch; H_E is Nei's genetic diversity; H_O is the observed heterozygosity; F_{IS} is the inbreeding coefficient; *nMLG* is the number of MLGs; *eMLG* is the rarefied number of MLGs; λ is Simpson's multilocus diversity index; *singleMLG* is the proportion of unique MLGs; and *MFMLG* is the frequency of the most frequent MLG.

Patch	N_{plant}	N	H_E	H_O	F_{IS}	n_{MLG}	*eMLG* (SD)	λ	*singleMLG*	*MFMLG*
1	29	221	0.31	0.01	0.96	13	7.03 (1.29)	0.72	0.01	0.48
3	28	212	0.59	0.06	0.91	64	14.85 (1.96)	0.93	0.19	0.18
6	41	320	0.59	0.08	0.87	114	16.52 (2.00)	0.95	0.26	0.14
7	19	156	0.49	0.08	0.84	55	13.65 (1.99)	0.92	0.28	0.19
8	27	212	0.57	0.08	0.83	77	14.48 (2.07)	0.90	0.26	0.27
9	44	300	0.42	0.04	0.90	60	10.49 (2.02)	0.79	0.11	0.42
11	3	23	0.55	0.00	1.00	3	3.00 (0.00)	0.65	0.00	0.39
12	30	285	0.44	0.04	0.90	39	7.78 (1.76)	0.72	0.08	0.46
Total	221	1729	0.49	0.05	0.90	415	-	0.96	0.16	0.17

Table S2: Single and multilocus genetic diversity within the experimental patches of the maternal genotypes as inferred by COLONY.

N_{plant} is the number of maternal plants sampled in the patch; H_E is Nei's genetic diversity; H_O is the observed heterozygosity; F_{IS} is the inbreeding coefficient; *nMLG* is the number of MLGs; *eMLG* is the rarefied number of MLGs; λ is Simpson's multilocus diversity index; *singleMLG* is the proportion of unique MLGs; and *MFMLG* is the frequency of the most frequent MLG.

Patch	N_{plant}	H_E	H_O	F_{IS}	*nMLG*	*eMLG* (SD)	λ	*singleMLG*	*MFMLG*
1	29	0.31	0.01	0.97	8	4.59 (1.02)	0.68	0.14	0.52
3	28	0.61	0.14	0.78	16	7.73 (1.09)	0.90	0.39	0.18
6	41	0.61	0.17	0.72	25	8.56 (1.01)	0.94	0.41	0.15
7	19	0.54	0.18	0.69	14	8.41 (0.92)	0.91	0.58	0.16
8	27	0.57	0.21	0.67	15	7.84 (1.04)	0.90	0.26	0.22
9	44	0.52	0.13	0.75	20	6.58 (1.30)	0.83	0.34	0.36
11	3	0.78	0.00	1.00	3	3.00 (0.00)	0.67	1.00	0.33
12	30	0.46	0.07	0.83	9	4.75 (1.09)	0.71	0.17	0.47
Total	221	0.54	0.11	0.79	98	-	0.96	0.29	0.15

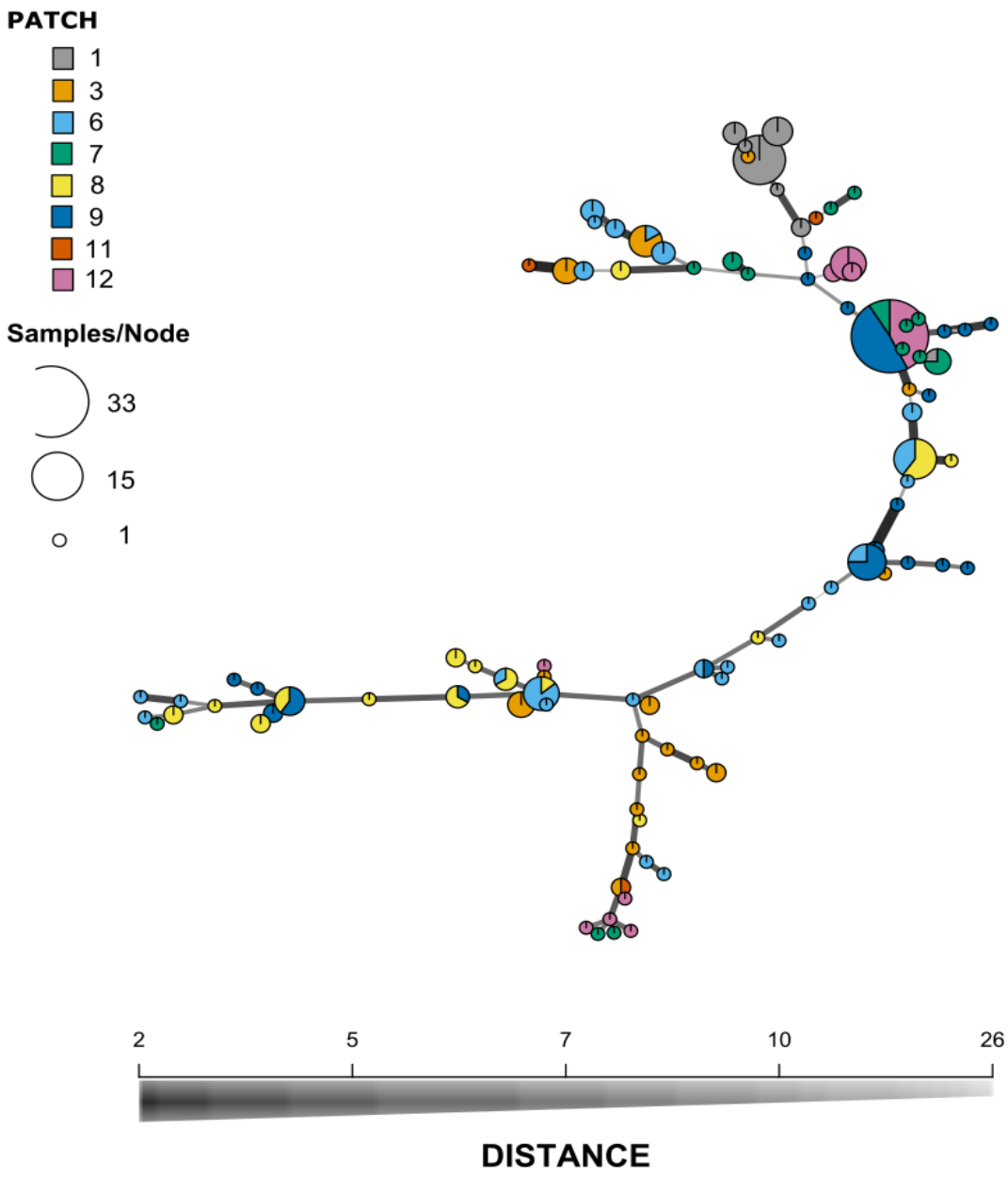


Figure S3: Minimum spanning network based on the maternal genotypes inferred by COLONY. Each circle represents a MLG and the diameter of the circle represents the frequency of the MLG in the overall sample. Multi-coloured circles represent MLGs found in several patches. The length and thickness of the lines linking the MLGs represents their genetic distance computed as the number of different alleles.

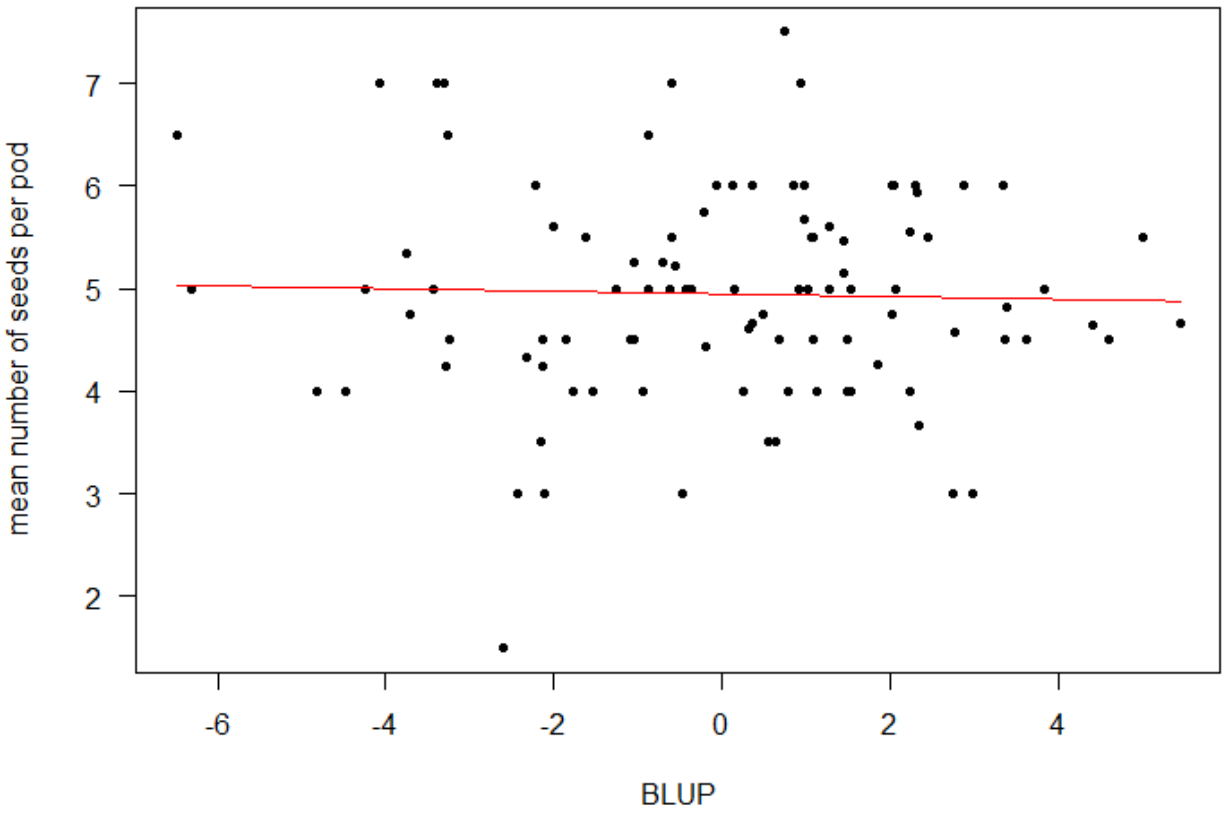


Figure S4: Best linear unbiased predictors (BLUP) of the maternal genotypes for the number of seeds per pod.

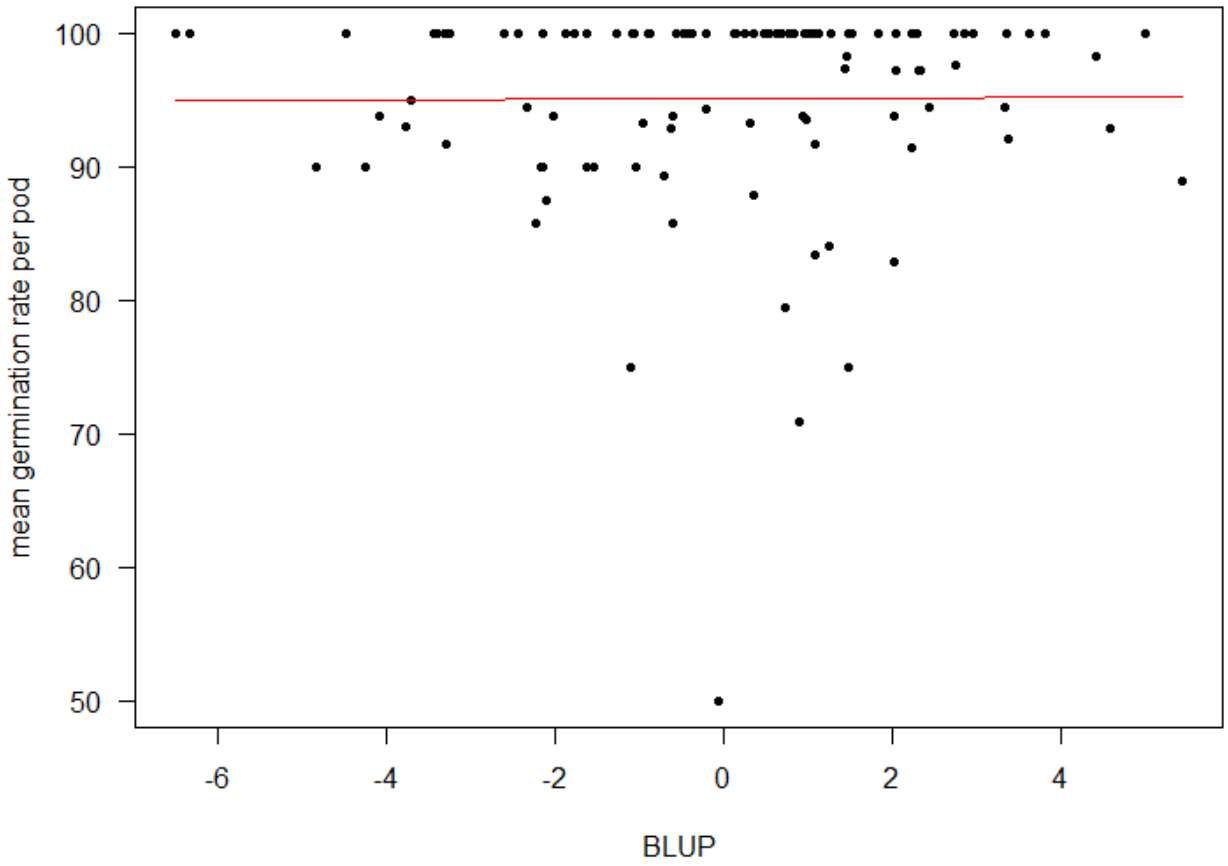


Figure S5: Best linear unbiased predictions (BLUP) of the maternal genotypes for the mean germination rate per pod.

CHAPITRE 2 :

STRUCTURE DE LA DIVERSITE MULTILOCUS EN REGIME AUTOGAME

Présentation générale

Nous avons pu voir en Introduction et dans le Chapitre 1 que les populations autogames ont une structure de diversité particulière avec quelques génotypes multilocus répétés en haute fréquence et plusieurs génotypes uniques (parfois nombreux). Allard (1975), puis Avise et Tatarenkov (2012), ont suggéré que cette organisation traduit les effets de la sélection qui fait monter en fréquence les génotypes les mieux adaptés localement. Cependant, nous ne disposons pas d'attendus analytiques neutres pour la diversité multilocus et on ignore si la dérive génétique, attendue forte, peut générer de tels patrons. La structure multilocus observée dans la littérature des populations autogames (Introduction) et chez *Medicago truncatula* (Chapitre 1), ainsi que les contraintes imposées par l'autofécondation (non indépendance entre locus), nous laissent supposer que la diversité multilocus est particulièrement informative chez les populations autogames. Pourtant, il n'existe pas d'attendus pour les effets sur la diversité multilocus d'évènements démographiques extrêmes de type goulots d'étranglement ou extinction-recolonisation, qui sont supposés fréquents dans les populations autogames. On ignore également quelle pourrait être leur influence sur l'estimation des tailles efficaces, en particulier s'ils pourraient expliquer les très faibles estimations rapportées par la littérature sur les populations autogames.

Dans ce chapitre, nous proposons une approche par simulations afin de (i) fournir des attendus neutres pour des indices de diversité multilocus et tester si des scénarios neutres peuvent générer les patrons de diversité caractéristiques des populations autogames ; (ii) évaluer la pertinence de combiner des indices mono- et multilocus pour distinguer les effets de l'autofécondation, de la taille de population et de scénarios plus complexes (migration, admixture, goulots d'étranglement, extinction-recolonisation) ; (iii) utiliser les variations de fréquences alléliques au cours du temps pour voir si l'on peut estimer des valeurs de taille efficace aussi faibles que dans la littérature. Nous testons la pertinence de notre approche de simulations en comparant nos résultats avec des données temporelles sur neuf populations naturelles de *Medicago truncatula*.

Nos simulations montrent que des génotypes multilocus répétés apparaissent en régime autogame et qu'il est donc possible d'expliquer la structure multilocus des populations autogames par des scénarios neutres de population isolée de petite taille avec de forts taux d'autofécondation (> 95%). Par ailleurs, nous montrons que si la taille efficace est un bon descripteur de la diversité monolocus, la diversité multilocus quant à elle est affectée plus fortement, probablement à cause de la non-indépendance entre locus qui s'installe avec l'autofécondation. Les résultats montrent aussi que la migration restaure la diversité monolocus plus vite que la diversité multilocus. Ainsi, les indices de

diversité multilocus basés sur la fréquence des MLGs ou sur leur similarité, évalués conjointement avec la diversité monolocus, permettent de différencier les effets de l'autofécondation et d'évènements démographiques (goulots d'étranglement et migration). Enfin, l'utilisation de données temporelles se révèle informative, notamment pour identifier de potentiels évènements d'extinction-recolonisation. La comparaison des données empiriques sur *M. truncatula* avec les scénarios démographiques simulés montre que nos simulations reproduisent bien certains des patrons de diversité observés dans les populations naturelles. Les indice de diversité multilocus examinés dans ce chapitre sont donc pertinents pour étudier l'histoire contemporaine des populations autogames.

Les résultats obtenus sont présentés et discutés de manière détaillée dans le manuscrit ci-dessous.

Heredity
https://doi.org/10.1038/s41437-019-0182-6

the geneticssociety

ARTICLE

Structure of multilocus genetic diversity in predominantly selfing populations

Margaux Jullien[1] · Miguel Navascués[2,3] · Joëlle Ronfort[1] · Karine Loridon[1] · Laurène Gay[1]

Received: 11 September 2018 / Revised: 28 December 2018 / Accepted: 8 January 2019

Abstract

Predominantly selfing populations are expected to have reduced effective population sizes due to nonrandom sampling of gametes, demographic stochasticity (bottlenecks or extinction–recolonization), and large scale hitchhiking (reduced effective recombination). Thus, they are expected to display low genetic diversity, which was confirmed by empirical studies. The structure of genetic diversity in predominantly selfing species is dramatically different from outcrossing ones, with populations often dominated by one or a few multilocus genotypes (MLGs) coexisting with several rare genotypes. Therefore, multilocus diversity indices are relevant to describe diversity in selfing populations. Here, we use simulations to provide analytical expectations for multilocus indices and examine whether selfing alone can be responsible for the high-frequency MLGs persistent through time in the absence of selection. We then examine how combining single and multilocus indices of diversity may be insightful to distinguish the effects of selfing, population size, and more complex demographic events (bottlenecks, migration, admixture, or extinction–recolonization). Finally, we examine how temporal changes in MLG frequencies can be insightful to understand the evolutionary trajectory of a given population. We show that combinations of selfing and small demographic sizes can result in high-frequency MLGs, as observed in natural populations. We also show how different demographic scenarios can be distinguished by the parallel analysis of single and multilocus indices of diversity, and we emphasize the importance of temporal data for the study of predominantly selfing populations. Finally, the comparison of our simulations with empirical data on populations of *Medicago truncatula* confirms the pertinence of our simulation framework.

Introduction

Most angiosperms are hermaphrodite (70%; Yampolsky and Yampolsky 1922). This co-occurrence of both male and female reproductive organs on the same individual allows self-pollination and, in the absence of self-incompatibility mechanisms, self-fertilization. Indeed, about 40% of flowering plants do self at various rates, and about 15% of them reproduce predominantly through selfing, with outcrossing rates lower than 10% (Igic and Kohn 2006). Such high selfing rates are expected to have strong consequences on the genetic diversity of natural populations and on its organization.

Theoretical studies have examined the consequences of selfing for the genetic diversity of populations, particularly in terms of effective population size N_e. The effective population size is defined as the size of an ideal Wright–Fisher population experiencing the same rate of genetic drift as the population under consideration (Crow and Kimura 1970). As reviewed in Charlesworth (2009), the effective population size can be affected by several factors, including the mating system. Self-fertilization reduces the number of independent gametes sampled for reproduction, which directly decreases N_e (Pollak 1987). Demographic events are also likely to affect the effective size of selfing populations where founder effects can be frequent, e.g., through the establishment of a new population by a single

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-019-0182-6) contains supplementary material, which is available to authorized users.

✉ Margaux Jullien
margauxjullien3@gmail.com

[1] AGAP, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

[2] CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Université de Montpellier, Montpellier, France

[3] Institut de Biologie Computationnelle IBC, Montpellier, France

Published online: 22 January 2019

Table 1 Ranges of temporal F_{ST} and estimates of effective population sizes in the literature of predominantly selfing populations

Reference	Species	Number of populations	τ	F_{ST}	$\widehat{N_e}$	H_E
Siol et al. (2007)	*Medicago truncatula*	4	3–6	–	9.8–153[a]	0.18–0.47
Gomaa et al. (2011)	*Arabidopsis thaliana*	9	1–4	0.033–0.264	1–12[a]	0.00–0.28
Frachon et al. (2017)	*Arabidopsis thaliana*	1	8	0.0215	91[b]	–
Bomblies et al. (2010)	*Arabidopsis thaliana*	14	1	0.03–0.129	2–8[b]	0.00–0.32
Lundemo et al. (2009)	*Arabidopsis thaliana*	11	1	0–0.626	1–250[b]	0.00–0.17
Meunier et al. (2004)	*Lymnea truncatula*	5	1–4	0.002–0.47	1–25[b]	0.05–0.49
Trouvé et al. (2005)	*Galba truncatula*	6	1–3	0.005–0.175	1–65[b]	0.00–0.54
Viard et al. (1997)	*Bulinus truncatus*	12	1–6	0.06–1	1–27[b]	0.10–0.75
Barrière and Félix (2007)	*Caenorabditis elegans*	7	2–72[c]	0.159–0.963	1–95[b]	0.00–0.88

The temporal estimate of N_e, $\widehat{N_e}$, is given as the number of diploid individuals. τ stands for the number of generations between the two samplings, H_E is the range of genetic diversity across populations. When not directly computed in the study, $\widehat{N_e}$ is computed from temporal F_{ST} values according to Frachon et al. (2017) (see Methods).

[a]Estimated in the literature using F_C

[b]Estimated from temporal F_{ST}

[c]Using generation time in laboratory conditions of 1 generation every 2 weeks

individual (Baker 1967). In addition, according to the "dead-end hypothesis" (Stebbins 1957), selfing populations are expected to accumulate deleterious mutations (Lynch et al. 1995; Abu Awad et al. 2014), and could lack the genetic diversity required to adapt to changing environmental conditions (Charlesworth and Charlesworth 1995; Lande and Porcher 2015; Abu Awad and Roze 2018). Therefore, we can expect frequent catastrophic demographic events, with strong bottlenecks or even extinctions followed by recolonization accompanied by strong founder effects (Schoen and Brown 1991). Such metapopulation dynamics (Ingvarsson 2002) are expected to further reduce N_e (Pannell and Charlesworth 2000). Finally, selective effects can also reduce the effective size in selfing populations. Indeed, the increase in homozygosity due to selfing (Caballero and Hill 1992) reduces the effective recombination (Golding and Strobeck 1980; Nordborg 2000). The reduction of N_e due to selective sweeps or background selection is thus expected to extend by hitchhiking to larger linked regions of the genome or, in some extreme cases, to the whole genome (Charlesworth 2009). These different effects of the mating system on the effective population size have been summarized by Glémin (2007) as

$$N_e = \frac{\alpha N}{1+F}, \quad (1)$$

where N is the census size, $F = \sigma/(2 - \sigma)$ is Wright's equilibrium fixation index with a selfing rate σ, and α summarizes the reduction of the effective population size due to demographic effects and hitchhiking ($\alpha \in [0; 1]$). Overall, this formalizes the predominant role of genetic drift in shaping genetic diversity within selfing populations.

Empirical observations confirm these theoretical expectations, notably through estimations of N_e and genetic diversity in natural predominantly selfing populations. The contemporary effective size of a population can be estimated based on the temporal changes in allele frequency (F_C, Waples 1989) as was done by Siol et al. (2007) and Gomaa et al. (2011) (Table 1). More recently, Frachon et al. (2017) extended the original method to use the temporal differentiation (F_{ST} instead of F_C). When we apply this method to published temporal F_{ST} values, we find that N_e estimates in predominantly selfing populations are generally lower than 100 (Table 1). For comparison, Palstra and Ruzzante (2008) reviewed temporal estimates of N_e in 83 studies concerning different taxa (including the aforementioned selfing species) and found a median N_e of 260. Published estimates of N_e (or temporal F_{ST}) therefore support theoretical predictions of a reduced effective population size in predominantly selfing populations compared to outcrossing ones. Reviews on allozyme data (Schoen and Brown 1991; Hamrick and Godt 1997), as well as on sequence polymorphism (Glémin et al. 2006), showed convincing evidence for lower genetic diversity (as measured by H_E) in predominantly selfing populations compared to outcrossing ones. Yet, Schoen and Brown (1991) also reported larger variability in levels of genetic diversity among selfing populations than among outcrossing populations. Empirical estimates for genetic diversity reported in the temporal studies we reviewed for estimates of effective size are consistent with these findings (Table 1), with some monomorphic populations ($H_E = 0$) along with highly diverse populations (up to $H_E = 0.88$). Substantial genetic diversity can therefore persist in some predominantly selfing populations, suggesting that evolutionary forces other

than genetic drift may play a significant role in shaping the genetic diversity of natural populations reproducing predominantly through selfing.

Besides the level of genetic diversity, the within-population genetic structure is also affected by selfing. Indeed, due to reduced effective recombination (Nordborg 2000), we expect populations to be organized in homozygous lineages, where some multilocus genotypes (thereafter called MLGs) can reach a high frequency (Hartfield et al. 2017). Empirical studies confirm this expectation, for example in the nearly obligate selfing species *Lobelia inflata* where substantial genetic differentiation was found between completely homozygous lineages co-occurring within populations (Hughes and Simons 2015). Similar population genetic structures have also been observed in several other predominantly selfing plant or animal species (e.g., Barrière and Félix 2007; Montesinos et al. 2009; Siol et al. 2008). Because this multilocus genetic structure is specific to predominantly selfing populations, we believe that the comparison of single and multilocus indices of diversity can be relevant to separate the effects of selfing and genetic drift due to small population sizes or demographic processes such as population size changes or migration.

Along with the effect of selfing, Allard (1975) interpreted this distinctive genetic structure of diversity in repeated genotypes as a result of selection favoring locally adapted MLGs, which can then reach high frequencies in the population. The reduction in gene flow through pollen dispersal in selfing populations could indeed promote local adaptation, as suggested by Hereford (2010), even if no significant effect of the mating system on local adaptation was found in his meta-analysis. However, given the strong incidence of genetic drift expected in populations undergoing high and recurrent selfing, the efficacy of selection is questionable and the role of local adaptation as opposed to genetic drift in shaping the MLGs composition of these populations remains to be assessed. We propose to test whether neutral processes alone can be responsible for the peculiar genetic structure observed in highly selfing populations (high-frequency MLGs) in the absence of selection. Answering this question requires analytical predictions for multilocus diversity indices such as those available for single locus diversity. Such predictions are lacking, and little is known about the expected range of values for multilocus diversity indices under high and recurrent selfing. Overall, a formal description of the multilocus genetic diversity expected in predominantly selfing populations evolving under neutral scenarios is still lacking and limits interpretations of empirical data.

In addition, the organization of predominantly selfing populations in MLG lineages offers the possibility to follow the changes in MLG frequencies through time. Such temporal surveys can give additional insight into the processes shaping diversity. In particular, they are useful to measure the strength of genetic drift through the estimation of the effective population size (e.g., Table 1). Although data gathered from time series are frequent in experimental populations evolving under artificial selection, they are more rarely available for natural populations (Bailey and Bataillon 2016), in particular predominantly selfing populations (Table 1). Temporal studies of natural selfing populations have found that MLGs can be maintained within population over time (Siol et al. 2007; Bomblies et al. 2010; Gomaa et al. 2011). Nonetheless, the last two studies have also found populations in which all the MLGs changed over time, which led the authors to propose extinction–recolonization dynamics to explain their observations. Yet, because there are no theoretical predictions for the trajectory of MLG frequencies over time in populations evolving neutrally, it is not clear how demographic events (from changes in population size to extinction–recolonization) affect the persistence of multilocus genotypes over time without selection.

Here, we propose to use simulations to explore how predominant selfing shapes single locus and multilocus genetic diversity in neutrally evolving populations. The goals of our study are threefold. First, we use simulations to provide neutral expectations for multilocus indices of diversity and determine whether neutral scenarios can explain the peculiar population genetic structure (with high frequency and persistent MLGs) observed in predominantly selfing species without selection. Second, we examine how combining single and multilocus indices of diversity may be insightful when studying the evolutionary trajectory of predominantly selfing populations to distinguish the effects of selfing, population size, and more complex demographic events such as bottlenecks, migration, admixture, or extinction–recolonization. Third, we use changes in allele frequency through time to examine whether we can estimate effective sizes as small as those reported in the literature, and we consider the influence of complex demographic scenarios such as bottlenecks, admixture, and extinction–recolonization on the trajectory of MLG frequencies through time. We compare our simulation results with observations from temporal data on nine populations of the highly selfing plant species *Medicago truncatula*. These nine temporal datasets for *M. truncatula* natural populations can be viewed as a reality check (independent iterations of evolution in a selfing population across 20 generations). As such, they validate the pertinence of our simulation framework as we find genetic diversity patterns similar to our simulations.

Material and methods

Simulation model and scenarios explored

We performed individual-based simulations of diploid hermaphroditic populations using SLiM 2.5 (Haller and Messer 2017). In order to be able to qualitatively compare the simulation results with our empirical data, we fixed some simulation parameters such as the type of genetic markers, the number of loci, and the time span between the temporal samples. We simulated the evolution of 20 independent loci (with a recombination rate of 0.5). SLiM output was processed in R (R Core Team 2018) in order to transform the mutations that occurred on each of the 20 predefined loci into microsatellite allele sizes following the stepwise mutation model (Ohta and Kimura 1973). Briefly, we randomly attributed an effect to each mutation (±1 repeat unit) and the effects of all the mutations occurring at a given locus in a given individual were summed in order to obtain microsatellite allele size. Mutations were neutral and occurred at a rate $\mu = 10^{-3}$ per generation and per locus, which is a realistic rate for plant microsatellites (Thuillet et al. 2002; Marriage et al. 2009). To produce the next generation, new zygotes were built as a combination of two gametes sampled either from two different individuals for outcrossing, or from the same individual for selfing, according to a fixed selfing rate (σ). Each simulation comprised two periods. A first period of 25 N generations (with N the demographic population size, measured as the number of diploid individuals) allowed the populations to reach the mutation-drift equilibrium. At this stage (time $t_0 = 0$), 100 diploid individuals were randomly sampled. Twenty generations later (t_{20}), a second sample of 100 individuals was drawn to obtain temporal sampling.

Five demographic scenarios were considered. In the first one, we simulated a single isolated population with a constant demographic size N. Four demographic population sizes were considered: $N \in [50; 100; 250; 1000]$ and combined to five different values of selfing rate (σ): 0 (completely outcrossing population), 0.5 (partially selfing population), 0.95, 0.98 (predominantly selfing population), and 1 (completely selfing population). To disentangle the effects of selfing from those of genetic drift, we also simulated populations of the same effective size with different selfing rates by setting $N = 2N_e/(2 - \sigma)$ for $\sigma \in [0; 0.5; 0.95; 0.98; 1]$ and $N_e \in [100; 250]$. To examine the impact of sampling effect, we reiterated the analysis for one of the simulations with $N = 100$ and $\sigma \in [0; 0.95; 1]$ after reducing the sample size at t_0 and t_{20} to 5, 10, 20, 30, or 50 individuals. Each sampling was repeated independently 100 times. In the following scenarios, we considered only predominantly selfing populations ($\sigma = 0.95$). In a second scenario, we explored the combined effects of predominant selfing and a bottleneck. To this aim, we simulated an isolated population of size $N = 250$ and a selfing rate $\sigma = 0.95$ undergoing at time t_{10} a drastic demographic size reduction (to $N' = 1$, 5, or 25 diploid individuals) for one generation. In a third scenario, we evaluated the effects of migration by simulating an island model with ten subpopulations of constant size $N \in [50; 100; 250]$ exchanging diploid migrants at a constant rate. Three values of migration rate (m) were simulated: 2×10^{-4}, 2×10^{-3}, and 2×10^{-2} per generation. Samples were taken from a single deme (the focal population) in these structured scenarios. The effects of more drastic migration events were investigated in a fourth scenario, the admixture scenario, where a fraction of the focal subpopulation was replaced by individuals from another single population. The metapopulation was again simulated with an island model with a migration rate $m = 2 \times 10^{-3}$ per generation. At time t_{10} a single admixture event was simulated, with an admixture rate of 50%, 75%, or 100%. Note that 100% admixture is equivalent to a local extinction and recolonization scenario without change in population size. The focal population was sampled at generations t_0 and t_{20}, as in the previous scenarios. Because extinction–recolonization events may be associated with founder events, we evaluated a final set of scenarios with a bottleneck concomitant with an extinction–recolonization event. During the bottleneck, the focal population size was reduced to 1, 5, or 25 diploid individuals. After one generation, the population size was restored to $N = 250$ individuals and 100 diploid individuals were sampled at t_{20}. For each simulation scenario described above (and summarized in Table S1), 1000 independent replicates were performed. SLiM simulation scripts for each of these scenarios as well as R scripts are available on the INRA dataportal. https://doi.org/10.15454/VYPXIJ.

Diversity indices

Diversity analyses were performed using the Hierfstat package in R (Goudet 2005). The genetic diversity of each simulated population was assessed on the t_{20} sample using the average gene diversity across loci (H_E, Nei 1973), the variance in allele size (V), the average number of alleles per locus (n_A), and the number of polymorphic loci (PL). In an isolated random mating population at mutation-drift equilibrium, H_E and V measured on microsatellite markers evolving under the stepwise mutation model are expected to vary with the effective population size N_e and the mutation rate μ as

$$H_E = 1 - \sqrt{\frac{1}{2\theta + 1}} \quad (2)$$

$$V = 2N_e\mu \quad (3)$$

where $\theta = 4N_e\mu$ (Kimmel et al. 1998).

