Génétique des populations subdivisées : théorie et applications

Renaud Vitalis

▶ **To cite this version:**

Renaud Vitalis. Génétique des populations subdivisées : théorie et applications. Ecologie, Environnement. Université Montpellier II - Sciences et Techniques du Languedoc, 2001. Français. NNT : . tel-00535513

HAL Id: tel-00535513
https://theses.hal.science/tel-00535513

Submitted on 11 Nov 2010

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

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

Académie de Montpellier
Université Montpellier II
Sciences et Techniques du Languedoc

THÈSE

présentée à l'Université Montpellier II des Sciences et Techniques du Languedoc pour obtenir le diplôme de DOCTORAT

Spécialité : Biologie, Physiologie des Organismes et des Populations
Formation doctorale : Biologie de l'Évolution et Écologie
École doctorale : Biologie Intégrative

Génétique des populations subdivisées : théorie et applications

par

Renaud VITALIS

Soutenue le 4 décembre 2001 devant le jury composé de

Patrick GRILLAS	Chercheur, Tour du Valat	Co-directeur
Jean-Dominique LEBRETON	Directeur de Recherche, CNRS Montpellier	Président
Yannis MICHALAKIS	Directeur de Recherche, CNRS Montpellier	Examinateur
Richard A. NICHOLS	Reader, QMW Université de Londres	Rapporteur
Isabelle OLIVIERI	Professeur, Université Montpellier II	Directeur
François ROUSSET	Chargé de Recherche, Université Montpellier II	Examinateur
Myriam VALERO	Directrice de Recherche, CNRS Roscoff	Rapporteur

Excusé :

Michael WHITLOCK	Professeur, Université de Vancouver	Rapporteur

Académie de Montpellier
Université Montpellier II
Sciences et Techniques du Languedoc

THÈSE

présentée à l'Université Montpellier II des Sciences et Techniques du Languedoc pour obtenir le diplôme de DOCTORAT

Spécialité : Biologie, Physiologie des Organismes et des Populations
Formation doctorale : Biologie de l'Évolution et Écologie
École doctorale : Biologie Intégrative

Génétique des populations subdivisées : théorie et applications

par

Renaud VITALIS

Soutenue le 4 décembre 2001 devant le jury composé de

Patrick GRILLAS	Chercheur, Tour du Valat	Co-directeur
Jean-Dominique LEBRETON	Directeur de Recherches, CNRS Montpellier	Président
Yannis MICHALAKIS	Directeur de Recherches, CNRS Montpellier	Examinateur
Richard A. NICHOLS	Reader, QMW Université de Londres	Rapporteur
Isabelle OLIVIERI	Professeur, Université Montpellier II	Directeur
François ROUSSET	Chargé de Recherches, Université Montpellier II	Examinateur
Myriam VALERO	Directrice de Recherches, CNRS Roscoff	Rapporteur

Excusé :

Michael WHITLOCK	Professeur, Université de Vancouver	Rapporteur

Table des matières

Remerciements **xi**

Avant propos **xv**

I Document de synthèse **1**

1 Introduction **3**

1.1 Biologie et génétique de la conservation 4

1.1.1 La biodiversité en crise 4

1.1.2 Quels outils pour quelles questions ? 9

1.2 Un exemple : *Marsilea strigosa* 11

1.2.1 Biologie d'une espèce rare 11

1.2.2 Un niveau de variation génétique faible 13

1.2.3 Une très forte structuration de la variation 14

1.3 L'œuf et la poule 18

2 L'identité génétique **21**

2.1 L'identité génétique à un locus 22

2.1.1 L'identité par descendance et l'identité par état 22

2.1.2 L'identité génétique et les temps de coalescence 24

2.1.3 Les F-statistiques 30

2.2 Structure en classes d'individus 37

2.2.1 Le cas des espèces animales à sexes séparés 37

2.2.2 Le cas des plantes annuelles avec banques de graines . 44

2.3 L'identité génétique à deux locus 49

2.3.1 Définitions 50
2.3.2 Influence du modèle de mutation 55
2.3.3 Influence du régime de reproduction 56
2.3.4 Influence de la dispersion limitée 57
2.4 Conclusion 60

3 La taille efficace 61
3.1 Définitions 62
3.1.1 Le modèle de Fisher et de Wright 62
3.1.2 Taille efficace et probabilités de coalescence 65
3.2 Influence de la structure des populations 66
3.2.1 Structure spatiale 66
3.2.2 Structure temporelle 70
3.3 Influence de la sélection 76
3.3.1 L'effet Robertson 76
3.3.2 Quelques exemples 78
3.3.3 L'effet Hill-Robertson 79
3.4 Conclusion 80

4 Applications 81
4.1 Estimation de la taille efficace 81
4.1.1 Pourquoi mesurer la taille efficace et les flux de gènes ? 81
4.1.2 Comment estimer la taille efficace ? 84
4.1.3 Estimer l'effectif efficace et le taux d'immigration à partir des déséquilibres d'identité 86
4.1.4 Limites de l'approche 98
4.2 Estimation des taux de dispersion par sexe 101
4.2.1 Caractérisation des biais de dispersion 101
4.2.2 Mesure des taux de migration mâles et femelles 105
4.2.3 Limites de l'approche 111
4.3 Influence de la sélection sur les marqueurs 112
4.3.1 Un modèle de divergence d'une paire de populations . . 116

4.3.2 Une nouvelle méthode pour détecter des marqueurs répondant à la sélection 123
4.3.3 Limites de l'approche 131

5 Conclusion **135**
5.1 Quelle variation génétique ? 135
5.2 Originalités et limites des approches utilisées 138
5.3 Evolution des traits d'histoire de vie 141

Bibliographie **143**

II Annexes **171**

A La biologie des extinctions **173**

B *M. strigosa* : une espèce menacée **183**

C Locus microsatellites chez *M. strigosa* **199**

D Structure multilocus chez *M. strigosa* **205**

E Le déséquilibre d'identité **237**

F Estimation de la taille efficace **255**

G Estim 1.0 **273**

H Différenciation sexe-spécifique **279**

I Détection de la sélection **295**

J Taille efficace et banques de graines **311**

Liste des Figures

1.1 Carte représentant les 13 localités où *M. strigosa* a été échantillonnée en France, au Maroc et en Espagne 13
1.2 Répartition géographique des génotypes microsatellites (multilocus) dans la réserve naturelle de Roque-Haute 16
1.3 Réseau des génotypes microsatellites dans la réserve naturelle de Roque-Haute 17

2.1 Un exemple de généalogie de 10 gènes 25
2.2 Distribution des probabilités de coalescence 29
2.3 F-statistiques et distribution des probabilités de coalescence . 36
2.4 Valeurs à l'équilibre des F_{ST} par sexe 40
2.5 Variation des valeurs de F_{ST} par sexe au cours du temps . . . 41
2.6 Distribution des probabilités de coalescence par sexe 43
2.7 Cycle de vie d'une plante annuelle et de sa banque de graines . 45
2.8 Distribution de l'âge des germinations, pour différents taux de germination 46
2.9 Influence de la taille de la population et du taux de germination sur F_{IS} 47
2.10 Valeurs de F_{IS} à l'équilibre, parmi les plantes en fleurs et parmi les graines de la banque 48
2.11 Définition des probabilités d'identité à deux locus lorsque les croisements se font au hasard 50
2.12 Définition des probabilités d'identité à deux locus pour un régime de reproduction mixte 52

2.13 Déséquilibre d'identité attendu sous différents modèles de mutation 55
2.14 Déséquilibre d'identité η' attendu pour différents régimes de reproduction 57
2.15 Influence de la dispersion limitée sur le déséquilibre d'identité 58
2.16 Effet de la dispersion limitée sur le déséquilibre d'identité pour deux régimes de reproduction contrastés 59

3.1 Un exemple de généalogie de gènes dans deux populations connectées 67
3.2 Taille efficace dans une population structurée d'une espèce dioïque 69
3.3 Taille efficace asymptotique d'une population de plantes avec une banque de graines 72
3.4 Taille efficace et différenciation génétique d'une population structurée de plantes avec une banque de graines 73
3.5 Taille efficace asymptotique d'une population de plantes monocarpiques pérennes 74
3.6 Démographie d'une population de plantes monocarpiques pérennes 75

4.1 Le déséquilibre d'identité en fonction de la taille efficace et du taux d'immigration 87
4.2 Déséquilibre d'identité η'_S attendu à l'équilibre, en fonction de la taille efficace et de F, dans un modèle à nombre infini d'allèles 94
4.3 Distribution des estimateurs de N_e and m dans un modèle à nombre d'îles infini pour différentes valeurs du produit des paramètres Nm et pour différents nombres de locus génotypés 96
4.4 Distribution des estimateurs de N_e and m dans un modèle à 20 îles 97
4.5 Distribution des estimateurs de N_e et m dans un modèle à nombre d'îles infini en situation de non-équilibre 99
4.6 Un exemple d'estimation des taux de migration par sexe . . . 107

4.7 Erreur quadratique moyenne relative sur les valeurs estimées des taux de migration par sexe 109
4.8 Un exemple de généalogie de $n = 10$ gènes dans un modèle de divergence de deux populations isolées 116
4.9 Distribution totale des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans un modèle de divergence de deux populations 121
4.10 Distribution conditionnelle des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans un modèle de divergence de deux populations 122
4.11 Valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ estimées pour 43 locus du jeu de données de *Drosophila simulans* dans la comparaison impliquant les populations de France ($n = 55$) et de Tunisie ($n = 52$) 125
4.12 Valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ estimées pour 43 locus du jeu de données de *Drosophila simulans* dans toutes les comparaisons impliquant la population du Congo ($n = 45$) 128
4.13 Application de la méthode de Beaumont et Nichols (1996) au jeu de données *D. simulans* décrit dans Singh *et al.* (1987) et Choudhary *et al.* (1992) 129

Liste des Tableaux

4.1 Propriétés des estimateurs des probabilités d'identité à un et deux locus 93
4.2 Résultats de simulations d'un modèle à nombre d'îles fini . . . 108
4.3 Application de la méthode d'identification de locus soumis à la sélection 132

Remerciements

La première personne à qui j'ai envie d'adresser ces remerciements est Isabelle Olivieri. Car s'il est vrai que cette thèse est à plus d'un titre une mosaïque de travaux, c'est elle qui a su créer les conditions dans lesquelles ces collaborations ont pu prendre place. Montpellier a été un port d'attache vers lequel je suis toujours revenu. On a souvent évoqué le côté un peu « chats perdus » des thésards d'Isabelle, qui sont souvent arrivés dans son labo parce que d'autres projets étaient tombés à l'eau. Je te remercie Isabelle de m'avoir permis d'*évoluer* (dans tous les sens du terme) dans cet univers et de m'avoir permis de participer pleinement à la vie de la recherche. Merci Isabelle pour les discussions, les conseils et même les coups de gueule ! Merci aussi pour ton amitié, pour tout ce qui fait que tu n'est pas une « chef » comme les autres.

Merci à Myriam Valero (« sans accent, c'est un nom espagnol »), Richard Nichols et Mike Whitlock d'avoir accepté d'être rapporteur de cette thèse. J'espère Myriam que tu nous pardonneras les effets de « douche écossaise » qu'Isabelle et moi t'avons fait subir à quelques jours du dépôt de ce manuscrit ! Merci également à Jean-Dominique Lebreton, Yannis Michalakis et François Rousset, qui ont accepté de participer à ce jury et de juger ce travail.

Je voudrais remercier la station biologique de la Tour du Valat qui m'a financé, à travers la fondation Sansouire, pendant ces trois dernières années. Merci à Luc Hoffmann et Jean-Paul Tarris de m'y avoir accueilli. J'ai également bénéficié d'un financement de la Commission des Communautés Européennes qui m'a permis de travailler avec Pierre Boursot et Kevin Dawson au laboratoire Génome, Populations et Interactions dirigé par François Bon-

homme. Mes travaux de recherches ont été également partiellement financés par le programme européen FRAGLAND, coordonné par Ilkka Hanski. La région Languedoc-Roussillon, le ministère des Affaires Étrangères (à travers le programme PICASSO) ont également contribué au financement de l'étude de *M. strigosa*. Qu'ils en soient remerciés.

Merci à Pierre Capy, avec qui j'ai fait mes premiers pas dans la recherche lors d'un stage « volontaire » à Gif-sur-Yvette ; merci également à Denis Couvet, avec qui ces premiers pas se sont concrétisés par un mémoire de DEA et quelques articles (*at last*) ; merci également à Kevin Dawson et Pierre Boursot avec qui j'ai pu aborder de nouvelles questions lors de mon passage au laboratoire GPI ; à Bruno Colas et Miquel Riba, qui m'ont initié aux joies du terrain (je sais que ça amuse tout le monde quand je dis que j'ai fait du terrain) ; à Sylvain Gandon et à François Rousset qui m'ont initié aux joies des modèles de sélection de parentèle (ça rigole moins, tout de suite...) ; merci à Patrick Grillas, qui m'a laissé une totale liberté durant ces trois dernières années : je pense que nous regrettons tous les deux que les projets initialement prévus n'aient pas pu aboutir.

Thomas Bataillon, Kevin Dawson, Hélène Fréville, Carlos Garza, Bernard Godelle, Jérôme Goudet, Olivier Hardy, Gordon Luikart, Sandrine Maurice, Isabelle Olivieri, Nicolas Perrin, Ophélie Ronce et François Rousset ont lu et relu des manuscrits et je les en remercie. En particulier, le document que vous tenez entre les mains a profité des critiques, conseils, et autres commentaires de Thomas Bataillon, Hélène Fréville, Sandrine Maurice (« Catastrôoophhe ! »), Isabelle Olivieri et Joëlle Ronfort qui ont bien voulu me relire (parfois plusieurs fois !) sans jamais faiblir (même si c'était parfois un peu « Kloug », *dixit* Toto Bataillon qui a toujours des citations trèz-à propos).

Ces années que j'ai passées à l'ISEM auraient été très différentes sans les discussions avec François Rousset, Ophélie Ronce, Virginie Ravigné, et bien d'autres encore. Merci François pour ta rigueur, pour tes « là je peux pas travailler comme ça, il faut écrire les équations », merci Ophélie pour ton regard critique et les discussions improvisées, merci enfin Virginie pour (là aussi) les discussions improvisées à la pause café.

Être dans l'équipe d'Isabelle, cela signifie aussi la présence de beaucoup

d'étudiants de DEUG, de licence, de maîtrise, de DEA, de grandes écoles, et j'en passe. Je voudrais remercier Isabelle de m'avoir donné l'occasion d'en encadrer quelques uns, parmi lesquels (dans l'ordre chronologique de leur passage à l'ISEM) Sylvain Glémin, Sylvie Muratorio, Vuokko Vanninen, Vassilis Geogiadis, Ambroise Dalecky et Nadir Alvarez. Je les remercie tous pour ce qu'ils m'ont apporté, et j'espère en retour leur avoir apporté quelque chose, aussi peu que ce soit.

Je remercie également toutes les personnes grâce auxquelles j'ai suivre et échantillonner les populations naturelles de *Marsilea strigosa* : Bruno « Qu'est ce j'ai foutu d'mes clés » Colas, et Miquel « (Ahh Poutain) Ma Lounette » Riba, grâce à qui j'ai découvert *M. strigosa* et traversé l'Espagne à la recherche de petites mares, d'étangs, ou de quoi que ce soit qui retienne un peu d'eau de temps à autre ; Josep Lluis Gradaille Tortella et Magdalena Vicens qui nous ont accompagnés sur les populations de Mallorca ; Laurent Desnouhes, Perrine Gautier, Nicolas Leclainche, de la Tour du Valat, qui m'ont aidé pour le suivi annuel de la population de Roque-Haute ; James Molina et Henri Michaud, qui m'ont fait partager leur découverte de la population des Pyrénées Orientales ; Laïla Rhazi et Patrick Grillas, qui ont échantillonné les dayas marocains ; Hassan Souheil, de l'association de gestion de la réserve, et Bernard De Clock, propriétaire, qui m'ont accordé leur autorisation pour que nous puissions travailler à Roque-Haute.

Une partie des travaux que je présente ici a été réalisée à la paillasse, à la Station de Génétique et Amélioration des Plantes (INRA Montpellier) à Mauguio. Merci à Jean-Marie Prosperi pour m'y avoir accueilli un temps et à Sylvie Roussel pour ses conseils et son aide lors des électrophorèses d'enzymes. A l'ISEM, je voudrais remercier Arnaud Berthomieu, Isabelle Bourrié, Marie-Pierre Dubois, Fabienne Justy et Jacques Lagnel non seulement pour l'aide qu'ils m'ont apportée mais également pour leur présence au côté des étudiants que j'ai encadrés. Merci Faby, pour ta tchatche et tes coups de main ; merci Jaco pour tes conseils au moment d'enrichir la banque (quel métier...). Parce que les données que j'ai analysées n'ont pas toutes été obtenues par moi, je remercie Rama Singh et Richard Morton, qui m'ont gentiment envoyé leurs jeux de données d'électrophorèse d'enzymes de *Drosophila*.

A GPI, un grand merci à François Bonhomme qui m'a accueilli dans son labo ; à Khalid Belkhir, à Pierre Boursot, qui m'a proposé ce contrat européen alors que je venais de quitter Lausanne ; à Kevin Dawson, Nicolas Galtier, Bernard Godelle et Nathalie Raufaste, pour les discussions et leur gentillesse.

Merci à tout l'ISEM et en particulier à Clothilde Bernard, Janice Britton-Davidian, Josette Catalan, Guila Ganem, Joëlle Lopez, Sandra Unal ; à celles et ceux du troisième étage Ariane Boyer, Isabelle Chuine, Danièle Duzer, Jacky Ferrier et Denis-Didier Rousseau ; enfin à tous ceux de l'équipe d'Isabelle dont je n'ai pas encore parlé, Mariane Alleaume-Benharira, Sophie Dubois, Eric Imbert, Lucile Lafuma, Eric Petit, aux anciens Yvain Dubois et Gilles Bena. Merci à tous pour votre gentillesse, et pour faire de cet Institut un endroit où il fait bon faire sa thèse. Merci à Anne-Marie Biglione, Brigitte Daniel, Viviane Delon, Anne-Marie Duffour, Nicole Germain, Philippe Kaczmarek, Jean-Marie Leroux et Evelyne Machetel pour régler nos petits tracas administratifs (et le mot est faible... à la fois pour les tracas, et pour les remerciements...)

L'intégralité du document de synthèse ainsi qu'une bonne partie des annexes ont été réalisés grâce au logiciel de formatage de texte LaTeX 2$_\varepsilon$. J'adresse un grand merci à François Rousset pour sa TeXpertise ! Je lui dois, là aussi, beaucoup de « trucs et astuces », de conseils et de « purisme » (ou de « dégraissage de mammouth », c'est selon...) qui m'ont aidé dans la réalisation de ce document.

Merci enfin à Ophélie et Axel (pour leur amitié, leur couscous, leur « boktof » ... et les cigares !), Agnès, Jean-Christophe et Madeleine, Sandrine et Jean-Luc, Bernard, Isabelle, Joëlle, Christophe et Duane, Isabelle et Bruno, Thomas et Bodil, Cécile et Sylvain, Jeff et tous les autres aussi pour toutes ces petites bouffes, les apéros, les cinoches... la vie quoi !

A mes parents, qui m'ont toujours laissé libre de mes choix, à ma grand-mère à Sumène, qui a fait partager ma soutenance à bon nombre de ses voisins et de ses voisines, à mon frangin Mathias : merci pour tout !

Les mots ne peuvent pas dire grand chose de ce que je te dois Hélène, de ces années passées ensemble. *Kiitos* !

Avant propos

Cette thèse résume l'essentiel des travaux de recherches que j'ai réalisés depuis mon DEA,[1] en 1995. Ces travaux ont été conduits principalement au laboratoire Génétique et Environnement (Institut des Sciences de l'Évolution de Montpellier-UMR 5554, Université Montpellier II), sous l'attention constante d'Isabelle Olivieri, puis en 1998 à l'Institut de Zoologie et d'Écologie Animale (Université de Lausanne, Suisse), en collaboration avec Nicolas Perrin et, depuis 1998, entre le laboratoire Génétique et Environnement et la Tour du Valat (Camargue), en collaboration avec Patrick Grillas. Du point de vue de sa forme, cette thèse est composée de deux parties. La première partie est un document de synthèse en français, qui reprend les principaux arguments théoriques qui ont servi de base à mes travaux. La seconde partie de la thèse est un recueil d'articles publiés, sous presse ou soumis, ou bien encore de manuscrits en cours de préparation. Ces deux parties peuvent être lues indépendamment l'une de l'autre. Cette thèse a été financée, en grande partie, par la Fondation Sansouire (Tour du Valat). A l'origine, je devais travailler en collaboration avec Patrick Grillas sur des aspects théoriques et empiriques de la biologie des populations de *Damasonium polyspermum* Cosson (Alismataceae), une plante annuelle rare que l'on trouve notamment sur le domaine de la Tour du Valat. Malheureusement, deux années consécutives de forte sécheresse printanière et estivale ont mis à mal les suivis démographiques, ainsi que le programme de recherches initialement prévu. J'ai donc repris l'étude de la fougère aquatique *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta) et j'ai entrepris de mener

[1] Diplôme d'Études Approfondies en Évolution et Écologie, intitulé « Information multigénique et structure des populations » et réalisé sous la direction de Denis Couvet.

à bien un ensemble de modèles dont les applications pratiques, si elles ont des liens évidents avec la biologie des petites populations, vont bien au delà des problèmes de conservation des espèces rares. Une partie de ces travaux théoriques a été financée dans le cadre du programme *Development, optimisation and validation of molecular techniques for the measurement of genetic diversity in domestic ungulates* de la Commission des Communautés Européennes[2] et a été réalisée au laboratoire Génome, Population et Interactions (UMR 5000, Université Montpellier II) en collaboration avec Pierre Boursot et Kevin Dawson. Je vous souhaite bonne lecture.

[2]Contrat numéro BIO4-CT96-1189 coordonné par Godfrey M. Hewitt (University of East Anglia, Norwich).

Première Partie

Document de synthèse

Chapitre 1

Introduction

La génétique des populations est une discipline qui est née de la synthèse des théories de Mendel, de Darwin et des biométriciens du début du XX$^{\text{ème}}$ siècle (Provine 1971). Ronald A. Fisher, Sewall Wright et John B. S. Haldane en sont les pères fondateurs. Ils ont, chacun à leur manière, posé les bases conceptuelles et une formalisation mathématique de l'évolution de la variation génétique dans les populations. C'est de leurs convictions quasi philosophiques, mais également de leurs oppositions personnelles (Provine 1971), qu'est né le débat sur l'importance relative des forces évolutives que sont la sélection naturelle, d'une part, et les effets stochastiques liés aux effets d'échantillonnage, d'autre part.

Si l'on mesurait l'importance d'une discipline à la façon dont elle est perçue dans la société, la génétique des populations ne souffrirait probablement pas d'une cote de popularité bien élevée. C'est malheureusement le cas des sciences de l'Évolution en France qui, pour des raisons essentiellement historiques (Monod 1970), n'ont pas beaucoup percé dans le grand public. A l'heure actuelle, les outils et les concepts de cette discipline diffusent dans le grand public à travers, me semble-t-il, deux champs assez distincts. Le premier exemple qui me vient à l'esprit peut sembler assez éloigné de nos préoccupations quotidiennes. Il s'agit de l'utilisation des concepts de la génétique des populations dans le domaine légal (Evett et Weir 1998). Les outils moléculaires qui y sont utilisés, tout comme les méthodes d'analyse,

font de plus en plus partie de l'univers médiatique où nous sont relatés les résultats d'enquêtes, les comptes-rendus d'audience lors de procès publics, etc. Plus proche de nos préoccupations vient la prise de conscience de la nécessité de conserver les espèces en général et les ressources génétiques d'espèces d'intérêt agronomique (ou des espèces qui leur sont apparentées) en particulier. En effet, les notions de diversité génétique, de transferts de gènes, de barrières entre espèces, etc. sont actuellement de mieux en mieux relayées vers le grand public dans le cadre des débats sur les risques associés aux organismes génétiquement modifiés, sur la nécessité de préserver la faune et la flore ou encore sur la constitution de banques de semences d'espèces ou de populations menacées. Bien entendu, la génétique des populations ne limite pas son champ d'investigation à ces deux exemples, tout comme il serait injuste de n'y voir que deux applications pratiques d'une discipline plus large. Car la biologie de la conservation, par les problèmes théoriques qu'elle soulève, participe également de façon importante aux avancées théoriques actuelles de la discipline des sciences de l'évolution.

Aussi, dans cette introduction à mes travaux, je m'appuierai sur les problèmes soulevés par l'étude des espèces rares pour présenter un certain nombre d'outils de mesure de la variation génétique et de sa distribution. Plus précisément, je poserai les principales questions auxquelles le biologiste des populations peut être confronté sur le terrain en retraçant ma propre démarche dans le cadre de l'étude d'une fougère aquatique rare, *Marsilea strigosa*. On retiendra, au travers des exemples qui serviront d'illustration à ces travaux, que les outils que j'ai développés au cours de ma thèse (chapitre 4) peuvent tout aussi bien s'appliquer dans un cadre plus général que celui de la biologie de la conservation.

1.1 Biologie et génétique de la conservation

1.1.1 La biodiversité en crise

Il est une crise majeure dont notre discipline est le témoin depuis quelques années. Cette crise ne touche pas les sciences de l'Évolution, comme on a pu

le lire ça et là dans des journaux « à scandales »[1] ou à travers les écrits de certains auteurs peu au fait des fondements de cette discipline (Denton 1988). Non, la crise à laquelle nous faisons face actuellement concerne l'augmentation du nombre d'espèces qui disparaissent quotidiennement (voir l'annexe A). Car si les extinctions en masse ont ponctué l'histoire de la vie sur Terre, la phase actuelle d'extinction, baptisée la « sixième extinction » par Leakey et Lewin (1995), est caractérisée par une extrême rapidité, sans précédent à l'échelle des temps géologiques (Dobson 1996; Eldredge 1998; Sih *et al.* 2000; Smith *et al.* 1993).

Quelles sont les causes actuelles de l'extinction des espèces ? L'homme, indubitablement, est le responsable des principales menaces qui pèsent aujourd'hui sur nombre d'espèces ou de populations (Channell et Lomolino 2000; Flannery 1999; Miller *et al.* 1999). Dans une étude réalisée sur quelques deux mille espèces en Amérique du Nord, Wilcowe *et al.* (1998) ont montré que les trois principales menaces exogènes qui pèsent sur les espèces sont la destruction et la dégradation de l'habitat, l'introduction d'espèces exotiques ainsi que la pollution. Bien sûr, il existe plusieurs niveaux de destruction des habitats : cela va de la réduction de la superficie totale (qui suit la destruction de fragments entiers d'habitat), à la fragmentation (qui accompagne, par exemple, la construction de routes et d'autoroutes), la détérioration de parcelles d'habitat (conséquence, par exemple, de la pollution de la nappe phréatique) jusqu'à la détérioration des espaces situés entre les habitats favorables (à cause, par exemple, de la mise en place de milieux interdisant le passage des individus migrateurs). L'ensemble de ces actions anthropiques entraîne une modification des connections entre les populations mais aussi une diminution de la taille des habitats favorables et donc des effectifs des populations naturelles. Or la réduction des effectifs constitue une réelle menace pour les espèces, comme nous allons le voir dans le § suivant.

[1] L'usage des guillemets s'impose, puisque parmi ces journaux « à scandales » figure en bonne place le magazine *La Recherche*. En 1996 en effet, une tribune était offerte au mathématicien M. -P. Schützenberger pour exposer son scepticisme vis-à-vis des mécanismes qui sous-tendent l'évolution des espèces (*La Recherche* **283** : 86-90)

Le paradigme des petites populations

Les populations de petite taille subissent des effets d'échantillonnages à plusieurs niveaux. Le premier niveau concerne la démographie, c'est-à-dire l'évolution de la composition des populations en classes d'individus ainsi que de leurs effectifs. Imaginons une population idéale, stable, composée d'individus tous rigoureusement identiques, asexués, chaque individu ne donnant naissance *en espérance* qu'à un seul descendant. La stochasticité démographique définit le fait que chaque individu ne donne en réalité pas naissance à un descendant, mais plutôt à un nombre de descendants, tiré dans une *loi* (ou *distribution*) de probabilités de moyenne égale à un[2]. A l'échelle de la population, il se peut que, par hasard, aucun individu ne se reproduise une année. Car si la probabilité que chaque individu ne donne aucun descendant est faible, elle n'est pour autant pas nulle. En réalité, on comprend bien que plus l'effectif de la population est faible, plus les effets stochastiques sont forts, c'est-à-dire plus la *moyenne* réalisée du nombre de descendants produits par individu, par exemple, est différente de la valeur du *paramètre* qui décrit la loi de probabilités. Les risques d'extinction associés à la stochasticité démographique sont donc d'autant plus forts que les populations sont de petite taille.

Le second niveau de stochasticité concerne l'évolution de la composition génétique des populations. Tout comme la stochasticité démographique, la stochasticité génétique, que l'on caractérise par le terme de *dérive*, décrit un phénomène analogue au précédent, mais à l'échelle des gènes. Dans une population de taille finie, chaque gène ne contribue pas par une copie et une seule à la composition de la génération suivante, mais par un nombre variable de copies qui suit une loi de probabilités de moyenne égale à un[3]. La probabilité que certaines lignées de gènes s'éteignent n'étant pas nulle, la conséquence de ce phénomène est donc une diminution du niveau de variation

[2]On dit alors que le nombre de descendants par individu est une *variable aléatoire*.
[3]La nature de cette loi de probabilités sera donnée au § 3.1.1, p. 62.

génétique, par le simple fait du hasard[4]. Si ce phénomène de dérive agit sur des gènes qui n'affectent ni la survie, ni la fécondité des individus qui les portent, il peut également toucher des gènes dont les effets sur la *valeur sélective* (survie, fécondité) des individus sont légèrement délétères. Dans ce cas, si leur effet est trop faible pour que la sélection naturelle ne les élimine, de tels gènes peuvent être fixés dans la population, par dérive, et diminuer ainsi la valeur sélective moyenne de la population. Il s'agit du *fardeau de mutation* qui s'exprime, sous certaines conditions, par ce que l'on appelle la *dépression de consanguinité*. Ainsi, le nombre d'individus viables et aptes à se reproduire peut diminuer, augmentant ainsi également les risques d'extinction.

Enfin, un troisième niveau est lié à la stochasticité environnementale. Ce processus est souvent divisé en effets dus aux changements cycliques mais imprévisibles du milieu ou du climat (années particulièrement sèches ou humides, intempéries) et en effets dus aux catastrophes naturelles (incendies, cyclones, etc.). On comprend aisément que cette forme de stochasticité a d'autant plus d'impact sur les populations qu'elles sont de petite taille.

Par conséquent, les populations de petite taille souffrent de problèmes liés à leur démographie, d'une part parce que le risque qu'aucun individu ne se reproduise une année n'est pas nul et d'autre part parce qu'il y est plus difficile de trouver un partenaire pour se reproduire : c'est l'*effet Allee* (Allee *et al.* 1950). Ces populations portent également un fardeau de mutations aux effets légèrement délétères ce qui peut entraîner une diminution de la valeur reproductive moyenne. Elles sont en outre d'autant plus sujettes aux changements plus ou moins brutaux de l'environnement que le nombre d'individus est faible. L'importance relative de ces trois niveaux de stochasticité dans les mécanismes d'extinction est mal connu et encore sujet à controverses (Lande 1988).

[4]Chez une espèce diploïde, en outre, chaque descendant ne recevant que la moitié des gènes de chacun de ses parents, il se trouve que, nécessairement, chaque individu ne transmet à son descendant qu'une copie (sur deux) de ses propres gènes.

« Évoluer » vers l'extinction

Il est important de s'intéresser à la biologie des espèces rares (annexe A) car leur étude nous éclaire, de façon plus générale, sur les processus évolutifs qui sont en jeu dans les populations de petite taille et sur les mécanismes impliqués dans l'extinction des espèces. Le potentiel évolutif d'une espèce représente sa capacité à s'adapter à un nouvel environnement, que ce soit parce que l'aire de répartition de cette espèce s'est déplacée, ou bien parce que l'environnement est partiellement détruit, dégradé ou bien changeant. L'évolution de l'aptitude de cette espèce à coloniser de nouveaux sites, de sa tolérance à une pollution environnementale, ou encore de son comportement, sont autant de facteurs qui peuvent éventuellement la sauver de l'extinction. La capacité d'une espèce à s'adapter vite aux changements rapides du milieu est donc un facteur crucial pour sa survie (Gomulkiewicz et Holt 1995). Récemment, Thomas *et al.* (2001) ont montré que l'agrandissement de l'aire de répartition de quatre espèces de papillon, suite au réchauffement du climat, s'accompagnait d'une évolution rapide des traits d'histoire de vie liés à la dispersion (par exemple, les formes à grandes ailes chez *Conocephalus discolor* et *Metrioptera roeselii* que l'on retrouve plus fréquemment dans les populations les plus récemment fondées) ou à la spécialisation écologique (par exemple, *Hesperia comma* a élargi le spectre de son habitat favorable, ce qui permet à cette espèce de coloniser plus de sites, qui lui sont plus facilement accessibles).

Mais s'il existe des cas où les espèces peuvent s'adapter rapidement aux changements du milieu et ainsi diminuer leur risque d'extinction, il arrive également que certaines s'engagent dans une spirale évolutive qui les condamne à l'extinction (voir l'annexe A). Tout ne va pas forcément pour le mieux dans le meilleur des mondes. Il existe ainsi un certain nombre de situations où le changement de structure génétique des populations peut menacer leur persistence (Holsinger *et al.* 1999). Il s'agit par exemple de la perte d'allèles au locus d'auto-incompatibilité. De Mauro (1993) a montré ainsi que la raison pour laquelle la population de l'Illinois de *Hymenoxys acaulis* (Asteraceae) ne produisait plus aucune graine, était qu'il n'existait plus qu'un allèle d'auto-

incompatibilité. Or les croisements ne peuvent être féconds qu'entre plantes portant des allèles différents à ce locus. Dans ce cas, donc, la réduction de la variation génétique a des conséquences directes, immédiates, sur la survie des populations[5]. Il a été montré, également, que la sélection s'exerçant sur les traits d'histoire de vie tels que l'âge à la maturité, la survie adulte ou bien la probabilité de recrutement, pouvait favoriser la mise en place de dynamiques démographiques non-linéaires des populations (Ferrière et Gatto 1993). Or de telles oscillations chaotiques des effectifs représentent un facteur important d'augmentation du risque d'extinction (Belovsky *et al.* 1999).

1.1.2 Quels outils pour quelles questions ?

Comme nous venons de le voir, l'effectif des populations est un paramètre crucial en ceci qu'il détermine l'intensité des forces stochastiques (démographiques et génétiques) qui sont en jeu. De même, les paramètres de dispersion sont tout aussi importants puisqu'ils déterminent également la dynamique des individus et des gènes (liés ou non à l'adaptation locale des individus à leur habitat) à l'échelle de la métapopulation[6]. Les capacités de dispersion des individus entre sites, qui leur permettent ou non de re-coloniser des sites vides, déterminent la susceptibilité des populations au risque d'extinction. Ainsi, la connaissance des effectifs des populations et de l'intensité avec laquelle les individus se déplacent ou bien dispersent leur progéniture entre sites sont autant de pré-requis nécessaires à une bonne compréhension des systèmes biologiques étudiés.

On distingue deux types d'approches pour aborder ces questions, que l'on qualifie d'approches *directes* et d'approches *indirectes*. Les approches directes reposent sur le suivi démographique des populations naturelles, sur l'estimation de paramètres liés au cycle de vie de l'espèce et sur des méthodes matricielles de projection des effectifs sous différents modèles stochastiques. Ces méthodes d'analyse matricielle ont été rendues accessibles grâce, notam-

[5]Toutefois on a également vu des cas où des mutants auto-compatibles ont pu apparaître par mutation et envahir des populations d'espèces auto-incompatibles menacées (Reinartz et Les 1994).

[6]Une métapopulation est définie comme une population de populations (Levins 1968).

ment, à l'ouvrage de Caswell (1989), récemment ré-édité et augmenté dans une nouvelle édition (Caswell 2001). Lorsque les suivis individuels sont possibles (c'est-à-dire lorsque l'on peut marquer les individus, ou bien les équiper de systèmes de radio-émission), les approches directes permettent également d'estimer les *paramètres d'échanges* d'individus entre sites (voir, par exemple, Grosbois 2001). Ces méthodes d'étude de la dispersion, plus adaptées pour les suivis d'espèces animales, peuvent également s'appliquer dans une moindre mesure aux espèces végétales (par exemple, grâce aux techniques de marquage du pollen par des poudres fluorescentes, ou bien par la construction de pièges à graines). Toutefois, les suivis démographiques ne sont pas toujours possibles, par exemple dans le cas de plantes annuelles vivant dans des habitats de mares temporaires.

Les approches indirectes concernent toutes les approches qui reposent sur l'analyse du polymorphisme génétique mesuré sur des marqueurs moléculaires. Comme nous le verrons dans le chapitre 3, puis dans le § 4.1, il est possible de définir et d'estimer les paramètres qui décrivent la dynamique de la perte de polymorphisme génétique dans des populations de taille finie. Ces approches nous renseignent donc, indirectement, sur les risques d'extinction liés à la stochasticité génétique que j'évoquais précédemment. Ces approches indirectes permettent également, à travers l'étude de la distribution spatiale du polymorphisme, d'inférer les mouvements[7] des gènes (voir, par exemple, Beerli et Felsenstein 1999, 2001; Slatkin 1987; Slatkin et Barton 1989). Je donnerai un exemple d'une telle approche dans le § 4.2.

Toutefois, les paramètres estimés par des approches indirectes ne sont pas nécessairement identiques ou équivalents aux paramètres estimés par les approches directes. Par exemple, les flux de gènes que je viens d'évoquer correspondent au mouvement d'individus qui se sont reproduits avec succès hors de leur lieu de naissance, et non pas au mouvement net d'individus entre les différents sites d'un lieu d'étude. D'autre part, l'échelle spatiale à laquelle l'estimation de la dispersion est pertinente dépend des taux de mutation aux locus considérés (Rousset 1997, 2001a,b). Enfin, les marqueurs moléculaires à partir desquels est mesurée la distribution du polymorphisme peuvent être

[7] J'utiliserai dans la suite du document le terme consacré de *flux* de gènes.

soumis à des effets sélectifs. C'est un point que j'aborderai dans les § 3.3 et 4.3. Pour autant, et contrairement à une idée répandue selon laquelle les approches démographiques permettent d'estimer les paramètres *actuels* tandis que les approches indirectes ne permettraient d'estimer que les valeurs *passées* des paramètres (Boileau *et al.* 1992; Koenig *et al.* 1996), il n'est pas certain que les échelles de temps considérées à travers ces deux types d'approches soient très différentes (voir Leblois 2000).

Ainsi, le propos de ma thèse sera de montrer comment des modèles de populations subdivisées peuvent permettre l'estimation de paramètres importants pour la compréhension des systèmes biologiques étudiés, à partir de marqueurs moléculaires. Dans un premier temps, j'illustrerai mon propos par l'étude que j'ai réalisée sur une fougère aquatique rare, *Marsilea strigosa*.

1.2 Un exemple : *Marsilea strigosa*

1.2.1 Biologie d'une espèce rare

Marsilea strigosa Willd. (Pterydophyta, Marsileaceae) est une fougère aquatique rare, inféodée aux habitats de mares temporaires, et que l'on trouve en France, en Espagne, en Algérie, au Maroc, en Egypte, en Italie et en Sardaigne[8] (Tutin *et al.* 1964). Cette espèce est citée dans le Livre Rouge des espèces menacées de France (Olivier *et al.* 1995), dans la convention de Berne pour la conservation des espèces sauvages et des habitats naturels (1979)[9] ainsi que dans la directive européenne *Habitats* (1992)[10]. Les zones humides méditerranéennes ont vu leur taille et leur nombre diminuer de façon importante au cours du XX$^{\text{ème}}$ siècle (Blondel et Aronson 1999). Le climat méditerranéen favorise la mise en eau éphémère de ces habitats. Les

[8] Tutin *et al.* (1964) décrivent également la présence de cette espèce au Nord du delta de la Volga.

[9] Il s'agit du traité *Convention on the Conservation of European Wildlife and Natural Habitats* signé à Berne le 19.IX.1979 (European Treaty Series no. 104).

[10] Il s'agit de la directive *COUNCIL DIRECTIVE 92/43/EEC* du 21.V.1992 et intitulée *conservation of natural habitats and of wild fauna and flora*.

pluies d'automne initient la mise en eau, qui dure en général d'octobre à juin. Puis vient l'assèchement estival, qui laisse le lit des mares, marais ou autres plans d'eau complètement sec. Ces mares inondées de façon temporaire offrent au biologiste l'occasion, assez rare, de pouvoir délimiter facilement les populations des espèces qui s'y trouvent (Holland et Jain 1981). Les espèces végétales que l'on recense dans ces habitats particuliers partagent un certain nombre de traits d'histoire de vie, particulièrement en ce qui concerne la durée du cycle de vie (annualité) et la dispersion dans le temps (dormance des graines) ou dans l'espace (Blondel et Aronson 1999). Les activités anthropiques sont la cause majeure de la dégradation de ces habitats et de leur disparition, entraînant la raréfaction, voire la disparition, d'un nombre important de taxons que l'on trouve dans les mares temporaires (Jain 1994).

Les frondes de *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta) ont l'aspect d'un trèfle à quatre feuilles. Comme tous les membres de la section des Marsileaceae, *M. strigosa* possède des structures reproductrices spécialisées appelées *sporocarpes*, au sein desquelles sont produites les spores. Les sporocarpes résistent à la sécheresse, et peuvent rester viables pendant de très nombreuses années, au moins dans les conditions de conservation d'herbier (Allsopp 1952; Johnson 1985; Soltis et Soltis 1986). La germination a lieu au début du printemps dans les mares inondées, au cours de laquelle les sporocarpes s'ouvrent, libérant ainsi les microspores et les mégaspores qui vont former, après quelques divisions cellulaires, les gamétophytes mâles et femelles, respectivement. La fécondation se produit dans les 48 heures[11]. La croissance des sporophytes (diploïdes) se fait au cours de deux phases successives. La première phase, aquatique, est caractérisée par des feuilles glabres flottant à la surface, la seconde phase, terrestre, est caractérisée par des feuilles poilues, plus petites. Durant la phase terrestre, les plantes ont une croissance clonale, et les sporocarpes (ayant pour origine ontogénétique des feuilles modifiées) sont produits à la base du rhizome, auquel ils sont attachés.