The deviation from Hardy-Weinberg proportions was measured using the inbreeding coefficient F_{IS} with the R package Hierfstat. The percentage of pairs of loci showing significant linkage disequilibrium (*LD%*) was calculated using Genepop (Rousset 2008) with a significance threshold of 0.05. The identity disequilibrium (*g2*), which is expected to depend on the selfing rate following $g_2 = \frac{1-\sigma}{\left(1-\frac{\sigma}{4}\right)\left(1-\frac{\sigma}{2-\sigma}\right)^2} - 1$, (David et al. 2007) was also computed using the R package inbreedR (Stoffel et al. 2016). The R package Poppr (Kamvar et al. 2014) was used to identify the number of private alleles (p_A), to group individuals with identical combinations of alleles (multilocus genotypes, MLG), compute the number of distinct MLGs, their frequency and their repartition over time in the two samples (t_0, t_{20}). The multilocus diversity was characterized by the Shannon's index, computed as $H = -\sum p_i ln(p_i)$, where p_{i_p} is the frequency of the ith MLG. The frequency of the most frequent MLG (*MFMLG*) was also computed and we analyzed the correlation between *MFMLG* and *H* through Spearman correlations using R.

We calculated the pairwise genetic distances between individuals at generation t_{20} as the number of allele differences (between 0 and 40) between each pair of individuals, regardless of the allele size. We used two indices to characterize the distributions of distances: the mean pairwise genetic distance (D_{mean}) and the maximum pairwise genetic distance (D_{max}). The correlation between some indices (D_{mean} and H_E, or D_{max} and *LD%*) was measured through Spearman correlations using R.

To summarize the trajectories of MLG frequencies through time, we considered the joint MLG frequency spectrum (*MLGFS*, by analogy with the allele frequency spectrum) as the matrix containing the proportion of MLGs found at the corresponding individual counts in each generation, averaged over simulation replicates. For K simulation replicates, we have $MLGFS[i,j] = \frac{\sum_{k=1}^{K} \frac{MLG(i,j)}{MLG_k}}{K}$, where MLG($i$, j) is the number of MLGs found in i individuals at t_0 and in j individuals at t_{20}, and MLG_k is the total number of different MLGs in replicate k. The *MLGFS* therefore allows to follow the evolution of the frequency of MLGs overtime.

The relative temporal differentiation between the two samples was assessed with Weir and Cockerham's F_{ST} (1984), estimated using the R package Hierfstat (Goudet 2005). The effective population size was estimated based on the temporal differentiation between samples (temporal F_{ST}) as outlined in Frachon et al. (2017):

$$\widehat{N_e} = \frac{\tau(1 - F_{ST})}{4F_{ST}}, \quad (4)$$

where $\widehat{N_e}$ is the estimate of the effective population size and τ is the number of generations separating the two sampling events. This method assumes that the population is isolated (no migration), of constant size and that no mutation occurs between samplings. We estimated the focal population effective size in our different simulation scenarios in order to examine whether the deviations from theoretical assumptions (e.g., admixture or bottlenecks) can lead to N_e estimates as small as those reported in the literature. N_e estimates in scenarios of isolated populations were compared with the theoretical expectations given by Eq. (1) (assuming $\alpha = 1$: $N_e = \frac{N}{1+\frac{\sigma}{2-\sigma}}$) for an isolated population with no change of population size, where N is the demographic size of the population and σ is the selfing rate (Pollak 1987); and $\frac{1}{N_e} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_e^t}$ for an isolated population undergoing bottlenecks, where N_e^t is the effective population size at generation t (Crow and Kimura 1970).

Medicago truncatula natural populations

Medicago truncatula is an annual, predominantly selfing species of the legume family (Fabaceae), found around the Mediterranean Basin. Maternal progeny analyses have shown very low levels of residual outcrossing (Siol et al. 2008). Between 1986 and 2014, nine natural populations located in Spain (SP1–SP3), Corsica (CO1–CO3), and southern France (FR1–FR3) were sampled two or three times each (locations of the different populations can be found on a map in Figure S1). In order to avoid oversampling the progeny of a single individual, pods were sampled along transects running across the populations, with at least 1-m distance between each collected pod. This sampling strategy also allows to limit spatial effects due to the very fine spatial structure observed in *M. truncatula* natural populations (Bonnin et al. 2001). Sample sizes varied between 31 and 232 individuals. Hereafter, each temporal sample will be denominated by its population code followed by the sampling year.

DNA was extracted from 50 mg of fresh leaves with the Chemagic DNA Plant Kit (Perkin Elmer), according to the manufacturer's instructions. The protocol is adapted to the use of the KingFisher Flex™ (Thermo Fisher Scientific) automated DNA purification workstation. Twenty microsatellite loci were used for genotyping. Eighteen of them have been described previously (Baquerizo-Audiot et al. 2001; Arrighi et al. 2006; Ronfort et al. 2006; Siol et al. 2007). Two new loci, 319 and DMI1-6, were developed in our team after identifying long and polymorphic simple sequence repeats in resequencing studies (319-F GTGGGATTTGAATAGGATTG, 319-R CGATATGGTCCACTTTTGTC, annealing temperature: 57 °C;

Table 2 Mean values for single locus and multilocus indices of genetic diversity for isolated populations with increasing selfing rates and a constant demographic size of $N = 250$

N	σ	N_e	F_{IS}	H_E	V	$nMLG$	H	$LD\%$	g_2	$NbLoc\_g2$
250	0	250	0.0 (0.0)	0.42 (0.00)	0.50 (0.39)	99.9 (0.0)	4.61 (0.00)	0.06 (0.00)	0.001 (0)	19.9 (0.4)
250	0.5	188	0.32 (0.00)	0.37 (0.00)	0.38 (0.25)	99.2 (1.0)	4.59 (0.00)	0.13 (0.00)	0.292 (0.004)	19.5 (0.7)
250	0.95	132	0.85 (0.00)	0.30 (0.00)	0.27 (0.14)	58.2 (35.7)	3.79 (0.03)	0.38 (0.01)	5.576 (3.607)	16.7 (1.8)
250	0.98	128	0.91 (0.00)	0.29 (0.00)	0.25 (0.14)	42.7 (33.0)	3.34 (0.05)	0.41 (0.02)	11.268 (49.45)	13.9 (2.8)
250	1	125	0.94 (0.00)	0.29 (0.01)	0.24 (0.11)	29.3 (19.9)	2.86 (0.06)	0.47 (0.04)	−0.04 (7.324)	6.2 (2.1)

Values in brackets show the variance over 1000 replicates. N_e is the effective size calculated using Eq. (1) with $\alpha = 1$; F_{IS} is the inbreeding coefficient. H_E is the estimated gene diversity, V is the variance of allele size, $nMLG$ is the number of MLGs, H is the Shannon's index, $LD\%$ is the percentage of loci with significant linkage disequilibrium, g_2 is the identity disequilibrium and $NbLoc\_g2$ is the number of loci used to compute $g2$. Expected values of F_{IS}, H_E, and V are reported in Table S2

DMI1-6-F1 TAGAAGATGAAGCGCAAACG, DMI1-6-R2 TTCACCTTAACGCGTCCAAC, annealing temperature: 60 °C). We followed the protocol of amplification reactions described in Siol et al. (2007). Samples were prepared by adding 3 µl of diluted PCR products to 16.5 µl of ultrapure water and 0.5 µl of the size marker AMM524. Amplified products were analyzed on an ABI prism 3130 Genetic Analyzer and genotype reading was performed using GeneMapper Software version 5. Individuals and loci with more than 10% missing data across all samples of a population were removed from the diversity analyses, as well as completely monomorphic loci.

For each population and year, we performed the same analyzes of single and multilocus diversity as those performed on our simulated populations. To account for variation in sample sizes, mean allelic richness per locus (Rs) and private alleles (p_A) were computed using the rarefaction method with the program ADZE (Szpiech et al. 2008). Selfing rates were estimated from F_{IS} using the classical relationship $F_{IS} = \sigma/(2 - \sigma)$ (Hartl and Clark 1998) and using a maximum-likelihood approach based on the identity disequilibrium (g_2), with the software RMES (David et al. 2007). MLG frequency spectra were computed for each population. Temporal F_{ST} estimates were used to estimate the effective population size using the method described previously (Eq. (3), Frachon et al. 2017), assuming a single generation per year. Approximate bootstrap confidence intervals for the temporal estimates of effective size were computed following DiCiccio and Efron (1996).

Results

Single-locus and multilocus genetic diversity in isolated populations

In the simulations of a single isolated population for different combinations of selfing rates and demographic population sizes, estimates of single locus indices (H_E and V) are in accordance with theoretical predictions at mutation-drift equilibrium (Table S2), showing a decrease in the neutral genetic diversity with increasing selfing rates (Table 2). When the demographic size is adjusted to keep the effective size constant while the selfing rate varies, single locus diversity indices remain around the expected value too (Table 3). These results are not new as they replicate the known effects of selfing on single locus diversity but they are helpful to validate our simulation framework.

As shown in Table 2, we found that the multilocus diversity ($nMLG$ and H) also decreases with selfing, while the homozygosity (F_{IS}) and the associations between loci ($LD\%$, $g2$) increase. For completely selfing populations ($\sigma = 1$), $g2$ is biased downwards due to extremely high homozygosity limiting the number of loci available for the estimation (as $g2$ measures the correlation of heterozygosity between loci). In completely outcrossing or low selfing populations (up to 50% selfing in our simulations), there are on average as many MLGs as individuals sampled. In contrast, $nMLG$ decreases to around two thirds of the sample in our simulations with 95% selfing and to less than one-third in completely selfing populations for $N = 250$. This loss of MLGs is even more dramatic when the population size is lower (e.g., $N = 50$, Table S2). In addition, contrary to single locus indices, multilocus diversity indices keep decreasing with increasing selfing rate even for a given N_e value (Table 3), in conjunction with the increase in linkage disequilibrium. Our analysis of the effect of sample size shows that, for a given selfing rate, both H_E and the frequency of the most frequent MLG ($MFMLG$) are biased and less precise for small sample sizes. The statistics approach the expected value with a smaller sample size for single locus compared to multilocus diversity ($N_{samp} = 20$ for H_E and $N_{samp} > 30$ for $MFMLG$, Fig. S7).

Table 3 Simulated populations with increasing selfing rates and demographic sizes adjusted to keep the effective size (N_e) constant and equal to 250

N	σ	N_e	F_{IS}	H_E	V	$nMLG$	H	$LD\%$	g_2	$NbLoc\_g2$
250	0	250	0.0 (0.00)	0.42 (0.002)	0.50 (0.39)	99.9 (0.0)	4.61 (0.00)	0.06 (0.00)	0.001 (0.000)	19.9 (0.4)
333	0.5	250	0.32 (0.00)	0.42 (0.002)	0.50 (0.44)	99.4 (0.7)	4.60 (0.00)	0.14 (0.00)	0.29 (0.004)	19.8 (0.4)
475	0.95	250	0.89 (0.00)	0.42 (0.002)	0.50 (0.40)	71.1 (26.2)	4.11 (0.01)	0.54 (0.01)	5.875 (3.138)	18.8 (1.2)
490	0.98	250	0.94 (0.00)	0.42 (0.002)	0.50 (0.42)	57.0 (27.5)	3.80 (0.02)	0.63 (0.01)	12.75 (45.025)	16.5 (2.6)
500	1	250	0.98 (0.00)	0.42 (0.002)	0.53 (0.63)	40.8 (22.5)	3.33 (0.03)	0.75 (0.03)	−0.151 (4.641)	6.4 (2.1)

The corresponding N is set according to Eq. (1) with $\alpha = 1$; the expected value for H_E is 0.42 (Eq. (2)), and 0.5 for V (Eq. (3))

Structure of multilocus genetic diversity in more complex scenarios

The frequency of the *MFMLG* summarizes the increase of repeated multilocus genotypic combinations, because it increases with the selfing rate, especially for $\sigma \geq 0.95$ (Fig. 1a, Table S2) and is highly correlated with Shannon's index H ($P < 2.2 \times 10^{-16}$, $r^2 = 0.91$, Fig. S2). In the following, we will use *MFMLG* as an indicator of multilocus diversity variations. Our simulations with lower population size, bottlenecks or metapopulation dynamics highlight that extreme patterns of MLG repetition ($MFMLG > 30\%$) are observed only for very low demographic population sizes ($N = 50$) combined with high selfing rates ($\sigma \geq 0.95$), or with strong bottlenecks (reduction to fewer than five individuals), associated with predominant selfing (Fig. 1b, Table S2). These extreme patterns of MLG repetition are nevertheless highly variable among simulation replicates. Figure 1b also illustrates that this increase of *MFMLG* with low population size or bottlenecks is associated with a decrease in single locus diversity (H_E). In addition, Fig. 1b shows that changes in single locus diversity due to migration are greater than changes in the frequency of the *MFMLG* (migration and admixture scenarios on Fig. 1b). Similar patterns can be observed when comparing Shannon's index (H) and H_E (Fig. S3).

The mean distance between two individuals within a population (D_{mean}) is highly correlated to H_E ($P < 2.2 \times 10^{-16}$, $r^2 = 0.98$, Fig. S4). We used the maximum distance found between two individuals in a population (D_{max}) to describe the genetic divergence accumulated between MLGs. D_{max} increases with the genetic diversity (with increasing N) and with the selfing rate (for populations with the same effective size, $N_e = 250$ or $N_e = 100$ in Table S2). For predominantly selfing populations ($\sigma \geq 0.95$), D_{max} is strongly correlated with $LD\%$ ($P < 2.2 \times 10^{-16}$, Fig. S5). As a consequence, D_{max} is the highest in scenarios involving migration, either at a constant rate or after drastic admixture (Fig. 2b). Another interesting result is that only very low demographic population sizes or very strong bottlenecks in isolated populations result in low D_{max} (<0.5).

Temporal changes

We measured the change in allele frequencies through time (samples separated by 20 generations) with the relative genetic differentiation between temporal samples using F_{ST}. We observed that F_{ST} estimates and their variance increase with selfing (Table S2). Whereas migration equilibrium (at rates ≤ 0.002) does not affect the temporal differentiation much, occasional large migration events (admixture) raise the differentiation through time. Genetic differentiation is particularly strong in extreme demographic scenarios such as extinction–recolonization because of population replacement.

We used the temporal F_{ST} to estimate the effective population size according to Eq. (4), ignoring the fact that some of our simulated scenarios do not meet the assumption of the underlying theoretical model (isolated populations). Figure 3a shows that, as expected in isolated populations of constant size (see Eq. (2)), $\widehat{N_e}$ estimates increase with H_E. Despite a large variance between replicates, average $\widehat{N_e}$ estimates in isolated populations are close to the theoretical expectations (see Table S2), except for large populations sizes or complete selfing ($\sigma = 1$). In these cases, the variance of $\widehat{N_e}$ is extreme, due to sampling variance and linkage disequilibrium, respectively. Under complex demographic scenarios involving strong migration events or extinction–recolonization, $\widehat{N_e}$ estimates are remarkably low compared to the simulated demographic population sizes. Those estimates disagree with the levels of genetic diversity (H_E) observed in these populations given the expectations from Eq. (2). This could be caused by departures from the model's assumptions (see discussion). Indeed, the effects of migration are visible through the increase in the number of private alleles at t_{20} (p_A, Table S2).

Figure 4 shows patterns of MLGs persistence over time (after 20 generations in our simulations) in the different scenarios we investigated. Under complete outcrossing, the conservation of a MLG after 20 generations is extremely rare, as expected due to recombination (Fig. 4a). In contrast, under predominant selfing (Fig. 4b), MLGs are frequently shared between temporal samples. Moreover, MLGs that reach a

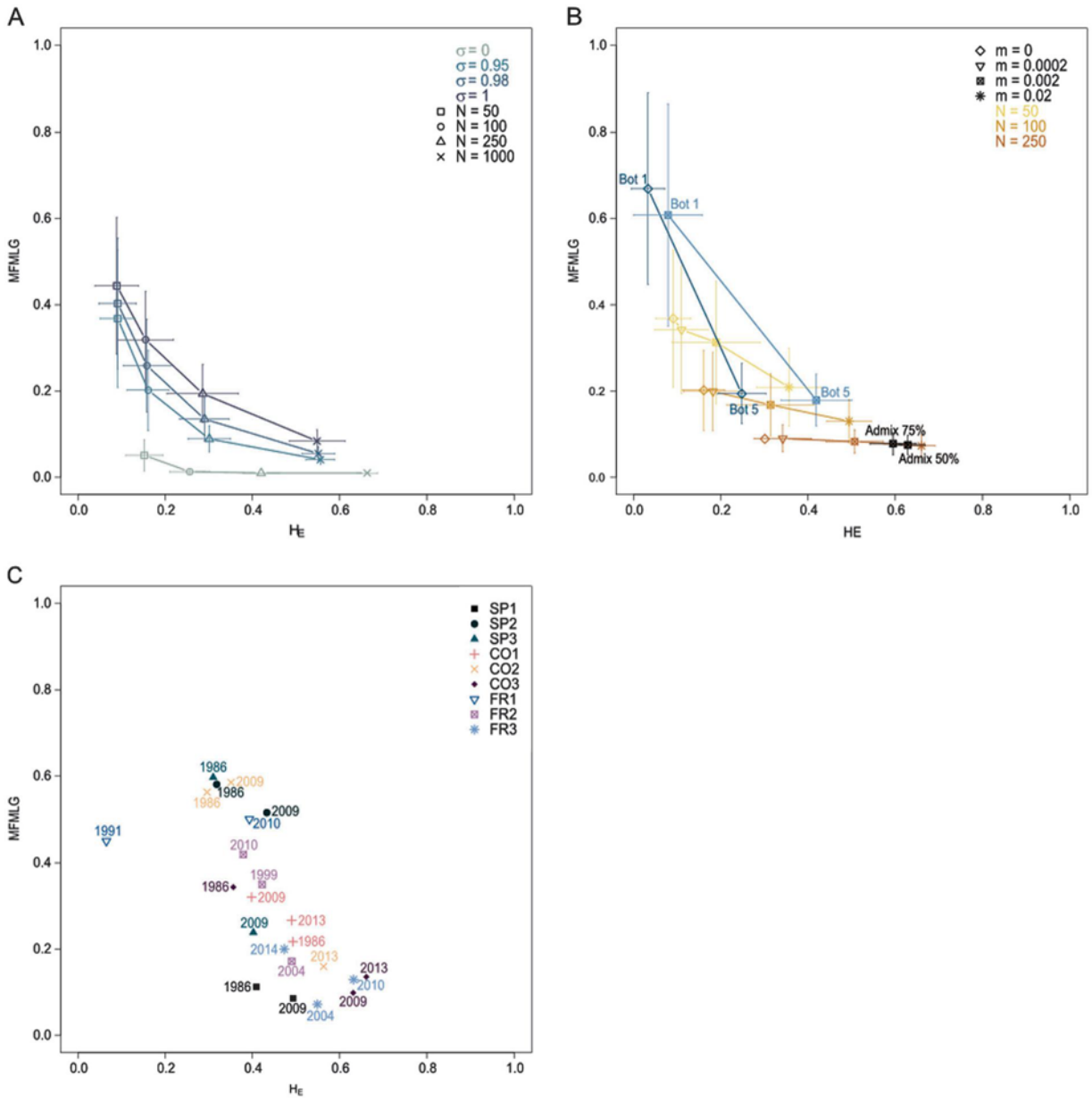


Fig. 1 Covariation between the frequency of the most frequent MLG (*MFMLG*) and the gene diversity (H_E) across simulated scenarios. **a** Scenarios of isolated populations with varying demographic sizes N and selfing rates σ; **b** scenarios of migration (orange), admixture (black), bottleneck (blue), and extinction–recolonization (light blue) with $\sigma = 0.95$; **c** natural populations of *Medicago truncatula* for each sampling date. For **a** and **b**, points indicate means and horizontal and vertical bars stand for the standard deviation across the 1000 replicates

high frequency within the first generation (measured by the abscissa for t_0), tend to remain at high frequency at t_{20}. This pattern is amplified when the population size is low (Fig. 4c). Strong bottlenecks (Fig. 4d) raise the frequency of some MLGs independently of their frequency in the first generation. Scenarios with migration and admixture slightly reduce the occurrence of conserved high-frequency MLGs (Fig. 4e, f) while scenarios including extinction–recolonization produce spectra with fewer MLGs conserved over time (Fig. 4g, h for extinction–recolonization with a bottleneck).

Empirical data

In the nine natural populations of *M. truncatula* studied, F_{IS} values are high, ranging between 0.88 and 1. This translates into very high-selfing rate estimates for all populations ($\sigma_{FIS} > 0.9$, Table 4). Selfing rate estimates with RMES are sometimes lower but remain well above 0.8 (Table 4). The number of MLGs is generally low and compatible with high selfing rates.

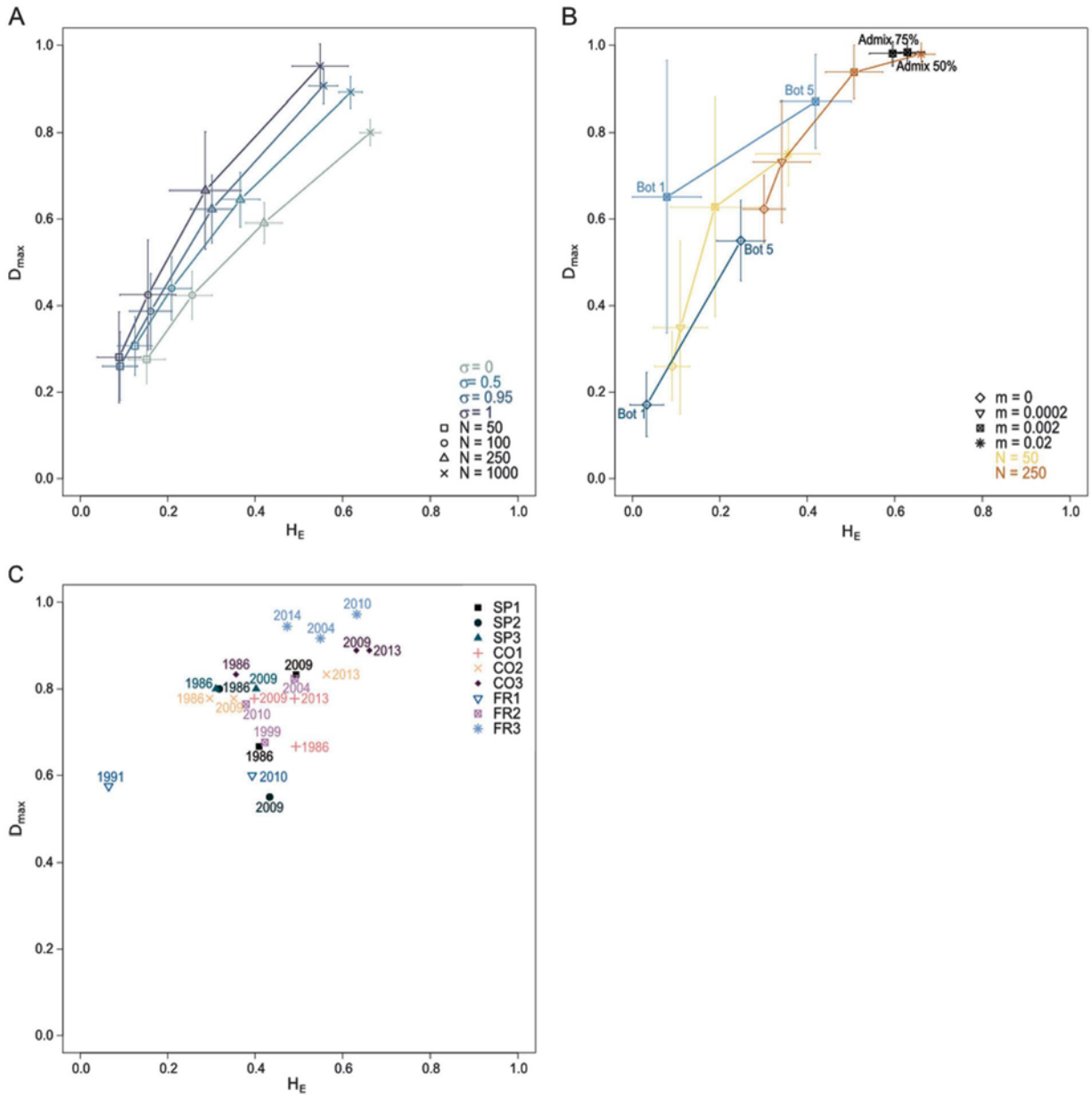


Fig. 2 Covariation between the maximum distance between pairs of individuals in a population (D_{max}) and H_E across simulated scenarios. **a** Scenarios of isolated populations with increasing demographic sizes N and selfing rates σ; **b** scenarios of migration (orange), admixture (black), bottleneck (blue) and extinction–recolonization (light blue) with $\sigma = 0.95$; **c** natural populations of *Medicago truncatula* for each sampling date. For **a** and **b**, points indicate means and horizontal and vertical bars stand for the standard deviation across the 1000 replicates

Single and multilocus diversity

The mean gene diversity within population and year (H_E) is remarkably high (higher than 0.3, Table 4). The maximum genetic distance between two individuals, D_{max}, is also always high (higher than 0.5, Fig. 2c). Most populations therefore seem to be distributed within a parameter space more limited than the one explored by our simulations (high H_E associated with high D_{max}). Only sample FR1_1991 presents both low single and multilocus diversity (Fig. 1c, Fig. 2c). *MFMLG* values are highly variable, with extreme patterns of MLG repetition (*MFMLG* higher than 30%) in nearly half of the populations studied. Accordingly, these populations also display the lowest values of Shannon's H (Table 4). However, such a combination of high single-locus diversity and extremely low multilocus diversity was not observed in any of our simulated scenarios (Fig. 1c).

Temporal dynamics of diversity

The MLG frequency spectra highlight two different types of dynamics of MLGs through time. In populations SP1, SP2,

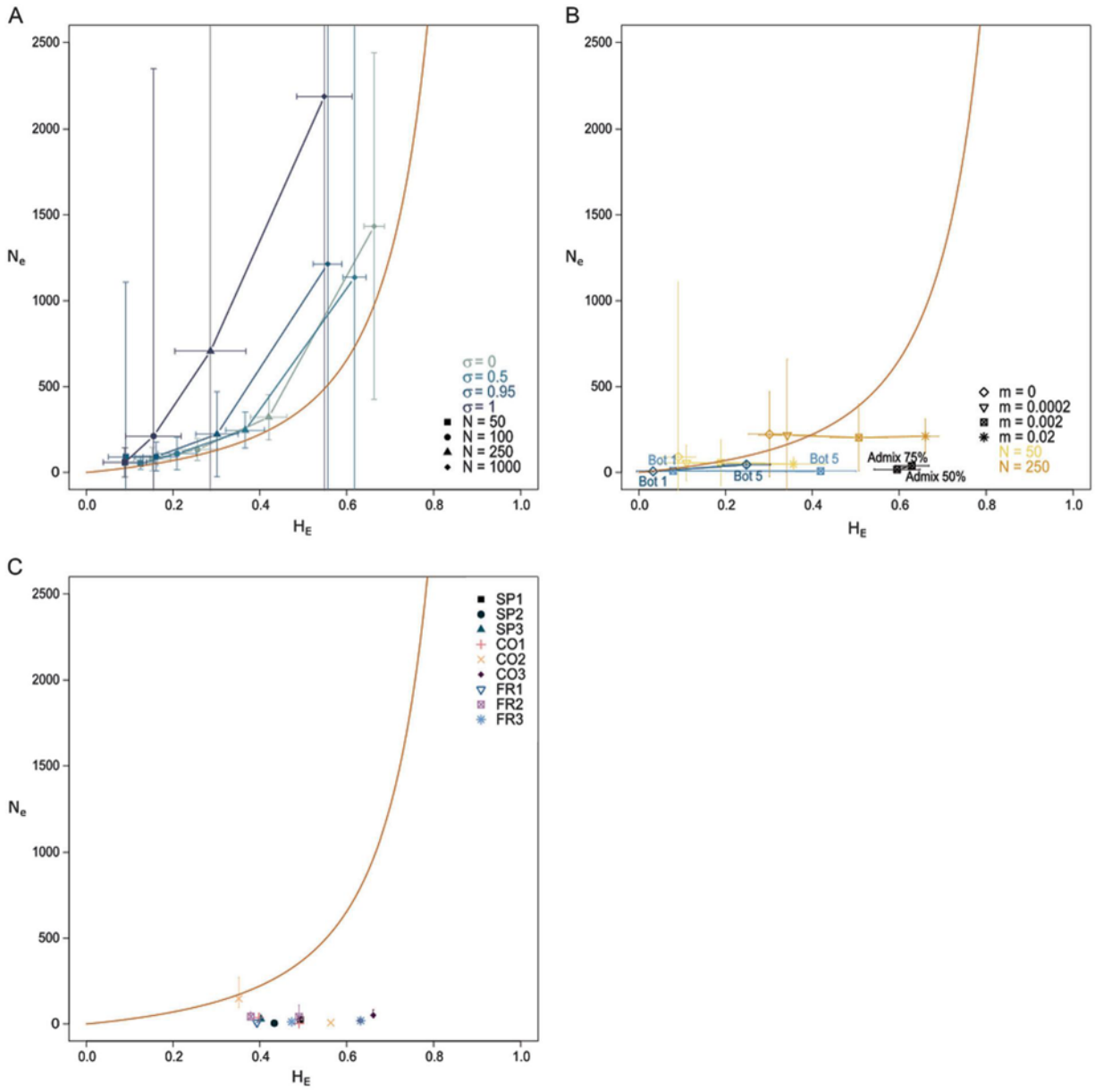


Fig. 3 Temporal estimation of effective population size ($\widehat{N_e}$) compared to gene diversity (H_E). The red curve corresponds to the expected relationship between H_E and N_e assuming a constant isolated population (Eq. (2)) and a mutation rate of 0.001. **a** Scenarios of isolated populations with varying selfing rates σ and demographic sizes N; **b** scenarios of migration, admixture, bottleneck and extinction–recolonization with $\sigma = 0.95$; **c** temporal $\widehat{N_e}$ estimates in natural populations of *Medicago truncatula*, with vertical bars representing the 95% confidence interval. For **a** and **b**, horizontal and vertical bars stand for the standard deviation across the 1000 replicates

and CO1, a single low-frequency MLG or none at all remain over time (Fig. S6). In our simulations, such dynamics of multilocus diversity are obtained with extinction–recolonization events only (Fig. 4g, h). In the other populations, several MLGs are conserved through time (SP3, CO3, FR3, CO2, FR2, and FR1, Fig. S6). Among these populations, FR3 and CO3 are the most diverse ($H_E > 0.5$) and present extreme patterns of multilocus diversity with *MFMLG* lower than 0.2 and D_{max} higher than 0.8 (Figs. 1c and 2c). Such patterns were also observed in our simulations with strong migration and admixture (Figs. 1b and 2b).

Temporal differentiation between sampling years, as estimated by temporal F_{ST}, is high for most of the populations studied (Table 4). The effective population sizes estimated using the temporal F_{ST} method are variable but consistently low: all estimates except one are lower than 100 (with a maximum of 150 for population CO_2). These estimates are too low compared with the observed single locus diversity given the expectations from eq. 2 (Fig. 3c).

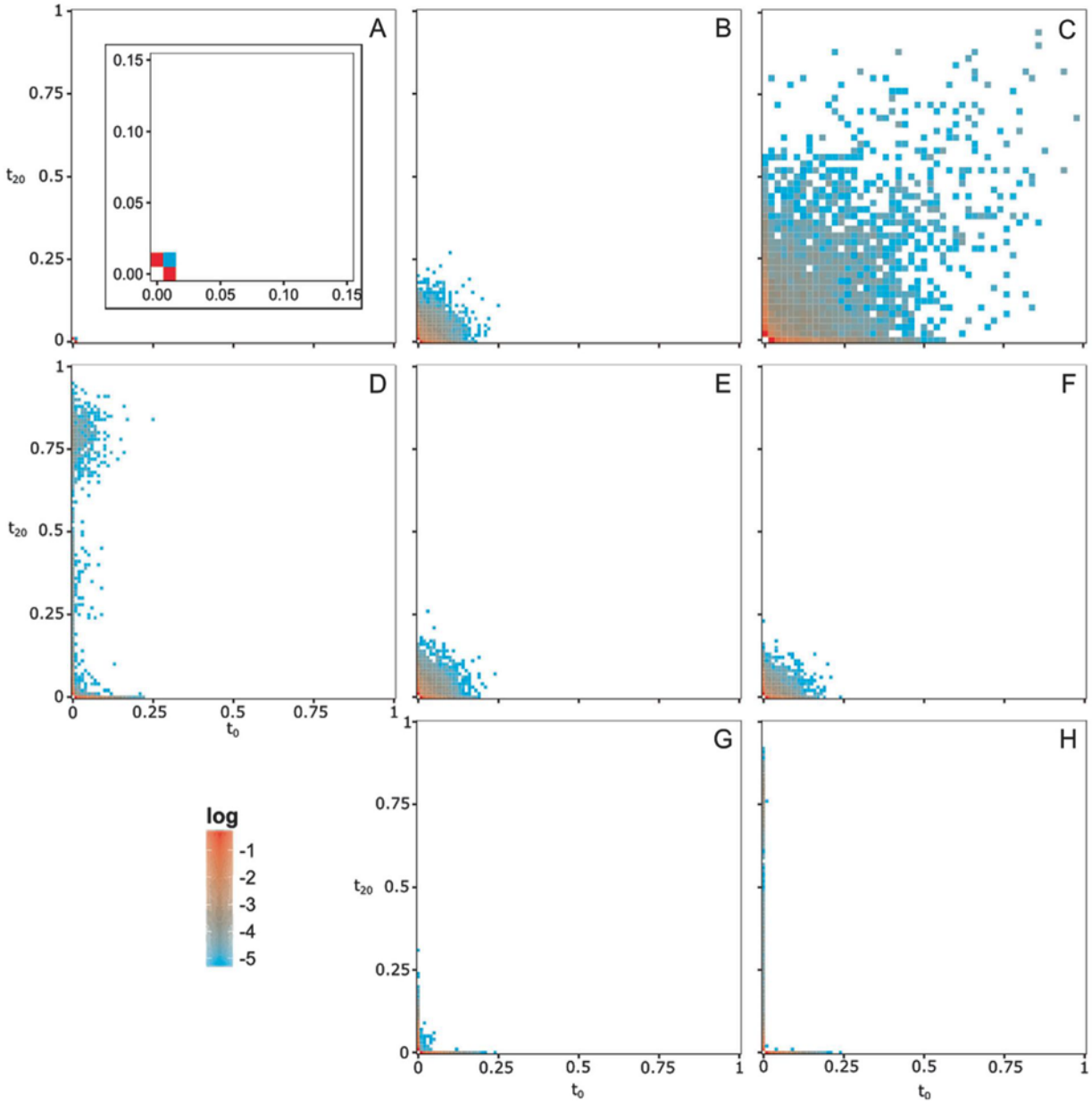


Fig. 4 Joint MLG frequency spectra for each demographic scenario. The horizontal axis represents the frequency at which a MLG is found in the first sample (t_0), the vertical axis represents the frequency of this same MLG in the second sample (t_{20}). The color gradient represents the $\log_{10}$ of the frequency at which each case is observed in 1000 simulation replicates. **a** Isolated outcrossing population ($\sigma = 0$) of 250 individuals. The inset is a zoom of frequencies between 0 and 0.15; **b** isolated predominantly selfing population ($\sigma = 0.95$; $N = 250$); **c** isolated predominantly selfing population ($\sigma = 0.95$; $N = 50$); **d** isolated predominantly selfing population ($\sigma = 0.95$; $N = 250$) undergoing a bottleneck of one individual at t_{10}; **e** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$); **f** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing 50% admixture with constant population size at t_{10}; **g** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing extinction–recolonization with constant population size at t_{10}; **h** predominantly selfing population in an island model ($\sigma = 0.95$; $m = 0.002$; $N = 250$) undergoing extinction–recolonization by one individual at t_{10}

In our simulations, we observed a similar mismatch between H_E and N_e values for migration and admixture scenarios. This resemblance with migration and admixture scenarios is also visible in the high number of private alleles we observe in recent compared to older temporal samples in our populations (Table 4, Table S2).