[11]Des observations *in vitro* ont montré qu'une matrice gélatineuse empêchait le déplacement libre des sporanges dans l'eau et que *M. strigosa* était auto-compatible, suggérant un possible régime de reproduction consanguin.

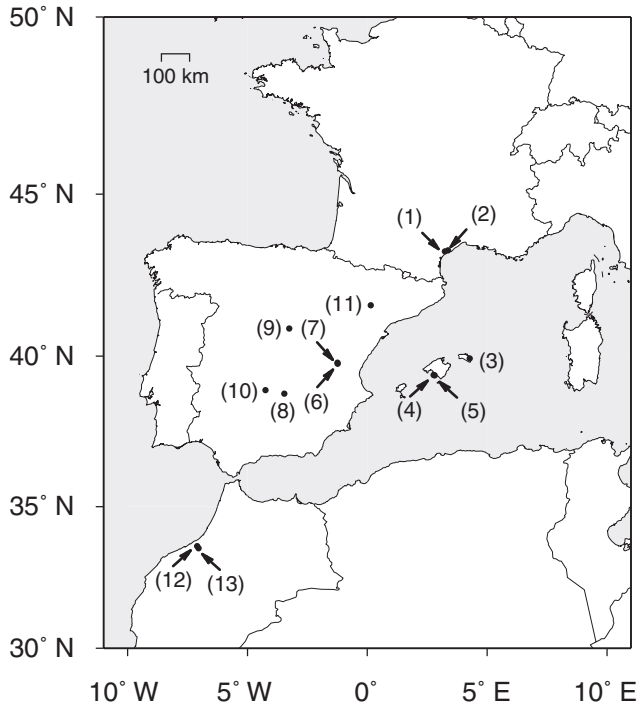


Figure 1.1: Carte représentant les 13 localités où *M. strigosa* a été échantillonnée en France, au Maroc et en Espagne. Les coordonnées géographiques (longitudes et latitudes) font référence au méridien international de Greenwich et au système géographique mondial WGS84.

1.2.2 Un niveau de variation génétique faible

Treize stations ont été échantillonnées en France, en Espagne et au Maroc, entre 1996 et 1999 (figure 1.1). Onze systèmes enzymatiques ont été analysés sur les populations de Roque-Haute et Vendres, qui ont permis d'identifier dix-sept locus présumés. Il n'existe aucune variation génétique à ces dix-sept locus, ni à l'intérieur de ces deux populations, ni entre elles. Sur les onze systèmes enzymatiques, sept ont été analysés pour l'ensemble des autres populations, exceptées les populations du Maroc. A cette échelle également, les douze locus présumés révélés par ces sept systèmes se sont avérés monomorphes au sein des populations ainsi qu'à l'échelle de l'ensemble des populations, et ont révélé la présence des mêmes allèles que ceux observés

dans les populations françaises (voir l'annexe B).

Les espèces rares présentent en général un polymorphisme génétique neutre réduit, bien que cela ne soit pas une règle absolue (Avise 1994; Young *et al.* 1993). De nombreuses espèces végétales endémiques se sont révélées monomorphes pour un nombre comparable de locus enzymatiques. C'est le cas, par exemple, de *Harperocallis flava* Mc Daniel (Liliaceae) (Godt *et al.* 1997), *Malacothamus fasciculatus* (Nutt.) Greene var. *nesioticus* (Rob.) Kearn. (Malvaceae) (Swensen *et al.* 1995), et *Bensoniella oregona* (Saxifragaceae) (Soltis *et al.* 1992). De même, *Pedicularis furbishiae* (Scrophulariaceae) (Waller *et al.* 1987), *Howellia aquaticus* (Campanulaceae) (Lesica *et al.* 1988), et *Trifolium reflexum* (Fabaceae) (Hickey *et al.* 1991), toutes des espèces rares ou menacées, se sont également révélées monomorphes à l'échelle intra- comme inter-populationnelle. D'autres espèces rares ou menacées ont montré un polymorphisme enzymatique très faible. C'est le cas de *Trifolium stoloniferum* (Hickey *et al.* 1991), *Helonias bullata* (Liliaceae) (Godt *et al.* 1995) ainsi que de *Geum radiatum* (Rosaceae), *Carex misera* (Cyperaceae), *Trichophorum cespitosum* (Cyperaceae) et *Calamagrostis cainii* (Poaceae) (Godt *et al.* 1996).

1.2.3 Une très forte structuration de la variation

S'il ne fait pas de doute que ce faible niveau de variation observé est dû aux faibles taux de mutation des marqueurs enzymatiques (de l'ordre de 10^{-6} par locus et par génération, voir Voelker *et al.* 1980), qu'en est-il de la distribution de la variation à d'autres locus, plus polymorphes ? Des marqueurs de type microsatellite[12], dont les taux de mutations sont beaucoup plus élevés (de deux à quatre ordres de grandeur par rapport aux marqueurs enzymatiques : voir, par exemple, Ellegren 2000; Goldstein et Schlötterer 1999; Jarne et Lagoda 1996) ont été mis au point chez *M. strigosa* (voir l'annexe C).

Une analyse de la distribution de la variation mesurée à sept locus mi-

[12]Les marqueurs microsatellites sont de courtes séquences d'ADN répétés que l'on trouve dans la plupart des génomes. Ces séquences sont constituée de répétitions plus ou moins fidèles d'un motif de base de quelques nucléotides (Goldstein et Schlötterer 1999).

crosatellites a montré que la différenciation génétique entre les populations méditerranéennes de *M. strigosa* était très prononcée et hautement significative. Autrement dit, les flux de gènes apparents entre ces populations distantes en moyenne de 576 km sont extrêmement réduits. D'autre part, il a été montré que la différenciation génétique à l'échelle de la réserve naturelle de Roque-Haute,[13] à cinq des sept locus mis au point (les deux autres locus n'étant pas polymorphes à cette échelle spatiale) était également très forte, signe que les flux de gènes sont tout aussi limités à une échelle spatiale beaucoup plus fine (la distance moyenne entre les mares de Roque-Haute étant d'environ 267 m).

Nous avons également montré que la structure génétique multilocus[14] était particulièrement marquée à l'échelle du bassin méditerranéen (voir l'annexe D). De façon tout aussi marquée, la distribution des génotypes multilocus sur les marqueurs microsatellites est extrêmement structurée dans l'espace à l'échelle, beaucoup plus fine, de la réserve naturelle de Roque-Haute (figure 1.2).

Afin de mieux rendre compte encore de la structure multilocus de la variation aux locus microsatellites, nous avons défini une distance génétique entre les génotypes i et j comme

$$D_{ij} = \sum_{l=1}^{L} \sum_{k=1}^{2} (a_{ikl} - a_{jkl})^2 \tag{1.1}$$

où L est le nombre de locus, a_{ikl} est le nombre de répétitions du motif de base de la $k$ième copie du génotype i, au locus l (les copies de gènes sont ordonnées dans l'ordre croissant du nombre de répétitions). Cette distance est donc d'autant plus grande que le nombre de répétitions des allèles microsatellites diffère d'un génotype à l'autre. Sous l'hypothèse que les mutations se tra-

[13] La réserve naturelle de Roque-Haute est située dans la commune de Portiragnes, près de Béziers (Hérault) sur un plateau basaltique d'environ 150 hectares, creusé de quelques 200 mares temporairement inondées.

[14] C'est-à-dire la structure génétique qui prend en compte les interactions statistiques entre les locus (l'alternative étant de considérer la variation estimée à chaque locus comme une mesure indépendante, résultat d'un tirage dans la même distribution). J'aborderai ce point plus en détail dans le § 2.3.

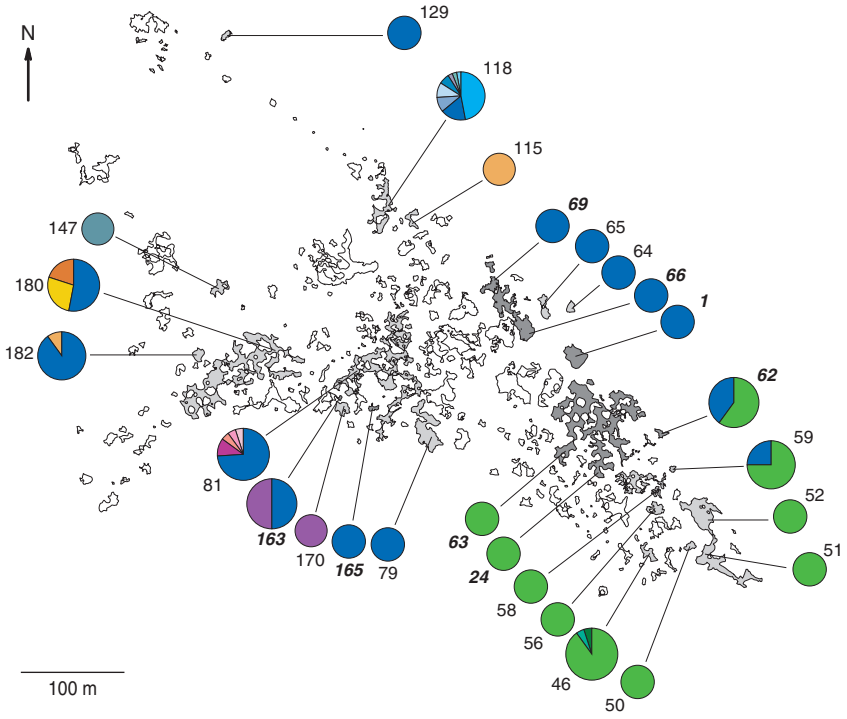


Figure 1.2: Répartition géographique des génotypes multilocus dans la réserve naturelle de Roque-Haute. Les 26 mares échantillonnées sont représentées en gris. La composition génotypique de chaque mare est représentée par un diagramme en camembert (les diagrammes des mares monomorphes sont de plus petite taille que les diagrammes de mares polymorphes). Chaque couleur représente un génotype multilocus. Deux génotypes multilocus sont distincts si un allèle présent à un locus dans l'un des génotype est absent de l'autre génotype à ce même locus. Les numéros des mares sont indiqués sur la figure. Les 26 mares échantillonnées sont représentées en gris.

duisent par l'ajout ou le retrait de quelques motifs de base[15], cette distance mesure la divergence entre les génotypes multilocus. Le réseau minimal[16] que l'on peut construire pour joindre les différents génotypes à partir de la matrice de distance calculée d'après (1.1) est représenté par la figure 1.3. A

[15]Il s'agit des modèles de mutation *par pas*, traduction de l'anglais *stepwise mutation model*.

[16]Il s'agit du *Minimum Spanning Network*, en anglais, construit d'après l'algorithme de Kruskal (1956) et implémenté par Schneider *et al.* (2000).

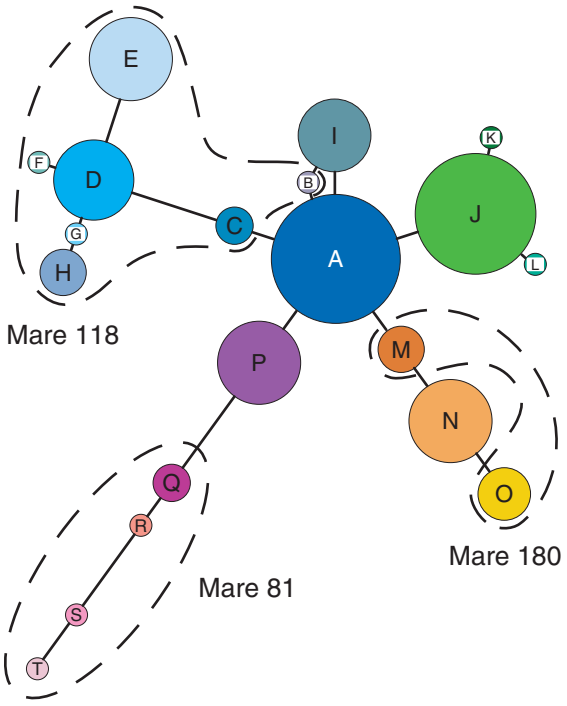


Figure 1.3: Réseau des génotypes microsatellites dans la réserve naturelle de Roque-Haute. La taille relative de chaque nœud (c'est-à-dire chaque génotype multilocus) du réseau est proportionnelle au logarithme de la fréquence du génotype en question sur l'ensemble de l'échantillon. Le réseau est construit à partir de la matrice des distances moyennes basées sur les différences de taille entre allèles (voir l'équation 1.1). La distance séparant deux nœuds est proportionnelle à la distance entre les deux génotypes que représentent ces nœuds. La gradation des couleurs figure la distance génétique entre les génotypes. Les génotypes notés sur fond blanc sont hétérozygotes à un locus au moins. Les lignes pointillées regroupent des génotypes présents dans une même mare polymorphe (à l'exception des génotypes A et J, très fréquents).

l'échelle de la réserve naturelle de Roque-Haute, les génotypes présents dans une même mare sont aussi les génotypes les plus proches d'après la distance (1.1) que nous avons construite (figure 1.3). Ce résultat (que l'on retrouve également à l'échelle du bassin méditerranéen), montre d'une part que les flux de gènes sont extrêmement réduits à toutes les échelles géographiques consi-

dérées dans cette étude et d'autre part que la mutation et la recombinaison[17] pourraient être les seules forces évolutives qui influencent la distribution de la variation génétique chez cette espèce.

1.3 L'œuf et la poule

L'étude de la distribution de la variation génétique chez *M. strigosa* à différentes échelles géographiques nous a permis de montrer que les flux de gènes étaient extrêmement réduits, quelle que soit l'échelle spatiale considérée. D'autre part, le faible niveau de polymorphisme à l'intérieur de chaque mare est la conséquence d'une dérive génétique très importante. Enfin, la particularité des modèles de mutation *par pas* nous a permis de proposer l'hypothèse que la mutation pourrait être la source de variation majeure qui explique la distribution observée de la diversité génétique (annexe D).

Mais le faible niveau de variation génétique, tel qu'observé chez *M. strigosa* par exemple, est-il la cause ou la conséquence de la rareté des espèces ?[18] Holsinger *et al.* (1999) considèrent qu'un faible niveau de variabilité génétique est le symptôme plutôt que la cause de la rareté des espèces. En réalité, cette question restera encore probablement longtemps sans réponse (du moins qui fasse office de généralité).

D'un point de vue plus méthodologique, je dirais également que l'intérêt de travailler sur des espèces rares est de mieux nous aider à comprendre comment la variation résiduelle est distribuée à différents niveaux hiérarchiques, afin de mieux cerner les forces évolutives qui sont en jeu dans les populations de petite taille. La dynamique de l'évolution de certains traits d'histoire de vie dépend notamment de la structure génétique des populations (voir Hamilton et May 1977, pour une des premières études mettant en évidence le rôle de la compétition entre individus apparentés sur l'évolution de la dispersion). Par conséquent, l'estimation de paramètres liés à la dynamique de la

[17] Il est également à noter que la recombinaison efficace pourrait être extrêmement limitée elle aussi, étant donné le fort taux d'autogamie inféré d'après la distribution de la variation génétique à l'intérieur des mares : voir l'annexe D.

[18] Question réminiscente du vieux problème : qui de la poule ou de l'œuf est venu le premier ?

variation génétique est un point important de l'étude du potentiel évolutif, ou potentiel adaptatif, des espèces.

Tout au long de cette thèse, je montrerai tout d'abord comment l'analyse des modèles de populations structurées peut nous permettre de mieux comprendre comment agissent les différentes forces évolutives dans les populations subdivisées et de taille finie. Mais je montrerai également que ces modèles de populations structurées peuvent nous permettre d'estimer des paramètres d'intérêt (comme des paramètres liés à la dynamique de la perte de diversité génétique ou bien à la dispersion des gènes).

La suite de ce document de synthèse est structurée comme suit. Dans le second chapitre je discuterai de la notion d'identité génétique. Je reprendrai les modèles classiques de la littérature et montrerai comment analyser des situations plus complexes, comme la structure des populations en classes d'individus distincts, ou bien la prise en considération d'interactions entre gènes à différents locus. Dans le troisième chapitre, je discuterai de la notion de taille, ou d'effectif efficace. Ceci m'amènera également à discuter de l'effet de la structure (spatiale ou temporelle) des populations sur le phénomène de *dérive* génétique. Enfin, dans le quatrième chapitre, je détaillerai les principales méthodes que j'ai développées au cours de ma thèse, et qui, toutes, s'appuient sur les éléments théoriques que j'aurai abordés dans les chapitres 2 et 3.

Chapitre 2

L'identité génétique

> « On dit que 2 individus d'une population sont *apparentés*[1] s'ils ont un ou plusieurs ancêtres communs. Leur différence génétique doit être moins grande en moyenne que celle de 2 individus quelconques puisque certains de leurs gènes descendent d'un même gène d'un même ancêtre commun et ne peuvent donc, si on néglige les mutations, être différents, alors qu'ils pourraient toujours l'être chez des individus non apparentés » (Malécot 1948).

Dans ce chapitre, je m'attacherai à définir et à discuter des concepts d'identité des gènes qui sont à la source des modèles que j'ai développés au cours de ma thèse. Je discuterai ici de l'extension à deux locus des modèles élaborés pour un locus, ainsi que de la prise en compte de la structure en classes (âge et sexe) dans les modèles à un locus. Dans un souci de précision, je commencerai tout d'abord par définir les termes qui reviendront le plus souvent tout au long de ce manuscrit.

Dans la suite de ce document, on appellera « gène » la copie d'une information génétique. Cette définition a un sens immédiat lorsque l'on s'intéresse à la (faible) partie du patrimoine génétique qui contribue à l'expression du phénotype des individus. La notion de variation (ou de polymorphisme) génétique implique que les différentes copies d'une même information (d'un même gène) ne sont pas nécessairement identiques, mais peuvent au contraire

[1] Malécot (1948) précise : « Nous réserverons le nom de *parents* au père et à la mère »

souffrir quelques différences. Pour certains marqueurs génétiques qui, par hypothèse, ne codent pas d'information génétique, on ne retiendra de cette définition du gène que la notion de copie. Un individu diploïde possède deux copies de la même information génétique. Chez une espèce où l'hérédité est biparentale, l'une de ces deux copies provient du père, tandis que l'autre copie provient de la mère. On appellera « locus » la classe d'homologie d'un gène, en ce sens que seuls deux gènes homologues peuvent *ségréger* (voir, par exemple, Gouyon *et al.* 1997) Cette définition élargit la notion selon laquelle le locus caractérise simplement l'emplacement du gène sur le chromosome. Enfin, un « allèle » (terme auquel on préférera celui d'« état allélique ») représentera une classe de gènes tous équivalents. Selon notre propre perception, donc, deux gènes seront dans le même état allélique, c'est-à-dire appartiendront à la même classe allélique, s'ils confèrent le même phénotype, si l'information qu'ils portent est codée par la même séquence d'ADN, ou bien s'ils sont la copie exacte d'un unique ancêtre commun.

2.1 L'identité génétique à un locus

A première vue, décrire les classes d'équivalence que sont les états alléliques ou bien déterminer l'exacte identité de deux gènes (au sens généalogique) ne semble pas tout à fait équivalent. En outre, seule la description des classes semble envisageable en pratique, tandis que la détermination de la vraie généalogie des gènes parait inaccessible. Il s'agit là de la différence entre l'identité par descendance[2] et l'identité par état.

2.1.1 L'identité par descendance et l'identité par état

On dira que deux gènes sont identiques par descendance[3] s'ils sont deux copies exactes de leur ancêtre commun le plus récent. Cette définition implique que deux gènes ne sont identiques que si aucune mutation n'est survenue le long de la généalogie qui les unit à leur ancêtre commun le plus récent.

[2]On parle aussi d'identité par *ascendance*.
[3]Terme traduit de l'anglais, *identical by descent, IBD*.

Imaginons par exemple une population constituée de N individus diploïdes. On suppose que ces individus produisent chacun un nombre infini de gamètes, dont une proportion μ a subit une mutation au locus que l'on considère. Les descendants de ces individus sont le fruit du croisement au hasard des gamètes dans la population. Chaque descendant a donc la même probabilité $1/N$ de descendre de n'importe lequel des N parents présents. Le nombre L de descendants qui survivent par parent suit donc une distribution binomiale de paramètres N (le nombre de tirages réalisés) et $1/N$ (la probabilité de tirer un parent en particulier), soit $L \sim \mathcal{B}\left(N, \frac{1}{N}\right)$. La probabilité que deux individus partagent un même parent est également $1/N$. D'autre part, deux individus qui partagent un même parent diploïde ont une probabilité égale à $1/2$ de recevoir chacun une copie du même gène parental. Par conséquent, la probabilité d'identité Q à $t+1$ de deux gènes pris au hasard dans la population s'exprime en fonction de l'identité de deux gènes à la génération précédente, comme

$$Q(t+1) = \gamma \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) Q(t) \right] \tag{2.1}$$

où $\gamma = (1-\mu)^2$ est la probabilité qu'aucun des deux gènes n'ait muté entre t et $t+1$. A l'équilibre,

$$Q = \frac{\gamma}{2N(1-\gamma)+\gamma} \tag{2.2}$$

On dira que deux gènes sont identiques par état[4] s'ils appartiennent à la même classe allélique. Bien sûr, cette appartenance à une classe donnée dépendra de notre perception : prenons le cas d'un locus microsatellite[5], par exemple. On dira que deux copies de ce marqueur sont identiques si les fragments d'ADN amplifié qui les constituent migrent à la même distance sur un gel d'électrophorèse. Ceci garantit que ces fragments d'ADN possèdent le même nombre de paires de bases, mais pas nécessairement la même séquence. C'est un premier niveau de ce que l'on appelle l'*homoplasie*, c'est-à-dire le

[4]Terme traduit de l'anglais, *identical in state, IIS*.
[5]Voir la note 12 au bas de la page 14.

fait que deux gènes sont identiques par état, mais pas par descendance. Il est toutefois à noter que deux copies de gènes rigoureusement identiques sur le plan de leur séquence peuvent ne pas être identiques par descendance, si par exemple une mutation qui survient après une autre rétablit l'état initial. L'homoplasie provient donc pour une part de notre perception des objets que l'on manipule (isoformes enzymatiques, produits d'amplification de l'ADN, séquences d'ADN, etc.) mais surtout de la nature des mutations et de leur effet sur le génotype. Si l'on considère que les mutations ne créent pas nécessairement de nouveaux allèles dans la population mais, au contraire, peuvent simplement générer des allèles déjà existants, alors seule l'identité par état est mesurable en pratique. En revanche, des modèles théoriques ont été développés pour décrire l'identité entre gènes comme une identité par descendance. Dans ces modèles (modèle à nombre d'allèles infini[6]), chaque mutation crée un allèle différent de tous les allèles présents avant la mutation. Ce modèle de mutation est implicite dans le raisonnement qui nous a conduit aux équations (2.1) et (2.2). Nous verrons par la suite que ces modèles sont utiles pour étudier les propriétés de certaines fonctions de probabilités d'identité, comme les F-statistiques.

2.1.2 L'identité génétique et les temps de coalescence

Il existe une relation étroite entre les probabilités d'identité de paires de gènes et les temps de *coalescence* (Slatkin 1991; Slatkin et Voelm 1991). La théorie de la coalescence, dont le lien avec les modèles de génétique des populations a été établie par Kingman (1982a,b), s'intéresse au calcul des probabilités de généalogies des gènes. On dira que deux gènes coalescent à l'instant t dans le passé s'ils ont un ancêtre commun à cet instant t. Pour bien comprendre cette définition, il faut se représenter le temps « à l'envers », c'est-à-dire allant du présent vers le passé, et se représenter graphiquement la coalescence de deux gènes comme un événement de « fusion » de ces deux gènes. Si l'on connaît la généalogie des gènes, c'est-à-dire toutes leurs relations de parenté, on dira donc que deux gènes coalescent lorsqu'ils se réunissent en

[6]Terme traduit de l'anglais, *infinite allele model, IAM.*

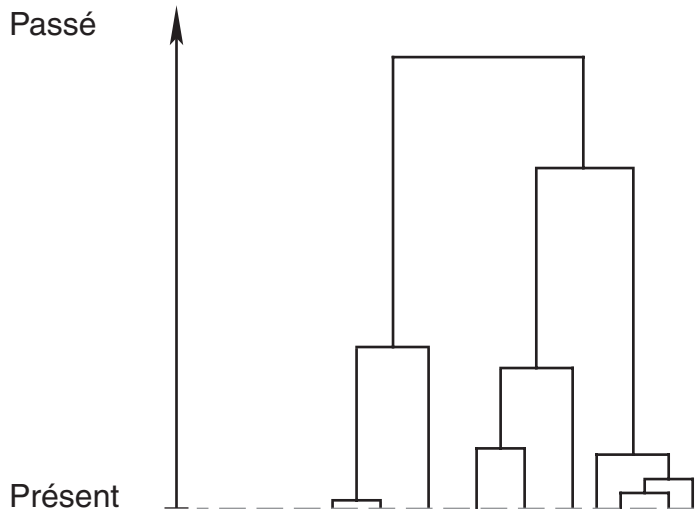


Figure 2.1: Un exemple de généalogie de 10 gènes. Cette figure représente une réalisation du processus de coalescence dans une population constituée de $N = 100$ individus haploïdes. Les temps de coalescence (donnés par l'âge de chaque nœud de la topologie) sont distribués de façon exponentielle.

un seul (voir la figure 2.1). L'intérêt d'étudier cette catégorie de modèles est que, sous l'hypothèse que les différents variants sont sélectivement neutres, les processus de mutation peuvent être découplés des processus généalogiques (Hudson 1990; Norborg 2001). Ainsi, la distribution des états alléliques de toutes les copies de gènes présentes dans une population est déterminée d'une part par la généalogie de ces gènes et d'autre part par la distribution des mutations qui se sont produites (Norborg 2001). D'une certaine manière, cela signifie que l'on sépare en deux processus distincts ce qui contribue à l'identité par descendance de ce qui contribue à l'identité par état. D'autre part, on peut calculer séparément les temps de coalescence de gènes pris dans différentes classes (Slatkin 1991). Dans ce qui suit, on ne s'intéressera qu'à un cas particulier de la théorie de la coalescence, où seules sont considérées des *paires* de gènes.

Considérons maintenant le modèle à nombre d'îles fini dont un traitement

exact a été donné par LATTER (1973)[7]. On raisonnera de la même manière que précédemment,[8] en considérant simplement qu'il existe un nombre fini n de populations de N individus diploïdes et que la probabilité pour un gène d'être immigrant est m. Les gènes sont dispersés indépendamment les uns des autres et ont une égale probabilité de venir de l'un des $(n-1)$ dèmes. Les gènes sont échantillonnés après la dispersion. En reprenant les notations de NAGYLAKI (1983) et de CROW et AOKI (1984) on notera la proportion des paires de gènes issues d'une sous-population unique à la génération précédente $a = (1-m)^2+m^2/(n-1)$ pour les gènes d'une même sous-population et $b = (1-a)/(n-1)$ pour les gènes de sous-populations distinctes[9]. Dans chaque sous-population, une proportion s des individus philopatriques[10] est produite par autofécondation. On note Q_0, la probabilité que les gènes d'un individu soient identiques par descendance, Q_1 la probabilité que deux gènes pris dans deux individus distincts dans une même sous-population soient identiques, et Q_2, la probabilité que deux gènes pris dans deux sous-populations distinctes soient identiques. On obtient les équations de récurrence suivantes

$$\begin{aligned} Q_0(t+1) &= \gamma\left[a\left(s\tfrac{Q_0(t)+1}{2}+(1-s)\,Q_1(t)\right)+(1-a)Q_2(t)\right] \\ Q_1(t+1) &= \gamma\left[a\left(\tfrac{1}{N}\tfrac{Q_0(t)+1}{2}+\left(1-\tfrac{1}{N}\right)Q_1(t)\right)+(1-a)Q_2(t)\right] \\ Q_2(t+1) &= \gamma\left[b\left(\tfrac{1}{N}\tfrac{Q_0(t)+1}{2}+\left(1-\tfrac{1}{N}\right)Q_1(t)\right)+(1-b)Q_2(t)\right] \end{aligned} \quad (2.3)$$

Ces équations peuvent s'écrire sous une forme matricielle. On note $\mathbf{Q} = (Q_0, Q_1, Q_2)^T$ le vecteur des probabilités d'identité de paires de gènes, où

[7]Le modèle à nombre d'îles infini ou « modèle en îles » a été développé par WRIGHT en 1951. Ce modèle considère une population subdivisées en un nombre infini de sous-populations de taille finie, échangeant des gamètes entre elles.

[8]Voir les équations (2.1) et (2.2), p. 23.

[9]Par définition, b est la proportion des paires de gènes pris dans des sous-populations distinctes et qui sont issues d'une sous-population unique à la génération précédente. Or avec la probabilité $2m(1-m)/(n-1)$ un seul gène a migré et provient de la même population que l'autre gène à la génération précédente. De même, avec la probabilité $(n-2)\left[m/(n-1)\right]^2$ les deux gènes ont migré et proviennent de la même sous-population à la génération précédente. Par conséquent, $b = 2m(1-m)/(n-1)+(n-2)\left[m/(n-1)\right]^2 = (1-a)/(n-1)$.

[10]Par définition, s est la probabilité conditionnelle que deux gènes soient issus d'un même parent, sachant que ces deux gènes proviennent d'une même sous-population avant dispersion.

l'exposant T indique qu'il s'agit du vecteur transposé (et donc que $\mathbf{Q}$ est un vecteur colonne). On peut ré-écrire les équations de récurrence (2.3) sous la forme

$$\mathbf{Q}(t+1) = \gamma\left[\mathbf{A}\mathbf{Q}(t) + \mathbf{D}\right] \tag{2.4}$$

où $\mathbf{A}$ est une matrice de transition et $\mathbf{D}$ un vecteur de tous les termes constants dans les récurrences.

$$\mathbf{A} = \begin{pmatrix} a\frac{s}{2} & a(1-s) & (1-a) \\ a\frac{1}{2N} & a\left(1-\frac{1}{N}\right) & (1-a) \\ b\frac{1}{2N} & b\left(1-\frac{1}{N}\right) & (1-b) \end{pmatrix} \quad \text{et} \quad \mathbf{D} = \begin{pmatrix} a\frac{s}{2} \\ a\frac{1}{2N} \\ b\frac{1}{2N} \end{pmatrix} \tag{2.5}$$

A l'équilibre, on obtient d'après l'équation (2.4) :

$$\mathbf{Q} = \gamma\left(\mathbf{I} - \gamma\mathbf{A}\right)^{-1}\mathbf{D} \tag{2.6}$$

où $\mathbf{I}$ est la matrice identité.

Si l'on revient à la définition de l'identité par descendance, on peut exprimer autrement la probabilité d'identité Q_h de deux gènes, avec $h = 0$ lorsque les deux gènes sont tirés dans un même individu, $h = 1$ lorsque les deux gènes sont tirés dans deux individus différents dans un même dème et $h = 2$ lorsque les deux gènes sont tirés au hasard dans des dèmes différents. On note $C_h(t)$ la probabilité que deux gènes aient leur ancêtre commun le plus proche t générations dans le passé. On dit également que $C_h(t)$ est la probabilité que ces gènes *coalescent* à t. L'identité par descendance Q_h a été définie comme la probabilité que deux gènes soient des copies exactes (sans mutation) de leur ancêtre commun le plus récent (Hudson 1990). Et puisque la probabilité que ces deux gènes n'aient pas muté pendant les t générations écoulées est γ^t, alors

$$Q_h = \sum_{t=1}^{\infty} \gamma^t C_h(t) \tag{2.7}$$

(Hudson 1990; Malécot 1975; Rousset 1996; Slatkin 1991; Slatkin

et VOELM 1991). L'équation (2.7) reflète bien la probabilité nette que deux gènes aient coalescé avant que l'un ou l'autre n'ait muté. Or l'équation (2.6) peut également s'écrire sous la forme

$$\mathbf{Q} = \sum_{t=1}^{\infty} \gamma^t \mathbf{A}^{t-1} \mathbf{D} \tag{2.8}$$

Par conséquent,

$$\mathbf{C}(t) = \mathbf{A}^{t-1} \mathbf{D} \tag{2.9}$$

donne le vecteur des probabilités de coalescence à t, en reprenant les mêmes indices que pour les probabilités d'identité, c'est-à-dire $\mathbf{C} = (C_0, C_1, C_2)^T$. Si la matrice $\mathbf{A}$ est diagonalisable, on peut l'exprimer sous la forme $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$ (voir, e.g. HORN et JOHNSON 1985, page 46), où $\mathbf{S}$ est la matrice composée des vecteurs propres à droite de $\mathbf{A}$, de telle façon que le vecteur propre à droite associé à la $k$ième valeur propre est le $k$ième vecteur-colonne de $\mathbf{S}$. $\mathbf{S}^{-1}$ est la matrice composée des vecteurs propres à gauche de $\mathbf{A}$, de telle sorte que le vecteur propre à gauche associé à la $k$ième valeur propre est le $k$ième vecteur-ligne de $\mathbf{S}$. $\mathbf{\Lambda}$ est la matrice diagonale qui contient les valeurs propres ordonnées $(\lambda_1, \ldots, \lambda_k)$ de $\mathbf{A}$ sur la diagonale principale. Par conséquent, l'équation (2.9) devient

$$\mathbf{C}(t) = \mathbf{S}\mathbf{\Lambda}^{t-1}\mathbf{S}^{-1}\mathbf{D} \tag{2.10}$$

où $\mathbf{\Lambda}^{t-1} = \text{diag}\left(\lambda_1^{t-1}, \ldots, \lambda_k^{t-1}\right)$ est facilement calculable numériquement et dépend des paramètres N, m, n et s du modèle. On peut donc facilement calculer la distribution des probabilités de coalescence à partir de l'équation (2.10). La figure 2.2 donne un exemple de la distribution des probabilités de coalescence pour différentes paires de gènes. Il est à noter que le nombre de générations en abscisse correspond au nombre de générations *écoulées* avant le temps présent (qui se trouve à l'origine de l'axe des abscisses). La distribution des probabilités de coalescence dans une classe de gènes [par exemple $C_1(t)$] est proportionnelle à la distribution des probabilités de coalescence dans une

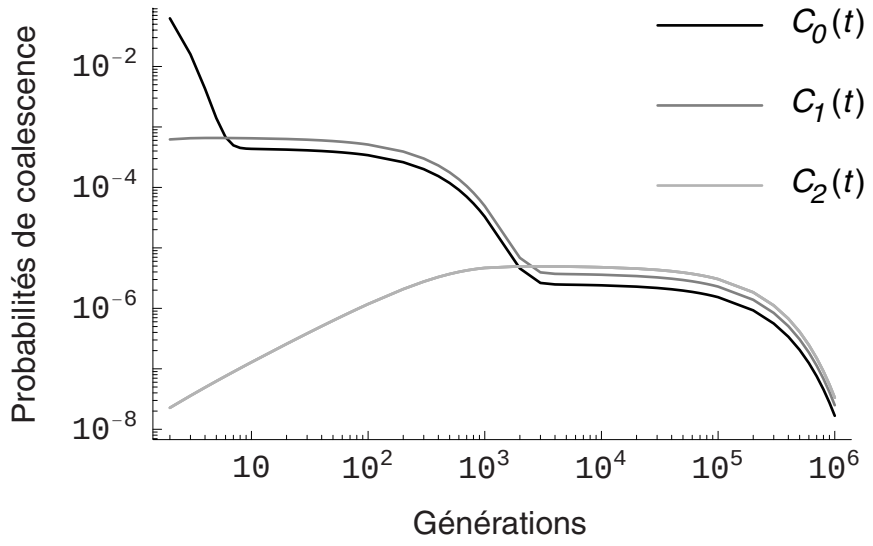


Figure 2.2: Distribution des probabilités de coalescence. Les probabilités $C_h(t)$ que deux gènes coalescent au temps t sont données pour différentes paires de gènes. On remarquera que la distribution des probabilités de coalescence à l'intérieur d'une classe devient proportionnelle à la distribution des probabilités de coalescence entre classes, pour des temps anciens. $N = 1000$, $m = 0.001$, $n = 100$ et $s = 0.5$. Noter la double échelle logarithmique (d'après ROUSSET 1996).

classe de gènes moins apparentés [par exemple $C_2(t)$] pour des temps anciens. ROUSSET (2001b) montre que la différence asymptotique entre les probabilités de coalescence est de l'ordre de F_{ST}. En revanche, dans la période des temps très récents les deux distributions ne sont pas proportionnelles l'une à l'autre et les probabilités que deux gènes coalescent quelques générations dans le passé sont très différentes, selon que ces gènes sont échantillonnés dans des individus distincts de la même sous-population ou de sous-populations différentes.

2.1.3 Les F-statistiques

Définitions

Nous avons vu précédemment que pour décrire un modèle de population structurée en sous-populations, il fallait définir des probabilités d'identités de paires de gènes pris à différents niveaux hiérarchiques. Ceci est vrai pour n'importe quel type de structure (voir, par exemple, RONFORT *et al.* 1998; ROUSSET 1999b; SLATKIN et VOELM 1991). Il est en effet toujours possible de définir des classes de gènes de telle sorte que le modèle est entièrement décrit par les probabilités d'identité de paires de gènes tirés à l'intérieur de ces classes et par les probabilités d'identité de paires de gènes tirés dans des classes distinctes. Dans l'exemple précédent, c'est-à-dire dans le modèle défini par le système d'équations (2.3), j'ai distingué les paires de gènes prises à l'intérieur des individus, des paires de gènes prises dans une même sous-population et des paires de gènes prises dans des sous-populations distinctes. Ces trois catégories de paires de gènes définissent donc les différentes classes dans le modèle précédent.

Dans ce contexte, on peut définir des corrélations des états alléliques pour les gènes pris à l'intérieur d'une classe par rapport aux gènes pris dans des classes distinctes, c'est-à-dire des corrélations génétiques intra-classes (COCKERHAM et WEIR 1987; ROUSSET 1996). Si les probabilités d'identité considérées sont des probabilités d'identité par état, ces corrélations sont mieux connues sous le nom de F-statistiques (WRIGHT 1951). Ainsi, les fameux paramètres F_{IS}, F_{ST} et F_{IT} introduits dans la littérature par WRIGHT (1951) sont-ils définis par

$$\begin{aligned} F_{IS} &= \frac{Q_0 - Q_1}{1 - Q_1} \\ F_{ST} &= \frac{Q_1 - Q_2}{1 - Q_2} \\ F_{IT} &= \frac{Q_0 - Q_2}{1 - Q_2} \end{aligned} \tag{2.11}$$

(voir, par exemple, COCKERHAM et WEIR 1987, 1993; ROUSSET 1996). De

façon plus « générique », on note

$$F_{IJ} = (Q_I - Q_J)/(1 - Q_J) \tag{2.12}$$

la corrélation des gènes dans la classe I (identiques par état avec la probabilité Q_I), relativement aux gènes de la classe J (identiques par état avec la probabilité Q_J).

Depuis Wright (1943, 1951) la définition, et donc l'estimation, des F-statistiques ont fait l'objet d'un vaste débat dans la littérature (Chakraborty et Danker-Hopfe 1991; Excoffier 2001; Rousset 2001b; Weir et Cockerham 1984). Wright (1951) a défini les F-statistiques comme des corrélations génétiques[11]. Les définitions de Wright (1951) ne sont valables que pour des locus bi-alléliques (c'est à dire des locus dont les gènes ne peuvent être que dans deux états alléliques). Nei (1973, 1977, 1986) et Nei et Chesser (1983) ont étendu ces définitions au cas, plus général, où l'on trouve plus de deux allèles à un locus. Le problème inhérent aux définitions de Nei (1973, 1977) tient essentiellement au fait qu'il ne considère pas explicitement un modèle d'évolution des fréquences alléliques (Nagylaki 1998). Ceci a pour conséquence que si l'on considère que les fréquences alléliques varient aléatoirement dans une population sous l'effet de forces évolutives telles que la mutation, la dérive génétique ou bien la migration, les indices de Nei (1973, 1977) sont des *variables aléatoires*[12]. Du point de vue des inférences que l'on peut faire à partir de ces indices, ou bien de l'étude de modèles théoriques de différenciation des populations, les définitions de Nei (1973, 1977) souffrent donc de leur dépendance aux fréquences des allèles et des génotypes dans la population considérée (voir, par exemple, Nagylaki 1989, 1998).

Une toute autre approche a été développée par Cockerham (1969, 1973). Elle repose sur l'analyse d'un modèle linéaire dans lequel la fréquence d'un gène est déterminée par son espérance dans la population totale ainsi que par des effets aléatoires (d'espérances nulles) inhérents aux processus de muta-

[11] « The correlation between gametes that unite to form an individual relative to the gametes of the total population will be represented by F_{IT} [...] » (Wright 1969, p. 294).

[12] C'est-à-dire des quantités qui prennent différentes valeurs selon une loi de probabilité définie par les paramètres du modèle.

tion, de dérive et de dispersion. Ces effets aléatoires se décomposent en composantes intra-individuelles, inter-individuelles au sein des sous-populations, et inter-individuelles entre sous-populations distinctes (Cockerham 1969, 1973). Dans ce cadre théorique (celui de l'ANOVA), les F-statistiques de Wright (1943, 1951) s'expriment en fonction de composantes de la variance, et sont donc bien définis comme des *paramètres* d'un modèle statistique[13].

S'il permet d'exprimer les F-statistiques dans le cadre d'un modèle où les fréquences alléliques observées (des *variables aléatoires*) sont des réalisations d'un processus stochastique dont on cherche à estimer les *paramètres*, ce formalisme souffre toutefois d'un problème de construction. Comme le reconnaissait en effet déjà Cockerham (1969), les « composantes de la variance » ainsi définies peuvent prendre des valeurs négatives (voir Chakraborty et Danker-Hopfe 1991; Excoffier 2001; Rousset 2001b). Cockerham et Weir (1987) ont tenté de surmonter cette difficulté en ajoutant un niveau hiérarchique supplémentaire à leur analyse (Cockerham et Weir 1987, 1993). Récemment, Rousset (2001b) a montré que la décomposition de la variance totale dans le modèle linéaire considéré par Cockerham (1969, 1973) n'était pas une somme de variances classique, si bien que les « composantes de la variance » développées dans Cockerham (1969, 1973), Weir et Cockerham (1984) et Cockerham et Weir (1987, 1993) sont en réalité des fonctions de covariances des fréquences génétiques (voir également Excoffier 2001). Cette confusion a pour origine l'hypothèse faite *a priori* par Cockerham (1969, 1973) que les effets aléatoires intra-individuels, inter-individuels au sein des sous-populations et inter-individuels entre sous-populations distinctes ne sont pas corrélés entre sous-populations, entre individus au sein des sous-populations ou entre gènes au sein des individus.