Discussion

Our work aimed at describing the consequences of high-selfing rates and metapopulation dynamics on the structure of genetic diversity in populations, and how it can change over time. We argue that the classical single locus diversity

Table 4 Genetic diversity in *M. truncatula* populations

Population	Year	n	F_{IS}	σ_{FIS}	σ_{RMES}	H_E	$nMLG$	p_A	H	$MFMLG$	F_{ST}
SP1	1986	71	0.96	0.98	0.93 (0.03)	0.41	31	–	3.10	0.11	0.186
	2009	93	0.99	0.99	0.92 (0.06)	0.49	39.2	1.87 (0.40)	3.55	0.09	
SP2	1986	31	1.00	1.00	–	0.32	4	–	0.90	0.58	0.536
	2009	66	0.97	0.98	0.98 (0.01)	0.43	9.3	2.31 (0.32)	1.76	0.52	
SP3	1986	67	0.97	0.98	0.99 (0.01)	0.31	12	–	1.51	0.60	0.261
	2009	88	0.98	0.99	0.99 (0.01)	0.40	24.7	1.67 (0.38)	2.71	0.24	
CO1	1986	46	1.00	1	–	0.49	18	–	2.43	0.22	
	2009	78	0.98	0.99	0.99 (0.01)	0.40	12.6	0.22 (0.10)	2.08	0.32	0.13
	2013	60	0.96	0.977	0.96 (0.02)	0.49	23.7	0.76 (0.12)	2.87	0.27	0.243
CO2	1986	64	0.95	0.97	0.96 (0.02)	0.30	20	–	1.89	0.56	
	2009	94	0.96	0.98	0.99 (0.01)	0.35	16.8	1.19 (0.27)	1.83	0.59	0.03
	2013	100	0.96	0.98	0.93 (0.02)	0.56	32.6	1.42 (0.45)	3.27	0.16	0.115
CO3	1986	64	0.94	0.97	0.97 (0.02)	0.36	20	–	2.16	0.34	
	2009	81	0.96	0.98	0.99 (0.01)	0.63	41.0	0.71 (0.24)	3.65	0.10	0.226
	2013	162	0.94	0.97	0.95 (0.01)	0.66	44.6	1.20 (0.45)	4.08	0.14	0.019
FR1	1991	91	0.88	0.94	0.97 (0.02)	0.07	8.7	–	1.29	0.45	0.337
	2010	82	0.99	0.99	–	0.39	11	1.75 (0.27)	1.34	0.50	
FR2	1999	60	0.98	0.99	0.99 (0.01)	0.42	11	–	1.77	0.35	
	2004	64	0.95	0.97	0.90 (0.04)	0.49	38.9	1.32 (0.33)	3.41	0.17	0.029
	2010	93	0.96	0.98	0.88 (0.06)	0.38	17.4	0.82 (0.26)	2.21	0.42	0.033
FR3	2004	97	0.98	0.99	0.97 (0.01)	0.55	48	–	3.57	0.07	
	2010	201	0.90	0.95	0.93 (0.01)	0.63	63.0	1.23 (0.26)	4.22	0.13	0.071
	2014	135	0.97	0.98	0.97 (0.01)	0.47	46.4	0.80 (0.18)	3.46	0.20	0.071

n is the sample size, F_{IS} is the inbreeding coefficient, σ_{Fis} is the selfing rate estimated from the F_{IS}, σ_{RMES} is the selfing rate estimated using RMES, H_E is the mean gene diversity, $nMLG$ is the MLG number (calculated using a rarefaction method), and p_A is the mean number of private alleles per locus found in the second or third temporal samples (calculated using a rarefaction method, with standard deviation in brackets), H is the Shannon's index, $MFMLG$ is the frequency of the most frequent MLG and F_{ST} is the temporal F_{ST} between successive temporal samples. In samples SP2_1986, CO1_1986 and FR1_1991, the lack of heterozygosity in the population prevented the estimation of σ_{RMES}

indices are not sufficient to fully understand the demographic history of predominantly selfing populations, which should benefit from multilocus indices. Because of the lack of analytical expectations for such indices, we proposed a simulation approach to address the question in a theoretical framework.

Neutral scenarios can explain the multilocus population genetic structure of predominantly selfing species

Our simulations of isolated and predominantly selfing populations (with selfing rates above 0.95), show a population genetic structure organized in repeated multilocus genotypes. As for single locus diversity, multilocus diversity (measured as the number of MLGs or the haplotypic diversity H), decreases with increasing selfing rates. In addition, the nonindependence between loci increases with selfing (Nordborg 2000), as seen through the elevated linkage and identity disequilibrium, and the multilocus diversity is further reduced. The maximum distance between two individuals, D_{max}, increases with the selfing rate, highlighting the fact that self-fertilizing populations are composed of differentiated lineages. The increase in D_{max} is caused by the reduced effective recombination in selfing populations, which constrains new mutations within only one genetic background. For predominantly selfing populations ($\sigma \geq 0.95$), D_{max} is also highly correlated to $LD\%$, because they both increase with within-population structure. The empirical data obtained on natural populations of *M. truncatula* strongly support our simulation results: our estimates of selfing rates confirm *M. truncatula* as a predominantly selfing species, and we find repeated MLGs and large D_{max} in every population. In *Arabidopsis thaliana*, studies in natural populations also showed that they are composed either of identical or highly differentiated individuals (Bakker et al. 2006; Montesinos et al. 2009). Overall, these results highlight the importance of multilocus analyses for the study of natural selfing populations. Yet, such analyses are often overlooked in the literature (e.g.,

Trouvé et al. 2005; Gow et al. 2007) or are limited to reporting the number of distinct MLGs (e.g., Bomblies et al. 2010; Gomaa et al. 2011). Furthermore, our results stress out that accurate estimates of MLG frequencies are essential, especially in the presence of rare MLGs, and require larger samples than for estimates of single locus diversity (above 30 individuals, Fig. S7).

The first aim of the present study was to verify if selectively neutral scenarios with high-selfing rates could lead to repeated MLGs, sometimes at high frequency and maintained through time. Indeed, Avise and Tatarenkov (2012) argued that this peculiar genetic structure in selfing populations provided evidence for the occurrence of selective processes promoting locally adapted MLGs. Our simulation results show that strong genetic drift induced by small population size or bottlenecks may be sufficient to explain the multilocus genetic structure observed in selfing populations, without any selection. It is, however, important to stress out that our results are not sufficient to rule out selective processes in a population but only present alternative hypotheses to explain the observed structure of genetic diversity. Testing Avise and Tatarenkov (2012)'s hypothesis would require reciprocal transplants, or at least measuring the fitness of the MLGs in order to see if the locally most frequent MLG has indeed the highest fitness in the local environment.

Combining single and multilocus indices of diversity is insightful when studying the demographic history of predominantly selfing populations

We analyzed the multilocus structure of the simulated populations and focused on indices describing MLG frequency (*MFMLG*) or MLG genetic similarity (D_{max}). Those indices are especially informative when analyzed conjointly with single locus diversity (H_E) and can help disentangle the effect of selfing and demographic events (such as bottlenecks or migration) on genetic diversity in selfing populations. Indeed, high *MFMLG* combined with low H_E is characteristic of small population (constantly small or due to a bottleneck) with a high selfing rate. On the other hand, low levels of multilocus diversity while single locus diversity is high were observed with strong migration or admixture scenarios. This highlights the fact that migration restores single locus diversity faster than multilocus diversity in predominantly selfing populations (Fig. 1b). This was also visible when analyzing conjointly H_E and D_{max}: in our migration scenarios, new alleles combined within migrant MLGs were introduced in the population, resulting in both high *LD%* and D_{max} values.

Even though our simulations explored only a restricted number of scenarios (in terms of population size, sample size, selfing rate, time span between sampling, etc.), they were able to replicate patterns observed in empirical data. Except for one population (FR1), the levels of genetic diversity (H_E) in most populations of *M. truncatula* were surprisingly high compared with theory and other studies of predominantly selfing populations (e.g., Gomaa et al. 2011; Lundemo et al. 2009; Stenøien et al. 2005). In addition, high single locus diversity (H_E) was combined with repeated MLGs (high *MFMLG*), which may be consistent with populations belonging to a metapopulation with strong migration or even admixture events. Genetic diversity measured by both *MFMLG* and H_E (or D_{max} and H_E) as well as the MLG frequency spectrum suggest likely extinction–recolonization events in three populations of *M. truncatula* (SP1, SP2, and CO1). Nevertheless, a fine analysis of the spatial genetic structure should be performed to ensure that the drastic changes in genetic structure observed in these populations are not due to a shift in the location of the sampling transect. However, in a set of populations, MLG repetition (*MFMLG*) and single-locus genetic diversity (H_E) were both very high (higher than 0.4). This combination of a low number of MLGs and high single-locus diversity was never observed in our simulations. Other studies also reported a similar pattern (e.g., Barrière and Félix 2007), which is probably associated with scenarios or combinations of parameter values that were not considered in our set of simulations. This highlights the need for a more systematic exploration of the parameter space if one intends to perform statistical inferences on empirical data.

Insights from the temporal analysis of predominantly selfing populations

The temporal dimension of our analyses is one of the main particularities of this study, and is rarely examined in natural populations with high selfing rates (we found less than a dozen studies, some being reported in Table 1). Yet temporal data are useful because they allow estimating the effective population size and thus give insight into the strength of genetic drift (Waples 1989). However, the decreased effective recombination in selfing populations reduces the number of independent loci, and after several generations of predominant selfing the whole genome tends to behave as a single "superlocus". F_{ST} estimates based on few or a single locus suffer from a large sampling variance (Weir and Hill 2002) and this is visible in our simulations in which the variability of F_{ST} estimates increased with the selfing rate. Indeed, measuring F_{ST} from linked loci is equivalent to measuring it from a lower number of loci. Moreover, if metapopulation dynamics are frequent in selfing populations, it may cause departures from the assumptions of the theoretical model underlying the estimation of N_e from temporal F_{ST} (i.e., isolated

population of constant size, Waples 1989). Thus, temporal estimates of N_e in highly selfing populations should be treated with caution.

Interestingly, our simulations showed that examining the trajectory of MLG frequencies through time gives insights into the demographic history of a predominantly selfing population. In particular, the MLG frequency spectrum (*MLGFS*) describes the upper and lower bounds for the trajectory of MLG frequencies between two generations for a given demographic scenario. For example, a strong increase in *MFMLG* between two generations suggests a bottleneck. Although *MLGFS* are more difficult to interpret on empirical data because of the absence of replicates (e.g., Fig. S6), we show that they can provide support for a hypothesis of extinction–recolonization, along with large values of temporal differentiation.

Finally, our study also highlights high temporal stochasticity in *M. truncatula* natural populations. Indeed, diversity often changed over time, and the temporal samples of a given population were not clustered together in joint analyses of diversity indices (Figs. 1c and 2c). This temporal variability was also described in *Bulinus forskalii* by Gow et al. (2007), who attributed it to "highly dynamic demographic systems, including bottleneck and extinction–recolonization events". Larger variance in diversity levels among selfing compared to outcrossing populations has also been reported before (Schoen and Brown 1991).

Conclusion

The comparison of our simulation results with data obtained in a highly selfing species (*Medicago truncatula*) highlighted the pertinence of our simulation approach. Yet, the number of scenarios and parameter values we explored were limited and this could limit the generalization of our results to other datasets (e.g., other molecular markers, other sampling tempo). We expect, however, that the general patterns highlighted here using microsatellites will remain unchanged with other molecular markers such as a large number of genome-wide SNPs. In addition, the scripts are available and easily amendable to fine tune the comparison with other empirical datasets (in terms of population size, sample size, selfing rate, time span between sampling, etc.).

The demographic scenarios examined here were sufficient to show that selection is not required to explain the prevalence of repeated MLGs in predominantly selfing populations and their persistence through time. If background selection or selective sweeps are expected to reduce the effective size and would reduce single and multilocus diversity concomitantly (Glémin 2007; Kamran-Disfani and Agrawal 2014), the effects of complex selection scenarios such as local adaptation are more difficult to predict. Simulating scenarios involving selection was beyond the scope of this study but would be useful to address the question of the threshold selfing rate beyond which selection will act at the MLG (or haplotype) level rather than at the locus level, due to the severely reduced effective recombination (Neher and Shraiman 2009). Another perspective to this work will be to develop an integrated method to infer parameters such as the effective size and the selfing rate, based on the summary statistics described here (using a likelihood-free inference method, e.g., Beaumont et al. 2002; Rousset et al. 2016).

Data archiving

Genotype data of *Medicago truncatula* natural populations are available on the INRA dataportal. https://doi.org/10.15454/VCZIMR

The scripts used for the simulations and the computation of diversity indicators are available on the INRA dataportal. https://doi.org/10.15454/VYPXIJ

Acknowledgments M.J.'s PhD fellowship is funded by the INRA (French National Institute of Agronomical Research) Department of Genetics and Plant Breeding and the INRA Metaprogram ACCAF. The authors thank J.M. Prosperi for the collection of seeds as well as C. Tollon, F. Mora, E. Figuet, and J. Terraillon who contributed to the production of the microsatellite dataset. We thank Mathieu Siol, John Pannell and three anonymous reviewers for comments on a previous version of the manuscript. Analyses were performed on the CIRAD—UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.southgreen.fr/). Funding was provided by the Agence Nationale de la Recherche (ANR SEAD-ANR-13-ADAP-0011).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abu Awad D, Roze D (2018) Effects of partial selfing on the equilibrium genetic variance, mutation load, and inbreeding depression under stabilizing selection. Evol Int J Org Evol 72:751–769

Abu Awad D, Gallina S, Bonamy C, Billiard S (2014) The interaction between selection, demography and selfing and how it affects population viability PLoS ONE 9:e86125

Allard RW (1975) The mating system and microevolution. Genetics 79:Suppl, 115–126

Arrighi J-F, Barre A, Amor BB, Bersoult A, Soriano LC, Mirabella R, Carvalho-Niebel F, de, Journet E-P, Ghérardi M, Huguet T et al. (2006) The medicago truncatula lysine motif-receptor-like kinase gene family includes NFP and new nodule-expressed genes. Plant Physiol 142:265–279

Avise JC, Tatarenkov A (2012) Allard's argument versus Baker's contention for the adaptive significance of selfing in a hermaphroditic fish. Proc Natl Acad Sci 109:18862–18867

Bailey SF, Bataillon T (2016) Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol Ecol 25:203–218

Baker HG (1967) Support for Baker's law-as a rule. Evolution 21:853–856

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J (2006) Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. Mol Ecol 15:1405–1418

Baquerizo-Audiot E, Desplanque B, Prosperi JM, Santoni S (2001) Characterization of microsatellite loci in the diploid legume *Medicago truncatula* (barrel medic). Mol Ecol Notes 1:1–3

Barrière A, Félix M-A (2007) Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. Genetics 176:999–1011

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

Bomblies K, Yant L, Laitinen RA, Kim S-T, Hollister JD, Warthmann N, Fitz J, Weigel D (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet 6:e1000890

Bonnin I, Ronfort J, Wozniak F, Olivieri I (2001) Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). Mol Ecol 10:1371–1383

Caballero A, Hill WG (1992) Effects of partial inbreeding on fixation rates and variation of mutant genes. Genetics 131:493–507

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205

Charlesworth D, Charlesworth B (1995) Quantitative genetics in plants: the effect of breeding system on genetic variability. Evol Int J Org Evol 49:911–920

Crow JF, and Kimura M (1970). An Introduction to Population Genetics Theory (Harper & Row, New York)

David P, Pujol B, Viard F, Castella V, Goudet J (2007) Reliable selfing rate estimates from imperfect population genetic data. Mol Ecol 16:2474–2487

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11:189–212

Frachon L, Libourel C, Villoutreix R, Carrère S, Glorieux C, Huard-Chauveau C, Navascués M, Gay L, Vitalis R, Baron E et al. (2017) Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. Nat Ecol Evol 1:1551

Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177:905–916

Glémin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc R Soc B Biol Sci 273:3011–3019

Golding GB, Strobeck C (1980) Linkage disequilibrium in a finite population that is partially selfing. Genetics 94:777–789

Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, Picó FX (2011) Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. Mol Ecol 20:3540–3554

Goudet J (2005) hierfstat, a package for R to compute and test hierarchical *F*-statistics. Mol Ecol Notes 5:184–186

Gow JL, Noble LR, Rollinson D, Tchuem Tchuenté L-A, Jones CS (2007) Contrasting temporal dynamics and spatial patterns of population genetic structure correlate with differences in demography and habitat between two closely-related African freshwater snails. Biol J Linn Soc 90:747–760

Haller BC, Messer PW (2017) SLiM 2: flexible, interactive forward genetic simulations. Mol Biol Evol 34:230–240

Hamrick JL, Godt MJW (1997) Allozyme diversity in cultivated crops. Crop Sci 37:26–30

Hartfield M, Bataillon T, Glémin S (2017) The evolutionary interplay between adaptation and self-fertilization. Trends Genet 33:420–431

Hartl D, and Clark AG (1998). Principles of Population Genetics (Sinauer Associates)

Hereford J (2010) Does selfing or outcrossing promote local adaptation? Am J Bot 97:298–302

Hughes PW, Simons AM (2015) Microsatellite evidence for obligate autogamy, but abundant genetic variation in the herbaceous monocarp *Lobelia inflata* (Campanulaceae). J Evol Biol 28:2068–2077

Igic B, Kohn JR (2006) The distribution of plant mating systems: study bias against obligately outcrossing species. Evol Int J Org Evol 60:1098–1103

Ingvarsson PK (2002) A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. Evol Int J Org Evol 56:2368–2373

Kamran-Disfani A, Agrawal AF (2014) Selfing, adaptation and background selection in finite populations. J Evol Biol 27:1360–1371

Kamvar, ZN, Tabima, JF, and Grünwald, NJ (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281

Lande R, Porcher E (2015) Maintenance of quantitative genetic variance under partial self-fertilization, with implications for evolution of selfing. Genetics 200:891–906

Lundemo S, Falahati-Anbaran M, Stenøien HK (2009) Seed banks cause elevated generation times and effective population sizes of *Arabidopsis thaliana* in northern Europe. Mol Ecol 18:2798–2811

Lynch M, Conery J, Bürger R (1995) Mutational meltdowns in selfing populations. Evol Int J Org Evol 49:1067–1080

Marriage TN, Hudman S, Mort ME, Orive ME, Shaw RG, Kelly JK (2009) Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). Heredity 103:310–317

Meunier C, Hurtrez-Bousses S, Durand P, Rondelaud D, Renaud F (2004) Small effective population sizes in a widespread selfing species, *Lymnaea truncatula* (Gastropoda: Pulmonata) Mol Ecol 13:2535–2543

Montesinos A, Tonsor SJ, Alonso-Blanco C, Picó FX (2009) Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments PLoS ONE 4:e7213

Neher RA, Shraiman BI (2009) Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proc Natl Acad Sci 106:6866–6871

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154:923–929

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 89:367–370

Palstra FP, Ruzzante DE (2008) Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence? Mol Ecol 17:3428–3447

Pannell JR, Charlesworth B (2000) Effects of metapopulation processes on measures of genetic diversity. Philos Trans R Soc B Biol Sci 355:1851–1864

Pollak E (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117:353–360

R Core Team (2018) R: The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi J-M (2006) Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. BMC Plant Biol 6:28

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

Rousset F, Gouy A, Martinez-Almoyna C, Courtiol A (2016) The summary-likelihood method and its implementation in the Infusion package. Mol Ecol Resour 17:110–119

Schoen DJ, Brown AH (1991) Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proc Natl Acad Sci USA 88:4494–4497

Siol M, Bonnin I, Olivieri I, Prosperi JM, Ronfort J (2007) Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual *Medicago truncatula*. J Evol Biol 20:2349–2360

Siol M, Prosperi JM, Bonnin I, Ronfort J (2008) How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. Heredity 100:517–525

Stebbins GL (1957) Self fertilization and population variability in the higher plants. Am Nat 91:337–354

Stenøien HK, Fenster CB, Tonteri A, Savolainen O (2005) Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. Mol Ecol 14:137–148

Stoffel MA, Esser M, Kardos M, Humble E, Nichols H, David P, Hoffman JI, Poisot T (2016) inbreedR: an R package for the analysis of inbreeding based on genetic markers. Methods Ecol Evol 7:1331–1339

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

Thuillet A-C, Bru D, David J, Roumet P, Santoni S, Sourdille P, Bataillon T (2002) Direct estimation of mutation rate for 10 microsatellite loci in Durum wheat, *Triticum turgidum* (L.) Thell. ssp durum desf. Mol Biol Evol 19:122–125

Trouvé S, Degen, Goudet J (2005) Ecological components and evolution of selfing in the freshwater snail *Galba truncatula*. J Evol Biol 18:358–370

Viard F, Justy F, and Jarne P (1997) Population dynamics inferred from temporal variation at microsatellite loci in the selfing snail *Bulinus truncatus*. Genetics 146:973–982

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121:379–391

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

Weir BS, Hill WG (2002) Estimating *F*-statistics. Annu Rev Genet 36:721–750

Yampolsky C, and Yampolsky H (1922). Distribution of sex forms in the phanerogamic flora (Gebrüder Borntraeger)

Supplementary Information

The following Supplementary Information is available for this article:

Supplementary Figure S1: Location of the *Medicago truncatula* populations

Supplementary Figure S2: Correlation between the frequency of the most frequent MLG (MFMLG) and Shannon's multilocus diversity index (H) in the simulated scenarios

Supplementary Figure S3: Covariation between Shannon's diversity index (H) and the gene diversity (H_E) across simulated scenarios

Supplementary Figure S4: Correlation between the mean distance between two individuals within a population D_{mean} and gene diversity (H_E) in the simulated scenarios

Supplementary Figure S5: Correlation between the maximum pairwise distance (D_{max}) and the percentage of loci under significant linkage disequilibrium (LD%) in simulated predominantly selfing populations ($\sigma \geq 0.95$)

Supplementary Figure S6: MLG frequency spectra of *Medicago truncatula* natural populations

Supplementary Figure S7: Estimates of single (H_E) and multilocus diversity ($MFMLG$) for different sample sizes (N_{samp})

Supplementary Table S1: Detail of the parameters used for each simulated scenario

Supplementary Table S2: Diversity indices for the simulated scenarios and the empirical data

Figure S1: Location of the *Medicago truncatula* populations.

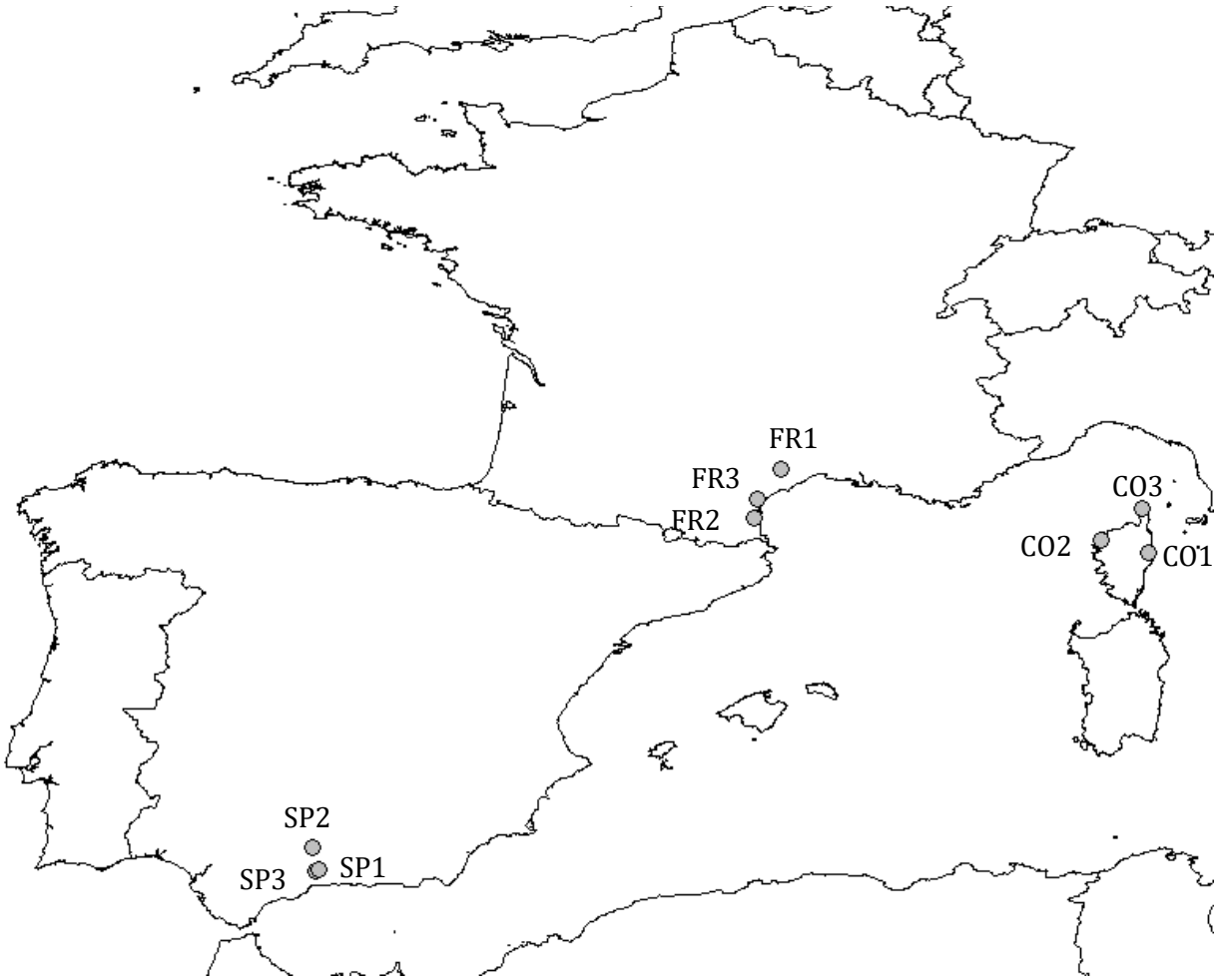


Figure S2: Correlation between the frequency of the most frequent MLG (*MFMLG*) and Shannon's multilocus diversity index (*H*) in the simulated scenarios.

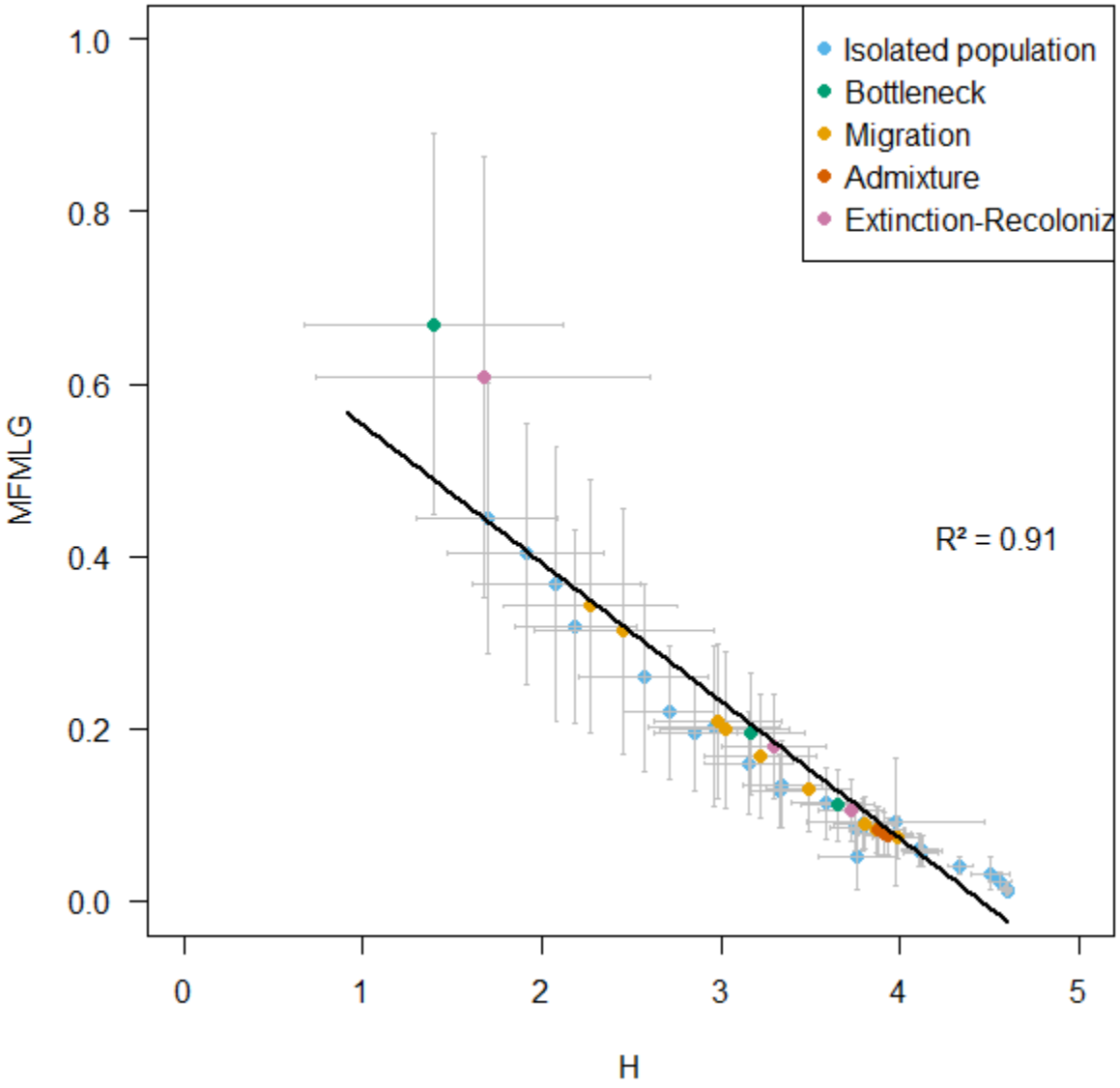


Figure S3: Covariation between Shannon's diversity index (*H*) and the gene diversity (H_E) across simulated scenarios.

(A) Scenarios of isolated populations with varying sizes *N* and selfing rates σ; (B) scenarios of migration (orange), admixture (black), bottleneck (blue) and extinction-recolonization (light blue) with $\sigma = 0.95$; (C) natural populations of *Medicago truncatula* for each sampling date. For (A) and (B), points indicate means, horizontal and vertical bars stand for the standard deviation across the 1000 replicates.