Le point important de ces développements est qu'une décomposition de la variance totale du modèle linéaire considéré par Cockerham (1969, 1973) conduit naturellement à l'expression des F-statistiques en termes de probabilités d'identités par état, c'est-à-dire dans les termes de l'équation (2.12)

[13] Long (1986) a proposé un développement multivarié de ce modèle pour le cas multiallélique.

(Ronfort *et al.* 1998; Rousset 2001b).

Estimation

Écrire les F-statistiques en termes de probabilités d'identité par état, c'est-à-dire dans les termes de l'équation (2.12) permet également de proposer une statistique de la forme

$$\widehat{F}_{IJ} = (\widehat{Q}_I - \widehat{Q}_J)/(1 - \widehat{Q}_J) \tag{2.13}$$

pour estimer le paramètre F_{IJ}. Rousset (2001b) montre que les estimateurs de la forme (2.13) sont exactement identiques à ceux de Weir et Cockerham (1984) (voir également Weir 1996).

La construction d'intervalles de confiance à partir des estimateurs de moment des F-statistiques se fait le plus souvent grâce à des techniques de ré-échantillonnage. Weir (1996) recommande l'utilisation du *bootstrap*[14] sur les locus. L'hypothèse sous-jacente est que chaque locus apporte une information indépendante de tous les autres ce qui, en toute rigueur, implique l'absence de liaison génétique et un régime de reproduction panmictique.

Au cours de ce travail (voir, par exemple, les annexes G et H), j'ai plus particulièrement utilisé la méthode analytique ABC[15] de DiCiccio et Efron (1996). Cette méthode est une approximation analytique d'un algorithme[16] qui permet de calculer un intervalle de confiance à partir d'une distribution *bootstrap*. L'avantage de la technique *ABC* est que l'on calcule analytiquement, par une approximation, les bornes de l'intervalle de confiance. Par conséquent, aucun ré-échantillonnage n'est nécessaire, ce qui permet de réduire de façon importante le temps de calcul.

[14] La technique du *bootstrap* consiste à réaliser beaucoup de nouveaux échantillons à partir du jeu de données de départ, en réalisant, pour chaque nouvel échantillon, un tirage avec remise dans le jeu de données initial.

[15] Terme traduit de l'anglais, *approximate bootstrap confidence interval*.

[16] Appelé, en anglais, *bias-corrected and accelerated bootstrap*, BC_a.

Valeurs à l'équilibre

On peut calculer les valeurs à l'équilibre des F-statistiques définies en (2.11) dans le modèle en îles fini (§ 2.1.2). A partir des solutions du système d'équations (2.6), on obtient

$$F_{IS} = \frac{\sigma a \gamma}{2 - \sigma a \gamma} \tag{2.14}$$

et

$$F_{ST} = \frac{\gamma d}{\gamma d + N(2 - \sigma a \gamma)(1 - \gamma d)} \tag{2.15}$$

où $\sigma = s - 1/N$, $d = a - b$ et, comme précédemment, $\gamma = (1 - \mu)^2$. (voir, par exemple, Rousset 1996). Il faut bien garder à l'esprit que le modèle décrit dans le § 2.1.2 donne les probabilités d'identité par descendance (modèle à nombre infini d'allèles). Toutefois, les valeurs à l'équilibre des F-statistiques ne dépendent pas du modèle de mutation, tant que les taux de mutation sont faibles (Rousset 1996). Pour des taux de mutation faibles, l'équation (2.15) se réduit à

$$F_{ST} \approx \frac{1}{1 + 4Nm\left[n/(n-1)\right]^2} \tag{2.16}$$

(Crow et Aoki 1984; Slatkin 1991).

Les F-statistiques ont fait et font encore l'objet d'une littérature très abondante. Il existe deux raisons à cela. La première est que certains paramètres « démographiques » des populations (comme par exemple leur taille, le nombre d'immigrants moyen par génération) peuvent être exprimés en fonction de F_{ST} (voir, par exemple, Slatkin 1987; Waples 1989). D'après l'équation (2.16) en effet, et si l'on considère que le nombre de sous-populations est grand, on peut écrire $Nm = \left(1/F_{ST} - 1\right)/4$. L'estimation du « nombre efficace d'immigrants par génération » Nm a été largement utilisée pour décrire la structure des populations naturelles, bien que les conditions d'application de cette formule (liées aux hypothèses peu réalistes du modèle à nombre d'îles infini) ne soient en général pas remplies (Whitlock et Mc-

Cauley 1999). La seconde raison est que les F-statistiques apparaissent également dans des modèles d'adaptation des populations structurées, comme la théorie des paysages adaptatifs[17] de Wright (1931, 1977) (voir aussi Hartl et Clark 1989, p. 323-326). Enfin, les F-statistiques apparaissent dans les modèles de sélection de parentèle en populations subdivisés (voir, par exemple, Gandon et Rousset 1999; Rousset et Billiard 2000).

F-statistiques et temps de coalescence

D'après la définition des probabilités d'identité donnée par l'équation (2.7) et la définition des F-statistiques (équation 2.11) on peut exprimer les F_{IJ} en fonction des temps moyens de coalescence pour différentes paires de gènes

$$\lim_{\mu \to 0} F_{IJ} = \frac{T_J - T_I}{T_J} \tag{2.17}$$

(Slatkin 1991; Slatkin et Voelm 1991) où $T_I = \sum_{t=1}^{\infty} t \cdot C_I(t)$ est le temps moyen de coalescence de paires de gènes dans la classe I[18]. On peut donc comprendre les propriétés des F-statistiques à travers l'étude des probabilités de coalescence de paires de gènes échantillonnées dans différentes classes. On se souvient que la figure 2.2 montrait que la distribution des probabilités de coalescence dans une classe de gènes [par exemple $C_1(t)$] était proportionnelle à la distribution des probabilités de coalescence dans une classe de gènes moins apparentés [par exemple $C_2(t)$] pour des temps anciens. En revanche, dans la période des temps très récents les deux distributions n'étaient pas proportionnelles l'une à l'autre. Cela signifie que la surface couverte par la distribution $C_0(t)$ peut être décomposée en la somme de la surface couverte par la distribution $C_1(t)$ et d'une surface initiale (représentée par la région gris-clair dans la figure 2.3). Cette surface initiale représente une « masse de probabilité », que Rousset (1996, 2001b) a montré être équivalente à F_{IS}. De

[17] Cette théorie est connue an anglais sous le nom de *shifting balance theory*.

[18] On note que l'équation (2.7) implique que $Q_I = 1 - 2\mu T_I + O(\mu^2)$. Par conséquent,

$$\lim_{\mu \to 0} \frac{Q_I - Q_J}{1 - Q_J} = \lim_{\mu \to 0} \frac{1 - 2\mu T_I - (1 - 2\mu T_J)}{1 - (1 - 2\mu T_J)} = \frac{T_J - T_I}{T_J}$$

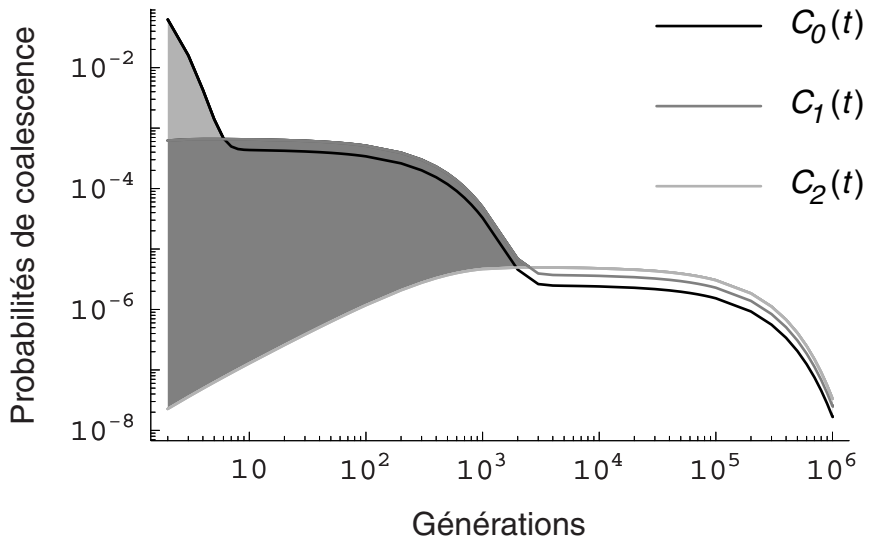


Figure 2.3: F-statistiques et distribution des probabilités de coalescence. Les surfaces grisées représentent les « masses de probabilités initiales » définies dans le texte. La surface gris-claire représente une masse de probabilité dont la valeur est F_{IS}. De la même manière, la surface gris-foncé représente une masse de probabilité dont la valeur est F_{ST}. Voir la légende de la figure 2.2 ROUSSET (d'après 1996)

.

la même manière, F_{ST} est équivalent à la masse de probabilité représentée par la région gris-foncé entre les distributions $C_1(t)$ et $C_2(t)$ (figure 2.3). D'après cette figure, on voit bien que F_{IS} ne dépend que des événements récents de coalescence (de l'ordre de dix générations, pour les valeurs des paramètres de la figure 2.3). Ce paramètre est donc très peu influencé par la mutation (ROUSSET 1996). F_{ST}, quant à lui, dépend des différences des probabilités de coalescence $C_1(t)$ et $C_2(t)$ sur une période de temps plus longue, et donc pour des temps plus anciens. Ce paramètre sera donc d'autant plus sensible à la mutation, que cette période de temps sera longue (c'est-à-dire pour de faibles taux de migration dans un modèle en îles, par exemple) (ROUSSET 2001b).

J'ai présenté ici un ensemble de définitions, d'hypothèses et de résultats qui ont servi de base à mon travail (voir les annexes E, F, G, H, I, J). Pour-

tant, si ces travaux ont pour point de départ les résultats présentés précédemment, je me suis exercé à étendre ces modèles classiques à des situations plus complexes en ajoutant tout d'abord à la structure spatiale un niveau de structuration des individus en classes (sexe, âge), en analysant ensuite la structure génétique des populations à plus d'un locus.

2.2 Structure en classes d'individus

2.2.1 Le cas des espèces animales à sexes séparés

F-statistiques

Je développerai ici une extension du modèle en îles fini (voir, par exemple, Latter 1973; Nagylaki 1983; Rousset 1996), adapté au cas des espèces dioïques. Par souci de simplicité, l'analogie sera faite avec des espèces animales, à sexes séparés (mâles et femelles). Je reprendrai ici les mêmes notations que celles utilisées précédemment (§ 2.1.2). En revanche, je distinguerai les probabilités notées Q_h (après dispersion) des probabilités notées Q_h^* (avant dispersion). La notation Q_h^{XY} sera utilisée pour définir la probabilité d'identité de deux gènes, l'un échantillonné dans un individu de sexe X, le second échantillonné dans un individu de sexe Y, et ce quels que soient $X \in \{♂, ♀\}$ et $Y \in \{♂, ♀\}$. De cette manière, deux individus, l'un de sexe X, l'autre de sexe Y peuvent être tous les deux des mâles, tous les deux des femelles ou bien l'un peut être un mâle et l'autre une femelle. Le modèle que je considère reprend les principales hypothèses du modèle à nombre fini d'îles (voir § 2.1.2), à l'exception des suivantes :

(i) Ce sont les individus qui migrent, plutôt que leurs gamètes (on notera m_X la probabilité qu'un *individu* de sexe X soit immigrant). Par conséquent, après la dispersion des individus, la fréquence des paires d'individus qui proviennent d'une même sous-population à la génération précédente est $a_{XY} = (1 - m_X)(1 - m_Y) + m_X \cdot m_Y/(n - 1)$ pour des individus tirés au hasard dans une seule sous-population et $b_{XY} = (1 - a_{XY})/(n - 1)$ pour des individus tirés dans des sous-

populations distinctes.

(ii) Il y a autant de mâles que de femelles dans chaque sous-population, et les $N/2$ mâles d'une sous-population ont chacun la même probabilité de se croiser avec leurs $N/2$ congénères de l'autre sexe. Le nombre de descendants est très grand (infini) et le nombre L_{XY} de jeunes de sexe X d'un parent de sexe Y qui survivent suit une loi binomiale de paramètres $\frac{N}{2}$ et $\frac{2}{N}$, soit $L_{XY} \sim \mathcal{B}\left(\frac{N}{2}, \frac{2}{N}\right)$ (Nagylaki 1995).

Après la dispersion, étant donnée la dispersion diploïde, la probabilité d'identité des deux gènes homologues d'un individu est égale à la probabilité d'identité de ces gènes avant la dispersion, c'est-à-dire

$$Q_0(t+1) = Q_0^*(t+1) \tag{2.18}$$

De même, les probabilités d'identité de paires de gènes tirées au hasard entre individus distincts après dispersion sont égales aux probabilités d'identités de paires de gènes avant dispersion, pondérées par les probabilités a_{XY} et b_{XY} que les paires de gènes proviennent d'une sous-population unique avant dispersion, soit

$$Q_1^{XY}(t+1) = a_{XY} Q_1^*(t+1) + (1 - a_{XY})\, Q_2^*(t+1) \tag{2.19}$$

et

$$Q_2^{XY}(t+1) = b_{XY} Q_1^*(t+1) + (1 - b_{XY})\, Q_2^*(t+1) \tag{2.20}$$

Avant la dispersion, en revanche, la probabilité d'identité des deux gènes homologues d'un individu est égale, à la mutation près, à la probabilité d'identité des gènes de ses parents, soit

$$Q_0^*(t+1) = \gamma\, Q_1^{♂♀}(t) \tag{2.21}$$

Les probabilités d'identité de paires de gènes entre individus distincts d'une même sous-population dépendent de leur origine paternelle ou maternelle. Dans un cas sur deux, les deux gènes sont des copies issues d'individus de

même sexe. Il peut d'ailleurs s'agir d'un seul individu (avec la probabilité $2/N$) ou de deux individus de même sexe (avec la probabilité $(1-2/N)$). Dans le premier cas, les deux copies sont identiques avec la probabilité $\gamma\left[Q_0(t)+1\right]/2$ tandis que dans le second cas, ils sont identiques avec la probabilité $\gamma\, Q_1^{\male\male}(t)$ si elles proviennent de gamètes mâles (en proportion 1/2), ou $\gamma\, Q_1^{\female\female}(t)$ si elles proviennent de gamètes femelles (en proportion 1/2). Dans un cas sur deux, également, l'un des gènes échantillonnés provient du père de l'individu qui le porte, tandis que l'autre gène vient de la mère de l'individu qui le porte. Ces deux gènes ne peuvent donc être la copie d'un seul gène à la génération précédente, et sont donc identiques avec la probabilité $\gamma\, Q_1^{\male\female}(t)$. On suit le même raisonnement pour décrire l'évolution en une génération des probabilités d'identité de paires de gènes entre individus appartenant à des sous-populations distinctes, soit

$$Q_1^*(t+1) = \gamma\left[\frac{Q_0(t)+1}{2N} + \left(1-\frac{2}{N}\right)\frac{Q_1^{\male\male}(t)+Q_1^{\female\female}(t)}{4} + \frac{Q_1^{\male\female}(t)}{2}\right] \quad (2.22)$$

et

$$Q_2^*(t+1) = \gamma\left[\frac{Q_2^{\male\male}(t)+2Q_2^{\male\female}(t)+Q_2^{\female\female}(t)}{4}\right] \quad (2.23)$$

En réunissant les équations (2.18) à (2.23), on peut établir les équations de récurrence pour les probabilités d'identité avant dispersion, d'une part, et après dispersion, d'autre part. En suivant les définitions de Cockerham et Weir (1987) et Rousset (1996), j'ai défini les F-statistiques conditionnelles au sexe des individus échantillonnés comme des rapports de fonctions des probabilités d'identité conditionnelles au sexe (voir également, Wang 1997a, 1999)

$$\begin{aligned} F_{IS}^{XY} &= \frac{Q_0 - Q_1^{XY}}{1-Q_1^{XY}} \\ F_{ST}^{XY} &= \frac{Q_1^{XY} - Q_2^{XY}}{1-Q_2^{XY}} \\ F_{IT}^{XY} &= \frac{Q_0 - Q_2^{XY}}{1-Q_2^{XY}} \end{aligned} \quad (2.24)$$

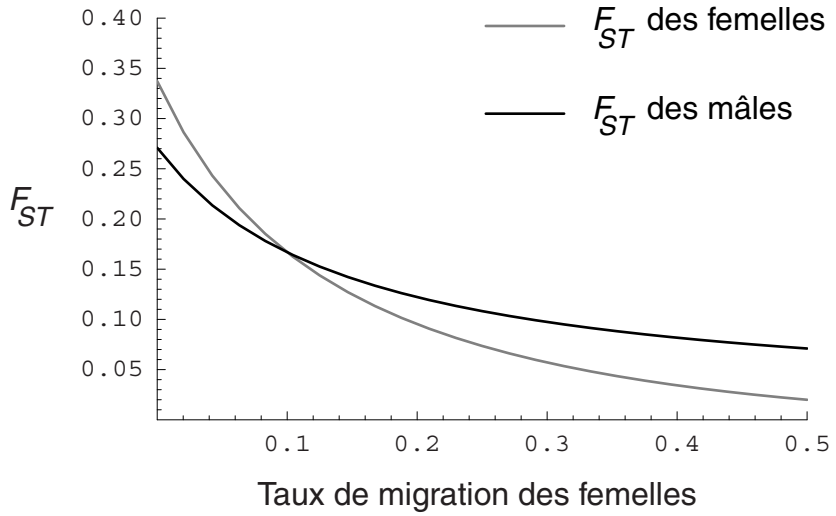


Figure 2.4: Valeurs à l'équilibre des F_{ST} par sexe. Les valeurs des $F_{ST}^{♂♂}$ et $F_{ST}^{♀♀}$ sont données en fonction du taux de migration des femelles, pour un taux de migration mâle fixé. $2N = 20$, $n = 20$, $Nm_{♂} = 1$ et $\mu = 10^{-6}$.

pour tout $(X, Y) \in \{♂, ♀\}^2$. A l'équilibre,

$$F_{ST}^{XY} = \frac{\gamma d_{XY}}{2N(1-\gamma\overline{d}) + \gamma\left(a_{♂♂} + a_{♀♀} - b_{XY}\right) - \gamma^2\left(a_{♂♀} - \overline{b}\right)} \tag{2.25}$$

où $\overline{a} = \left(a_{♂♂} + 2a_{♂♀} + a_{♀♀}\right)/4$, $\overline{b} = \left(b_{♂♂} + 2b_{♂♀} + b_{♀♀}\right)/4$ et $\overline{d} = \overline{a} - \overline{b}$. Si la dispersion n'est pas biaisée en faveur de l'un ou l'autre sexe, c'est-à-dire si $m_{♂} = m_{♀} = m$, l'équation (2.25) se simplifie en

$$F_{ST} = \frac{\gamma d}{2N\left(1-\gamma d\right) + \gamma\left[b + (2-\gamma)d\right]} \tag{2.26}$$

c'est-à-dire une version légèrement différente[19] de l'équation (2.15) (voir aussi Rousset 1996). La figure 2.4 montre les valeurs d'équilibre des F_{ST} mâles et femelles. Le taux de migration des mâles ($m_{♂}$) est fixé et égal à 0.1 et les valeurs d'équilibre des F_{ST} mâles et femelles sont données en fonction du taux de migration des femelles. On voit que les deux courbes se croisent au

[19]La différence tient au fait que dans l'équation (2.15) la migration est haploïde.

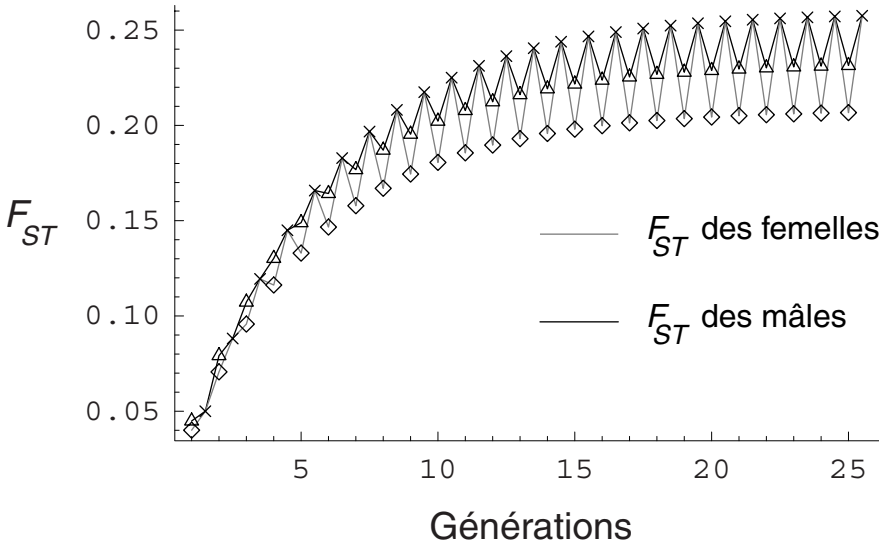


Figure 2.5: Variation des valeurs de F_{ST} par sexe au cours du temps. Les valeurs de F_{ST} mesurées chez les mâles avant ($\times$) ou après ($\triangle$) dispersion sont indiquées en noir. Les valeurs de F_{ST} mesurées chez les femelles avant ($\times$) ou après ($\diamond$) dispersion sont indiquées en gris. $2N = 20$, $n = 20$, $Nm_{\male} = 0.5$, $Nm_{\female} = 1$ et $\mu = 10^{-6}$.

point où les taux de migration sont égaux ($m_{\female} = m_{\male} = 0.1$). Si les femelles dispersent moins que les mâles, $F_{ST}^{\female\female} > F_{ST}^{\male\male}$ et si les femelles dispersent plus que les mâles, $F_{ST}^{\female\female} < F_{ST}^{\male\male}$. Il est important de noter ici que que cette différence entre les valeurs de F_{ST} mesurées chez les mâles et chez les femelles dépend du moment où l'échantillonnage est réalisé. Si l'on échantillonne les gènes immédiatement après la reproduction (c'est-à-dire avant la dispersion juvénile), alors on n'observe aucune différence[20] (Chesser 1991a,b; Wang 1997a). Avant dispersion, on peut montrer (équations 2.18-2.23) qu'à l'équilibre,

$$F_{ST}^{*} = \frac{\gamma}{2N(1-\gamma\overline{d}) + \gamma\left(a_{\male\male} + a_{\female\female}\right) - \gamma^2\left(a_{\male\female} - \overline{b}\right)} \tag{2.27}$$

[20]Il me faut préciser que cela dépend également de la distribution du nombre de descendants par parent : voir par exemple la figure 5 de Wang (1997a).

La figure 2.5 montre les variations des valeurs de F_{ST} par sexe au cours du temps, pour des gènes échantillonnés avant et après la dispersion des individus. Les femelles dispersent plus que les mâles et donc $F_{ST}^{\female\female} < F_{ST}^{\male\male}$. En outre, la différence entre les valeurs de F_{ST} avant et après la dispersion est toujours plus grande chez les femelles que chez les mâles.

On notera également que la valeur de F_{ST}, lorsque l'on échantillonne les individus indépendamment de leur sexe est donnée par

$$F_{ST} = \frac{\gamma\overline{d}}{2N(1-\gamma\overline{d}) + \gamma\left(a_{\male\male} + a_{\female\female} - \overline{b}\right) - \gamma^2\left(a_{\male\female} - \overline{b}\right)} \tag{2.28}$$

En négligeant les termes du second ordre des taux de migration, on retrouve l'expression « classique » du F_{ST} dans le modèle à nombre d'îles fini (équation 2.16), avec un taux de migration égal à la moyenne des taux de migration des mâles et des femelles (voir également Berg *et al.* 1998).

Temps de coalescence

De la même manière que précédemment (§ 2.1.2), on peut examiner les distributions des probabilités de coalescence (Rousset 1996, 2001b) pour des paires de gènes échantillonnées à différents niveaux hiérarchiques, mais également à l'intérieur et entre les classes définies ici (c'est-à-dire les mâles et les femelles). On note $\mathbf{C} = \left(C_0, C_1^{\male\male}, C_1^{\female\female}, C_1^{\male\female}, C_2^{\male\male}, C_2^{\female\female}, C_2^{\male\female}\right)^T$ le vecteur des probabilités de coalescence, par analogie aux probabilités d'identité Q_h^{XY}. La figure 2.6 montre la distribution des probabilités de coalescence, obtenues à partir des équations (2.9) et (2.10) pour une matrice $\mathbf{A}$ et un vecteur $\mathbf{D}$ définis à partir des équations (2.18-2.23). Sur cette figure, les femelles dispersent à un taux cent fois supérieur à celui des mâles. On observe que les distributions des probabilités de coalescence de paires de gènes échantillonnées à l'intérieur des sous-populations sont proportionnelles quel que soit le sexe des individus échantillonnés. En revanche, les distributions (conditionnelles au sexe des individus) des probabilités de coalescence de paires de gènes échantillonnées entre des sous-populations distinctes ne sont proportionnelles que pour des temps anciens ($t > 50$ générations). Pour des temps

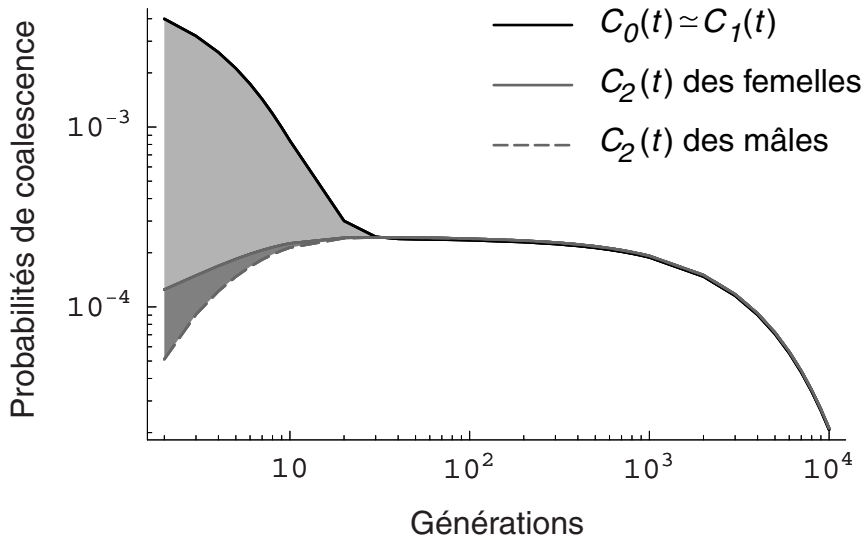


Figure 2.6: Distribution des probabilités de coalescence par sexe. $2N = 100$, $n = 20$, et $m_{♀} = 0.2 = 100 \times m_{♂}$. La surface entre $C_1(t)$ and $C_2^{♀♀}(t)$ (F_{ST} femelle) est en gris-clair. La surface entre $C_1(t)$ et $C_2^{♂♂}(t)$ (F_{ST} mâle) est représentée par la réunion des surfaces gris-clair et gris-foncé.

plus récents, les probabilités de coalescence de paires de gènes échantillonnées dans deux sous-populations distinctes chez les femelles sont supérieures à ces mêmes probabilités chez les mâles. Ceci parce que la probabilité que deux gènes échantillonnés dans deux sous-populations distinctes proviennent d'une seule sous-population à la génération précédente (une condition nécessaire pour que deux gènes coalescent) est plus grande si ces gènes sont échantillonnés dans des individus femelles. Par conséquent, le temps moyen de coalescence de paires de gènes entre sous-populations chez les femelles $\left[T_2^{♀♀} = \sum_{t=1}^{\infty} t \cdot C_2^{♀♀}(t)\right]$ est inférieur au temps moyen de coalescence chez les mâles $\left[T_2^{♂♂} = \sum_{t=1}^{\infty} t \cdot C_2^{♂♂}(t)\right]$.

En reprenant les représentations graphiques de Rousset (1996, 2001b) on note que la surface définie par la différence entre les distributions $C_1^{XX}(t)$ et $C_2^{XX}(t)$ est plus petite pour les femelles ($X = ♀$) que pour les mâles ($X = ♂$) (voir la figure 2.6). Or dans la limite des faibles taux de mutation,

les densités de probabilités définies par ces surfaces ont pour valeurs $F_{ST}^{\varphi\varphi}$ et $F_{ST}^{\sigma\sigma}$, respectivement. Lorsque la dispersion des individus est conditionnelle à leur sexe, les différences observées entre les valeurs des F_{ST} conditionnels sont donc la conséquence d'une plus grande probabilité de coalescence des paires de gènes échantillonnées dans des sous-populations distinctes, parmi les individus qui dispersent le plus.

2.2.2 Le cas des plantes annuelles avec banques de graines

Nous avons vu jusqu'à présent les conséquences d'une structuration « spatiale » des populations[21]. Or les populations naturelles peuvent être structurées à la fois dans le temps et dans l'espace. C'est par exemple le cas lorsque la survie des individus adultes dépasse la durée d'une seule année. On parle alors de générations *chevauchantes*, ce qui traduit le fait que des individus d'âges différents peuvent se croiser entre eux. Le modèle en nombre fini d'îles (voir, par exemple, Latter 1973; Nagylaki 1983; Rousset 1996) suppose, lui, un cycle de vie annuel. Dans la suite de cette section, je vais donc développer les bases d'un modèle de populations subdivisées, structurées en âges. Précisément, je m'attacherai à décrire le cas d'espèces annuelles de plantes dont les graines peuvent ne pas germer l'année de leur production. On dit de telles graines qu'elles sont *dormantes*. Ce phénomène existe également chez les insectes sous le nom de *diapause*. Le modèle que je considère reprend les principales hypothèses du modèle à nombre fini d'îles (voir § 2.1.2), à l'exception des suivantes :

(i) Chaque plante en fleur produit un grand nombre f de graines à chaque génération.

(ii) Les flux de gènes sont assurés par les grains de pollen (fonction de dispersion mâle, haploïde) et par les graines (dispersion diploïde).

(iii) Une fraction z des graines est dispersée, parmi lesquelles une fraction c ne survit pas. Parmi les graines restées dans leur sous-population

[21] Mais la structure que nous avons considérée n'est pas spatialement explicite (voir, par exemple, Rousset 1997; Slatkin 1993, pour des modèles d'isolement par la distance).

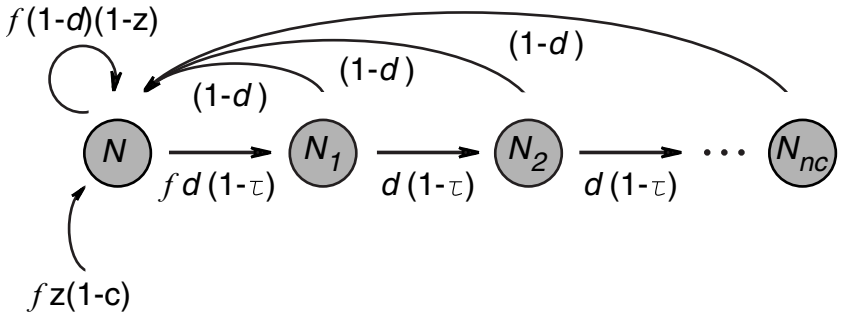


Figure 2.7: Cycle de vie d'une plante annuelle et de sa banque de graines. Chaque cercle représente un stade (adulte ou graine) et les flèches déterminent les probabilités de transition d'un stade à l'autre.

d'origine (les graines philopatriques), une fraction d entre en dormance et survivent avec la probabilité $(1-\tau)$ (on dit aussi que les graines dormantes *paient* un coût τ). Aucune graine dispersée n'entre en dormance. Chaque génération, une fraction $(1-d)$ des graines dormantes germe (la probabilité de germination est la même, quel que soit l'âge de la graine). Toutes les graines d'âge n_c qui ne germent pas meurent.

(iv) La régulation par la densité ne s'exerce que sur les germinations. Cela signifie que le nombre de graines dans la banque est infini si la fécondité est infinie. Après régulation, le nombre de germinations qui survivent pour devenir des adultes et fleurir est N (il s'agit de la taille, supposée constante, de la population en « surface »).

Le lecteur intéressé par plus de détails sur ce modèle pourra se référer à l'annexe J. Le cycle de vie peut être résumé par le graphe de la figure 2.7, où une seule sous-population est représentée. On fait l'hypothèse qu'il n'y a pas de stochasticité démographique ni d'événements catastrophiques, si bien que les populations ne s'éteignent pas et que la structure en âges à l'équilibre est la même pour toutes les sous-populations. Étant données les hypothèses du modèle, le nombre de graines d'âge i nous est donné par

$$N_i = fN(1-z)\left[(1-\tau)d\right]^i \tag{2.29}$$

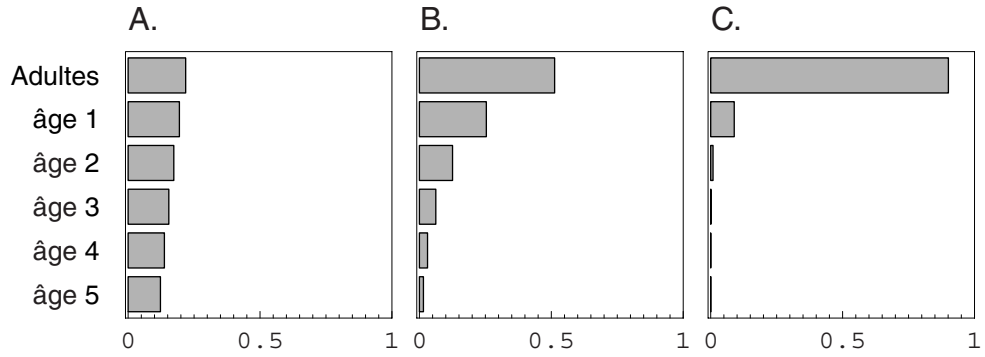


Figure 2.8: Distribution de l'âge des germinations, pour différents taux de germination. A. $g = 0.1$. B. $g = 0.5$. C. $g = 0.9$. Chaque barre horizontale donne la proportion de graines qui germent qui sont issues de parents présents la génération précédente (« Adultes »), ou bien de parents présents plusieurs générations auparavant (« âge i »).

et la proportion de graines d'âge i parmi les graines qui germent est $p_i = N_i / \left(\sum_{j=1}^{n_c} N_j \right)$. L'effet du taux de germination $g = (1 - d)$ sur la distribution en âge stable des graines qui germent chaque année est illustré par la figure 2.8. Plus le taux de germination est faible, plus le nombre de classes d'âge qui contribuent au recrutement est important. Quand le taux de germination tend vers 1, le modèle converge vers un modèle de plantes annuelles avec dispersion de graines et de pollen dans un modèle à nombre d'îles fini (voir Maruyama et Tachida 1992).

L'effet Wahlund temporel

L'influence du nombre d'individus en fleurs et du taux de germination sur la valeur attendue à l'équilibre du paramètre F_{IS} est illustrée par la figure 2.9. Il faut noter que ce paramètre est mesuré parmi les seuls individus de surface et non pas parmi l'ensemble des individus (graines et plantes) présents dans la population. Dans une situation où les gamètes se croisent au hasard dans une population (panmixie), la valeur attendue de F_{IS} est nulle (équation 2.14 avec $\sigma = 0$). Pourtant la figure 2.9A montre que pour des faibles taux de germination et des faibles effectifs de plantes en fleur, F_{IS} est posi-

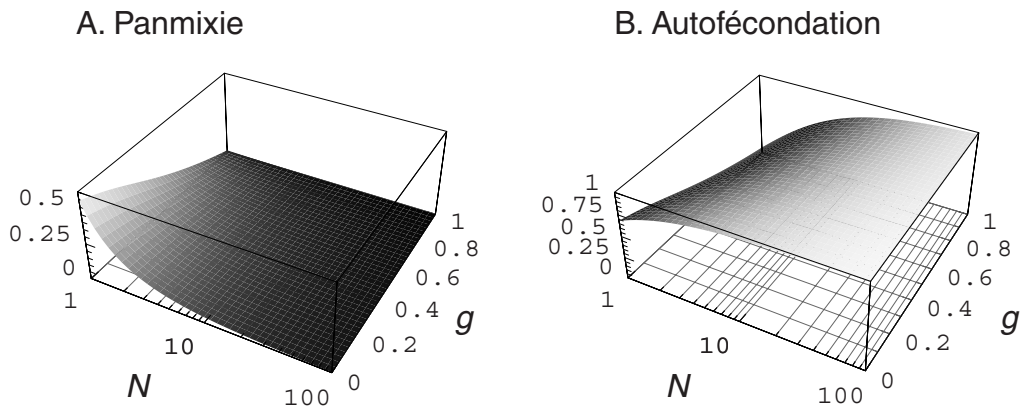


Figure 2.9: Influence de la taille de la population et du taux de germination sur F_{IS}. A. Le cas d'un régime de reproduction panmictique. B. Le cas d'un régime de reproduction par autofécondation stricte. $n_c = 10$, $\tau = 0.1$ et $\mu = 10^{-4}$.

tif. Ceci indique un déficit en hétérozygotes parmi les individus de surface, par rapport aux proportions attendues à l'équilibre de Hardy-Weinberg. Ces résultats montrent que, dans le cas où le nombre d'individus en surface est faible, l'action de la dérive suffit à faire fluctuer les fréquences alléliques d'une génération à l'autre et donc à créer un effet Wahlund temporel, par analogie à l'effet Wahlund spatial[22] (Wahlund 1928). D'ailleurs, l'effet Wahlund est d'autant plus important que le taux de germination est faible, c'est-à-dire lorsque le nombre de classes d'âge qui participent au recrutement augmente (voir figure 2.8). Dans le cas de l'autofécondation complète (figure 2.9B), deux forces s'opposent. Le faible effectif de la population de surface tend à diminuer la valeur de F_{IS} par rapport à celle attendue sous un régime d'autofécondation totale dans une population de taille infinie (voir § 2.1.3, équation 2.14), car la part d'autofécondation due à la panmixie ($1/N$) aug-

[22]L'effet Wahlund caractérise la diminution de la proportion des génotypes homozygotes qui suit le mélange de deux sous-populations panmictiques isolées (voir, par exemple, Crow et Kimura 1970, pp. 54-55). Le corollaire de cette définition est que l'on s'attend à détecter un déficit en génotypes hétérozygotes (par rapport aux proportions attendues à l'équilibre de Hardy-Weindberg) dans un échantillon unique de plusieurs populations en mélange.

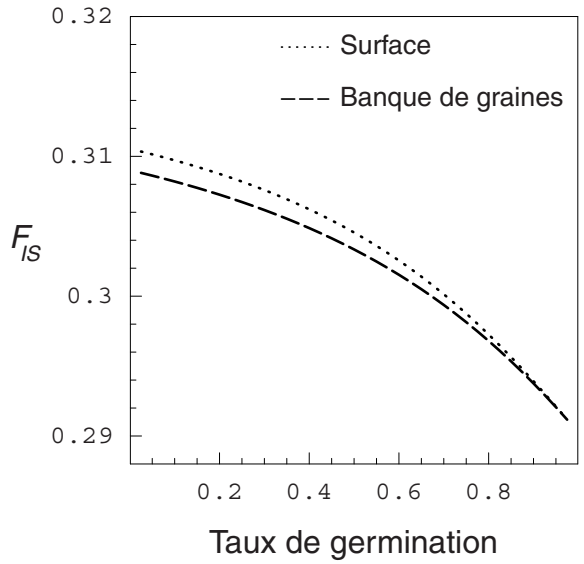


Figure 2.10: Valeurs de F_{IS} à l'équilibre, parmi les plantes en fleurs et parmi les graines de la banque. Il y a 2 classes d'âge dans la banque de graines, $N = 20$, $s = 0.5$, $\tau = 0.1$ et $\mu = 10^{-4}$.

mente lorsque N diminue. Mais ce faible effectif favorise également un effet Wahlund temporel qui, lui, tend à augmenter la valeur de F_{IS}. Pourtant, il faut bien reconnaître que cet effet n'est sensible que lorsque l'effectif de la population de surface est très faible.

La plupart des études visant à comparer la structure génétique de plantes à différents stades de leur cycle de vie (adultes, juvéniles, graines de la banque) montrent que la valeur du paramètre F_{IS} diminue entre les stades « graine » et « adulte » (Alvarez-Buylla *et al.* 1996; Cabin *et al.* 1998; Del Castillo 1994; McCue et Holtsford 1998; Tonsor *et al.* 1993). Nos résultats ne permettent pas d'expliquer ces observations. En effet, le modèle que j'ai présenté ne permet pas d'expliquer de différences de valeur des F-statistiques mesurées sur des adultes et sur des graines de la banque (figure 2.10). Une raison intuitive que l'on peut avancer est que les individus en surface sont le fruit d'un échantillonnage régulier des graines de la banque. Or si l'on comprend bien comment ce modèle permet de définir un effet Wahlund temporel pour des effectifs très faibles, on ne peut guère avancer

d'hypothèse qui nous permettrait d'expliquer des différences entre la composition génotypique de la banque de graines et celle de la surface, alimentée chaque génération par cette dernière. Il est possible toutefois que des effets sélectifs (comme l'expression de la dépression de consanguinité, par exemple) qui agiraient à des stades différents du cycle de vie puissent expliquer ces différences observées.

Nous avons étudié deux exemples de structure (autre que spatiale) des individus dans une population. Dans le premier exemple (§ 2.2.1), des classes d'individus (ici qualifiées par leur sexe) adoptaient des comportements différents de dispersion. Nous avons montré que ces différences de comportement se reflétaient à travers les probabilités de coalescence conditionnelles au sexe des individus échantillonnés dans des sous-populations distinctes. Dans le second exemple (§ 2.2.2), les individus participaient plus ou moins tardivement à la reproduction. Nous avons montré que cette structure en âges favorisait l'établissement d'une structure génétique analogue à celle observée dans le cas où les populations sont subdivisées dans l'espace.