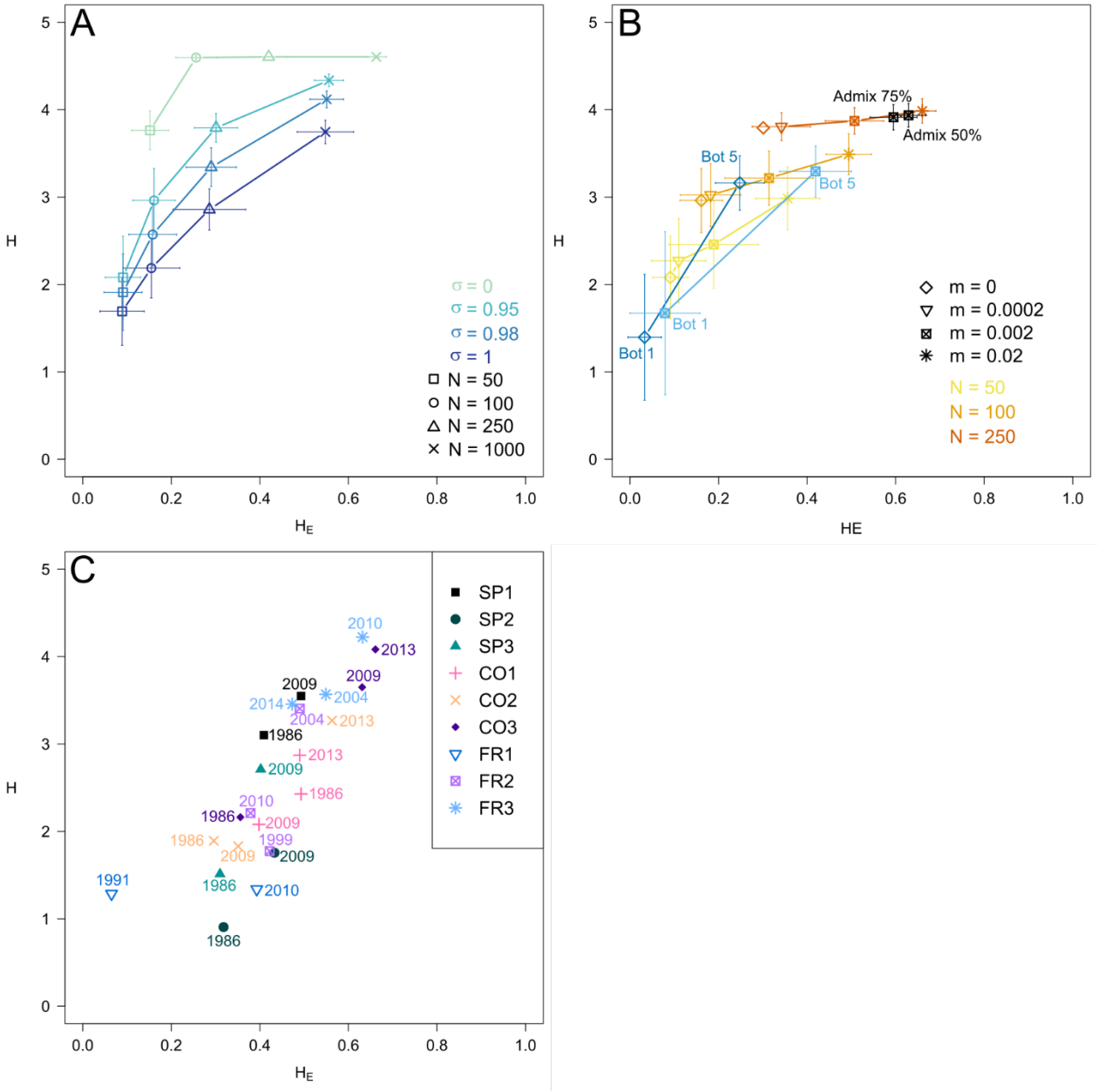


Figure S4: Correlation between the mean distance between two individuals within a population D_{mean} and gene diversity (H_E) in the simulated scenarios.

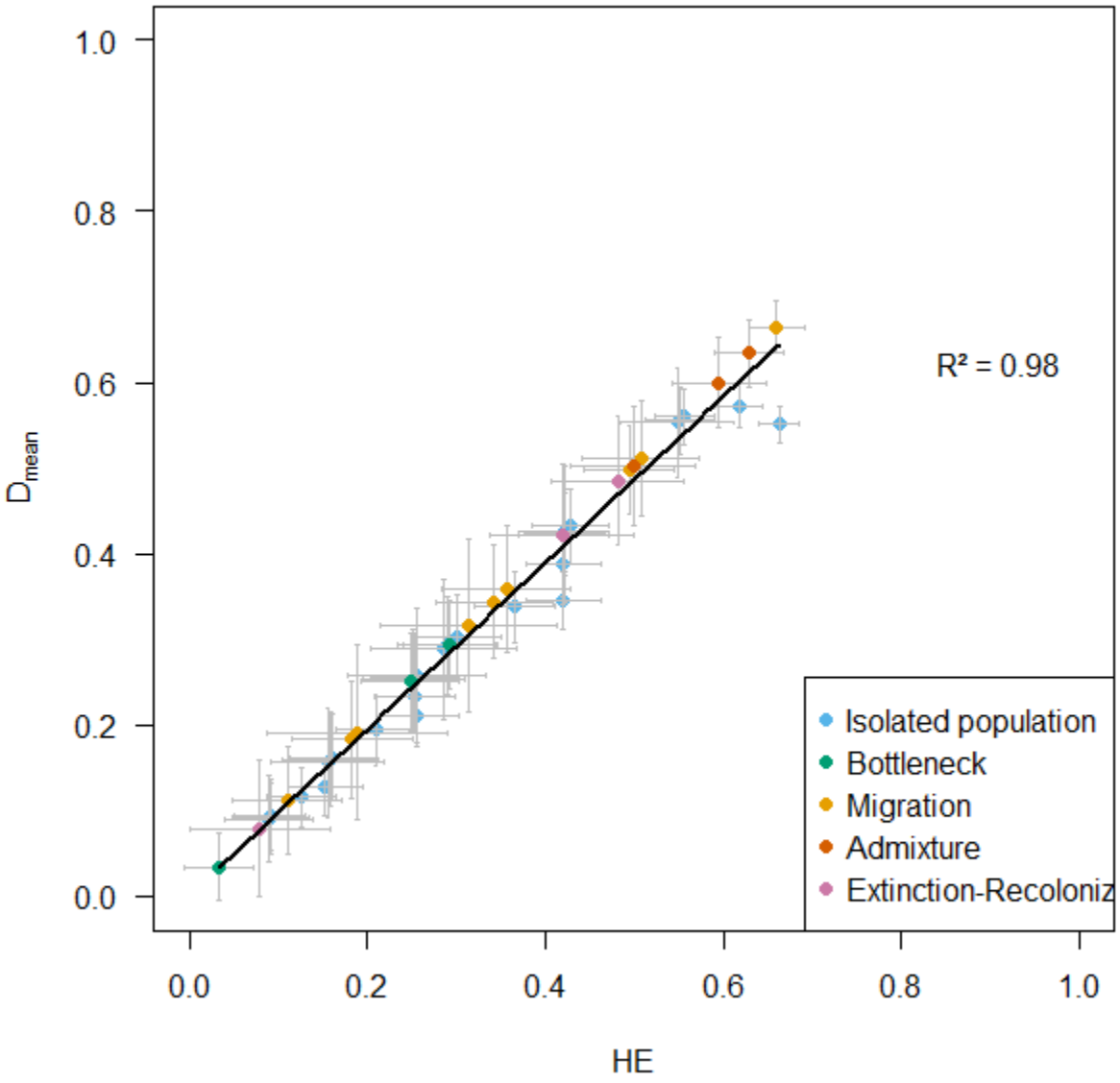


Figure S5: Correlation between the maximum pairwise distance (D_{max}) and the percentage of loci under significant linkage disequilibrium ($LD\%$) in simulated predominantly selfing populations ($\sigma \geq 0.95$).

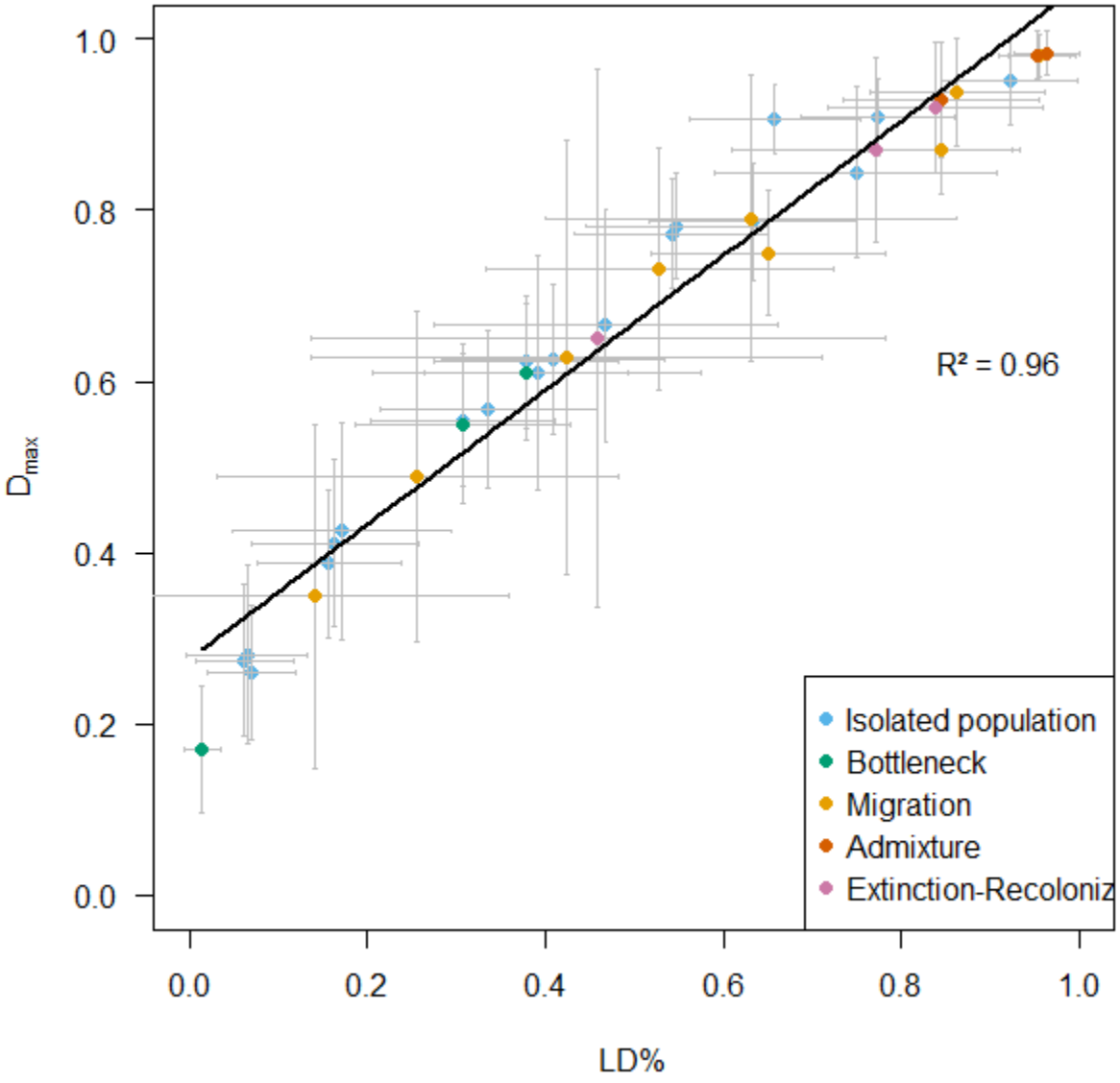
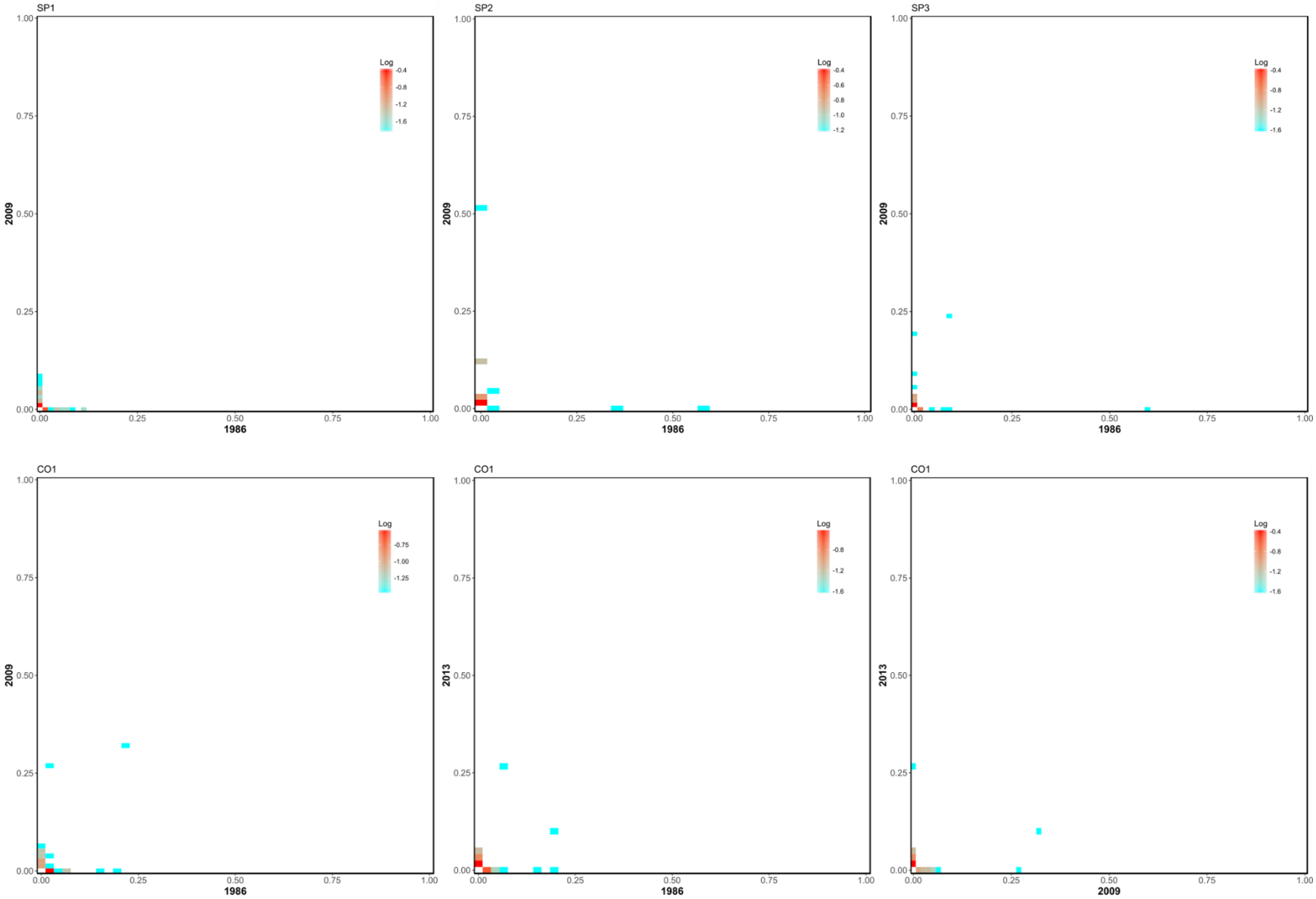
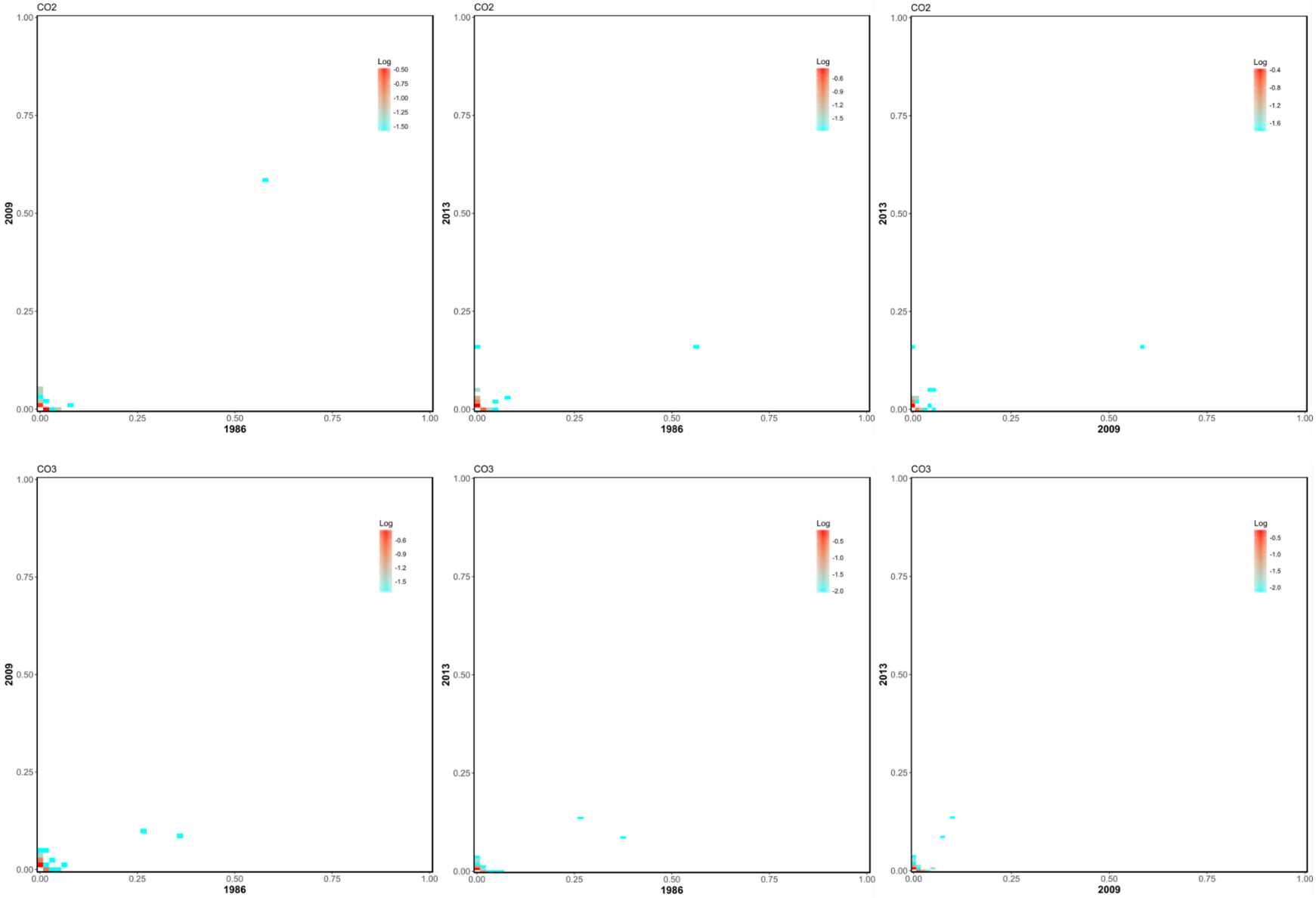
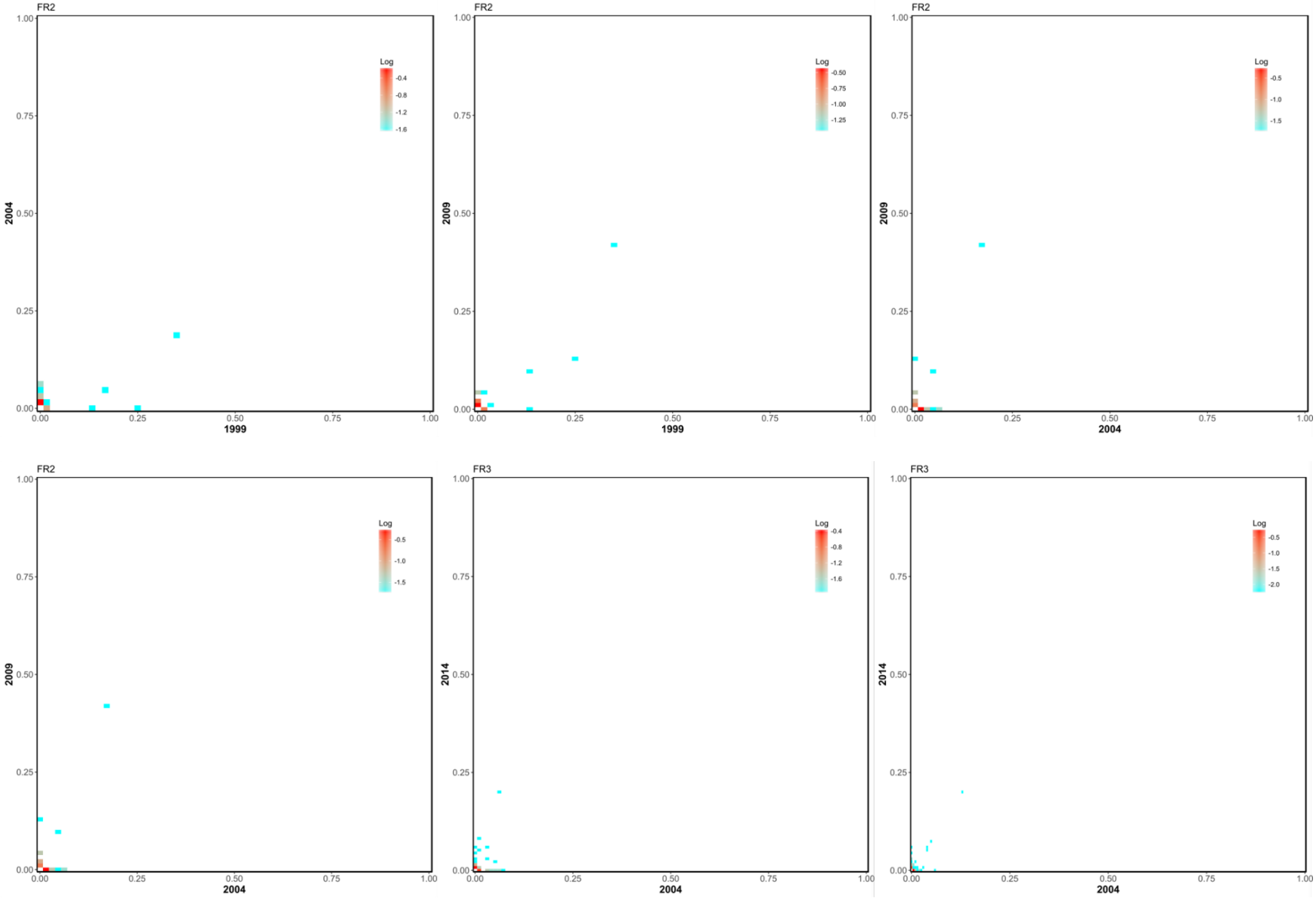


Figure S6: MLG frequency spectra of *Medicago truncatula* natural populations.




CO2
Log
-0.50
-0.75
-1.00
-1.25
-1.50
2009
1986
CO2
Log
-0.6
-0.9
-1.2
-1.5
2013
1986
CO2
Log
-0.4
-0.8
-1.2
-1.6
2013
2009
CO3
Log
-0.6
-0.9
-1.2
-1.5
2009
1986
CO3
Log
-0.5
-1.0
-1.5
-2.0
2013
1986
CO3
Log
-0.5
-1.0
-1.5
-2.0
2013
2009
0.00
0.25
0.50
0.75
1.00


FR2
Log
-0.4
-0.8
-1.2
-1.6
2004
1999
FR2
Log
-0.50
-0.75
-1.00
-1.25
2009
1999
FR2
Log
-0.5
-1.0
-1.5
2009
2004
FR2
Log
-0.5
-1.0
-1.5
2009
2004
FR3
Log
-0.4
-0.8
-1.2
-1.6
2014
2004
FR3
Log
-0.5
-1.0
-1.5
-2.0
2014
2004

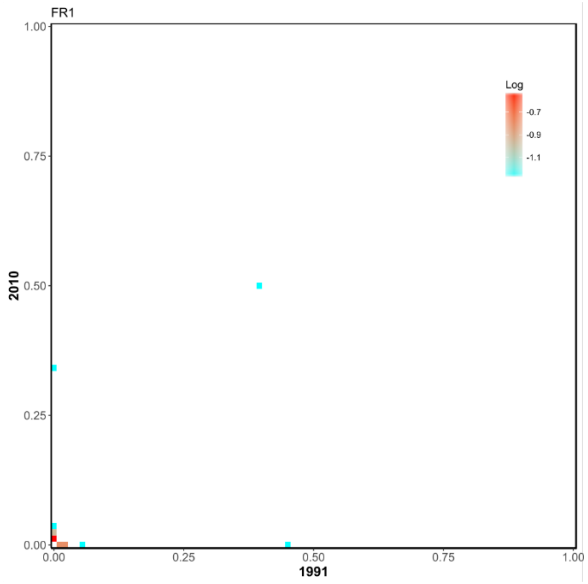

FR1
Log
-0.7
-0.9
-1.1
1.00
0.75
0.50
0.25
0.00
2010
0.00
0.25
0.50
0.75
1.00
1991

Figure S7: Estimates of single (H_E) and multilocus diversity (*MFMLG*) for different sample sizes (*Nsamp*). The sampling (N_{samp} = 5, 10, 20, 30, 50 or 100) was reiterated independently 100 times in a single simulation with N = 100 and σ = 0 (black dots), 0.95 (orange dots), or 1 (blue dots). The error bars represent the standard deviation between sampling replicates.

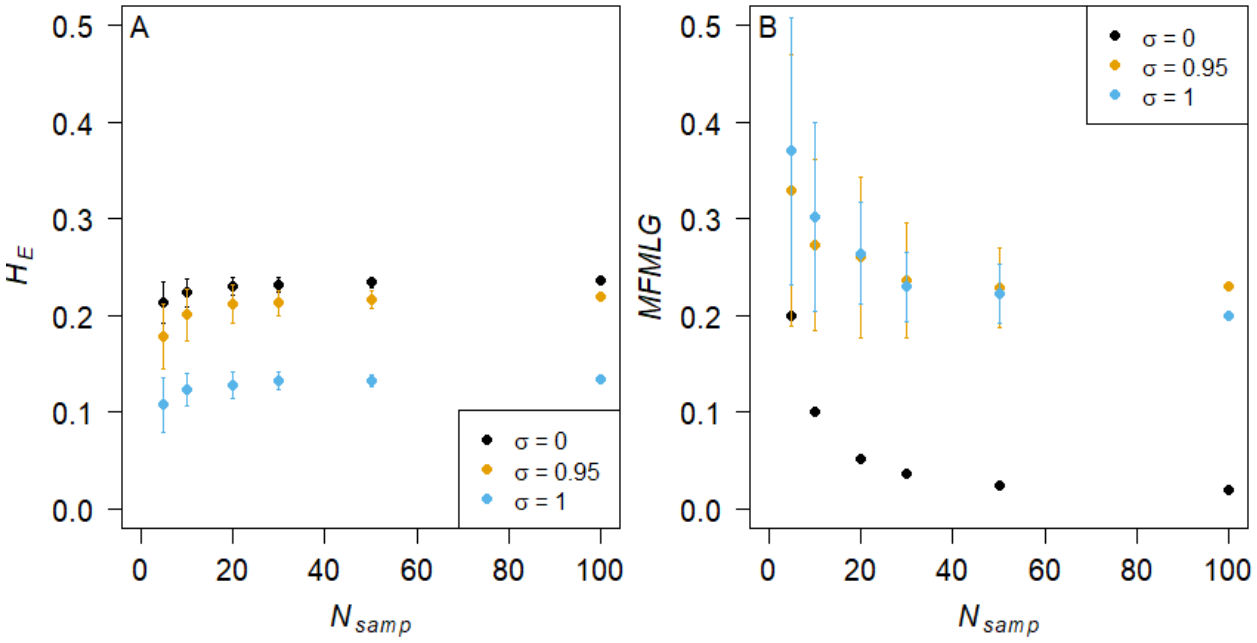


Table S1: Detail of the parameters used for each simulated scenario. σ is the selfing rate; N is the census size; m is the migration rate.

Model	σ	N	m	% admixture	bottleneck
Outcrossing	0	[50; 100; 250; 1000]	0	0	0
Mixed Mating	0.5	[50; 100; 133; 250; 333; 1000]	0	0	0
Predominantly selfing 0.95	0.95	[50; 100; 190; 250; 475; 1000]	0	0	0
Predominantly selfing 0.98	0.98	[50; 100; 196; 250; 490; 1000]	0	0	0
Selfing	1	[50; 100; 200; 250, 500; 1000]	0	0	0
Bottleneck	0.95	250	0	0	[1; 5; 25]
Migration	0.95	[50; 100; 250]	[0.0002; 0.002; 0.02]	0	0
Admixture	0.95	250	0.002	[50; 75; 100]	0
Extinction-recolonization	0.95	250	0.002	100	[1; 5; 25]

Chapitre 3 :

Inférence conjointe du taux d'autofécondation, de la taille efficace et de la migration chez des populations naturelles de *Medicago truncatula* à partir de données temporelles

Introduction

L'histoire démographique d'une population naturelle contribue à façonner la diversité génétique ainsi que l'organisation de cette diversité (Wright, 1949). Les processus qui caractérisent l'histoire démographique sont principalement la migration et les changements de taille démographique. Une des questions centrales en génétique des populations est l'estimation de l'histoire démographique des populations naturelles afin de mieux comprendre les processus qui ont abouti à la diversité actuelle. La littérature scientifique est donc très riche en études cherchant à inférer l'histoire démographique de populations ou d'espèces à partir de données génétiques et les développements méthodologiques associés à ces questions sont très nombreux (voir Pool *et al.* 2010 ; Beichman *et al.* 2018 pour des revues de méthodes et exemples d'applications). Nous avons pu voir au chapitre précédent que le système de reproduction, et notamment l'autofécondation, a également une influence forte sur le niveau de diversité et sur la structure génétique des populations. L'autofécondation peut amplifier les effets de différents processus démographiques, tels que la migration ou les goulots d'étranglement, sur la diversité ; ou produire des effets similaires à ces processus. Il est donc parfois difficile de distinguer les effets respectifs de l'autogamie de ceux liés à la démographie des populations. En effet, nous avons vu dans le Chapitre 2 que l'autofécondation a des effets similaires à ceux de goulots d'étranglement sur la fréquence du génotype multilocus majoritaire (*MFMLG*) et à ceux de la migration sur la distance génétique maximale entre paire d'individus (D_{max}). Il est donc important de prendre en compte le régime de reproduction lors de l'estimation de l'histoire démographique des populations.

Par ailleurs, l'utilisation de séries temporelles est particulièrement intéressante pour estimer les processus démographiques contemporains. En effet, les données temporelles permettent de regarder les fluctuations récentes de la diversité et donnent donc un accès indirect aux processus démographiques à l'origine de ces fluctuations.

Comme décrit en Introduction, la taille efficace d'une population (N_e) est un paramètre d'intérêt utilisé très régulièrement en génétique des populations, en biologie évolutive et en biologie de la conservation. Elle traduit l'intensité de la dérive génétique à l'œuvre dans une population. Elle est particulièrement intéressante dans le cas de populations autogames pour lesquelles on s'attend à des tailles efficaces réduites du fait de l'autofécondation (Glémin 2007, Introduction). Cependant, l'estimation de N_e se révèle souvent compliquée et peu précise car la plupart des populations naturelles, et surtout les populations autogames, ne sont pas conformes aux hypothèses des modèles d'estimation. De nombreuses méthodes d'estimation à partir de données génétiques existent, avec

des performances variables (pour une revue, voir Wang 2005 ; Wang *et al.* 2016 et l'Introduction). La méthode dite temporelle, qui est basée sur les changements de fréquences alléliques dans une population au cours du temps, est l'une des plus populaires (Krimbas and Tsakas, 1971 ; Waples, 1989). Elle repose cependant sur des hypothèses simplificatrices (panmixie, population isolée de taille constante, absence de sélection), qui sont rarement respectées dans les populations naturelles et peuvent résulter en des estimations biaisées ou peu précises. En particulier, la méthode temporelle ne prend pas en compte l'excès d'homozygotie dû à l'autogamie (Bécheler, 2014). De plus, outre l'écart à la panmixie, les populations autogames présentent souvent des dynamiques en métapopulations avec des taux de migration variables ainsi que des variations de taille démographique au cours du temps (Ingvarsson, 2002). On s'attend donc à ce que la méthode temporelle soit peu adaptée pour estimer la taille efficace de populations autogames et ceci a été validé par nos simulations (Chapitre 2). Gilbert and Whitlock (2015) ont comparé les performances de plusieurs estimateurs de N_e parmi les plus utilisés dans la littérature avec des scénarios incluant, ou non, de la migration (mais sans autofécondation). Ils formulent des recommandations mais notent que la performance d'un estimateur donné dépend du scénario démographique considéré. Ainsi, ils préconisent de connaître au préalable le scénario démographique sous-jacent aux données considérées afin d'adapter la méthode d'estimation. Waples (2016) suggère de combiner différents estimateurs indépendants, tels qu'un estimateur temporel (Waples, 1989) et un estimateur basé sur le déséquilibre de liaison (Waples and Do, 2008), pour améliorer la précision des estimations de N_e. Cependant, comme le notent Gilbert and Whitlock (2015), cela n'est pas suffisant lorsque les estimations sont biaisées. Une prise en compte explicite de la migration et du taux d'autofécondation au sein du modèle d'estimation pourraient améliorer l'estimation.

Un autre paramètre important pour décrire l'histoire démographique d'une population est le taux de migration. Une estimation indirecte de la migration efficace (N_em, le nombre efficace d'individus migrants par population par génération) peut être obtenue à partir de la variance des fréquences alléliques entre populations grâce à l'approximation $F_{ST} \simeq \frac{1}{1+4N_em}$ (Wright, 1949), où F_{ST} est le rapport des probabilités d'identité au sein et entre les populations (Wright, 1949), N_e est la taille efficace de chaque population et m le taux de migration entre les populations. Cette approximation repose cependant sur les hypothèses contraignantes du modèle de populations en îles (Wright, 1931). En particulier, on suppose que les populations sont à l'équilibre migration-dérive et que la taille de population et le taux de migration sont uniformes entre les populations et constants au cours du temps. Whitlock and McCauley (1999) présentent une revue des violations d'hypothèses couramment rencontrées en conditions naturelles. Viard *et al.* (1996) ont utilisé cette méthode sur des populations

naturelles de *Bulinus truncatus*, un escargot aquatique majoritairement autogame, mais soulignent que l'estimation de N_em est particulièrement peu adaptée en régime autogame. De nombreuses autres méthodes d'estimation de la migration ont été proposées. Beerli et Felsenstein (2001) ont notamment développé un estimateur basé sur le coalescent qui accommode des modèles plus réalistes avec des taux de migrations asymétriques et des tailles de populations variables au sein du programme MIGRATE (Beerli, 2006). Cependant, les estimations de taux de migration avec MIGRATE sont peu fiables (Abdo *et al.*, 2004). D'autre part, ces différentes méthodes mesurent des flux de gènes historiques et sont insensibles à de faibles changements de fréquences alléliques. Elles ne sont donc pas appropriées pour l'étude des flux de gènes contemporains. Les flux de gènes contemporains peuvent être estimés directement à l'aide de tests d'assignation (Paetkau *et al.* 1995). Cette méthode suppose l'absence de déséquilibre de liaison entre les locus considérés, ce qui peut être problématique lorsqu'on considère des espèces autogames. En effet, comme expliqué en Introduction, l'autogamie augmente le déséquilibre de liaison à travers l'augmentation de l'homozygotie (Caballero et Hill 1992 ; Nordborg 2000) et les locus ne sont donc pas forcément indépendants. De plus, elle nécessite d'échantillonner toutes les populations sources de migrants, ce qui peut demander un effort d'échantillonnage trop important (temps, ressources, etc.). En populations autogames, la plupart des études n'infèrent pas un taux de migration. Elles testent l'hypothèse de la présence ou de l'absence de migration en comparant les compositions des différentes populations étudiées, notamment les génotypes partagés entre populations (ex : Bakker *et al.* 2006 ; Montesinos *et al.* 2009 ; Bomblies *et al.* 2010).