Dans ces deux exemples, je ne me suis intéressé qu'à la distribution de la variation génétique à un seul locus, isolé du reste du génome. Or les locus ne sont pas indépendants les uns des autres. Comme nous l'avons vu d'ailleurs dans le § 1.2 (chapitre 1), l'analyse de distribution de la variation à plusieurs locus peut nous en apprendre plus que l'analyse de la variation *en moyenne* sur les locus (c'est-à-dire lorsque l'on considère chaque locus comme apportant une information indépendante et distribuée de la même façon à tous les locus). Dans la section suivante, je m'intéresserai donc à l'analyse d'un modèle qui décrit explicitement la distribution des probabilités d'identité à plusieurs locus.

2.3 L'identité génétique à deux locus

Dans ce qui va suivre je vais poser les bases d'un modèle à nombre d'îles infini (Wright 1951), dans lequel je décrirai les changements des probabilités d'identité à deux locus simultanément. Je reprendrai les mêmes notations que précédemment. Les populations sont de taille N et échangent des gènes à un

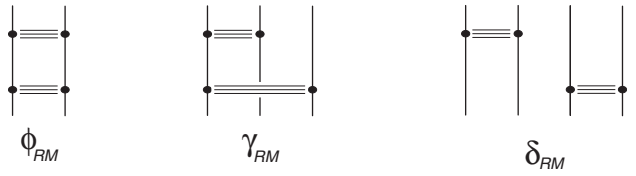


Figure 2.11: Définition des probabilités d'identité à deux locus lorsque les croisements se font au hasard. Les lignes verticales représentent les haplotypes échantillonnés, où les locus sont figurés par un cercle noir. Le symbole $\equiv$ définit l'identité entre gènes homologues.

taux m. Le taux de recombinaison entre deux locus est noté r. Le lecteur intéressé par plus de détails pourra se référer à l'annexe E.

2.3.1 Définitions

Panmixie

Lorsque les croisements se font localement au hasard, dans un modèle à nombre d'îles infini, seules trois probabilités d'identité sont nécessaires pour décrire l'ensemble des équations de transition des probabilités d'identité à deux locus (Whitlock *et al.* 1993). Dans la suite de ce chapitre, j'utiliserai le terme *haplotype* pour décrire un ensemble de deux gènes échantillonnés dans un seul individu à deux locus distincts, qui provient d'un même et unique gamète à la génération précédente. On notera que les deux locus ne sont pas nécessairement localisés sur le même chromosome. ϕ_{RM} est défini comme la probabilité que deux haplotypes tirés au hasard dans une population soient identiques aux deux locus[23]. γ_{RM} est la probabilité que, lorsque l'on tire trois haplotypes au hasard, deux soient identiques au premier locus, tandis que le troisième haplotype est identique à l'un des deux autres haplotypes au second locus. δ_{RM} est la probabilité que, parmi quatre haplotypes, deux soient identiques au premier locus, tandis que les deux autres sont identiques au second locus. Ces différentes probabilités d'identité sont schématiquement

[23]L'indice RM vaut pour *Random Mating* qui signifie que les croisements se font localement au hasard. Voir l'annexe E.

représentées sur la figure 2.11.

En quoi ces trois probabilités d'identité sont-elles nécessaires pour décrire l'ensemble des relations d'identité de deux paires de gènes à deux locus distincts ? Prenons, par exemple, la probabilité ϕ_{RM}. Avec la probabilité $(1-m)^2(1-1/N)$ une paire d'haplotypes échantillonnée dans une sous-population est issue de parents distincts à la génération précédente. Mais ces deux haplotypes peuvent être tous les deux de type recombinant[24] ou bien tous les deux de type parental[25], ou bien encore l'un peut être de type recombinant et l'autre de type parental. Ceci implique que le nombre d'haplotypes à considérer à la génération précédente est de 4, 2 ou 3, respectivement. Par conséquent, ces haplotypes sont identiques avec une probabilité égale à $[(1-r)^2\phi_{RM} + 2r(1-r)\gamma_{RM} + r^2\delta_{RM}]$. Avec la probabilité $(1-m)^2/N$ une paire d'haplotypes échantillonnée dans une sous-population est issue d'un unique parent à la génération précédente. Ces deux haplotypes sont tous les deux de type recombinant ou tous les deux de type parental en proportion $[(1-r)^2 + r^2]$. Dans les deux cas, les haplotypes échantillonnés (c'est-à-dire les gamètes produits par le parent) peuvent être deux copies d'un seul haplotype ou bien d'haplotypes distincts et sont donc identiques avec la probabilité $(\phi_{RM}+1)/2$. Avec la probabilité $[2r(1-r)]$, en revanche, l'un des deux haplotypes échantillonnés est de type recombinant et l'autre est de type parental. Ils sont donc nécessairement des copies distinctes des haplotypes parentaux et ne peuvent être identiques qu'à un locus, avec la probabilité Q_{RM}. Le raisonnement peut être poursuivi pour écrire les équations de récurrence pour les probabilités γ_{RM} et δ_{RM}.

Autofécondation partielle

Lorsque les croisements ne se font pas au hasard, par exemple chez des espèces dioïques ou bien des espèces qui s'autoféconden, ces trois probabilités d'identité ne suffisent pas à décrire complètement l'ensemble des relations

[24] On dit d'un haplotype qu'il est de type *recombinant* s'il est le produit de la recombinaison des deux haplotypes du parent dont il est issu.

[25] On dit d'un haplotype qu'il est de type *parental* s'il est la copie de l'un des deux haplotypes du parent dont il est issu.

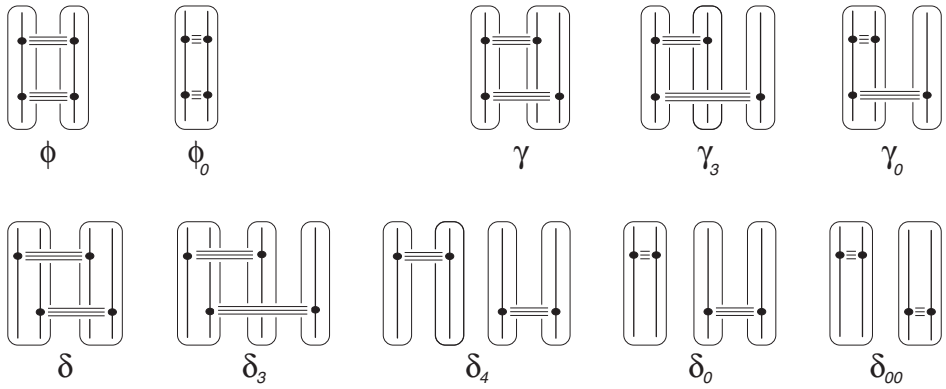


Figure 2.12: Définition des probabilités d'identité à deux locus pour un régime de reproduction mixte. Les lignes verticales représentent les haplotypes échantillonnés, où les locus sont figurés par un cercle noir. Les individus sont représentés par un rectangle. Le symbole ≡ définit l'identité entre paires de gènes homologues.

d'identité entre paires de gènes à deux locus. En d'autres termes, cela signifie que chacune de ces probabilités d'identité prend une valeur différente, selon qu'elle est évaluée parmi des paires, des triplets ou bien des quadruplets d'individus (Weir *et al.* 1980; Weir et Cockerham 1969; Weir et Hill 1980b). Les dix probabilités d'identité à deux locus qui sont nécessaires et suffisantes pour décrire l'ensemble des relations d'identité entre paires de gènes à deux locus sont représentées schématiquement dans la figure 2.12. Je n'en détaillerai ici qu'une partie. ϕ, γ et δ correspondent exactement aux probabilités ϕ_{RM}, γ_{RM} et δ_{RM} définies plus haut, lorsque les haplotypes sont échantillonnés dans des paires d'individus. δ_4 est la probabilité que, parmi quatre haplotypes échantillonnés dans quatre individus distincts, deux soient identiques au premier locus, tandis que les deux autres soient identiques au second locus.

Il n'est pas possible en pratique, à partir des seules données génotypiques, d'estimer chacune de ces probabilités d'identité. La raison en est que l'on ne distingue en général pas les cas où les gènes à deux locus sont en *phase*, ou bien en *répulsion*. On dit en général de deux gènes pris à deux locus distincts dans un même individu qu'ils sont en phase si les deux locus sont position-

nés sur le même chromosome. On étendra ici cette définition au cas où deux gènes, échantillonnés dans un individu à deux locus distincts, appartiennent à un unique haplotype (c'est-à-dire sont hérités d'un seul parent). De même on dira ici que deux gènes, échantillonnés dans un individu à deux locus distincts, sont en répulsion s'ils proviennent de gamètes parentaux distincts. Par exemple, lorsque deux individus portent des allèles identiques à deux locus distincts, il n'est en général pas possible de déterminer si les allèles d'un même individu proviennent d'un seul gamète parental ou de deux. En d'autres termes, il n'est pas possible de déterminer si les probabilités d'identité sont de type ϕ, γ ou δ. Par conséquent, nous avons défini la probabilité d'identité composite Φ comme étant la probabilité qu'à deux locus, les gènes échantillonnés dans deux individus distincts soient identiques. Deux paires de gènes échantillonnées à deux locus distincts peuvent être portées par 2, 3 ou 4 haplotypes. Ces trois cas correspondent aux probabilités d'identité ϕ, γ ou δ (voir la figure 2.12). Lorsque l'on échantillonne toutes les paires possibles de gènes à deux locus distincts, ces trois « configurations » sont en proportion 1/4, 1/2 et 1/4, respectivement. Par conséquent,

$$\Phi = \frac{\phi + 2\gamma + \delta}{4} \tag{2.30}$$

Le déséquilibre d'identité

Ohta (1980) a défini une mesure d'association génétique à plusieurs sites dans des séquences de protéines. Cette mesure est définie comme l'excès d'identité conjointe à deux sites, par rapport à l'attendu sous l'hypothèse d'indépendance des probabilités d'identité à deux sites. Cette quantité est équivalente à la covariance des probabilités de non-identité à deux locus (Avery et Hill 1979; Hedrick 1987). Nous avons défini une mesure analogue, que l'on nommera « déséquilibre d'identité », noté η_S, comme la différence entre la probabilité conjointe d'identité à deux locus (entre deux individus de la même sous-population) et le produit des probabilités d'identité à chacun de ces locus soit, pour deux locus i et j

$$\eta_{S,ij} = \Phi_{ij} - \delta_{4ij} \tag{2.31}$$

(voir l'annexe E). Cette quantité, qui peut être définie comme la covariance des probabilités d'identité entre locus, est une fonction simple des probabilités d'identité à deux locus. Il s'agit donc bien d'un paramètre, lui même fonction de certains paramètres du modèle de populations subdivisées (nombre d'individus, taux de migration) mais également du modèle de mutation. Une définition alternative du déséquilibre d'identité nous est donnée par

$$\eta'_{S,ij} = \frac{\Phi_{ij} - \delta_{4ij}}{(1 - Q_{2i})(1 - Q_{2j})} \tag{2.32}$$

(voir, par exemple, Hedrick 1987; Ohta 1980; Takahata 1982). Nous verrons par la suite l'intérêt d'une telle définition.

Relation au déséquilibre de liaison

Si l'on note P_v^u la fréquence des gamètes qui portent l'allèle u au premier locus et l'allèle v au second locus, ainsi que P^u et P_v les fréquences des allèles u et v au premier et au second locus, respectivement, alors le déséquilibre de liaison est défini par $D_v^u = P_v^u - P^u \cdot P_v$. Ainsi, le déséquilibre de liaison entre deux locus est défini comme l'excès des gamètes portant les allèles u et v par rapport à ce que l'on attend sous l'hypothèse que l'association des allèles aux différents locus se fait au hasard.

Le déséquilibre de liaison $\mathcal{E}(D)$ entre des locus neutres tend vers zéro dans des populations de taille infinie (Bennett 1954) ou bien de taille finie (Hill et Robertson 1966), quel que soit le nombre de locus considéré (Hill 1974a) et le régime de reproduction (Bennett et Binet 1956; Weir et Cockerham 1973). Dans les populations subdivisées, le déséquilibre de liaison ne peut exister que de façon transitoire (Nei 1973; Slatkin 1975) et finit par disparaître (Nei 1973; Ohta 1982a,b). En revanche, bien que le déséquilibre de liaison converge vers la valeur zéro dans tout modèle neutre la variance du déséquilibre de liaison entre les lignées qui ségrègent peut être très grande, en particulier dans les populations de petite taille (Hill 1974b; Hill et Robertson 1968; Ohta et Kimura 1969; Sved 1968; Weir et Hill 1980b). La variance du déséquilibre de liaison $\mathcal{E}(D^2)$ peut être décom-

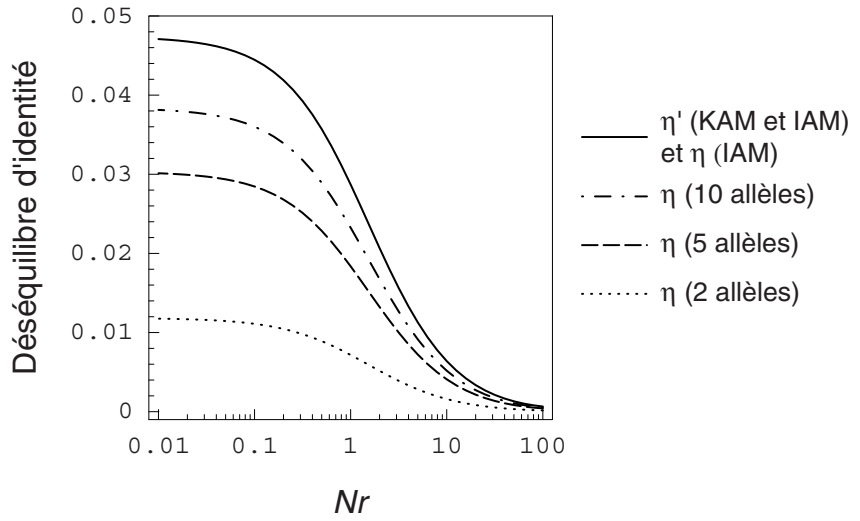


Figure 2.13: Déséquilibre d'identité attendu sous différents modèle de mutation. Les croisements se font au hasard dans chaque sous-population ($s = 1/N$). Noter l'échelle logarithmique en abscisses. $\mu = 10^{-6}$, $N = 200$ et $Nm = 1$. L'intensité de la recombinaison est donnée par la valeur du produit Nr.

posée pour tenir compte de la structure d'une population subdivisée (Ohta 1982a,b). Ces différentes composantes peuvent s'exprimer en fonction des probabilités d'identité à deux locus simultanément (Tachida et Cockerham 1986), telles que celles définies dans la figure 2.12. Ainsi, le déséquilibre d'identité défini par les équations (2.31) ou (2.32) est une mesure analogue à la variance du déséquilibre de liaison $\mathcal{E}(D^2)$.

2.3.2 Influence du modèle de mutation

La figure 2.13 montre les valeurs attendues de η_S et η'_S sous différents modèles de mutation (modèle à nombre d'allèles infini, ou *IAM*, et modèle à K allèles, ou *KAM*[26]). Lorsque le produit de la taille des sous-populations et du taux de recombinaison est grand, à la fois η_S et η'_S tendent vers zéro, quel que soit le modèle de mutation. Ceci montre que lorsque la taille des sous-populations est grande, de même lorsque le taux de recombinaison est fort,

[26]Terme traduit de l'anglais, *Infinite Allele Model* et *K Allele Model*.

les probabilités d'identité à deux locus tendent vers la valeur attendue du produit des probabilités d'identité à chacun des locus. Au contraire, pour des faibles valeurs de Nr, le déséquilibre d'identité est positif, et prend des valeurs d'autant plus grandes que le modèle de mutation s'approche du modèle à nombre d'allèles infini. D'autre part, plus la liaison génétique est forte plus le déséquilibre d'identité dépend du modèle de mutation. Les associations de gènes à différents locus sont donc créées d'autant plus facilement sous l'effet de la dérive que le nombre d'états alléliques possibles est grand. La figure 2.13 montre en outre que η'_S ne dépend pas du modèle de mutation, et ce quelles que soient les valeurs des autres paramètres du modèle (taux d'autofécondation, taux de recombinaison, taille des sous-populations, taux de migration). Ohta (1980) et Takahata (1982) ont trouvé des résultats analogues pour une population unique, isolée. η'_S dépend donc des paramètres d'*intérêt* du modèle (ici, le produit de la taille des sous-populations par le taux de recombinaison), sans toutefois dépendre de paramètres de *nuisance*[27] (tel que le modèle de mutation). Ce résultat justifie l'utilisation de la définition (2.32) du déséquilibre d'identité, plutôt que de la définition (2.31).

2.3.3 Influence du régime de reproduction

La figure 2.14 montre l'effet de la recombinaison sur la valeur attendue du déséquilibre d'identité pour une gamme de valeurs du taux d'autofécondation. Comme précédemment, le déséquilibre augmente à mesure que la valeur du produit Nr diminue. Mais il est d'autant plus fort que le taux d'autofécondation est élevé. Ces résultats sont en accord avec ceux de Golding et Strobeck (1980) dans le cas d'une population isolée. Golding et Strobeck (1980) ont montré que, lorsque $N\mu < 1$, la variance du déséquilibre de liaison augmente avec le taux d'autofécondation si $Nr > 1$ ou bien diminue avec lui si $Nr < 1$. En revanche, lorsque $N\mu > 1$ la variance du déséquilibre de liaison augmente avec le taux d'autofécondation quelle que soit la valeur de Nr. Dans le cas d'une sous-population non isolée du reste de la popula-

[27] On parle de paramètres de nuisance pour des paramètres nécessaires à la définition du modèle, mais dont l'estimation n'est pas intéressante en soi dans le cadre du problème posé.

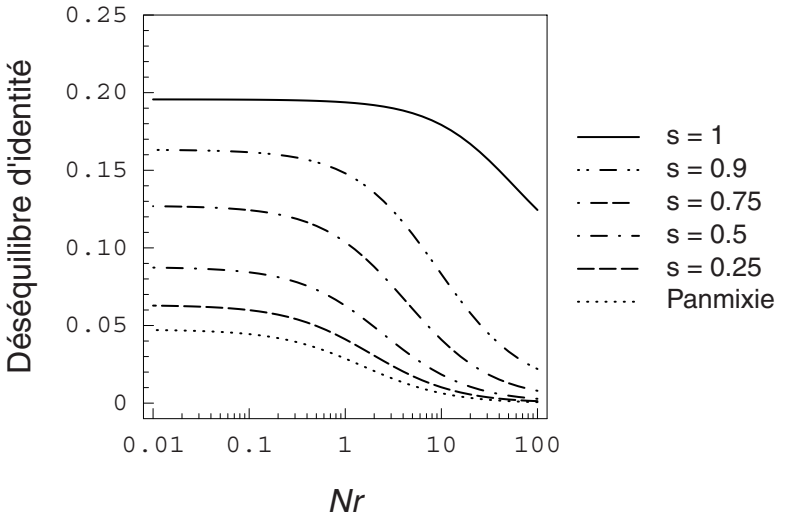


Figure 2.14: Déséquilibre d'identité attendu pour différents régimes de reproduction. Les valeurs du déséquilibre d'identité sont données dans le cas de croisements au hasard (panmixie) et pour différentes valeurs du taux d'autofécondation. Noter l'échelle logarithmique en abscisse. Le modèle de mutation est l'IAM et $\mu = 10^{-6}$. $N = 200$ et $Nm = 1$. L'intensité de la recombinaison est donnée par la valeur du produit Nr.

tion, nos résultats confirment ceux de Golding et Strobeck (1980), dans le cas où $Nm < N\mu < 1$. En revanche, dès lors que $Nm > N\mu$, le déséquilibre d'identité augmente avec le taux d'autofécondation quelle que soit la valeur du produit Nr (voir l'annexe E).