A notre connaissance, il n'existe pas de méthode d'inférence démographique basée sur des données temporelles et prenant en compte à la fois le taux d'autofécondation, la migration et les variations de taille efficace. Les données temporelles peuvent fournir des informations importantes pour comprendre l'histoire démographique des populations. Par ailleurs, les différents paramètres démographiques et leurs effets sont difficiles, voire impossible, à séparer en n'utilisant qu'un seul type d'information (ex : déséquilibre de liaison, hétérozygotie, etc.). Une méthode statistique basée sur le maximum de vraisemblance combinant tous les paramètres qui nous intéressent et utilisant plusieurs descripteurs des données génétiques demanderait des développements théoriques très extensifs. Les méthodologies d'inférence par calcul Bayésien approché (ou ABC, de l'anglais *approximate Bayesian computation* ; Tavaré *et al.* 1997 ; Beaumont *et al.* 2002) semblent fournir une bonne alternative. En effet, l'ABC permet de tester des scénarios complexes à l'aide d'un grand nombre de simulations sans avoir besoin d'écrire une fonction de vraisemblance. Une description générale des principes de l'ABC se trouve dans l'Encadré 2. Des développements récents s'appuyant

sur des méthodes d'apprentissage automatique (*machine learning*) comme les forêts aléatoires (ABC-RF, Pudlo *et al.* 2016, méthode décrite dans l'Encadré 3), ont permis d'améliorer l'efficience du processus et donc de réduire le nombre de simulations nécessaires. L'ABC avec forêts aléatoires a été utilisée avec succès pour inférer l'histoire démographique de différentes espèces, en testant des scénarios parfois complexes (ex : Fraimout *et al.* 2017 ; Estoup *et al.* 2018 ; Smith *et al.* 2018 ; Lombaert *et al.* 2018).

Dans ce chapitre, nous cherchons à développer un cadre statistique permettant d'inférer un modèle démographique et d'estimer conjointement les paramètres qui le caractérisent à partir de données temporelles chez des populations autogames. Nous cherchons en particulier à déterminer le rôle de la migration et des variations de taille démographique au cours du temps dans des populations naturelles autogames et ainsi à distinguer l'effet de l'autogamie de celui de l'histoire démographique « vraie » des populations. Nous avons choisi d'utiliser les méthodes de calcul Bayésien approché avec des forêts aléatoires pour tester des scénarios démographiques contrastés. La méthode développée est appliquée sur les données de cinq populations naturelles de *Medicago truncatula*. Ces populations présentent des profils de diversité génétique contrastés qui suggèrent des histoires démographiques différentes (Jullien *et al.*, 2019). Les résultats obtenus sur les populations de *M. truncatula* permettront d'évaluer les performances de notre méthode et de définir des pistes d'amélioration potentielles.

Encadré 2 : Méthodologie d'inférence par calcul Bayésien approché (ABC)

De nombreuses méthodes statistiques ont été développées afin d'inférer des processus démographiques à partir de données moléculaires. Ces méthodes reposent principalement sur le calcul de la vraisemblance, qui décrit la probabilité des données sachant un ou plusieurs paramètres. La motivation principale derrière le développement et l'utilisation des méthodes ABC repose sur la difficulté, voire l'impossibilité, de calculer la vraisemblance dans le cas de modèles complexes. La méthode ABC permet en effet de s'affranchir du besoin de calculer la vraisemblance (*likelihood-free inference method*). Cette méthode a été introduite en génétique des populations par Tavaré *et al.* (1997) et a ensuite été généralisée par Beaumont *et al.* (2002).

Algorithme de réjection

Le principe de base en ABC repose sur un algorithme de réjection (Algorithme 1), qui commence par la simulation de N_{ref} jeux de données x selon un modèle m dont les paramètres θ sont tirés dans des distributions de densité de probabilité *a priori* π. Les jeux de données simulés sont ensuite synthétisés à l'aide d'un vecteur de statistiques résumées $S(x)$ (par exemple des indices de diversité). Cela permet la construction d'une table de référence de taille N_{ref} regroupant les valeurs des statistiques résumées pour les N_{ref} simulations. Les simulations sont sélectionnées en fonction de la distance d entre les statistiques résumées calculées sur les données observées $S(x^0)$ et les statistiques résumées sur les données simulées $S(x)$, en fonction d'un seuil d'acceptation ε. Les simulations conservées permettent soit d'estimer la probabilité *a posteriori* du modèle m, soit d'estimer les distributions de probabilité *a posteriori* des paramètres du modèle.

Algorithme 1 : Algorithme de réjection

1. Générer une table de référence à partir de N_{ref} simulations selon le modèle m et les distributions a priori $\pi(\theta|m)$
2. Conserver les simulations telles que $d(S(x^0), S(x)) \leq \varepsilon$

L'algorithme de réjection de base décrit ci-dessus présente plusieurs problèmes d'efficience et de précision des inférences : (i) le nombre de simulations à réaliser est très important et peut donc se révéler limitant dans le cas de données génomiques où plusieurs dizaines de milliers de marqueurs sont analysés ; (ii) les estimations sont sensibles à la sélection d'un vecteur de statistiques résumées informatives et parcimonieuses, qui est donc cruciale et délicate ; (iii) la fixation du seuil de distance ε permettant un bon compromis entre précision et efficacité reste arbitraire. De nombreux développements ont été proposés pour répondre à ces limites et améliorer la méthode. Ils sont brièvement présentés ici.

Encadré 2 (suite)

Vecteur de statistiques résumées

L'utilisation de statistiques résumées implique nécessairement une perte d'information par rapport aux données complètes. Il est donc crucial de choisir un vecteur de statistiques résumées les plus informatives possibles relativement à l'inférence que l'on souhaite réaliser. Par ailleurs, il parait naturel d'augmenter le nombre de statistiques résumées afin de représenter le plus précisément les données. Cependant, cela peut conduire à une diminution de la précision des inférences (Beaumont *et al.*, 2002). En effet, les algorithmes d'ABC souffrent du « fléau de la dimension » (*curse of dimensionality*) : plus la dimension du vecteur de statistiques résumées augmente, plus la probabilité qu'une simulation soit acceptée diminue.

Afin de résoudre ce problème, des méthodes de sélection des statistiques résumées et de réduction de dimension ont été proposées. Par exemple, Joyce and Marjoram (2008) ont développé une méthode permettant de classer les statistiques résumées en fonction de l'amélioration de l'inférence permise par leur inclusion dans l'analyse. Wegmann *et al.* (2009) ont proposé une technique de moindres carrés partiels (partial least-square, PLS) afin de réduire la dimension du vecteur de statistiques résumées en tenant compte des corrélations potentielles entre les nombreuses statistiques initiales. Blum *et al.* (2013) comparent plusieurs méthodes de réduction de dimension.

Calibration de l'algorithme ABC

Le seuil de distance ε utilisé pour rejeter les simulations trop éloignées des données observées est choisi de manière arbitraire. Il contraint le nombre de simulations à réaliser. En effet, pour ε proche de zéro, la plupart des simulations seront rejetées et il faudra donc beaucoup de simulations pour pouvoir estimer une distribution a posteriori à partir des simulations acceptées. Ce phénomène est exacerbé par le fléau de la dimension : plus il y a de statistiques résumées, plus ε doit être grand pour qu'une simulation soit acceptée.

Afin d'optimiser l'algorithme ABC, trois types de méthodes ont été proposés. Beaumont *et al.* (2002) ont développé une méthode de régression linéaire locale permettant d'élargir l'intervalle de tolérance pour le même taux d'erreur qu'avec un algorithme de réjection simple. Marjoram *et al.* (2003) ont proposé la méthode ABC Markov-Chain Monte Carlo (MCMC-ABC), qui permet d'échantillonner les distributions a priori de manière plus efficace : les valeurs de paramètres produisant des simulations proches des données observées sont échantillonnées préférentiellement. Cependant, l'algorithme MCMC-ABC devient peu efficace lorsqu'il échantillonne dans une zone de l'espace des paramètres de faible probabilité. Le troisième type de méthode, sequential Monte Carlo

(SMC-ABC), proposé par Sisson *et al.* (2007), permet d'éviter que l'algorithme reste « coincé » dans une zone de faible probabilité. La distribution a posteriori est approximée en utilisant des paramètres (dénommés particules) tirés dans une distribution a priori. Les particules qui génèrent des données simulées trop éloignées des données observées sont rejetées, alors que les autres permettent d'échantillonner des distributions intermédiaires jusqu'à atteindre la distribution a posteriori.

Toutes ces méthodes, ainsi que leurs extensions, sont décrites de manière détaillée dans la littérature, par exemple par Beaumont (2010).

Encadré 3 : ABC random forest

L'introduction de la méthodologie de forêt aléatoire, ou random forest (RF), au sein de l'ABC a été proposée par Pudlo *et al.* (2016) à partir de la méthode développée par Breiman (2001). Cette méthodologie permet de réaliser des inférences ABC à partir d'un nombre de simulations relativement restreint, sans avoir besoin de choisir entre différentes statistiques résumées et ne nécessite pas la définition d'un seuil de distance ε.

Qu'est-ce qu'un arbre de classification ou de régression ?

Une forêt aléatoire de taille N_{tree} est composée de T arbres de classification ou de régression (CART). Un arbre est un outil d'aide à la décision qui est constitué de nœuds internes et de feuilles terminales (Figure 12). L'arbre est construit de la racine jusqu'aux feuilles et chaque nœud ν correspond à une séparation de l'espace des données selon une règle binaire où l'on compare une covariable X_j à un seuil s_j. Prédire la valeur de Y revient à parcourir l'arbre depuis la racine en appliquant la succession de règles binaires jusqu'à atteindre une feuille terminale. La covariable X_j et le seuil s_j permettant une séparation optimale des données dans les nœuds ν_1 et ν_2 sont sélectionnés en minimisant un critère de divergence : $N(\nu_1)Q(\nu_1) + N(\nu_2)Q(\nu_2)$; avec $N(\nu_i)$ le nombre d'observations dans le nœud ν_i et $Q(\nu_i)$ le critère de divergence au nœud ν_i (par exemple l'indice de Gini, Gastwirth 1972).

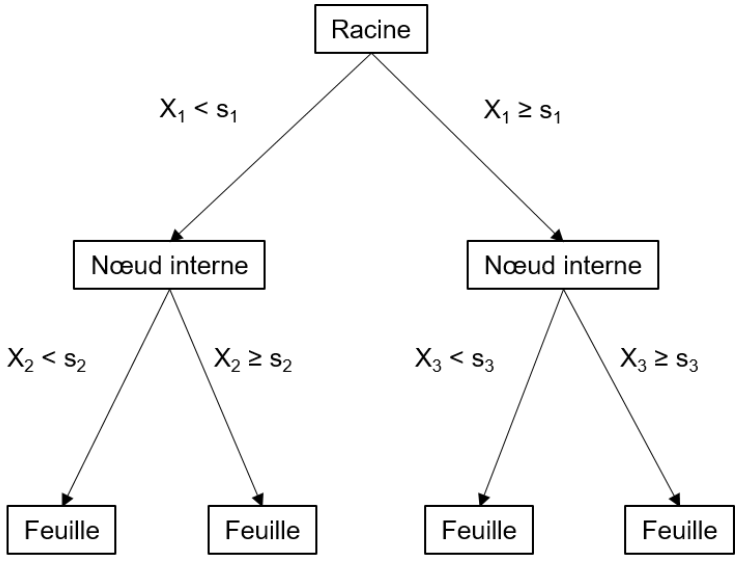


Figure 12 : Exemple d'arbre de classification et de régression.

Encadré 3 (suite)

La construction d'un arbre aléatoire peut être représentée par l'algorithme suivant :

Algorithme 2 : Construction d'un arbre aléatoire de classification ou de régression (adapté de Pudlo *et al.* 2016)

1. Initialiser l'arbre avec une racine unique
2. Répéter
   Choisir un nœud externe ν non homogène tel que $Q(\nu) > 0$ (classification) ou $N(\nu) > 5$
   Ajouter à ν deux nœuds fils ν_1 et ν_2
   Tirer aléatoirement n_{try} covariables
      Pour chaque covariable X_j : chercher le seuil t_j tel que $X_j < t_j$ minimise $N(\nu_1)Q(\nu_1) + N(\nu_2)Q(\nu_2)$
   Chercher la combinaison $X_j < t_j$ qui minimise $N(\nu_1)Q(\nu_1) + N(\nu_2)Q(\nu_2)$ et l'appliquer à ν
   Jusqu'à ce que tous les nœuds externes soient homogènes (classification) ou regroupent 5 observations (régression)
3. Définir la valeur de chaque nœud externe

Qu'est-ce qu'une forêt aléatoire ?

L'algorithme 3 décrit comment les arbres construits via l'algorithme 2 sont agrégés via la technique de « bagging » (*bootstrap aggregating*) au sein de la forêt aléatoire. L'algorithme de RF produit N_{tree} arbres construits à partir de N_{boot} échantillons issus de bootstrapping sur la table de référence. Chacun des N_{tree} arbres est construit à partir de N_{boot} observations et $N_{boot,stat}$ statistiques résumées selon l'algorithme 2. La technique de bagging permet de décorréler les arbres et ainsi de réduire la variance de prédiction (Hastie *et al.*, 2009). Une fois les N_{tree} arbres construits, on peut obtenir une prédiction pour le point x (Algorithme 3).

Encadré 3 (suite)

Algorithme 3 : Random forest pour régression ou classification (adapté de Hastie *et al.* (2009))

1. Générer une table de référence à partir de N_{ref} simulations selon le modèle m et les distributions a priori $\pi(\theta|m)$
2. Construire N_{tree} arbres aléatoires qui prédisent m à partir de $S(x)$ (algorithme 2)

   Pour $b = 1$ à N_{tree} :

   Echantillonner par bootstrap N_{boot} observations dans la table de référence

   Construire un arbre aléatoire T_b
3. Générer l'ensemble d'arbres $\{T_b\}_1^B$

Pour réaliser une prédiction à un nouveau point x :

- Régression : moyenne des réponses
- Classification : vote majoritaire sur tous les arbres de la forêt aléatoire

L'utilisation d'ABC RF pour la sélection de modèle et pour l'inférence de paramètres est motivée par le fait que les forêts aléatoires ne sont pas sensibles aux statistiques résumées fortement corrélées, ni aux statistiques peu ou pas informatives (Raynal *et al.*, 2017). En particulier, Raynal *et al.* (2017) ont montré la meilleure performance d'ABC RF par rapport aux algorithmes d'ABC classiques sur un jeu de données exemple en incluant vingt statistiques résumées non informatives (bruit). L'ABC RF permet donc d'éviter l'étape préliminaire de choix des statistiques résumées. En outre, la méthodologie RF permet de diminuer la taille des tables de référence, dont la génération est une étape particulièrement coûteuse en temps de calcul, notamment dans le cas de jeux de données de grande taille.

Matériel et méthodes

Génération des tables de référence

Construction des modèles démographiques

Pour chacun des modèles présentés par la suite, les caractéristiques du plan d'échantillonnage telles que le nombre d'échantillons temporels, la taille des échantillons (N_s), le nombre de locus microsatellites simulés (N_{loc}) et les intervalles de temps entre chaque échantillon (Δt) sont fixes et déterminés par les données empiriques (Tableau 3). Les autres paramètres sont tirés dans des lois de distribution de probabilité *a priori* et sont inférés par la suite.

Tableau 3 : Paramètres fixés par les données empiriques.
$n_{s,i}$ est la taille de l'échantillon i ; n_{loc} est le nombre de locus polymorphes simulés ; Δt est le nombre de générations entre les échantillons temporels

Population	n_s	n_{loc}	Δt
CO1	$n_{s,1} = 46$ $n_{s,2} = 78$ $n_{s,3} = 60$	15	$\Delta t_1 = 23$ $\Delta t_2 = 4$
CO3	$n_{s,1} = 64$ $n_{s,2} = 81$ $n_{s,3} = 162$	16	$\Delta t_1 = 23$ $\Delta t_2 = 4$
SP1	$n_{s,1} = 71$ $n_{s,2} = 93$	17	$\Delta t_1 = 23$
FR1	$n_{s,1} = 91$ $n_{s,2} = 82$	16	$\Delta t_1 = 19$
FR3	$n_{s,1} = 97$ $n_{s,2} = 201$ $n_{s,3} = 135$	18	$\Delta t_1 = 6$ $\Delta t_2 = 4$

Un taux de mutation arbitraire (mais réaliste pour des marqueurs microsatellites, voir Thuillet *et al.* 2002; Marriage *et al.* 2009), $\mu = 0.001$ par génération et par locus a été choisi pour générer la diversité initiale des populations simulées (selon le stepwise mutation model, de la même manière qu'au chapitre 2). Par ailleurs, nous supposons le taux de mutation (μ) suffisamment faible pour que la probabilité d'apparition d'une mutation entre les échantillons temporels soit négligeable. Les simulations sont réalisées avec le logiciel SLiM (version 2.6.0, Haller and Messer 2017).

Modèle de population isolée

Le premier modèle simulé considère une population isolée dont la taille démographique (N) et le taux d'autofécondation (σ) peuvent varier au cours du temps (**Figure 13**). Une nouvelle taille démographique et un nouveau taux d'autofécondation sont donc fixés après chaque échantillonnage.

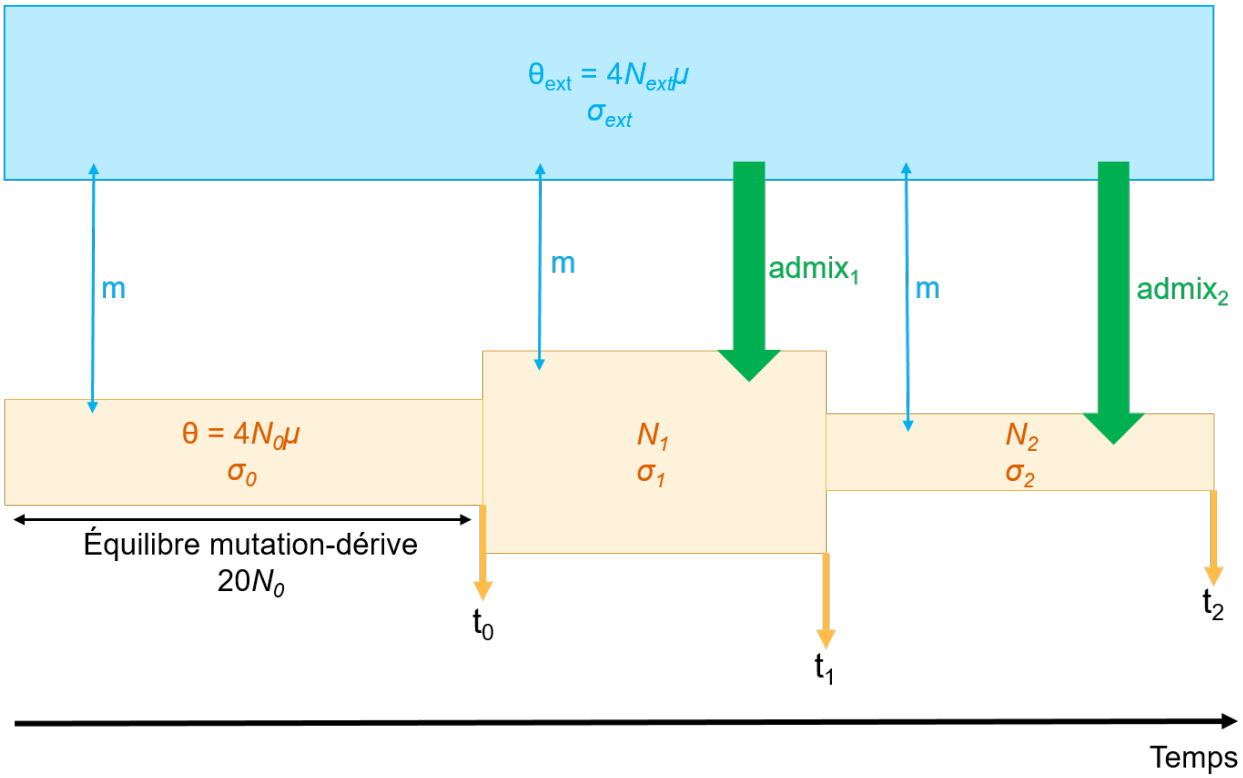


Figure 13 : Modèles simulés pour l'analyse ABC (cas avec 3 échantillons temporels).
Les paramètres du modèle 1 sont représentés en orange, les nouveaux paramètres introduits par le modèle 2 sont représentés en bleu et les paramètres du modèle 3 en vert. La population focale est représentée en orange et la population externe en bleu. θ et θ_{Ext} sont les diversités ancestrales de la population focale et de la population externe respectivement. N_0, N_1, N_2 et N_{Ext} sont les tailles démographiques simulées au cours du temps ; σ_0, σ_1, σ_2 et σ_{Ext} sont les taux d'autofécondation des différents segments temporels et de la population externe ; μ est le taux de mutation ; m est le taux de migration ; $admix_1$ et $admix_2$ sont les taux d'admixture précédant le deuxième et le troisième échantillonnage temporel respectivement.

Modèle de migration

Pour ce deuxième modèle, la population focale est connectée par un taux de migration symétrique (m) à une population externe dont la taille démographique (N_{ext}) et le taux d'autofécondation (σ_{ext}) sont tirés dans des distributions *a priori* (**Figure 13**). Contrairement au taux d'autofécondation et à la taille de la population focale, le taux de migration est constant au cours du temps.

Modèle de remplacement partiel

Le dernier modèle évalué est un modèle avec remplacement partiel (ou total) de la population focale par la population externe (**Figure 13**). La fréquence des génotypes issus d'admixture au sein de la population focale à un instant t dépend des évènements qui ont eu lieu entre le véritable moment où l'admixture se produit (t-x) et t, en particulier de la dérive génétique. Nous avons donc choisi de

contraindre les évènements d'admixture pour qu'ils aient lieu juste avant l'échantillonnage. Les génotypes migrants ne subissent donc pas d'évènement de reproduction, et donc pas de recombinaison via l'allofécondation résiduelle avant l'échantillonnage.

Choix des distributions a priori

Distribution a priori des taux d'autofécondation (σ)

Les taux d'autofécondation de la population focale au cours du temps (σ_0, σ_1, σ_2) et de la population externe (σ_{ext}) sont tirés dans une distribution Beta(α, β) avec α = 10 et $\beta = \alpha(1\text{-}S)/S$, où S est la moyenne de la distribution de probabilité a priori. De cette manière, les taux d'autofécondation sont corrélés et S peut être interprété comme le taux d'autofécondation moyen de l'espèce. Comme *Medicago truncatula* est une plante majoritairement autogame, nous n'explorons pas les faibles valeurs de taux d'autofécondation et la distribution de probabilité a priori de S est uniforme entre 0.7 et 1.

Distribution a priori des tailles démographiques (N)

La taille démographique initiale (N_0) est déterminée en fonction de l'indice de diversité $\theta = 4N_0\mu$ où θ représente la diversité génétique de la population ancestrale. μ est fixé et égal à 0.001. La valeur de θ est tirée dans une distribution log-uniforme entre 0.01 et 10, ce qui correspond à des valeurs de N_0 à simuler variant entre 1 et 2500 individus diploïdes. L'analyse RF permettra d'inférer $\log_{10}(\theta)$. L'utilisation de distributions log-uniformes permet de limiter l'incertitude autour des valeurs estimées lorsque la valeur réelle est petite (échantillonnage équivalent de tous les ordres de grandeur, Parag et Pybus 2019). On simule $t_{eq} = 20N_0$ générations avant le premier échantillon pour s'assurer d'atteindre l'équilibre mutation-dérive. Comme les simulations effectuées sont des simulations *forward*, la durée de la simulation peut fortement augmenter avec la valeur de N_0. Pour conserver un temps de calcul raisonnable, nous avons donc choisi de borner N_0 à 2500 individus diploïdes. Les tailles démographiques des autres segments temporels (N_1 et N_2) sont également tirées dans des distributions log-uniformes entre 1 et 5000. Il est important de noter que le paramètre N_0 n'est pas la taille démographique au premier échantillon temporel car il est déterminé par $\theta = 4N_0\mu$ et dépend donc du taux de mutation (fixé arbitrairement ici). Θ et N_1 ne sont donc pas directement comparables et on ne peut pas inférer de changements de taille démographique entre t_0 et t_1. De la même manière que pour N_0, la taille démographique de la population externe (N_{ext}) est déterminée par θ_{ext}, tiré dans une distribution log-uniforme entre 0.01 et 10. Cela correspond à des valeurs de N_{ext} variant entre 1 et 2500. Ainsi, dans les modèles 2 et 3, la durée t_{eq} pour atteindre l'équilibre mutation-migration-dérive (t_{eq}) dépend de la valeur maximale entre N_0 et N_{ext} (t_{eq}=$20N_0$ si $N_0 > N_{ext}$; t_{eq}=$20N_{ext}$

si $N_{ext} > N_0$). Comme pour N_0, la borne maximale de N_{ext} est fixée à 2500 individus afin de limiter le temps de calcul.

Distribution a priori du taux de migration (m)

Le taux de migration *m* est constant au cours du temps et est tiré dans une loi log-uniforme U(10^{-4} - 0.5). Comme pour la taille démographique, l'analyse RF permettra d'inférer $\log_{10}(m)$.

Distribution a priori du taux d'admixture (admix)

La distribution de probabilité a priori du taux d'admixture (*admix*) est une loi uniforme U(0, 1). Les cas où *admix* tend vers 0 correspondent à une absence de migration depuis la population externe (dans ces cas-là, cela revient à une simulation selon le modèle 2). Lorsque *admix* tend vers 1, il y a un remplacement complet de la population focale par des individus issus de la population externe.

Statistiques résumées

Les tables de référence de chaque modèle sont construites à partir de statistiques résumées calculées avec R sur les échantillons temporels des simulations. Les statistiques, pour la plupart décrites dans le Chapitre 2, peuvent être classées en trois catégories : statistiques de diversité monolocus (**Tableau 4**), statistiques de diversité multilocus (**Tableau 5**), et statistiques temporelles (**Tableau 6**). La distance génétique moyenne intra échantillon ($within\_D_{mean}$) est groupée avec les statistiques monolocus du fait de sa très forte corrélation avec le H_E (cf. chapitre 2). De même, la distance génétique maximale intra échantillon ($within\_D_{max}$) est groupée avec le pourcentage de locus en déséquilibre de liaison (*%LD*). Les statistiques résumées sont calculées pour chaque échantillon temporel, sur les échantillons groupés deux à deux et sur tous les échantillons groupés ensemble. Les statistiques temporelles, telles que le F_{ST} temporel, le nombre d'allèles privés (p_A) ou les distances génétiques par paire inter-échantillon, sont calculées pour toutes les combinaisons d'échantillons temporels. De plus, Pudlo *et al.* (2016) ont montré que l'inclusion des axes d'une analyse discriminante linéaire (LDA) sur l'ensemble de la table de référence en tant que statistique résumée diminuait l'erreur a priori lors du choix du modèle. Les deux axes d'une LDA sont donc également utilisés comme statistiques résumées additionnelles pour inférer le modèle le plus probable. Les statistiques résumées sont calculées à l'aide du logiciel R (R Core Team, 2018).

Tableau 4 : Statistiques résumées monolocus.

Statistique		Package
H_E (v_H_E)	Hétérozygotie selon Nei (1973) et sa variance	Hierfstat (Goudet, 2005)
*within_*D_{mean}	Distance génétique moyenne par paire d'individus intra échantillon	
n_A (V_n_A)	Nombre moyen d'allèles par locus (et variance)	
p_A (V_p_A)	Nombre moyen d'allèles privés par locus (et variance)	
n_{ML}	Nombre de locus monomorphes	
F_{IS} (v_F_{IS})	Coefficient de consanguinité $F_{IS} = 1 - \frac{H_O}{H_E}$	Hierfstat
σ_{Fis}	Taux d'autofécondation estimé à partir de F_{IS} selon $\sigma_{Fis} = \frac{2F_{IS}}{1+F_{IS}}$	

Tableau 5 : Statistiques résumées multilocus.

Statistique		Package
n_{MLG}	Nombre de génotypes multilocus	poppr (Kamvar *et al.*, 2014)
MFMLG	Fréquence du MLG majoritaire	
single_MLG	Nombre de MLGs uniques	
H	Indice de diversité multilocus de Shannon, $H = \sum_{i=1}^{S} p_i log_{10}(p_i)$, où p_i est la fréquence du MLG i	poppr
λ	Indice de diversité multilocus de Simpson, $\lambda = 1 - \sum_{i=1}^{S} p_i^2$	poppr
$E5$	Evenness, $E5 = \frac{\left(\frac{1}{\lambda}\right)-1}{e^H-1}$	poppr
G	Indice de diversité multilocus de Stoddart et Taylor, $G = \frac{1}{\sum_{i=1}^{S} p_i^2}$	poppr
%LD	Proportion de locus en déséquilibre de liaison significatif	genepop (Rousset, 2008)
*within_*D_{max}	Distance génétique maximale par paire d'individus intra échantillon	
I_a	Indice d'association, $I_a = \frac{V_O}{V_E} - 1$, où V_O est la variance du nombre de différences entre individus et V_E est la somme des variances pour chaque locus	poppr
$\overline{r_D}$	Indice d'association standardisé (correction pour le nombre de locus échantillonnés)	poppr
$g2$ ($n_{loc,g2}$)	Déséquilibre d'identité : covariance d'hétérozygotie entre locus au sein des individus (Nombre de locus disponibles pour estimer $g2$)	InbreedR (Stoffel *et al.*, 2016)
*within_*$FreqD_0$	Fréquence d'individus de génotypes identiques intra-échantillon	

$within\_D_{median}$	Distance génétique médiane par paire d'individus intra-échantillon	

Tableau 6 : Statistiques résumées temporelles.

Statistique		Package
F_{ST}	F_{ST} temporel estimé selon Weir et Cockerham (1984)	Hierfstat
$between\_D_{mean}$	Distance génétique moyenne par paire d'individus inter échantillon	
$between\_D_{median}$	Distance génétique médiane par paire d'individus inter échantillon	
$between\_D_{max}$	Distance génétique maximale par paire d'individus inter échantillon	
$between\_FreqD_0$	Fréquence d'individus identiques inter échantillon	

Analyse ABC avec forêts aléatoires

Inférence d'un scénario démographique

Chaque modèle est simulé 50000 fois afin de construire des tables de référence de taille suffisamment importante (Pudlo *et al.*, 2016). L'analyse ABC-RF pour choisir le modèle le plus probable est ensuite réalisée en comparant tous les modèles deux à deux, ainsi que les trois modèles en même temps à l'aide du package R abcrf (Pudlo *et al.* 2016, voir Box 2). Les comparaisons de modèles deux à deux permettent de calculer le facteur de Bayes (Jeffreys, 1998), $K = \frac{P(M_1|D)}{P(M_2|D)}$ pour choisir entre les modèles M_1 et M_2 à partir des données D. Si $K > 1$, cela signifie que les données supportent plus M_1 que M_2. Une échelle d'interprétation des valeurs de K est donnée dans le **Tableau 7**. Les forêts aléatoires utilisées pour choisir le modèle décrivant le mieux les données sont constituées de 2000 arbres.

Tableau 7 : Échelle d'interprétation du facteur de Bayes (Jeffreys, 1998).

K	Force de la preuve
$K < 1$	Négative (supporte M_2)
$1 < K < 10^{1/2}$	Faible
$10^{1/2} < K < 10^1$	Substantielle
$10^1 < K < 10^{3/2}$	Forte
$10^{3/2} < K < 10^2$	Très forte
$K > 10^2$	Décisive

Evaluation de la performance de la méthode

Il est possible d'estimer facilement un taux d'erreur de l'inférence réalisée à l'aide des données « out-of-bag » (*out-of-bag error rate*). Les données « out-of-bag » sont les simulations qui n'ont pas été utilisées dans l'échantillon de bootstrap lors de la création des arbres qui composent la forêt, soit un tiers de la table de référence pour chaque arbre. On peut utiliser les arbres aléatoires qui n'ont pas

été construits à partir de ces simulations pour réaliser une prédiction et comparer avec la vraie valeur (du modèle ou du paramètre) pour obtenir un taux d'erreur. Ce taux d'erreur est global sur l'ensemble de l'espace de paramètres défini par les distributions a priori.

La méthode RF permet également d'évaluer quelles sont les statistiques résumées informatives pour inférer le modèle ou la valeur du paramètre d'intérêt. En effet, une statistique résumée sera d'autant plus informative qu'elle est utilisée souvent lors de la construction des arbres de la forêt aléatoire. Ainsi, lors de la création de la forêt aléatoire, il est possible d'enregistrer le nombre d'utilisations de chaque statistique résumée.

Inférence des paramètres du scénario démographique sélectionné

Une fois le scénario démographique inféré, les paramètres démographiques peuvent être inférés à leur tour. Pour chaque paramètre, des forêts aléatoires de 2000 arbres ont été construites et l'inférence réalisée selon le processus détaillé en Box 2.

De la même manière que pour l'inférence du modèle démographique, la qualité de l'estimation peut être évaluée à l'aide des données « out-of-bag ». Une interprétation graphique consiste à tracer les estimations issues des données « out-of-bag » en fonction de la vraie valeur du paramètre considéré (la valeur utilisée pour réaliser la simulation). De même, les statistiques résumées les plus informatives peuvent également être examinées.

Données empiriques

Choix des populations

Les inférences sont réalisées sur cinq populations naturelles de *Medicago truncatula* parmi les neuf analysées dans le Chapitre 2 : SP1, FR1, FR3, CO1, et CO3. En effet, nous avons vu que ces populations avaient des patrons de diversité contrastés (voir Figures 1 et 2, Chapitre 2). En particulier, FR1 présente une faible diversité mono- et multilocus et change peu au cours du temps, ce qui est compatible avec un scénario de petite population isolée. Au contraire, aucun des MLGs de la population SP1 n'est conservé au cours du temps, ce qui évoque plutôt un scénario d'extinction-recolonisation. Les populations FR3 et CO3 présentent quant à elles beaucoup de diversité et des patrons de déséquilibre de liaison et de différenciation entre MLG compatibles avec des scénarios de forte migration ou d'admixture. La population CO1 ne présente pas de patron clairement compatible avec les scénarios évalués dans le Chapitre 2.

Analyses de diversité

En plus des statistiques résumées utilisées pour l'ABC, des réseaux de génotypes multilocus (*minimum spanning networks*, Bandelt *et al.* 1999) ont été construits pour chaque population à l'aide

du package poppr (Kamvar *et al.*, 2014). Les réseaux sont construits à partir de la distance génétique, estimée comme le pourcentage d'allèles différents, entre chaque MLGs deux à deux.

Les statistiques résumées qui n'ont pas été évaluées dans le Chapitre 2 (et présentées dans les Tableaux 2, 3, et 4) reflètent les caractéristiques de ces réseaux. En particulier, les distances génétiques (moyennes, médianes et maximales, **Tableau 6**) entre MLGs de différents échantillons temporels reflètent les distances entre cercles de couleurs différentes, la fréquence de génotypes identiques entre échantillons temporels ($FreqD_0$, **Tableau 6**) représente la fréquence de cercles partagés, le nombre de MLGs uniques (*single_MLG*, **Tableau 5**) représente le nombre de petits cercles. De plus, l'evenness (*E5*, **Tableau 5**) traduit la variance de la taille des cercles. En effet, *E5* traduit la manière dont les génotypes multilocus sont répartis au sein de la population. Si un petit nombre de génotypes est en fréquence élevée, *E5* est proche de 0. Au contraire, si tous les génotypes sont en fréquence équilibrée, *E5* est maximal (*E5* = 1). Des indices dérivés du déséquilibre de liaison sont également calculés (**Tableau 5**) : l'indice d'association (*Ia*, Brown *et al.* 1980) et l'indice d'association standardisé ($\overline{r_D}$, Agapow and Burt 2001).

Résultats

Diversité génétique au cours du temps observée chez cinq populations de *M. truncatula*

L'analyse des réseaux semble confirmer les hypothèses émises dans le Chapitre 2. La population SP1 n'a aucun MLG conservé au cours du temps (**Figure 14**), ce qui correspond à un $FreqD_0$ temporel nul (**Tableau supplémentaire 1**). De plus, les branches du réseau (et le D_{max} temporel) indiquent que les MLGs sont fortement différenciés au cours du temps (**Figure 14** et **Tableau supplémentaire 1**), compatible avec un scénario d'extinction-recolonisation. Au contraire, les autres populations, et plus particulièrement FR1, partagent des MLGs en haute fréquence. Les réseaux des populations CO3 et FR3 sont caractérisés par un grand nombre de MLGs uniques et quelques MLGs en forte fréquence (**Figure 14**) avec des branches indiquant des distances génétiques importantes, suggérant des apports de migrants. De plus, ces populations ont des valeurs d'*E5* faibles (**Tableau supplémentaire 1**). La population CO1 montre une forte différenciation au cours du temps, surtout entre 2009 et 2013. En effet, les MLGs de 2013 sont presque tous groupés ensemble sur le réseau (**Figure 14**), ce qui suggère un fort évènement de migration. FR1 a une diversité qui augmente au cours du temps, suggérant un évènement de fondation récent et une expansion de la population.

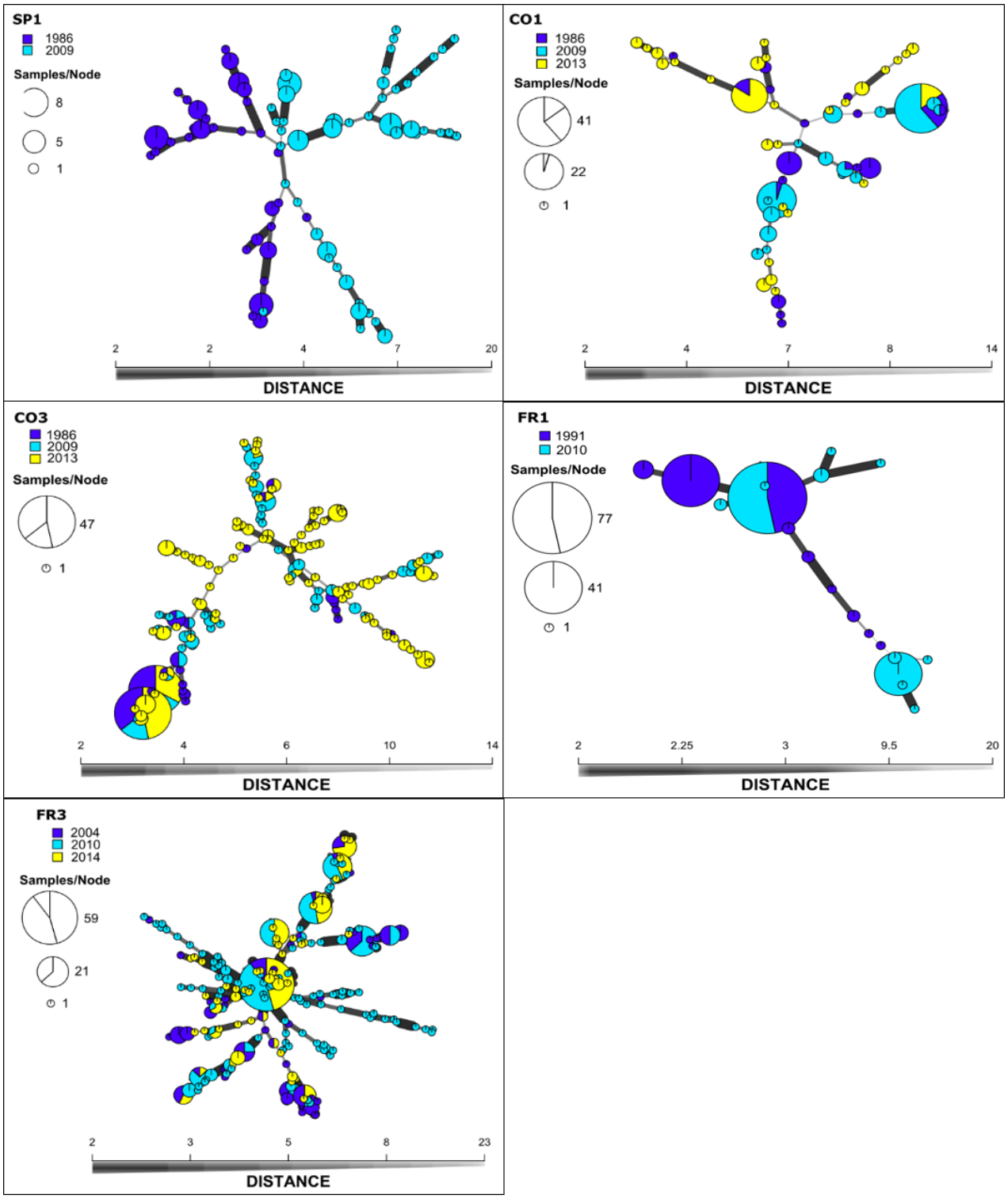


Figure 14 : Réseaux de génotypes multilocus (MLGs) au cours du temps chez cinq populations de *Medicago truncatula*. Chaque cercle représente un génotype multilocus (MLG), la taille des cercles est proportionnelle au nombre d'individus qui possèdent le MLG. L'épaisseur du trait reliant les cercles est inversement proportionnelle à la différenciation génétique entre MLGs, mesurée comme le nombre d'allèles différents.

Choix du modèle démographique

Les erreurs *a priori* sont relativement élevées (> 0.13, **Tableau 8**), ce qui indique des difficultés à discriminer les trois modèles démographiques de manière précise à l'aide de notre méthode. Les erreurs sont plus importantes lorsqu'on compare les trois modèles en même temps plutôt que deux à deux. Les modèles 1 et 3 sont les plus faciles à discriminer, en effet c'est entre ces deux modèles que l'erreur *a priori* est la plus faible (et la probabilité *a posteriori* la plus forte, **Tableau 8**).

Pour toutes les populations, le modèle avec admixture (modèle 3) présente les probabilités *a posteriori* les plus fortes (**Tableau 8**). Le modèle 1 (population isolée), est le moins probable à chaque fois, sauf dans le cas de la population CO3 pour laquelle c'est le modèle 2 qui est le moins supporté. Pour les populations CO1 et FR3, le choix du modèle 3 plutôt que le modèle 2 n'est pas fortement appuyé (facteur de Bayes $K < 3$, **Tableau 8**). En effet, les probabilités *a posteriori* associées au modèle 3 sont faibles (0.67 et 0.64, respectivement). Ce résultat est également visible à travers la projection des données simulées et empiriques sur les deux axes de l'analyse discriminante linéaire (LDA, **Figure 15**). Le modèle 2 est difficile à distinguer des deux autres modèles avec des probabilité *a posteriori* faibles lorsque les modèles sont comparés deux à deux.

Tableau 8 : Taux d'erreur a priori, probabilités a posteriori et facteur de Bayes (*K*) des modèles démographiques pour quatre populations de *M. truncatula*.

* Preuve substantielle, ** preuve forte.

Population		Modèle 1 vs modèle 2	Modèle 1 vs modèle 3	Modèle 2 vs modèle 3	Trois modèles
SP1	Erreur a priori	0.31	0.18	0.26	0.39
	Probabilité a posteriori	0.66 (Modèle 2)	0.90 (Modèle 3)	0.91 (Modèle 3)	0.87 (Modèle 3)
	K	1.90	8.73 *	9.68 *	-
CO1	Erreur a priori	0.29	0.15	0.23	0.35
	Probabilité a posteriori	0.75 (Modèle 2)	0.82 (Modèle 3)	0.70 (Modèle 3)	0.67 (Modèle 3)
	K	3.02	4.67 *	2.29	-
CO3	Erreur a priori	0.26	0.14	0.22	0.32
	Probabilité a posteriori	0.85 (Modèle 1)	0.97 (Modèle 3)	0.84 (Modèle 3)	0.90 (Modèle 3)
	K	5.53 *	29.22 **	5.09 *	-
FR1	Erreur a priori	0.31	0.17	0.27	0.39
	Probabilité a posteriori	0.89 (Modèle 2)	0.98 (Modèle 3)	0.84 (Modèle 3)	0.79 (Modèle 3)
	K	8.14 *	47.37 **	5.09 *	-
FR3	Erreur a priori	0.28	0.13	0.19	0.33
	Probabilité a posteriori	0.69 (Modèle 2)	0.88 (Modèle 3)	0.63 (Modèle 3)	0.64 (Modèle 3)
	K	2.25	7.09 *	1.71	-

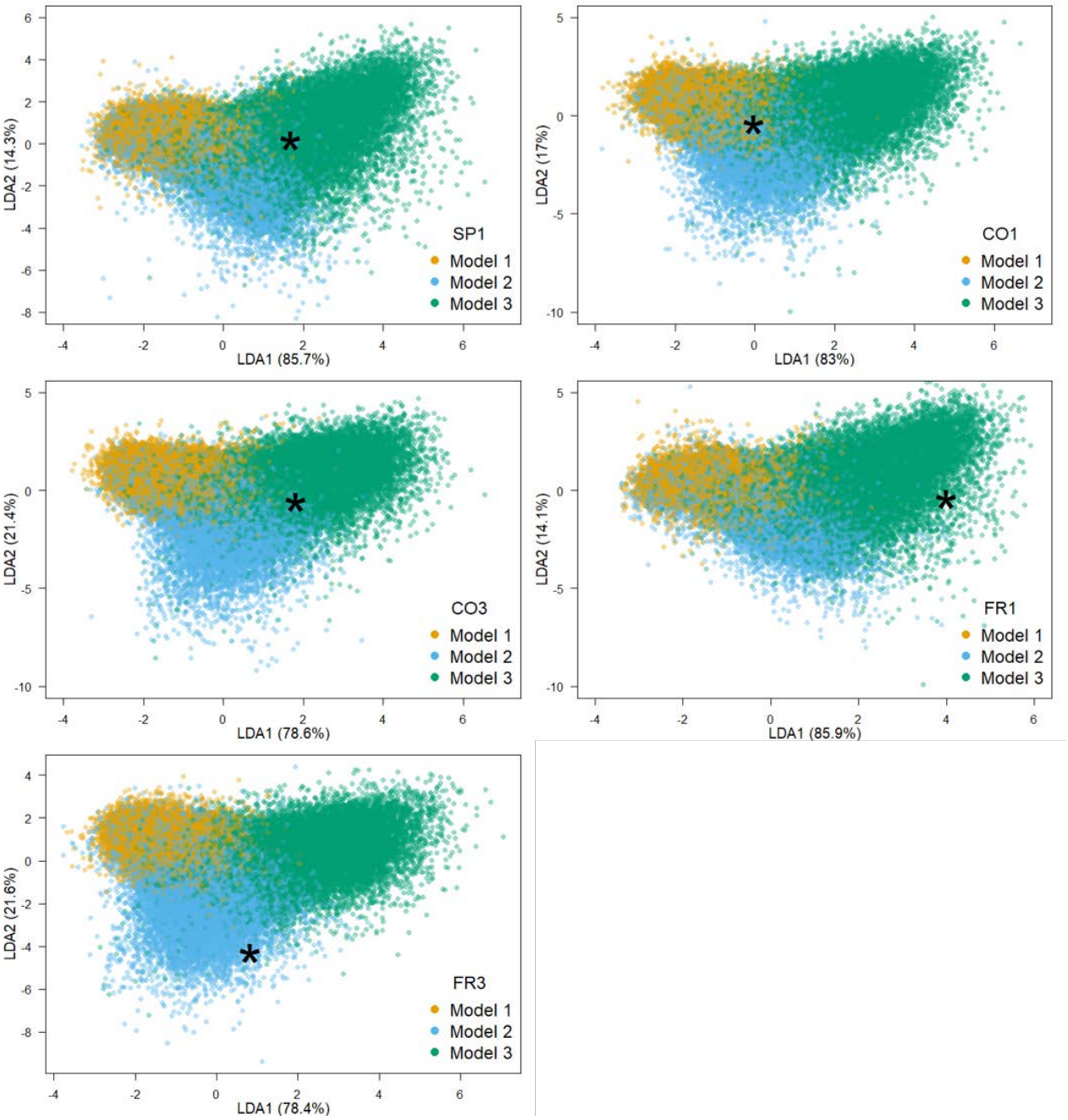


Figure 15 : Projection des trois modèles démographiques sur les deux axes de l'analyse discriminante linéaire.

La projection des données empiriques est représentée par l'étoile noire.

Statistiques les plus informatives

Les statistiques les plus utilisées par l'algorithme de RF pour choisir le modèle démographique reflètent surtout les effets de la migration. En effet, les statistiques les plus informatives sont principalement des statistiques dérivées du déséquilibre de liaison (*LD*, *Ia*, et $\overline{r_D}$) et des statistiques temporelles (F_{ST}). Pour les populations avec seulement deux échantillons temporels (SP1 et FR1), des statistiques de diversité monolocus (n_A, H_E et $within\_D_{mean}$ au deuxième échantillon) sont aussi utilisées pour différencier le modèle 3 des autres modèles. Ces résultats sont assez intuitifs étant donné que le F_{ST} temporel est affecté par la migration, et plus particulièrement par les évènements d'admixture (voir Chapitre 2). De plus, l'indice d'association est apparenté au déséquilibre de liaison, qui est lui aussi affecté par la migration.

Estimation des distributions des paramètres du modèle

Pour chacune des populations, nous avons utilisé le scénario démographique le plus probable (modèle 3) pour estimer les distributions *a posteriori* des paramètres qui le caractérisent : les taux d'autofécondation, les tailles démographiques (nombre d'individus), les taux de migration et les taux d'admixture.

Taux d'autofécondation

On estime des taux d'autofécondation moyens (*S*) supérieurs à 0.95 (**Tableau 9**). Les taux d'autofécondations estimés pour chaque échantillon temporel (σ_0, σ_1, σ_2) sont très similaires à ceux estimés à partir du F_{IS} (table 4, chapitre 2). En effet, la statistique résumée la plus informative pour toutes les inférences de taux d'autofécondation est le F_{IS}. Les autres statistiques utilisées sont le déséquilibre d'identité (*g2*) et des statistiques de diversité multilocus telles que la fréquence du MLG majoritaire (*MFMLG*), la fréquence de génotypes identiques au sein d'une génération ($within\_FreqD_0$) et l'indice de Simpson (λ). On observe peu de variabilité temporelle, sauf pour les populations CO3 en 2013 et FR3 en 2010 qui montrent une augmentation de leur taux d'allofécondation résiduelle. Les distributions des probabilités a posteriori des différents taux d'autofécondation estimés sont bien distinctes des distributions a priori (**Erreur ! Source du renvoi introuvable.**) et les intervalles de crédibilité sont étroits (**Tableau 9**), ce qui témoigne d'une bonne qualité des inférences. De plus, les taux d'erreur *out-of-bag* sont inférieurs à 0.01 pour toutes les inférences.

Tableau 9 : Taux d'autofécondation inférés par ABC-RF [95% CI].

Population	S	σ_0	σ_1	σ_2	σ_{ext}
SP1	0.96 [0.88 - 0.99]	0.98 [0.93 - 0.99]	0.96 [0.72 - 1]	-	0.99 [0.90 - 1]
CO1	0.97 [0.91 - 0.99]	0.99 [0.98 - 0.99]	0.99 [0.91 - 1]	0.98 [0.77 - 1]	0.98 [0.91 - 1]
CO3	0.95 [0.86 - 0.99]	0.98 [0.93 - 0.99]	0.98 [0.76 - 0.99]	0.95 [0.73 - 0.99]	0.98 [0.89 - 0.99]
FR1	0.96 [0.79 – 0.99]	0.97 [0.86 – 1]	0.99 [0.79 – 1]	-	0.99 [0.78 – 1]
FR3	0.95 [0.87 - 0.99]	0.99 [0.97 - 0.99]	0.91 [0.71 - 0.99]	0.98 [0.68 - 0.99]	0.97 [0.80 - 0.99]

Tailles démographiques et diversité ancestrale

Les diversités ancestrales (θ) sont bien inférées (distributions de probabilité a posteriori différenciées des distributions a priori, **Figure supplémentaire 2**), et sont variables d'une population à l'autre. Comme suggéré par les analyses de diversité, les inférences de diversité ancestrale confirment que FR1 est la population la moins diverse alors que FR3 a la plus grande diversité (**Tableau 10**). Les estimations des tailles démographiques N_1 et N_2 sont moins précises, en particulier N_2. En effet, les distributions de probabilité *a posteriori* sont souvent peu ou pas différenciées des distributions *a priori* (**Figure supplémentaire 2Erreur ! Source du renvoi introuvable.**) et les taux d'erreur *out-of-bag* sont élevés (avec des valeurs prédites parfois peu corrélées avec les valeurs réelles, **Figure supplémentaire 5**). Les inférences de N_1 pour CO1, FR1 et FR3 sont les plus précises, avec une faible taille démographique pour la population FR1 et une grande taille pour FR3. La qualité des inférences des autres tailles démographiques n'est pas suffisante et il est difficile d'identifier de potentiels changements de tailles au cours du temps. Les statistiques résumées les plus informatives pour inférer la diversité ancestrale θ sont des statistiques de diversité monolocus (nombre d'allèles n_A, H_E), et multilocus (nombre de MLGs uniques *singleMLG*, la distance génétique moyenne intra-échantillon D_{mean}, le nombre de MLGs n_{MLG}). Pour l'inférence de la taille démographique des autres échantillons temporels, ce sont plutôt des statistiques temporelles telles que le F_{ST} temporel ou des statistiques de diversité au sein du deuxième (ou troisième) échantillon.

Pour toutes les populations, on estime une forte diversité ancestrale de la population externe (réservoir de migrants, **Tableau 10**). Les distributions de probabilité *a posteriori* sont décalées vers la borne maximale des distributions *a priori*, notamment pour les populations CO3 et FR3 (**Erreur ! Source du renvoi introuvable.**). Cela suggère que les distributions *a priori* ne sont pas suffisamment étendues et que les véritables réservoirs de migrants peuvent avoir des diversités plus importantes

que celles inférées. La population externe associée à FR1 est la moins diverse, mais présente néanmoins plus de diversité que FR1.

Tableau 10 : Tailles démographiques et diversités ancestrales inférées par ABC-RF [95% CI].

Population	θ	N_1	N_2	θ_{Ext}
SP1	1.15 [0.36 - 3.52]	27 [1 - 2918]	-	2.40 [0.06 - 8.87]
CO1	0.86 [0.29 - 5.41]	22 [1 - 1502]	39 [1 - 4315]	1.97 [0.03 - 8.02]
CO3	0.89 [0.18 - 2.87]	171 [2 - 4847]	82 [2 - 4225]	6.75 [1.95 - 9.79]
FR1	0.62 [0.09 - 7.64]	10 [1 - 1192]	-	1.08 [0.02 - 7.20]
FR3	2.98 [0.82 - 9.32]	22 [2 - 3565]	47 [1 - 3651]	5.13 [0.02 - 9.82]

Taux de migration et taux d'admixture

On estime des taux de migration faibles pour SP1, CO3 et FR1 (**Tableau 11**, **Figure supplémentaire 3**) alors que CO1 et FR3 reçoivent plus de migrants. Les taux d'admixture inférés sont tous forts (> 30%, **Tableau 11**). Les populations SP1, CO1 et CO3 subissent les évènements d'admixture les plus importants (> 60%).

Les statistiques résumées les plus informatives pour inférer le taux de migration sont l'indice d'association (*Ia*) et le F_{ST} temporel. Pour les taux d'admixture, ce sont des statistiques de différenciation temporelle telles que le F_{ST}, la distance génétique moyenne entre deux échantillons temporels ($between\_D_{mean}$) ou la fréquence de génotypes identiques entre deux échantillons ($between\_FreqD_0$).

Tableau 11 : Taux de migration et taux d'admixture inférés par ABC-RF [95% CI].

Population	m	$admix_1$	$admix_2$
SP1	8.8×10^{-4} [1.2×10^{-4} - 0.13]	0.71 [0.08 - 0.99]	-
CO1	6.7×10^{-3} [2.2×10^{-4} - 0.32]	0.30 [0.02 - 0.90]	0.61 [0.14 - 0.98]
CO3	3.7×10^{-4} [1.1×10^{-4} - 0.04]	0.62 [0.12 - 0.96]	0.45 [0.04 - 0.90]
FR1	4.6×10^{-4} [1.1×10^{-4} - 0.01]	0.44 [0.16 – 0.91]	-
FR3	2.1×10^{-3} [1.2×10^{-4} - 0.41]	0.36 [0.04 - 0.90]	0.46 [0.03 - 0.95]

Discussion

L'objectif de ce chapitre était de développer une méthode permettant d'inférer l'histoire démographique d'une population autogame à partir de données temporelles, en prenant en compte la migration. Nous avons souhaité tirer parti de la grande flexibilité permise par la méthodologie de calcul Bayésien approché pour définir trois scénarios démographiques et les tester sur cinq populations naturelles de *Medicago truncatula*.

Des populations connectées par de la migration ?

Les scénarios démographiques testés ici permettent d'évaluer les rôles de plusieurs forces de manière conjointe : les variations temporelles du taux d'autofécondation, les variations temporelles de la taille démographique, la possibilité de migration et les variations du taux d'admixture au cours du temps le cas échéant. En comparant les modèles deux à deux, nos résultats montrent que les cinq populations de *M. truncatula* considérées ne sont pas isolées (rejet du modèle 1) et que la migration joue un rôle important. Ce résultat est particulièrement surprenant dans le cas de la population FR1 pour laquelle les analyses de diversité préalables suggéraient plutôt un scénario de population isolée de petite taille soumise à une forte dérive génétique (Chapitre 2 et résultats des analyses de diversité). Cependant, un nouveau MLG apparait en haute fréquence dans le deuxième échantillon temporel. Ce MLG est fortement différencié des autres (**Figure 14**) et notre méthode accommode cette structure de données avec un évènement d'admixture car la mutation ne peut pas expliquer l'apparition d'un MLG en si forte fréquence. Une hypothèse alternative pour expliquer l'apparition de ce nouveau génotype est que la zone d'échantillonnage a été décalée et, du fait de la forte structure spatiale caractéristique des populations autogames, la composition du deuxième échantillon était différente. Pour toutes les populations, l'adéquation entre le modèle et les données est toujours meilleure quand le taux d'admixture peut varier au cours du temps (Modèle 3). En effet, cela permet de lever une contrainte forte associée au modèle 2 qui est que le taux de migration est constant au cours du temps. Cependant, les probabilités *a posteriori* pour les choix de modèles sont faibles par rapport à ce que l'on peut obtenir avec des approches ABC-RF (par exemple dans Fraimout *et al.* 2017 ; Estoup *et al.* 2018). Il est surprenant qu'un seul modèle soit choisi préférentiellement pour les cinq populations analysées malgré des profils de diversité contrastés. Le modèle 2 est contraint par le fait que le taux de migration est constant au cours du temps or, on s'attend à une forte stochasticité temporelle des populations autogames avec des dynamiques en métapopulation (Ingvarsson, 2002). Le modèle 3 permet de mieux accommoder la stochasticité temporelle de la migration, ce qui pourrait expliquer pourquoi il est choisi à chaque fois. En effet, ce modèle permet des variations temporelles à la fois du

taux d'autofécondation, de la taille démographique, et des évènements de migration ponctuelle depuis la population externe parfois extrêmes (jusqu'à 60% d'admixture). Par ailleurs, les trois modèles démographiques testés ici sont emboîtés, ce qui cause des erreurs d'attribution, notamment pour le modèle 2 qui est intermédiaire entre les deux autres modèles. Cela peut expliquer les forts taux d'erreur *a priori* mesurés (> 0.30).

Des taux de migration variables au cours du temps ?

Pour toutes les populations, on infère un faible taux de migration continue. Cependant, on infère également de forts taux de migration ponctuelle (entre 30 et 70%). On obtient des inférences de taux de migration et d'admixture relativement précises avec notre modèle.

Malgré le fait qu'elle ne soit pas directement échantillonnée, les estimations de paramètres de la population externe semblent assez précises (taux d'autofécondation et diversité ancestrale). Les seules sources d'information proviennent des individus migrants retrouvés dans la population focale. Par ailleurs, les estimations de la diversité ancestrale θ_{ext} mettent en avant une limite éventuelle de notre modèle de simulation. En effet, la migration et les évènements d'admixture ne peuvent provenir que d'une seule population. Or les distributions a posteriori de θ_{ext} sont à la limite supérieure des distributions a priori (en particulier pour la population FR3). Cela suggère que les individus migrants sont très divers et fortement différenciés. La seule manière d'accommoder une telle diversité avec le modèle actuel est donc de supposer une très grande population externe, notamment dans le cas de la population FR3. Une hypothèse alternative serait que la population considérée est connectée avec plusieurs autres populations différenciées les unes des autres dans un système de métapopulation.

Inférence de la taille démographique

Variations de taille démographique au cours du temps

Les tailles démographiques au cours du temps sont les paramètres les moins bien estimés pour l'ensemble des populations. Les estimations sont particulièrement difficiles lorsque l'intervalle de temps entre deux échantillons successifs est court (Δt = 4 générations), ce qui suggère que dans ces cas-là, le signal sur les données n'est pas suffisamment fort. Par ailleurs, les variations de taille démographique au cours du temps ne peuvent être inférées par notre méthode que dans les cas où l'on dispose d'au moins trois échantillons temporels (populations CO1, CO3 et FR3). Des méthodes existent pour estimer des goulots d'étranglement récents, ou ancestraux, à partir d'un seul échantillon, notamment à partir des distributions de fréquences alléliques (Luikart *et al.*, 1998) ou de

l'hétérozygotie attendue (Piry *et al.*, 1999). Ces méthodes supposent cependant la panmixie et l'absence de migration récente, ce qui n'est pas le cas de nos populations.

Taille démographique versus taille efficace

Le cadre de simulations que nous avons développé ici permet d'estimer une taille démographique (nombre d'individus dans la population), et non pas une taille efficace. Afin de comprendre le fonctionnement des populations, il serait préférable d'avoir une estimation de la taille efficace. Cependant, N_e n'est pas estimable à partir des simulations actuelles. En effet, même si on peut corriger la taille efficace pour le taux d'autofécondation avec l'attendu $N_e = \frac{(2-\sigma)}{2}N$, ce dernier n'est plus valable en présence de migration. Afin de tester si notre méthode pourrait permettre d'inférer des tailles efficaces, on pourrait regarder les résultats de l'inférence au sein du modèle 1, pour lequel on se trouve dans les hypothèses de l'attendu. Une perspective des modèles de simulation actuels serait de sauvegarder la variance du succès reproducteur à chaque génération afin d'accéder à la taille efficace. Dans ce cas, la taille efficace des populations simulées peut être calculée comme la moyenne harmonique de la variance du succès reproducteur sur le pas de temps considéré. Des travaux de simulations en populations allogames sous sélection montrent que les valeurs de N_e estimées de cette manière sont très proches de N (Pavinato, pers. comm).

Inférence des taux d'autofécondation

Les taux d'autofécondation sont les paramètres les mieux inférés dans notre analyse. Ces derniers sont en accord avec les estimations via le déficit en hétérozygotes ou le déséquilibre d'identité, et confirment que les populations de *M. truncatula* sont fortement autogames. Les taux d'autofécondation sont peu variables au cours du temps. La seule population pour laquelle on infère des variations est FR3 qui présente un taux d'allofécondation résiduelle relativement élevé en 2010. Cette estimation est concordante avec les estimations précises réalisées à partir de descendances maternelles dans le Chapitre 1.

Conclusions et perspectives

Nous avons développé une méthode permettant d'inférer de manière conjointe plusieurs paramètres démographiques : le taux d'autofécondation, la diversité ancestrale de la population considérée, le taux de migration et des évènements d'admixture. Contrairement aux méthodes d'estimation disponibles dans la littérature, notre méthode ne repose pas sur des hypothèses de panmixie ou d'absence de migration et est très flexible. En effet, elle autorise des variations au cours du temps de tous les paramètres considérés (à l'exception du taux de migration). Cela la rend particulièrement adaptée à l'étude des populations autogames qui ont des dynamiques très variables dans le temps. Grâce à ce cadre d'inférences, il nous est désormais possible d'aller plus loin que dans le Chapitre 2 dans notre compréhension du fonctionnement des populations de *Medicago truncatula*. En effet, on peut maintenant associer les populations à un scénario démographique de manière plus rigoureuse. Par ailleurs, notre étude confirme l'utilité des statistiques multilocus et temporelles pour décrire le fonctionnement des populations autogames.

Plusieurs axes d'amélioration peuvent être développés dans le futur, notamment pour améliorer les inférences de taille démographique. Nous avons pu voir que la précision du signal détecté semblait plus faible lorsque l'intervalle de temps entre les échantillons était réduit. Il est possible de réutiliser les simulations réalisées ici sur les populations avec trois échantillons temporels en ne considérant que le premier et le troisième afin de voir si les estimations sont plus précises (en considérant que la taille démographique du nouveau segment temporel est la moyenne harmonique de N_1 et N_2). Une autre piste d'amélioration serait d'estimer le paramètre combiné Nm. Des pistes d'amélioration plus contraignantes (car nécessitant de refaire des simulations), seraient de modifier notre modèle 2 pour permettre des variations de taux de migration au cours du temps afin que celui-ci soit plus à même d'accommoder la stochasticité des populations autogames. Par ailleurs, nous pouvons envisager de changer le modèle de mutation utilisé en passant du stepwise mutation model au generalized mutation model, qui est généralement considéré comme plus réaliste. De plus, Pudlo *et al.* (2016) ont comparé les performances de l'ABC RF selon que les marqueurs utilisés sont des microsatellites ou des SNP et ont montré que les SNP sont plus informatifs. Cependant, cela nécessiterait de nouveaux jeux de données empiriques.

Bibliographie

Abdo Z, Crandall KA, Joyce P (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology* **13**: 837–851.

Agapow P-M, Burt A (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes* **1**: 101–102.

Bakker EG, Stahl EA, Toomajian C, Nordborg M, Kreitman M, Bergelson J (2006). Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* **15**: 1405–1418.

Bandelt HJ, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Beaumont MA (2010). Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst* **41**: 379–406.

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

Bécheler A (2014). Limite des génome-scans pour la détection de locus sous sélection chez les espèces autogames : une étude par simulations. Université Montpellier 2.

Beerli P (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**: 341–345.

Beerli P, Felsenstein J (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *PNAS* **98**: 4563–4568.

Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annu Rev Ecol Evol Syst* **49**: 433–456.

Blum MGB, Nunes MA, Prangle D, Sisson SA (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science* **28**: 189–208.

Bomblies K, Yant L, Laitinen RA, Kim S-T, Hollister JD, Warthmann N, *et al.* (2010). Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet* **6**: e1000890.

Breiman L (2001). Random Forests. *Machine Learning* **45**: 5–32.

Brown AHD, Feldman MW, Nevo E (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **96**: 523–536.

Caballero A, Hill WG (1992). Effects of partial inbreeding on fixation rates and variation of mutant genes. *Genetics* **131**: 493–507.

Estoup A, Raynal L, Verdu P, Marin J-M (2018). Model choice using Approximate Bayesian Computation and Random Forests: analyses based on model grouping to make inferences about the genetic history of Pygmy human populations. *Journal de la Société Française de Statistique* **159**: 167–190.

Fraimout A, Debat V, Fellous S, Hufbauer RA, Foucaud J, Pudlo P, *et al.* (2017). Deciphering the routes of invasion of Drosophila suzukii by means of ABC random forest. *Mol Biol Evol* **34**: 980–996.

Gastwirth JL (1972). The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics* **54**: 306–316.

Gilbert KJ, Whitlock MC (2015). Evaluating methods for estimating local effective population size with and without migration. *Evolution* **69**: 2154–2166.

Glémin S (2007). Mating systems and the efficacy of selection at the molecular level. *Genetics* **177**: 905–916.

Goudet J (2005). hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**: 184–186.

Haller BC, Messer PW (2017). SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol Biol Evol* **34**: 230–240.

Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning: data mining, inference, and prediction, Second Edition*, 2nd edn. Springer-Verlag: New York.

Ingvarsson PK (2002). A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* **56**: 2368–2373.

Jeffreys H (1998). *The Theory of Probability*, 3rd edn. Oxford University Press: Oxford.

Joyce P, Marjoram P (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7**.

Jullien M, Navascués M, Ronfort J, Loridon K, Gay L (2019). Structure of multilocus genetic diversity in predominantly selfing populations. *Heredity*: 1.

Kamvar ZN, Tabima JF, Grünwald NJ (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**.

Krimbas CB, Tsakas S (1971). The Genetics of Dacus Oleae. V. Changes of Esterase Polymorphism in a Natural Population Following Insecticide Control—Selection or Drift? *Evolution* **25**: 454–460.

Lombaert E, Ciosi M, Miller NJ, Sappington TW, Blin A, Guillemaud T (2018). Colonization history of the western corn rootworm (Diabrotica virgifera virgifera) in North America: insights from random forest ABC using microsatellite data. *Biol Invasions* **20**: 665–677.

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J Hered* **89**: 238–247.

Marjoram P, Molitor J, Plagnol V, Tavaré S (2003). Markov chain Monte Carlo without likelihoods. *PNAS* **100**: 15324–15328.

Marriage TN, Hudman S, Mort ME, Orive ME, Shaw RG, Kelly JK (2009). Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). *Heredity* **103**: 310–317.

Montesinos A, Tonsor SJ, Alonso-Blanco C, Picó FX (2009). Demographic and genetic patterns of variation among populations of *Arabidopsis thaliana* from contrasting native environments. *PLoS ONE* **4**: e7213.

Nei M (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* **70**: 3321–3323.

Nordborg M (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–929.

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

Parag KV, Pybus OG (2019). Robust Design for Coalescent Model Inference. *Syst Biol*.

Piry S, Luikart G, Cornuet J-M (1999). BOTTLENECK: a computer program for detecting recent reductions in the effective size using allele frequency data. *J Hered* **90**: 502–503.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010). Population genetic inference from genomic sequence variation. *Genome Research* **20**: 291–300.

Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP (2016). Reliable ABC model choice via random forests. *Bioinformatics* **32**: 859–866.

R Core Team (2018). *R: The R project for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria.

Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A (2017). ABC random forests for Bayesian parameter inference. *Peer Community in Evolutionary Biology*: 100036.

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

Sisson SA, Fan Y, Tanaka MM (2007). Sequential Monte Carlo without likelihoods. *PNAS* **104**: 1760–1765.

Smith CCR, Flaxman SM, Scordato ESC, Kane NC, Hund AK, Sheta BM, *et al.* (2018). Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. *Molecular Ecology* **27**: 4200–4212.

Stoffel MA, Esser M, Kardos M, Humble E, Nichols H, David P, *et al.* (2016). inbreedR: an R package for the analysis of inbreeding based on genetic markers. *Methods in Ecology and Evolution* **7**: 1331–1339.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**: 505–518.

Thuillet A-C, Bru D, David J, Roumet P, Santoni S, Sourdille P, *et al.* (2002). Direct estimation of mutation rate for 10 microsatellite loci in Durum wheat, Triticum turgidum (L.) Thell. ssp durum desf. *Mol Biol Evol* **19**: 122–125.

Viard F, Bremond P, Labbo R, Justy F, Delay B, Jarne P (1996). Microsatellites and the genetics of highly selfing populations in the freshwater snail *Bulinus truncatus*. *Genetics* **142**: 1237–1247.

Wang J (2005). Estimation of effective population sizes from data on genetic markers. *Phil Trans R Soc B* **360**: 1395–1409.

Wang J, Santiago E, Caballero A (2016). Prediction and estimation of effective population size. *Heredity* **117**: 193–206.

Waples RS (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**: 379–391.

Waples RS (2016). Making sense of genetic estimates of effective population size. *Molecular Ecology* **25**: 4689–4691.

Waples RS, Do C (2008). LDNe: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**: 753–756.

Wegmann D, Leuenberger C, Excoffier L (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**: 1207–1218.

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

Whitlock MC, McCauley DE (1999). Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity* **82**: 117–125.

Wright S (1931). Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright S (1949). The genetical structure of populations. *Annals of Eugenics* **15**: 323–354.

Matériel supplémentaire

Figure supplémentaire 1 : Distributions de probabilités *a posteriori* des taux d'autofécondation chez cinq populations de *Medicago truncatula*.

Figure supplémentaire 2 : Distributions de probabilité *a posteriori* des inférences de tailles démographiques chez cinq populations de *Medicago truncatula*.

Figure supplémentaire 3 : Distributions de probabilité *a posteriori* des inférences de taux de migration et d'admixture chez cinq populations de *Medicago truncatula*.

Figure supplémentaire 4 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les taux d'autofécondation.

Figure supplémentaire 5 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les diversités ancestrales et les tailles démographiques.

Figure supplémentaire 6 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les taux de migration et d'admixture.

Tableau supplémentaire 1 : Statistiques résumées complémentaires du chapitre 2

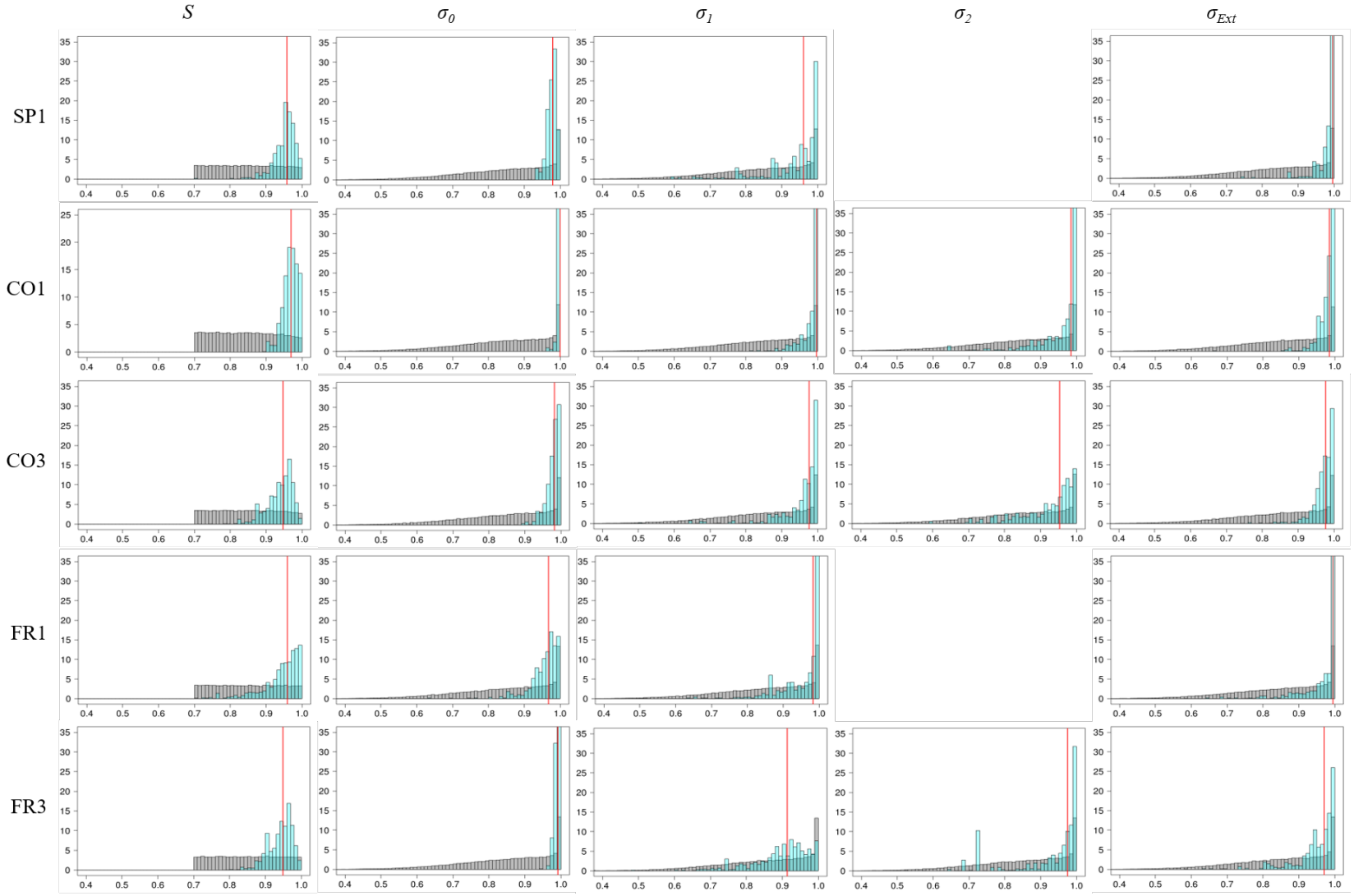


Figure supplémentaire 1 : Distributions de probabilités a posteriori des taux d'autofécondation chez cinq populations de *Medicago truncatula*. Les barres grises représentent les distributions a priori ; les barres bleues représentent les distributions de probabilité a posteriori ; la ligne verticale rouge correspond à la valeur médiane inférée pour le paramètre.

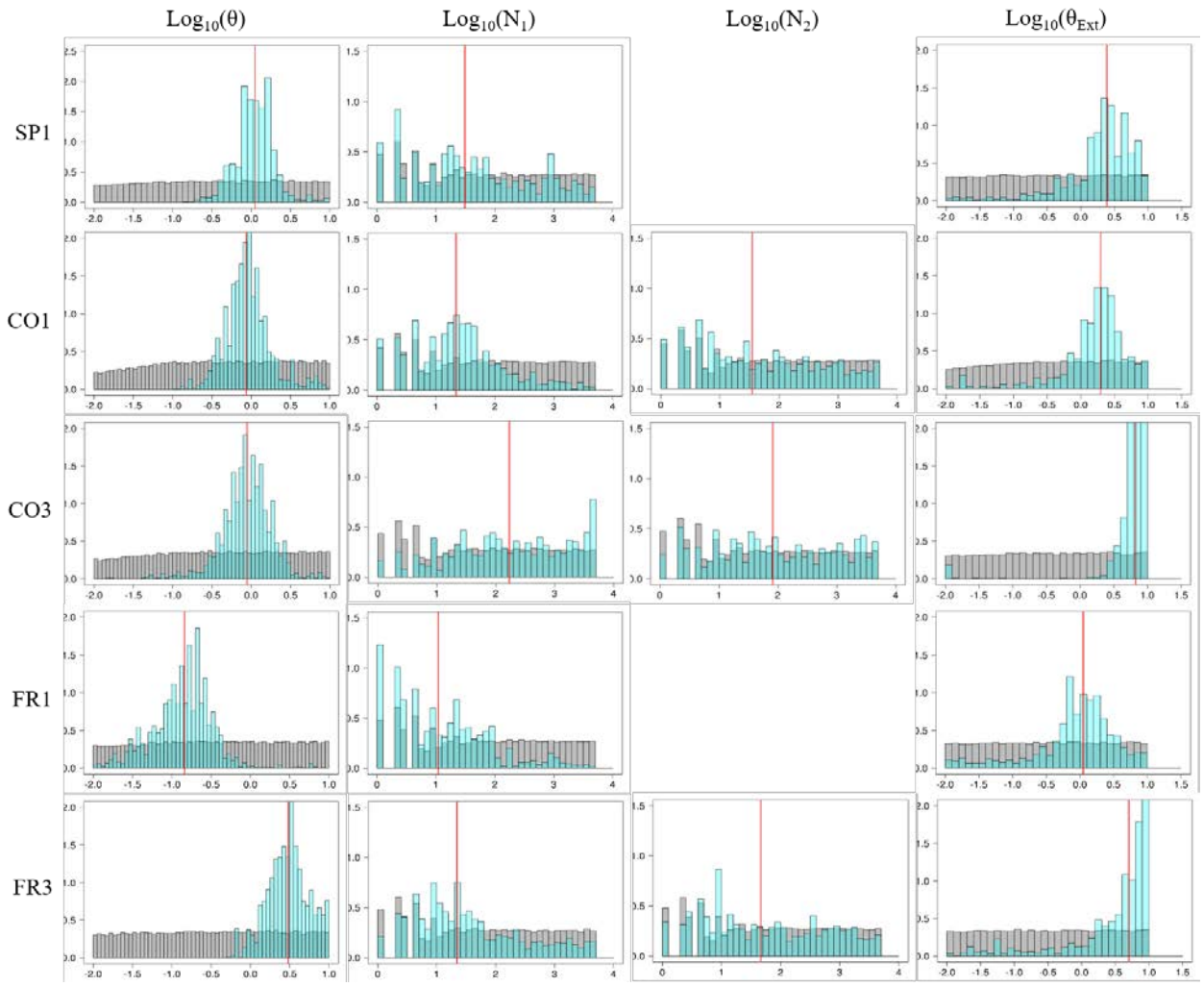


Figure supplémentaire 2 : Distributions de probabilité a posteriori des inférences de tailles démographiques chez cinq populations de *Medicago truncatula*.

Les barres grises représentent les distributions a priori ; les barres bleues représentent les distributions de probabilité a posteriori ; la ligne verticale rouge correspond à la valeur médiane inférée pour le paramètre.

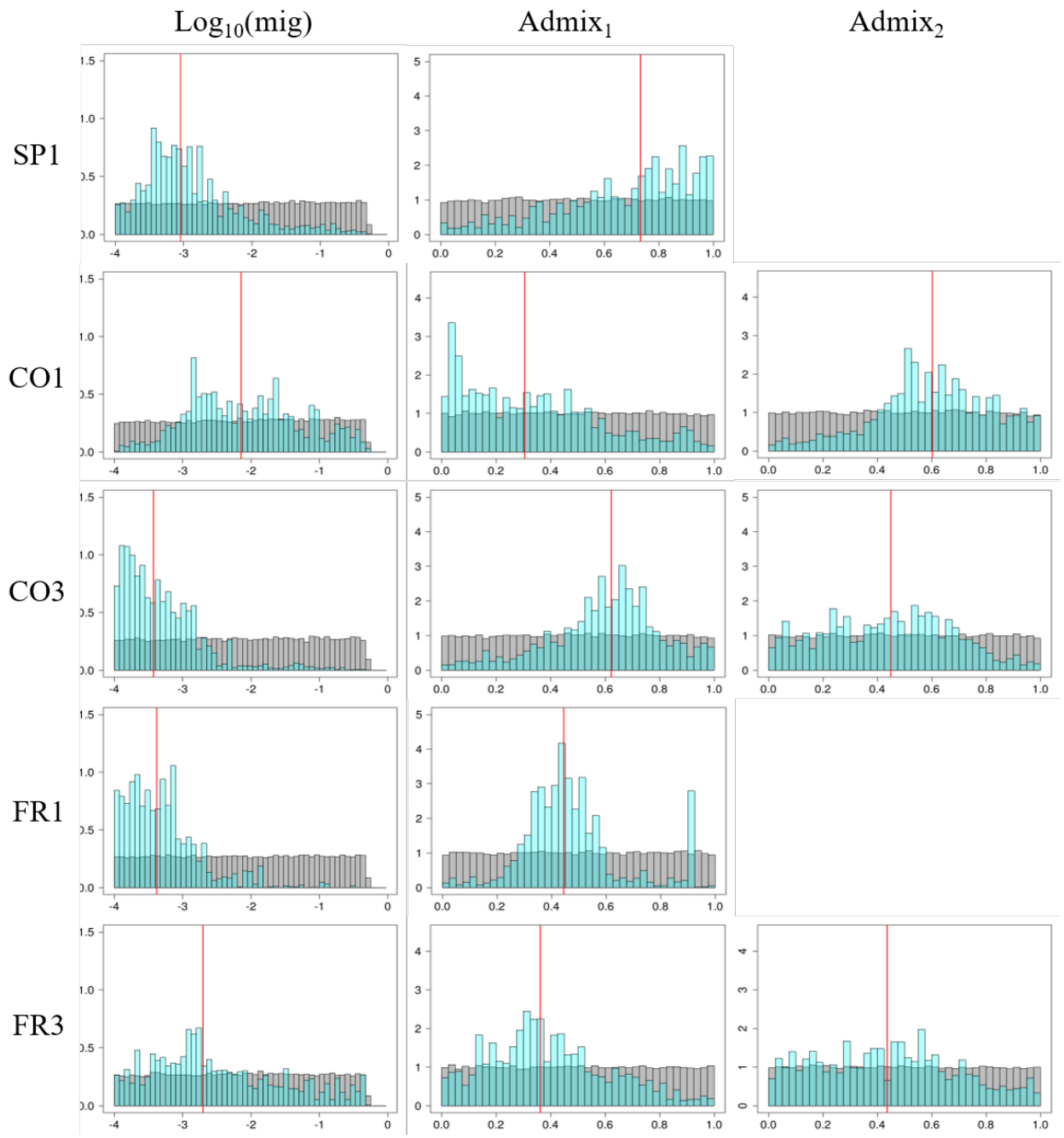


Figure supplémentaire 3 : Distributions de probabilité a posteriori des inférences de taux de migration et d'admixture chez cinq populations de *Medicago truncatula*.

Les barres grises représentent les distributions a priori ; les barres bleues représentent les distributions de probabilité a posteriori ; la ligne verticale rouge correspond à la valeur médiane inférée pour le paramètre.

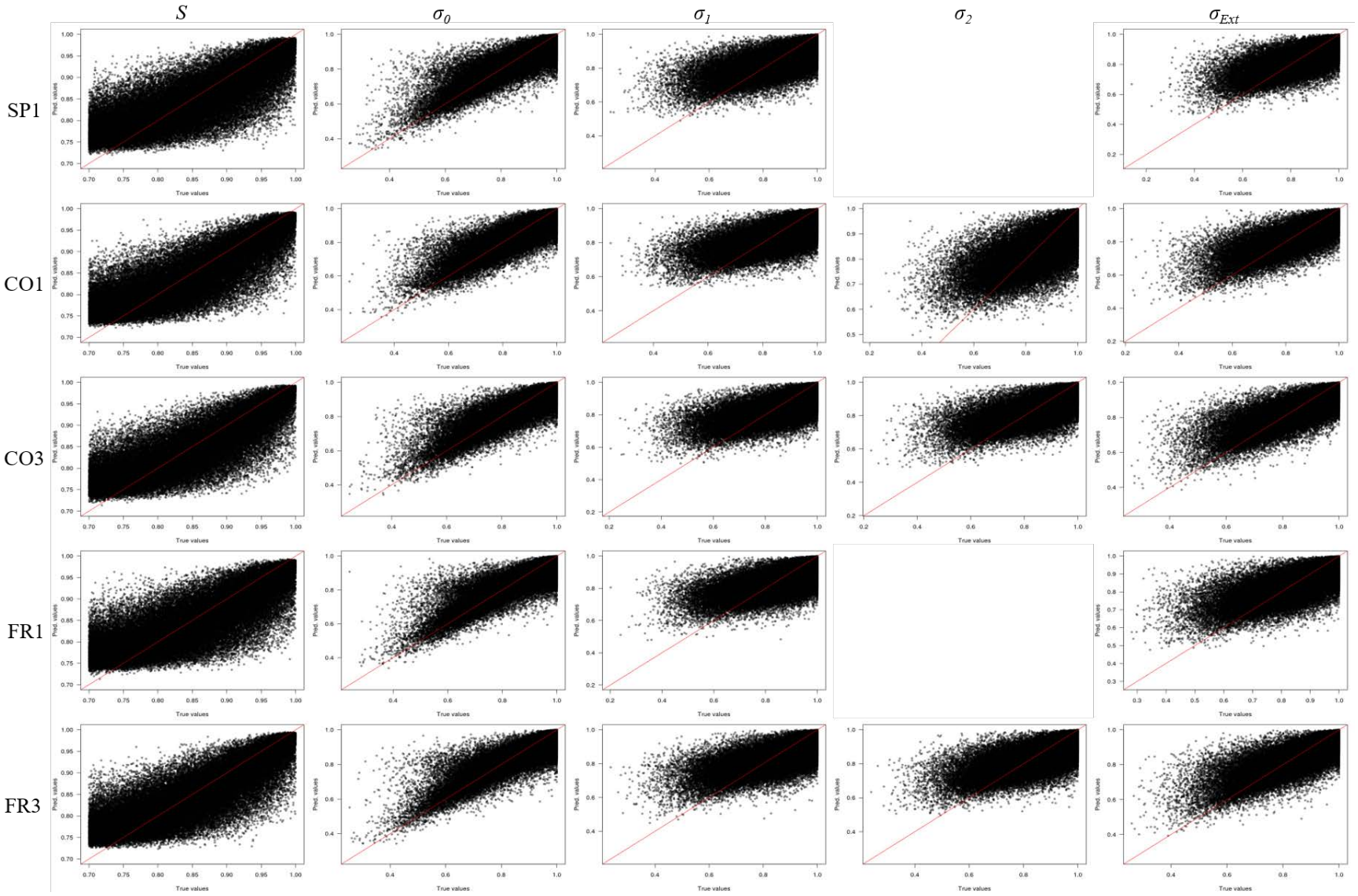


Figure supplémentaire 4 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les taux d'autofécondation.

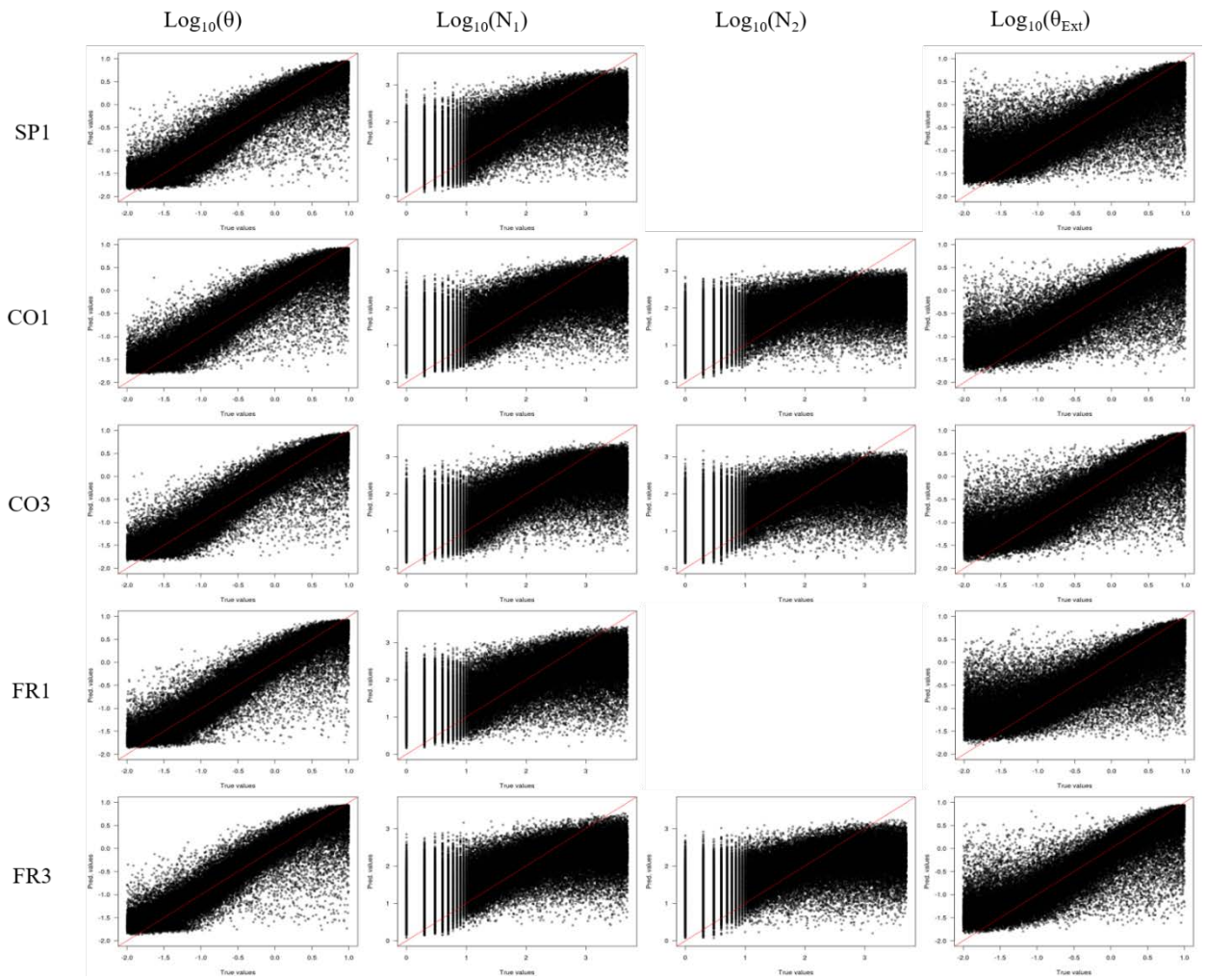


Figure supplémentaire 5 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les diversités ancestrales et les tailles démographiques.

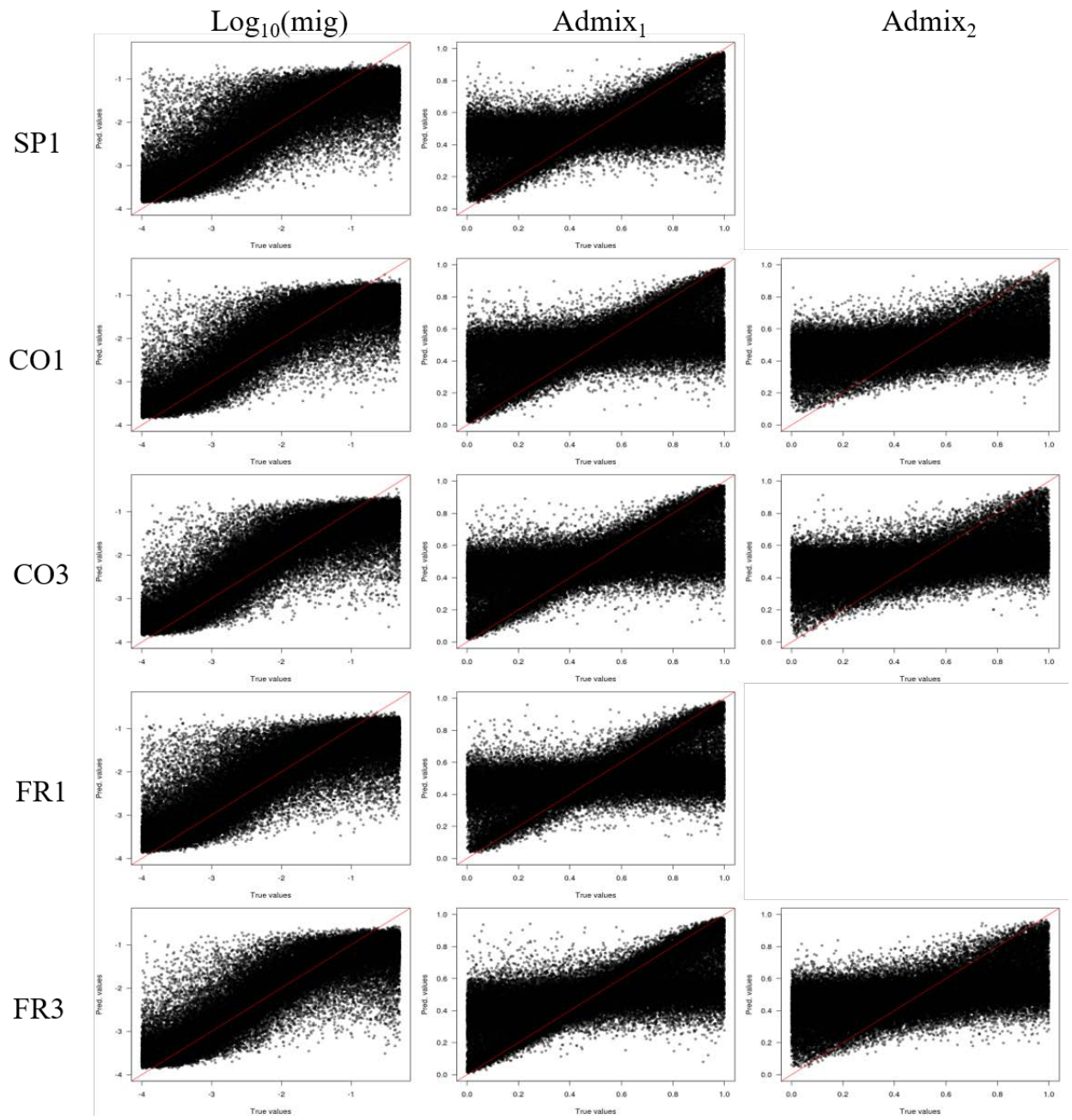


Figure supplémentaire 6 : Valeurs prédites vs vraies valeurs à partir des données out-of-bag pour les taux de migration et d'admixture.

Tableau supplémentaire 2 : Statistiques résumées complémentaires du chapitre 2

n_{ML} est le nombre de locus monomorphes ; λ est l'indice de Simpson ; *E5* est l'indice d'evenness ; I_a est l'indice d'association ; *G* est l'indice de diversité de Stoddart et Taylor ; $\overline{r_D}$ est l'indice d'association standardisé ; *singleMLG* est le nombre de MLG uniques ; D_{mean}, D_{median} et D_{max} sont les distances génétiques moyenne, médiane et maximale entre générations ; $FreqD_0$ est la fréquence de génotypes identiques entre générations.

Population	Year	n_{ML}	λ	*E5*	I_a	*G*	$\overline{r_D}$	*singleMLG*	D_{mean}	D_{median}	D_{max}	$FreqD_0$
SP1	1986	4	0.954	0.732	2.595	21.635	0.217	28				
	2009	4	0.970	0.743	2.812	32.886	0.178	45	0.541	0.529	0.882	0
CO1	1986	1	0.884	0.680	2.682	8.602	0.207	13				
	2009	1	0.844	0.547	3.770	6.404	0.306	12	0.505	0.533	0.867	0.077
	2013	4	0.901	0.530	2.602	10.056	0.194	19	0.586	0.6	0.9	0.032
CO3	1986	2	0.831	0.488	6.069	5.902	0.407	20				
	2009	2	0.970	0.747	2.553	33.646	0.171	43	0.626	0.75	1	0.062
	2013	1	0.966	0.456	3.290	29.823	0.220	84	0.644	0.688	1	0.021
FR1	1991	3	0.664	0.550	3.587	2.980	0.363	11				
	2010	1	0.646	0.564	9.557	2.823	0.749	8	0.339	0.125	0.875	0.192
FR3	2004	1	0.966	0.756	2.151	29.681	0.141	38				
	2010	1	0.969	0.435	4.726	32.713	0.294	106	0.627	0.611	0.944	0.016
	2014	1	0.941	0.443	3.681	16.891	0.236	52	0.586	0.611	0.944	0.035

Synthèse et perspectives

Rappel des objectifs de la thèse

Si les processus démographiques façonnent la diversité des populations, la caractérisation des patrons de diversité populationnels, réciproquement, permet d'inférer l'influence de certaines des forces évolutives. Cet aller-retour entre les prédictions théoriques, basées sur des modèles analytiques de génétique des populations, ou sur des attendus de simulations et les données est la démarche adoptée durant ma thèse. L'objectif de ce travail de thèse était de comprendre finement comment l'autofécondation affecte les attendus théoriques pour la diversité des populations naturelles autogames et pour la variation de cette diversité au cours du temps, puis de confronter ces attendus à des données sur des populations naturelles de la plante modèle des Légumineuses, *Medicago truncatula*. Dans un premier temps, j'ai cherché à caractériser le taux d'autofécondation, ses variations et ses déterminants (génétiques, environnementaux) dans une population naturelle. Dans un deuxième temps, j'ai étudié comment l'autofécondation combinée à la démographie (taille de population et migration) affecte la diversité génétique et son organisation au cours du temps. Comme l'autogamie entraîne des écarts aux hypothèses classiques de la génétique des populations, cette caractérisation s'est appuyée sur des simulations, et avait pour but d'identifier des statistiques informatives adaptées à l'étude des populations autogames. Je me suis plus particulièrement intéressée au comportement de descripteurs de la diversité génétique multilocus dans différents scénarios démographiques afin de définir des attendus neutres, ainsi que sur l'apport des données temporelles pour mieux comprendre l'histoire démographique récente de populations autogames. Enfin, j'ai cherché à développer une méthode permettant d'inférer conjointement le taux d'autofécondation, la taille des populations et la migration à partir de données temporelles.

Résumé des résultats principaux

Nous avons confirmé dans le chapitre 1 que *Medicago truncatula* n'est pas strictement autogame et qu'il peut exister des taux d'allofécondation résiduelle relativement importants (~10% dans la population étudiée). De plus, l'étude a montré que le taux d'allofécondation est variable au cours de la saison de floraison, avec une augmentation de l'allogamie en fin de saison (t_m = 0.08 en début de saison contre 0.14 en fin de saison). Notre étude suggère un effet génétique, qui reste à confirmer, sur la maintenance de ce taux d'allogamie résiduelle. La structure en génotype répétés de la population a également permis de mettre en évidence une héritabilité de l'allofécondation résiduelle, mais celle-ci reste faible (9%) et les effets environnementaux semblent être prédominants. L'occurrence d'allofécondation résiduelle va impacter la diversité génétique de la population en créant des génotypes multilocus hétérozygotes uniques, pouvant former à long terme des lignées

recombinantes. Il serait intéressant d'évaluer la valeur sélective des individus issus d'allofécondation afin de tester si l'allofécondation résiduelle confère un avantage sélectif via la recombinaison.

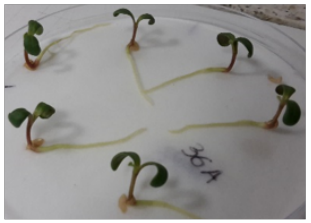
L'étude par simulations du chapitre 2 a permis de mettre en évidence que la combinaison de forts taux d'autofécondation avec de petites tailles de population résulte en une structure de population en génotypes multilocus répétés qui peuvent se maintenir au cours du temps, et ce sans impliquer de sélection. De plus, la combinaison de statistiques de diversité monolocus avec des statistiques multilocus montre que si la taille efficace est un bon prédicteur de la diversité monolocus, la diversité multilocus est quant à elle également affectée par le déséquilibre de liaison qui s'installe entre les locus. Les résultats montrent que la migration restaure la diversité monolocus plus vite que la diversité multilocus. Sur les simulations réalisées ici, la combinaison de statistiques mono- et multilocus permet ainsi de distinguer des scénarios incluant des goulots d'étranglement de scénarios contenant des évènements de migration ou d'admixture. La comparaison des patrons neutres définis par les différents scénarios simulés avec des données temporelles sur neuf populations de *M. truncatula* a également permis de montrer des comportements contrastés des populations considérées. On observe généralement des patrons qui ne sont pas compatibles avec des scénarios de petites populations isolées. Au contraire, la migration semble avoir un rôle important dans les populations de *M. truncatula*. Les attendus neutres développés par simulations ici peuvent servir de base pour explorer des scénarios plus complexes impliquant de la sélection.

Le cadre de simulations développé dans le chapitre 2 a été réutilisé afin d'inférer conjointement le taux d'autofécondation, la taille des populations et le taux de migration de cinq populations de *M. truncatula* à l'aide de la méthode de calcul Bayésien approché. Cette analyse confirme, et généralise à l'ensemble des populations étudiées, que les populations considérées ne sont pas des populations isolées mais qu'elles reçoivent des migrants, et que les taux de migration sont variables au cours du temps et peuvent être massifs (remplacements partiels de populations). Notre méthode ne permet cependant pas d'estimer correctement la taille démographique des populations ni ses changements au cours du temps. En particulier, outre les problèmes de précision de l'estimation rencontrés, on ne peut pas comparer la taille démographique ancestrale (estimée à partir de la diversité θ sur le premier échantillon) avec la taille démographique contemporaine (inférée sur l'échantillon suivant). Trois échantillons temporels sont donc nécessaires au minimum pour estimer des variations de taille démographique. Cela implique alors de réfléchir au pas de temps suffisant pour observer un signal.

Tableau 12 : Récapitulatif des principaux résultats obtenus dans les trois chapitres.

Analyse temporelle de la diversité en régime autogame : approches théorique et empirique

Objectifs :
- Comprendre comment l'autofécondation affecte les attendus théoriques sur la diversité génétique et ses variations au cours du temps en population naturelle
- Confronter les attendus développés à des données empiriques
- Développer des outils d'analyse adaptés aux populations autogames.

Chapitre 1 : Allofécondation résiduelle chez *M. truncatula*	**Chapitre 2** : Diversité génétique multilocus en population autogame	**Chapitre 3** : Inférence conjointe de l'autofécondation, de la taille démographique et de la migration
 Descendances maternelles sur population naturelle de *M. truncatula*	 Simulations Données temporelles sur 9 populations de *M. truncatula*	Calcul Bayésien approché Données temporelles sur 5 populations de *M. truncatula*
Résultats principaux : ➢ ~10% d'allofécondation résiduelle ➢ Augmentation de l'allofécondation résiduelle au cours de la saison de floraison ➢ 9% d'héritabilité	Résultats principaux : ➢ Des scénarios neutres avec de l'autogamie et des tailles de population réduites peuvent générer la structure en MLGs répétés ➢ L'analyse conjointe de statistiques mono- et multilocus permet de différencier des scénarios démographiques ➢ Les 9 populations ont des profils contrastés, mais certains patrons ne sont pas retrouvés par nos simulations	Résultats principaux : ➢ Les populations considérées ne sont pas isolées ➢ On infère bien les taux d'autofécondation, la diversité ancestrale et les taux de migration ➢ Difficultés pour inférer la taille démographique

Des dynamiques de métapopulations

Un des résultats marquants des analyses de diversité que j'ai conduites durant ma thèse sur les populations de *Medicago truncatula* est qu'elles maintiennent une diversité génétique importante par rapport à d'autres populations naturelles de plantes autogames (voir tableau 1 en introduction). Ces résultats sont surprenants compte tenu des attendus théoriques qui prédisent une forte réduction de la diversité génétique en régime autogame. De tels niveaux de diversité peuvent être attribués à des dynamiques de métapopulation, avec de la migration importante (par exemple Chauvet et al. 2004). De fait, les données temporelles ainsi que les comparaisons avec des attendus en simulation et l'inférence de modèle démographique en ABC nous confirment que les populations de *M. truncatula* sont influencées par de la migration, à un taux variable dans le temps. Il semblerait donc que les populations considérées fassent partie d'un ensemble plus large.

L'approche temporelle que nous avons utilisée ici est un outil pertinent pour identifier et préciser ces dynamiques de métapopulation à travers les variations des fréquences au cours du temps. Ces données temporelles sont particulièrement riches dans le cas de populations autogames car on peut s'intéresser à deux types de fréquences : les fréquences alléliques, mais aussi les fréquences des génotypes multilocus (MLG). Le spectre de fréquence de MLGs, développé durant ma thèse et présenté dans le chapitre 2, est un nouvel outil permettant de visualiser ces variations temporelles de fréquence de MLGs. Ces variations sont informatives, notamment pour faire des inférences sur les évènements d'extinction/recolonisation qui affectent les populations (dans la population SP1 par exemple).

Inférer précisément l'histoire démographique récente d'une population autogame à partir de données temporelles nécessite cependant un intervalle de temps suffisamment long entre les échantillons. En effet, nous avons vu dans le chapitre 3 que les inférences sont de meilleure qualité lorsque le signal des différents facteurs démographiques a eu le temps de devenir plus visible. Regarder seulement deux générations successives permet de voir surtout les effets de la dérive génétique qui est attendue forte dans les populations autogames. Une étude plus poussée de notre méthode d'estimation pourrait permettre d'identifier l'intervalle de temps suffisant pour avoir un signal net, ce qui permettrait de fournir des recommandations quant aux plans d'échantillonnage des populations naturelles.

Quels effets de la sélection ?

Au cours de cette thèse, nous n'avons pas considéré les effets de la sélection. Nous avons cependant montré que l'organisation en génotypes multilocus répétés caractéristique des populations autogames est attendue même sans effet de la sélection. Ce résultat vient nuancer l'hypothèse d'Allard (1975), discutée notamment par Avise et Tatarenkov (2012), qui suggère que les génotypes les plus fréquents dans une population sont les mieux adaptés localement. Tester cette hypothèse de manière plus formelle nécessiterait des dispositifs expérimentaux spécifiques tels que des transplantations réciproques des MLGs majoritaires dans toutes les populations, afin de vérifier si le plus fréquent localement est effectivement celui qui a la meilleure valeur sélective dans l'environnement local. Cependant, l'utilisation de données temporelles pourrait également apporter des informations sur la sélection. En effet, du fait de la non-indépendance entre locus, on s'attend à ce que la sélection agisse à l'échelle des MLGs (ou haplotypes) plutôt que des locus (Neher et Shraiman 2009). On peut donc envisager de visualiser les effets de la sélection avec les variations des fréquences de MLGs au cours du temps. En effet, si elles ne sont affectées que par la dérive génétique, on s'attend à observer des variations stochastiques des fréquences de MLGs. Au contraire, si la sélection est directionnelle et homogène au cours du temps, on s'attend plutôt à observer des variations dirigées des fréquences des MLGs sous sélection. Des séries temporelles assez longues et avec plusieurs échantillons temporels seraient intéressantes pour tester cet effet.

Autogames, mais avec un peu d'allofécondation

Un autre résultat marquant de ma thèse est que les populations majoritairement autogames réalisent quand même toutes de l'allofécondation, à un taux variable dans le temps et l'espace. Si l'autofécondation a un impact majeur sur le niveau et la structuration de la diversité des populations autogames, comme détaillé finement dans cette thèse, l'allofécondation résiduelle joue également un rôle en donnant naissance à des génotypes recombinants, qui constituent un deuxième « compartiment » à côté des lignées homozygotes. Les génotypes recombinants re-brassent la diversité génétique des lignées homozygotes à partir desquelles ils sont formés. Ainsi, contrairement à la migration, l'allofécondation résiduelle n'affecte pas la diversité monolocus mais augmente la diversité multilocus en créant des MLGs uniques. Certains de ces MLGs peuvent ensuite augmenter en fréquence pour former des lignées recombinantes lorsque l'hétérozygotie a suffisamment baissé avec les générations d'autofécondation (en F1 ou F2, la totalité et la moitié des locus polymorphes sont à l'état hétérozygote respectivement). De telles lignées ont notamment pu être identifiées dans des populations naturelles de *M. truncatula* (Bonnin et al. 2001 ; Siol et al. 2008).

L'allofécondation résiduelle, en formant ces nouvelles combinaisons alléliques, peut avoir des conséquences importantes sur le potentiel adaptatif des populations autogames. En effet, une fraction importante de la variance génétique des populations autogames peut être cachée par des associations entre locus. En effet, la réduction de la recombinaison efficace permet le maintien d'interactions bénéfiques entre locus qui améliorent la valeur sélective (Lande and Porcher 2015 ; Abu Awad et Roze 2018). L'allofécondation résiduelle libère cette variance et pourrait ainsi contribuer au potentiel adaptatif et donc améliorer la réponse évolutive à un changement environnemental (Clo et al. 2019). Toutefois, certaines des associations entre allèles présentes chez les génotypes homozygotes parentaux peuvent être des coadaptations positives, où l'effet délétère d'un allèle est compensé par l'effet d'un allèle à un autre locus. Ces coadaptations seront cassées par la recombinaison, ce qui peut entrainer une diminution de valeur sélective des individus issus de croisements par rapport à leurs parents, appelée dépression en croisement (« *outbreeding depression* »). De telles relations épistatiques négatives ont été mises en évidence par des nécroses au stade rosette au sein de 2% des croisements d'accessions de l'espèce autogame *Arabidopsis thaliana* (Bomblies et al. 2007). Si la valeur sélective des individus F1 est trop mauvaise, les descendants du croisement ne survivront pas en nombre suffisamment important pour voir apparaître une ou des lignées recombinantes dans la population. On s'attend à ce que la fréquence de relations épistatiques négatives augmente avec la distance génétique entre les parents (Lynch 1991). Or, on a vu que les populations autogames pouvaient être composées de lignées fortement différenciées (comme montré par les fortes valeurs de D_{max} dans le chapitre 2), qui peuvent donc avoir accumulé des incompatibilités. Cependant, on s'attend en parallèle à ce que les populations autogames accumulent des mutations délétères récessives sous l'effet de la dérive génétique accrue (Hartfield and Glémin 2014, 2016). Dans ce cas, l'allofécondation résiduelle entre MLGs ayant fixé des mutations délétères différentes peut engendrer de l'hétérosis, c'est-à-dire une augmentation de la valeur sélective des descendants de croisement par rapport à leurs parents. En effet, les mutations délétères sont masquées par l'état hétérozygote chez les individus F1. De l'hétérosis, ainsi que de la dépression de croisement, ont été observées chez des descendances issues de croisements entre des MLGs provenant de populations naturelles d'*A. thaliana* en Suède et en Italie (Oakley et al. 2015). Enfin, la recombinaison dans la descendance des F1 peut permettre de casser d'éventuelles associations entre allèles bénéfiques et mutations délétères, et ainsi favoriser la sélection sur les allèles bénéfiques.

Ainsi, les associations entre allèles maintenues par l'autofécondation vont fortement influencer la réponse à la sélection. En effet, la sélection n'agit plus à l'échelle du gène mais à l'échelle du génotype multilocus, ce qui accentue les effets de sélection d'arrière-plan ou de balayage sélectif, et

peut créer de l'interférence sélective, qui va ralentir la réponse à la sélection (Felsenstein 1974). Le résidu d'allofécondation peut donc avoir un rôle très important en créant de la variance génétique (via le cassage de combinaisons alléliques favorables, la réassociation d'allèles favorables, etc.). Cet effet sera d'autant plus important si l'allofécondation résiduelle a une composante génétique, comme suggéré par notre étude, et pourrait répondre à la sélection.

Conclusion

Qu'est-ce qu'une population autogame ?

Le concept de population est relativement flou et différentes définitions existent selon le domaine dans lequel on se place (Waples and Gaggiotti 2006). Dans le cas d'une espèce autogame, ce concept semble pouvoir se décliner à plusieurs niveaux. En effet, on peut considérer que l'unité d'évolution est la métapopulation, la population locale, les sous-populations locales, ou même le génotype multilocus. Une caractérisation fine de ces différents niveaux est donc cruciale pour déterminer à quelle échelle les forces évolutives vont agir.

Quel potentiel adaptatif ?

On considère généralement que les populations autogames ont un faible potentiel adaptatif du fait de la perte de diversité génétique entraînée par l'autofécondation. Noël et al. (2017) ont démontré expérimentalement une telle diminution du potentiel adaptatif chez des populations autogames isolées. Nous avons cependant vu que les populations autogames sont rarement isolées et qu'elles peuvent recevoir des migrants de façon régulière. Ce fonctionnement en métapopulations pourrait ainsi contrer les effets de la diminution de diversité et limiter les extinctions, allant à l'encontre du l'hypothèse du cul-de-sac évolutif.

Bibliographie

Abu Awad, D., and D. Roze. 2018. Effects of partial selfing on the equilibrium genetic variance, mutation load, and inbreeding depression under stabilizing selection. Evolution 72:751–769.

Allard, R. W. 1975. The mating system and microevolution. Genetics 79 Suppl:115–126.

Avise, J. C., and A. Tatarenkov. 2012. Allard's argument versus Baker's contention for the adaptive significance of selfing in a hermaphroditic fish. PNAS 109:18862–18867.

Bomblies, K., J. Lempe, P. Epple, N. Warthmann, C. Lanz, J. L. Dangl, and D. Weigel. 2007. Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLOS Biology 5:e236.

Bonnin, I., J. Ronfort, F. Wozniak, and I. Olivieri. 2001. Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). Molecular Ecology 10:1371–1383.

Chauvet, S., M. V. D. Velde, E. Imbert, M. L. Guillemin, M. Mayol, M. Riba, M. J. M. Smulders, B. Vosman, L. Ericson, R. Bijlsma, and B. E. Giles. 2004. Past and current gene flow in the selfing, wind-dispersed species *Mycelis muralis* in Western Europe. Molecular Ecology 13:1391–1407.

Clo, J., L. Gay, and J. Ronfort. 2019. How does selfing affect the genetic variance of quantitative traits? An updated meta-analysis on empirical results in angiosperm species. Evolution (accepted).

Felsenstein, J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Hartfield, M., and S. Glémin. 2014. Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. Genetics 196:281–293.

Hartfield, M., and S. Glémin. 2016. Limits to adaptation in partially selfing species. Genetics 203:959–974.

Lande, R., and E. Porcher. 2015. Maintenance of quantitative genetic variance under partial self-fertilization, with implications for evolution of selfing. Genetics 200:891–906.

Lynch, M. 1991. The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622–629.

Neher, R. A., and B. I. Shraiman. 2009. Competition between recombination and epistasis can cause a transition from allele to genotype selection. PNAS 106:6866–6871.

Noël, E., P. Jarne, S. Glémin, A. MacKenzie, A. Segard, V. Sarda, and P. David. 2017. Experimental evidence for the negative effects of self-fertilization on the adaptive potential of populations. Current Biology 27:237–242.

Oakley, C. G., J. Ågren, and D. W. Schemske. 2015. Heterosis and outbreeding depression in crosses between natural populations of *Arabidopsis thaliana*. Heredity (Edinb) 115:73–82.

Siol, M., J.-M. Prosperi, I. Bonnin, and J. Ronfort. 2008. How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in *Medicago truncatula*. Heredity 100:517–525.

Waples, R. S., and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15:1419–1439.

Annexe : Caractérisation de l'allofécondation résiduelle au sein d'un dispositif expérimental contrôlé

Introduction

Dans le chapitre 1, nous avons mis en évidence un déterminisme génétique du taux d'allofécondation résiduelle dans une population naturelle de *Medicago truncatula*. Cependant, l'effet génétique expliquait une très faible proportion de la variance. Ceci peut être une conséquence des forts effets environnementaux également détectés, mais aussi d'un problème d'identifiabilité dans les modèles statistiques dû au manque de répétitions de certains génotypes maternels. Nous cherchons donc ici à mieux caractériser le déterminisme génétique de l'allofécondation résiduelle chez *M. truncatula* en contrôlant mieux le design expérimental, c'est-à-dire les effets environnementaux et les génotypes maternels. De plus, nous cherchons à évaluer l'effet de l'allofécondation sur les traits d'histoire de vie des graines tels que la germination.

Matériels et méthodes

Dispositif expérimental

Afin de décortiquer le déterminisme génétique de l'allofécondation résiduelle chez une espèce fortement autogame, nous avons utilisé une expérience mise en place pour tester l'adaptation locale dans trois populations de *Medicago truncatula*. L'expérience, mise en place à l'automne 2014 comprenait deux dispositifs : un jardin commun (dispositif « inter-population ») et une expérience de transplantations réciproques (dispositif « intra-population »). Pour le dispositif « intra-population », des transplantations réciproques entre les génotypes majoritaires de trois zones de la population FR3, située à côté de Narbonne, ont été effectuées (Figure 16). Trois génotypes ont été sélectionnés dans chacune des zones et ont été semés de manière randomisée dans deux carrés de 50x50 cm par zone, à raison de quatre répétitions par génotype. Pour le dispositif « inter-populations » ou jardin commun, onze génotypes de *M. truncatula* provenant de trois populations ont été mélangés : six des neuf génotypes de la population FR3, trois génotypes d'une population corse (CO3) et deux d'une population espagnole (SP3). Ces onze génotypes ont été semés de manière randomisée dans six carrés

de 50x50 cm sur un terrain expérimental de l'INRA LBE à Narbonne, avec quatre répétitions par génotype par carré. Les génotypes diffèrent notamment par leur date de floraison, les génotypes de CO3 et SP3 étant plus précoces que les génotypes de FR3. Dans les deux dispositifs, la terre des carrés a été préalablement tamisée afin de retirer les gousses de *M. truncatula* naturellement présentes. Les mélanges de génotypes ont été semés sous forme de gousses. Les germinations sur le terrain ont été très hétérogènes, les génotypes espagnols notamment ont très mal germé et l'on dispose de peu de répétitions (seulement 4 plantes mères, Tableau 13).

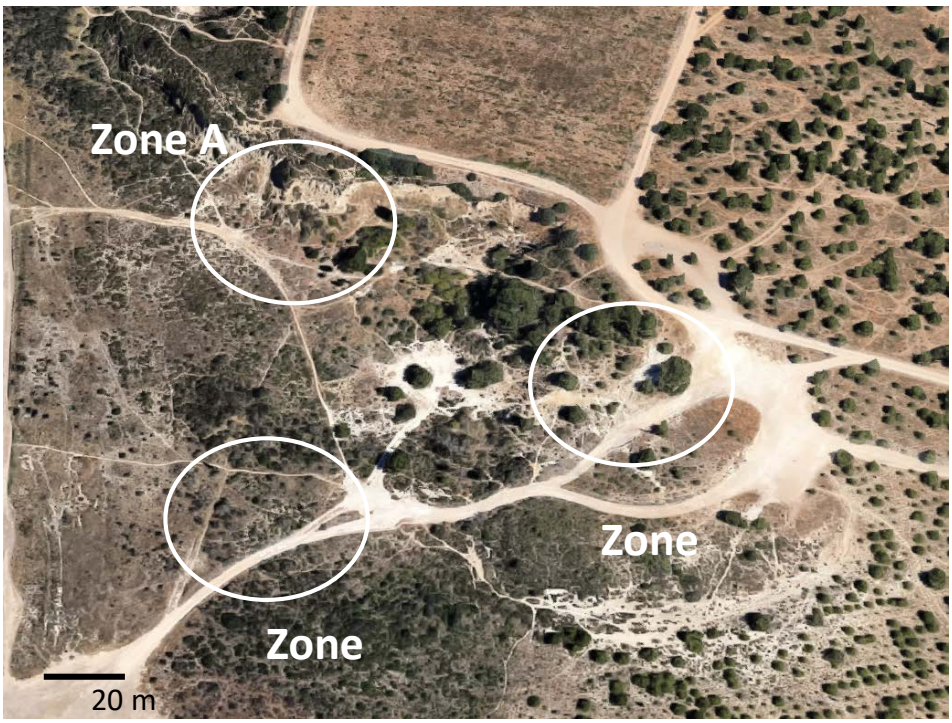


Figure 16 : Zones de la population FR3 utilisées pour les transplantations réciproques

Les dates de floraison de chaque plante mère ont été notées et la fécondation s'est faite librement. En fin de saison (été 2015), les gousses produites dans les deux dispositifs ont été récoltées pour chaque plante mère, puis battues pour pouvoir récupérer les graines (Tableau 13). Au total, 5411 graines ont été récoltées.

Tableau 13 : Effectifs des dispositifs expérimentaux.
$N_{plantes}$ est le nombre de plantes mères ayant germé, $N_{gousses}$ est le nombre de gousses collectées et $N_{graines}$ est le nombre de graines. Les génotypes sont identifiés par leur population d'origine (FR3, CO3 ou SP3) et l'identifiant du génotype.

Génotype	$N_{plantes}$		$N_{gousses}$		$N_{graines}$	
	Intrapop	Interpop	Intrapop	Interpop	Intrapop	Interpop
FR3_H1	7	13	11	150	46	957
FR3_H11.5	4	-	17	-	81	-
FR3_H2	11	9	21	82	104	485
FR3_H4	5	8	21	82	70	494
FR3_H5	14	8	27	50	103	266
FR3_H50	2	4	2	91	8	551
FR3_H6	2	-	3	-	11	-
FR3_H8	7	6	18	23	76	128
FR3_H96	4	-	7	-	19	-
CO3_H1	-	11	-	121	-	680
CO3_H4	-	12	-	83	-	470
CO3_H5	-	12	-	61	-	352
SP3_H2	-	3	-	54	-	330
SP3_H3	-	1	-	25	-	180
Total	56	87	127	822	518	4893
Intrapop + Interpop	143		949		5411	

Germination et tests de paternité

Afin de mesurer les taux d'allofécondation résiduelle, nous avons effectué des tests de paternité en génotypant les descendances maternelles récoltées à l'aide d'un quadruplex de quatre marqueurs microsatellites préalablement défini pour discriminer les génotypes maternels (TPC95G, TA34, TPC63A, DMI1-6, Baquerizo-Audiot *et al.*, 2001). Les génotypes maternels sont tous homozygotes sur ces quatre locus. Les graines ont été mises à germer sur papier filtre dans des boîtes de Petri. Chaque boîte contenait les graines d'une même gousse avec un maximum de 6 graines par boîte. La germination a eu lieu en chambre de culture (Aralab) en trois séries consécutives (soit trois blocs d'expérimentation). Dans une première étape, les graines ont été mises à gonfler par imbibition d'eau à l'obscurité et à 5°C pendant 3 jours. Les graines n'ayant pas gonflé au bout de 24h subissent une

dormance tégumentaire. Elles ont été scarifiées à l'aide de papier verre à grain fin pour lever la dormance. La chambre de culture était alors réglée à 20°C pour une alternance 12h/12h lumière/obscurité. Chaque jour, les stades de germination suivants étaient relevés : graine gonflée, sortie de radicule et cotylédons déployés (Figure 17). Les prélèvements pour le génotypage ont été réalisés au stade cotylédons déployés et la viabilité de la plantule était évaluée par la note de 0 (plantule non viable) ou 1 (plantule viable). Pour éviter le génotypage individuel de chaque plantule, le génotypage a été réalisé par mélange de plantules provenant d'une même gousse. Un cotylédon par plantule a été prélevé, le reste de la plantule étant conservé à -20°C. On s'attend donc à ce que, en l'absence d'allofécondation, tous les descendants soient de génotype identique et identique à la mère, ce qui se reflète par un seul allèle à chaque locus. Nous avons détecté les mélanges montrant de l'hétérozygotie à un ou plusieurs locus et les plantules correspondantes ont été décongelées afin de faire une extraction d'ADN et un génotypage individuels.

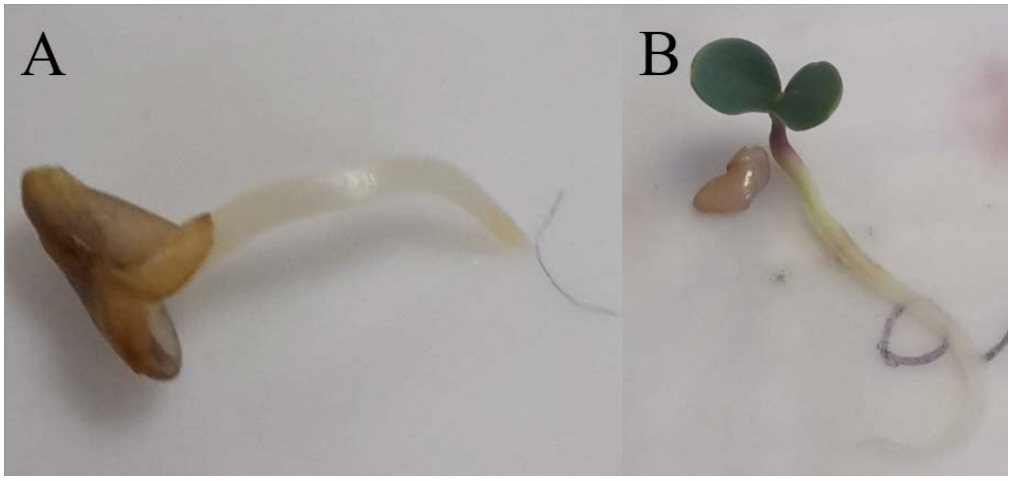


Figure 17 : Stades de germination de graines de *M. truncatula*.
(A) Stade radicule sortie ; (B) Plantule au stade cotylédons déployés.

Résultats

Taux de germination

Seulement 13% des graines ont gonflé sans scarification, ce qui signifie que 87% des graines subissent une dormance tégumentaire. La vitesse moyenne d'émergence de la radicule est de 1.78 jours après gonflement de la graine. La vitesse moyenne de déploiement des cotylédons est de 5.36 jours (Figure 18). Au total, 73% des plantules obtenues après germination ont été considérées comme viables.

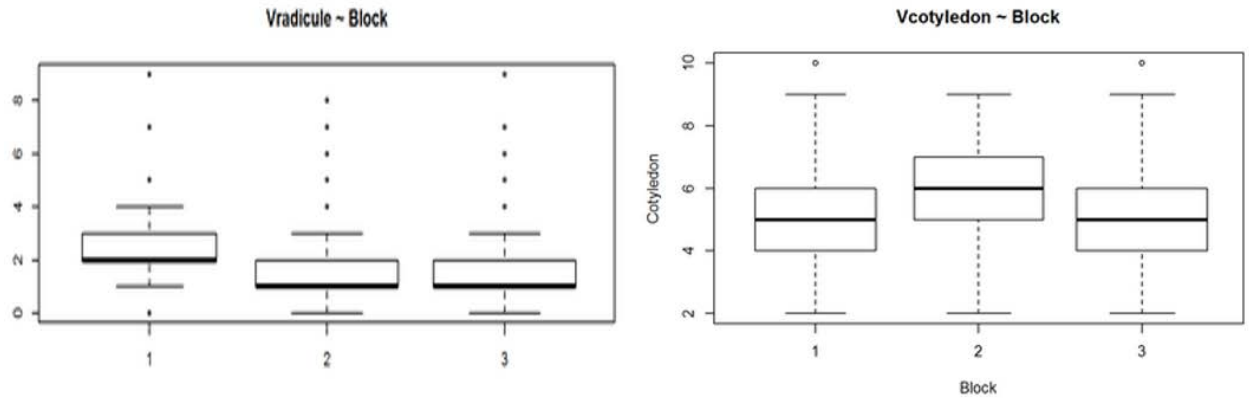


Figure 18 : Vitesses d'émergence de la radicule et des cotylédons dans les trois blocs expérimentaux.

Taux d'allofécondation résiduelle

Des évènements d'allofécondation ont été détectés chez seulement quatre plantules issues du génotype maternel FR3_H1, correspondant à un taux d'allofécondation résiduelle de seulement 0.7%.

Discussion

Le taux d'allofécondation résiduelle observé ici est particulièrement faible, même pour une plante majoritairement autogame telle que *Medicago truncatula*. Il n'y a pas suffisamment de variabilité pour tester le déterminisme génétique de l'allofécondation résiduelle. Ce résultat montre un taux d'autofécondation inférieur aux estimations obtenues dans le chapitre 1 (autour de 10%), le chapitre 2 (estimations avec le programme RMES pour les populations FR3, CO3 et SP3 : taux d'autofécondations compris entre 0.93 et 0.99) ou le chapitre 3 (estimation ABC pour les populations FR3 et CO3 : taux d'autofécondations compris entre 0.91 et 0.98). La comparaison de ces estimations souligne la forte variabilité du taux d'autofécondation, ce qui suggère que les effets environnementaux ont un effet très important sur le taux d'allofécondation résiduelle. Dans ce dispositif expérimental, on peut supposer en particulier un effet de la densité. En effet, les faibles taux de germination des plantes mères sur le terrain ont résulté en des densités de plantes faibles par rapport aux densités observées dans les populations naturelles. Or, plusieurs études lient la densité à la dispersion du pollen (Karron *et al.*, 1995). Par ailleurs, l'effet de la densité sur la dispersion du pollen et l'allofécondation chez *M. truncatula* a déjà été suggéré par Bonnin et al. (2001). Ce résultat souligne l'importance du contrôle des conditions environnementales lorsque l'on cherche à analyser le taux d'autofécondation.

Bibliographie

Baquerizo-Audiot E, Desplanque B, Prosperi JM, Santoni S (2001). Characterization of microsatellite loci in the diploid legume *Medicago truncatula* (barrel medic). *Molecular Ecology Notes* **1**: 1–3.

Bonnin I, Ronfort J, Wozniak F, Olivieri I (2001). Spatial effects and rare outcrossing events in *Medicago truncatula* (Fabaceae). *Molecular Ecology* **10**: 1371–1383.

Karron JD, Thumser NN, Tucker R, Hessenauer AJ (1995). The influence of population density on outcrossing rates in *Mimulus ringens*. *Heredity* **75**: 175–180.

Analyse temporelle de la diversité en régime autogame : approches théorique et empirique

La diversité génétique des populations est affectée à la fois par les processus démographiques (taille de population et migration) et par le système de reproduction. Chez les plantes, de nombreuses espèces sont préférentiellement autogames et se reproduisent presque exclusivement par autofécondation. Les attendus théoriques prédisent que l'autogamie va réduire drastiquement la diversité génétique et la taille efficace des populations. Les données empiriques confirment ces prédictions mais montrent une variabilité des niveaux de diversité génétique plus importante chez les autogames que chez les allogames. Par ailleurs, les populations autogames présentent une organisation spécifique avec quelques génotypes multilocus répétés et plusieurs génotypes uniques. Ce travail de thèse vise à mieux comprendre le fonctionnement des populations autogames et à développer des attendus pour différents scénarios démographiques afin de pouvoir inférer les processus en jeu dans des populations naturelles. Dans un premier temps, je montre que malgré un régime autogame, *Medicago truncatula* réalise régulièrement de l'allofécondation à des taux variables au cours de la saison de floraison et entre génotypes. Ensuite, j'utilise une approche de simulations pour définir des attendus sur la diversité multilocus dans des scénarios démographiques contrastés. Pour finir, je développe un cadre d'inférence permettant de considérer à la fois l'autofécondation, la taille démographique et la migration, et je l'applique à des données temporelles sur des populations de *M. truncatula*. Mes analyses montrent que l'organisation en génotypes multilocus caractéristique des populations autogames peut être générée par des processus démographiques neutres et que les indices de diversité multilocus sont informatifs pour différencier les effets de la taille de population, de la migration et d'évènements d'extinction-recolonisation. De plus, j'ai mis en évidence des dynamiques de métapopulation importantes dans les populations de *M. truncatula*, qui sont fortement influencées par la migration, parfois très importante.

Mots clés : diversité génétique, autofécondation, systèmes de reproduction, suivi temporel, *Medicago truncatula*

Temporal analysis of diversity and selfing using theoretical and empirical approaches

Genetic diversity of natural populations is affected by both demographic processes (population size and migration), and mating system. A large number of plant species reproduce almost exclusively through selfing. Theoretical expectations predict that selfing will drastically decrease genetic diversity and the effective size of populations. Empirical data confirm those expectations but show a higher variability of diversity levels in selfing compared to outcrossing species. Moreover, selfing populations are characterized by a few repeated multilocus genotypes and several unique genotypes. The purpose of this PhD thesis is to better understand the dynamics of selfing populations and to provide expectations under various demographic scenarios in order to infer the processes at play in natural populations. First, I show that despite a selfing regime, *Medicago truncatula* regularly outcrosses at different rates over the flowering season and between genotypes. I then use a simulation approach to provide multilocus diversity expectations under contrasted demographic scenarios. Finally, I develop an inference method taking into account selfing, population size and migration, and apply it on temporal data from *M. truncatula* natural populations. My analyses show that neutral demographic processes can result in the organization in repeated multilocus genotypes typical of selfing populations and that multilocus diversity indices are informative to distinguish the effects of population size, migration and extinction-recolonization events. Moreover, I show the prevalence of metapopulation dynamics in *M. truncatula* natural populations, with a strong influence of (sometimes strong) migration.

Key words: genetic diversity, selfing, mating systems, temporal survey, *Medicago truncatula*