2.3.4 Influence de la dispersion limitée

La figure 2.15 montre l'effet de la dispersion limitée sur le déséquilibre d'identité intra-population. Le déséquilibre n'est pas une fonction monotone du taux d'immigration. En effet, le déséquilibre atteint une valeur maximale pour des taux de migration intermédiaires. Pour de très faibles taux de migration, la migration introduit de nouvelles combinaisons de gènes dans des sous-populations quasiment monomorphes et crée le déséquilibre d'identité au sein des sous-populations locales. Dès lors, le déséquilibre d'identité aug-

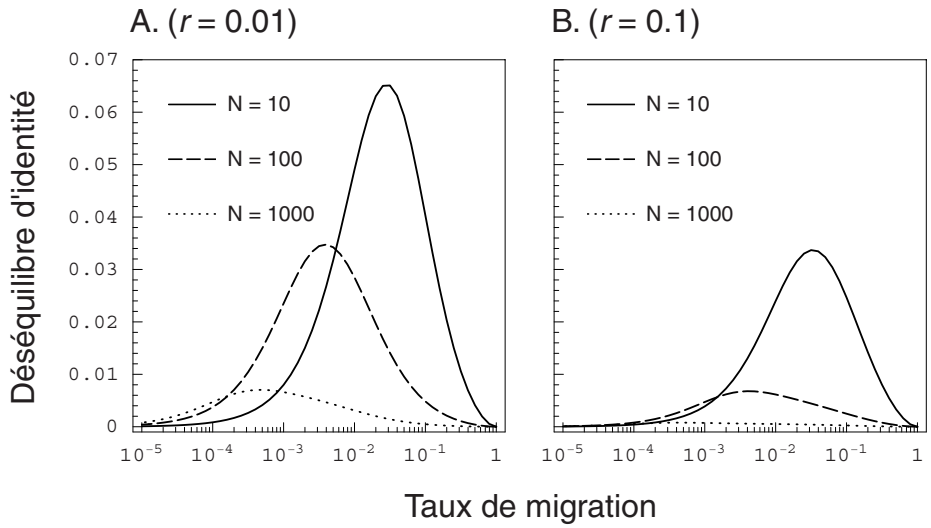


Figure 2.15: Influence de la dispersion limitée sur le déséquilibre d'identité. Les valeurs attendues du déséquilibre d'identité sont données pour différentes tailles des sous-populations ($N = 10$, 100 and 1000) et pour différents taux de recombinaison. A. $r = 0.01$. B. $r = 0.1$. Les croisements se font au hasard dans chaque sous-population. Le modèle de mutation est à nombre infini d'allèles, et $\mu = 10^{-6}$.

mente à mesure que le taux de migration augmente. Toutefois, l'augmentation du taux de migration fait tendre la population totale vers une situation de panmixie, et le déséquilibre d'identité intra-population converge vers la valeur attendue dans la population totale, de taille infinie, c'est-à-dire zéro. Des résultats tout à fait similaires ont été obtenus par Tachida et Cockerham (1986), qui ont développé le modèle de décomposition de variance du déséquilibre de liaison introduit par Ohta (1982a,b). On note également que le mode des courbes définies par les valeurs du déséquilibre d'identité dépend de la valeur du produit Nm plutôt que de m. De plus, les valeurs maximales atteintes par le déséquilibre d'identité augmentent à mesure que la taille des sous-populations diminue. Enfin, l'effet de la migration est d'autant plus prononcé que les locus sont liés physiquement sur le chromosome (comparer les figures 2.15A et 2.15B).

La figure 2.16 montre la valeur attendue du déséquilibre d'identité en

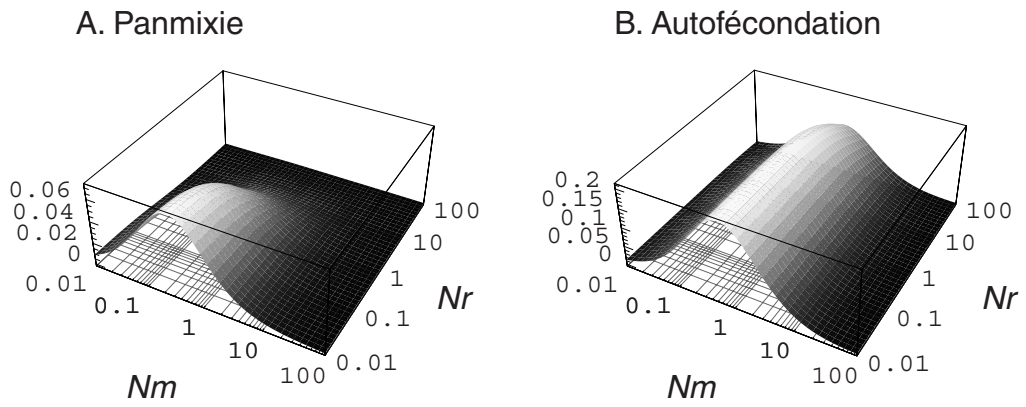


Figure 2.16: Effet de la dispersion limitée sur le déséquilibre d'identité pour deux régimes de reproduction contrastés. Les valeurs du déséquilibre d'identité sont données en fonction de Nm et de Nr. A. Les croisements se font au hasard dans chaque sous-population ($Ns = 1$). B. Les individus se reproduisent par autofécondation stricte ($s = 1$). Le modèle de mutation est à nombre d'allèles infini et $\mu = 10^{-6}$. Noter la double échelle logarithmique.

fonction des produits Nr et Nm. Il existe des valeurs de Nm pour lesquelles le déséquilibre d'identité intra-population est toujours très faible. On peut déduire de cette figure que lorsque $Nm < 10^{-2}$ ou bien $Nm > 10^2$, aucun déséquilibre ne peut se maintenir à l'équilibre. Dans le cas où les croisements se font au hasard dans chaque sous-population, le déséquilibre se maintient pour des valeurs de Nr inférieure à 1. Lorsque le régime de reproduction est l'autofécondation totale, le déséquilibre d'identité est toujours maintenu, pour des valeurs du taux de migration telles que $10^{-2} < Nm < 10^2$. Dans cette gamme de valeurs, le déséquilibre d'identité intra-population est maintenu lorsque $Nr(1 - s) < 1$ (voir l'annexe E), comme l'est la variance du déséquilibre de liaison dans une population isolée (voir GOLDING et STROBECK 1980).

2.4 Conclusion

Le déséquilibre d'identité, tel qu'il a été défini ici (équation 2.32), dépend des paramètres d'intérêt du modèle (la taille des populations, les taux d'immigration, le régime de reproduction) sans toutefois dépendre des paramètres de nuisance (voir la note de bas de page 27, p. 56). Contrairement au déséquilibre de liaison (voir p. 54), sa valeur n'est pas nulle à l'équilibre entre les forces de mutation, de migration et de dérive et peut donc être utilisée pour estimer l'intensité de ces forces.

Les modèles que je viens de présenter permettent de comprendre l'effet de différents types de structure (en âges, par sexe, etc.) sur l'identité génétique dans les populations. Dans la dernière partie j'ai également montré que la notion d'identité, étendue à plusieurs locus, nous permettait de comprendre la mise en place et le maintien des déséquilibres que l'on peut observer entre différents locus, en raison de la dérive, des flux de gènes limités ou d'un régime de reproduction consanguin. Dans les populations de taille finie, la dérive joue un rôle primordial pour expliquer la distribution des probabilités d'identité génétique à différents niveaux. Comment peut-on caractériser au mieux cette force évolutive ? Cette question est abordée dans le prochain chapitre.

Chapitre 3

La taille efficace

> « There remains one factor of the greatest importance in understanding the evolution of a Mendelian system. This is the size of the population. The constancy of gene frequencies in the absence of selection, mutation or migration cannot for example be expected to be absolute in populations of limited size. Merely by chance one or the other of the allelomorphs may be expected to increase its frequency in a given generation and in time the proportions may drift[1] a long way from the original values. » (Wright 1931).

Je m'attacherai dans ce second chapitre à discuter de la notion d'*effectif* ou de *taille* efficace, dont le concept a été introduit dans la littérature au début des années 1930 par Sewall Wright (1931). Sewall Wright fut l'un des premiers à reconnaître l'importance des effets stochastiques liés à la taille finie des populations.

[1]Il s'agit d'une des premières utilisations du terme *drift* (qui se traduit en français par « dérive ») pour faire référence aux effets cumulés des processus aléatoires liés aux variations d'échantillonnage lors de la gamétogénèse (voir Wright 1977, p.444).

3.1 Définitions

3.1.1 Le modèle de Fisher et de Wright

Le modèle connu sous le nom de modèle de Wright-Fisher est un modèle très simple d'évolution des fréquences alléliques dans une population de taille finie (Fisher 1930; Wright 1931)[2]. Dans ce modèle, on considère une population diploïde de taille constante N. Les individus sont hermaphrodites et il n'existe aucune différence sélective entre les différents allèles à un locus donné. Les individus se reproduisent de façon simultanée et les croisements se font au hasard. Ce modèle suppose que les gènes à la génération $t+1$ sont le résultat d'un tirage avec remise des gènes présents à la génération t (soit un tirage sans remise dans une *urne gamétique* contenant un nombre infini de copies des gènes présents à la génération t) et que les tirages sont réalisés indépendamment les uns des autres (voir Gale 1990, pp. 21-22).

Changement de fréquence des gènes en une génération

Traduits en termes statistiques, les hypothèses du modèle de Wright-Fisher impliquent que le nombre $X(t+1)$ de gènes appartenant à une classe allélique donnée est une variable aléatoire tirée dans une loi binomiale de paramètres $2N$ et $X(t)/(2N)$. Donc, la probabilité que le nombre de copies $X(t+1)$ de gènes d'une classe allélique soit égal à j, conditionnellement au nombre de copies de gènes dans cette classe à t est donnée par

$$P\left[X(t+1)=j|X(t)=i\right]=\binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j} \tag{3.1}$$

(voir, par exemple, Ewens 1969, 1979). Par conséquent, le nombre de copies qu'un gène en particulier transmet à la génération suivante, c'est-à-dire $P\left[X(t+1)=j|X(t)=1\right]$ suit une distribution binomiale (équation 3.1) de moyenne 1 et de variance $1-1/(2N)$. Pour de grandes valeurs de N, cette distribution converge vers une distribution Poisson de moyenne (et donc de variance) égale à 1 (voir, par exemple, Feller 1968). Notons que le raison-

[2]Voir Gale (1990, p. 21), Ewens (1979, p. 16).

nement est valable pour n'importe quel gène à la génération t. Ceci implique donc que chaque gène transmet, en espérance, le même nombre de copies à la génération suivante. En d'autres termes, un gène pris au hasard à $t+1$ a la même probabilité d'être la copie de n'importe quel gène à t. D'autre part, la distribution conjointe du nombre de copies de gènes transmis à leur descendance par les N individus présents à la génération t est une distribution multinomiale symétrique[3].

Propriétés du modèle

D'après le modèle de Wright-Fisher défini en (3.1), $X(\cdot)$ est une chaîne de Markov[4], dont la matrice de transition est donnée par

$$\mathbf{P} = P\left[X(t+1) = j \mid X(t) = i\right] \tag{3.2}$$

Les deux plus grandes valeurs propres de cette matrice sont égales à un. Si les hypothèses du modèle de Wright-Fisher sont respectées, alors le modèle est caractérisé par les propriétés suivantes :

(i) La plus grande valeur propre (différente de l'unité) de la matrice de transition des fréquences alléliques d'une génération à la suivante, est $[1 - 1/(2N)]$ (voir Ewens 1979, annexe B, p. 305).

(ii) La probabilité que deux gènes tirés au hasard dans la population soient des copies du même gène est $1/(2N)$.

(iii) Conditionnellement à la fréquence x d'un gène à t, la variance de la fréquence de ce gène à $t+1$ est $x(t)\left[1 - x(t)\right]/(2N)$ (variance binomiale : voir l'équation 3.1).

[3]Il est intéressant de constater que la distribution conjointe d'un ensemble de lois de Poisson indépendantes, conditionnelle à la somme de ces lois, est également une distribution multinomiale.

[4]On note $X(\cdot)$ la suite de variables aléatoires $X(1)$, $X(2)$, ..., $X(t)$ telle que la loi conditionnelle de $X(t)$ sachant $X(t-1)$, $X(t-2)$, ... est la même que la loi conditionnelle de $X(t)$ sachant $X(t-1)$ (voir, par exemple, Robert 1996, pp.91-123).

Plusieurs définitions de tailles efficaces

Il ne fait aucun doute que les hypothèses du modèle de Wright-Fisher sont simplistes et ne peuvent prétendre décrire la réalité de façon pertinente. La variance de succès reproducteur, par exemple, cause des écarts significatifs au modèle de Wright-Fisher. Tout d'abord les gènes n'ont plus les mêmes probabilités de transmettre des copies d'eux-mêmes à la génération suivante, et donc la distribution conjointe du nombre de copies de gènes transmis à leur descendance par les N individus présents à la génération t devient une distribution multinomiale asymétrique. D'autre part, si ces variations du succès reproducteur sont héritables, alors les fréquences alléliques sont corrélées entre générations. Mais il existe d'autres effets (non sélectifs) qui peuvent également causer des écarts aux hypothèses du modèle. Il s'agit, par exemple, de l'existence d'une structure sociale au sein des populations (Chesser 1991b), de biais de sexe-ratio (Wright 1931), de systèmes de reproduction complexes (voir, pour synthèse, Wright 1969), de la dispersion limitée (Wright 1931), ou bien encore de la mise en place d'une structure en âges au sein des populations (Charlesworth 1994b). Pourtant, ce modèle a permis de définir le concept de *taille efficace*[5], introduit dans la littérature par Wright (1931). Il est possible en effet de définir la taille efficace d'une population comme étant la taille d'une population de Wright-Fisher où la dérive génétique aurait la même intensité que dans la population (ou bien le modèle de population) qui nous intéresse (voir Hartl et Clark 1989). Or par quelle propriété mieux vaut-il caractériser la dérive génétique ? On voit bien qu'il est possible de définir autant de tailles efficaces que l'on peut compter de propriétés décrivant le phénomène de dérive. Plus précisément donc, si l'on choisit l'une des propriétés du modèle de Wright-Fisher, il est possible de définir la taille efficace d'une population comme étant la taille de la population de Wright-Fisher qui possède la même propriété (voir, par exemple, Ewens 1979). On distingue en général trois définitions de la taille efficace.

[5]Wright utilise alors le terme de *population number* ou d'*effective number* (Wright 1931).

(i) La taille efficace de consanguinité[6] est définie comme $N_e^{(i)} = 1/(2\pi)$, où π est la probabilité que deux gènes tirés au hasard dans la population soient des copies du même gène à la génération précédente (Wright 1931).

(ii) La taille efficace de variance[7] est définie comme $N_e^{(v)} = \frac{x(t)[1-x(t)]}{(2\ \mathrm{var}[x(t+1|x(t)])}$, où $x(t)$ est la fréquence du gène A à t (Crow et Denniston 1988).

(iii) La taille efficace de valeur-propre[8] est définie comme $N_e^{(e)} = \frac{1}{2}(1-\lambda)^{-1}$, où λ est la plus grande valeur propre (différente de 1) de la matrice de transition des fréquences alléliques, d'une génération à l'autre (Ewens 1979).

Ces différentes définitions montrent qu'il n'existe, en général, pas une seule valeur de la taille efficace. Car si ces définitions donnent la même valeur de $N_e = N$ dans le cas d'un modèle de Wright-Fisher strict, il n'en va pas de même dans des modèles de populations plus complexes où les effets de la dérive génétique ne sont pas nécessairement les mêmes que dans le cas d'une population idéale de Wright-Fisher.

3.1.2 Taille efficace et probabilités de coalescence

Une manière intuitive de se représenter les conditions pour lesquelles les définitions de N_e convergent vers une seule valeur vient de la relation entre le taux de coalescence et la taille efficace. En effet, le taux asymptotique auquel deux gènes coalescent (c'est-à-dire leur probabilité de coalescence) est égal à $1/(2N)$ dans une population idéale, composée de N individus diploïdes (Hudson 1990). Étant donné le lien étroit qui existe entre les probabilités d'identité de gènes et les temps de coalescence (Malécot 1975; Slatkin 1991, voir également le § 2.1.2), la plus grande valeur propre de la matrice de transition qui décrit le gain d'identité entre paires de gènes donne le taux asymptotique de coalescence des lignées de gènes. Il en est de même de la plus grande valeur propre de la matrice de transition qui décrit les changements de

[6]Terme traduit de l'anglais, *Inbreeding effective population size.*
[7]Terme traduit de l'anglais, *Variance effective population size.*
[8]Terme traduit de l'anglais, *Eigenvalue effective population size.*

distribution des fréquences alléliques d'une génération à l'autre (Whitlock et Barton 1997). Dans un modèle de populations structurées en classes, Rousset (1999a, équation 48) définit la taille efficace comme la moyenne harmonique des effectifs des différentes classes, pondérés par les probabilités que les gènes d'une classe coalescent, à la génération précédente, dans une autre classe. A travers cette définition, $1/(2N_e)$ est le taux asymptotique de coalescence d'une paire de gènes dans une population. C'est également le taux asymptotique du gain d'identité génétique, celui de la variance des fréquences génétiques et enfin le taux attendu de changement dans la distribution des états alléliques (Whitlock et Barton 1997). Ainsi, il existe une taille efficace asymptotique vers laquelle convergent les différentes tailles efficaces définies en § 3.1 (voir également Chesser *et al.* 1993).

3.2 Influence de la structure des populations

Si l'on considère donc une population structurée en classes d'individus (ces classes pouvant représenter des sous-populations distinctes, des âges différents, ou bien toute autre catégorie sociale), on imagine facilement que les lignées de gènes peuvent coalescer à des taux différents selon que ces lignées sont prises à l'intérieur des classes, ou bien entre elles (voir, pour une illustration, la figure 3.1). Par conséquent la structure des populations, qu'elle soit spatiale (populations subdivisées) ou temporelle (populations structurées en âges), influence nécessairement l'intensité de la dérive génétique.

3.2.1 Structure spatiale

Par structure spatiale, j'entends ici la « subdivision des populations en groupes d'individus, échangeant des gènes de façon homogène dans l'espace ». Cette structure n'est donc pas spatialement explicite (voir la remarque faite au début du § 2.2.2).

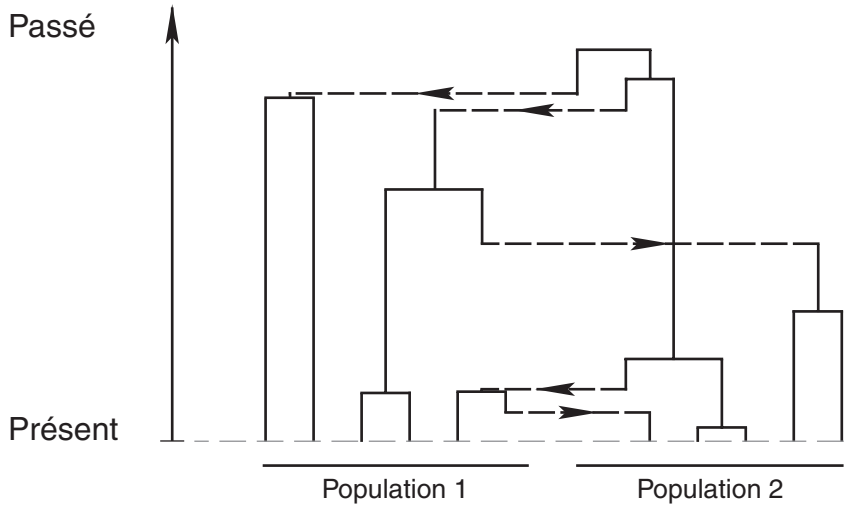


Figure 3.1: Un exemple de généalogie de gènes dans deux populations connectées. Cette figure représente une réalisation du processus de coalescence pour un échantillon de 2×5 gènes, dans deux populations constituées chacune de $N = 100$ individus haploïdes, qui échangent une proportion $m = 0.01$ de gènes, par génération. Chaque événement de migration est représenté par une ligne pointillée et fléchée.

Espèces monoïques

Dans un modèle à nombre n d'îles, on peut montrer que $N_e = Nn/(1 - F_{ST})$ (voir, par exemple, Barton et Whitlock 1997; Whitlock et Barton 1997). Par conséquent, l'effet de la subdivision des populations dans ces modèles est de diminuer l'effet de la dérive (la taille efficace est augmentée d'un facteur $1/(1 - F_{ST})$ par rapport à une population continue, de même taille totale Nn). Les lignées de gènes diminuent en nombre dans chaque sous-population, mais les allèles fixés dans chaque sous-population ne sont pas nécessairement identiques et la dérive génétique sur l'ensemble de la population est donc minimisée, par rapport à une population continue de même taille (Caballero 1994). Ce résultat est également valable pour un modèle de migration par pas[9], dans une ou deux dimensions, à condition que

[9]Terme traduit de l'anglais *stepping stone*. Il s'agit d'une famille de modèle où la dispersion des individus (ou des gènes) ne peut se faire qu'entre les sous-populations les plus proches.

la taille des sous-populations soit constante dans le temps et dans l'espace (Whitlock et Barton 1997). Wang (1997b) obtient des résultats tout à fait similaires dans le cas d'une espèce végétale avec migration du pollen et dispersion des graines.

En revanche, ces modèles ne prennent pas en compte d'éventuelles variations dans la contribution des sous-populations à la composition génétique future de la population dans son ensemble. Or des différences de taille des sous-populations, des taux d'immigration ou d'émigration variables, ainsi que les processus d'extinction et de recolonisation des sous-populations entraînent de telles variations. Whitlock et Barton (1997) et Barton et Whitlock (1997) ont montré que l'effet de la subdivision des populations, lorsque l'on tient compte de dynamiques démographiques explicites, est de réduire la taille efficace par rapport au nombre total d'individus reproducteurs dans la métapopulation. Ceci contredit donc les résultats attendus dans un simple modèle à nombre d'îles fini. Whitlock et Barton (1997) distinguent plusieurs effets pour interpréter leurs résultats. Tout d'abord, d'un point de vue dynamique, les effets conjugués des extinctions locales et de la stochasticité démographique dans des sous-populations de petite taille auront tendance à réduire l'effectif total (et donc la taille efficace) de la métapopulation, en comparaison avec une population unique non subdivisée qui occuperait le même habitat. D'autre part, la variance des contributions des différentes sous-populations à la composition génétique future de la métapopulation réduit, elle aussi, la taille efficace asymptotique (voir également Wakeley 2001). Ces deux facteurs, qui ne sont pas pris en compte dans le modèle à nombre d'îles fini, peuvent contre-balancer l'augmentation de la taille efficace due à la différenciation génétique entre sous-populations (Whitlock et Barton 1997).

Espèces dioïques

Wang (1997a) a développé un modèle de population structurée pour une espèce dioïque, pour une distribution arbitraire du nombre de descendants par individu reproducteur, et des taux de dispersion conditionnels au sexe des

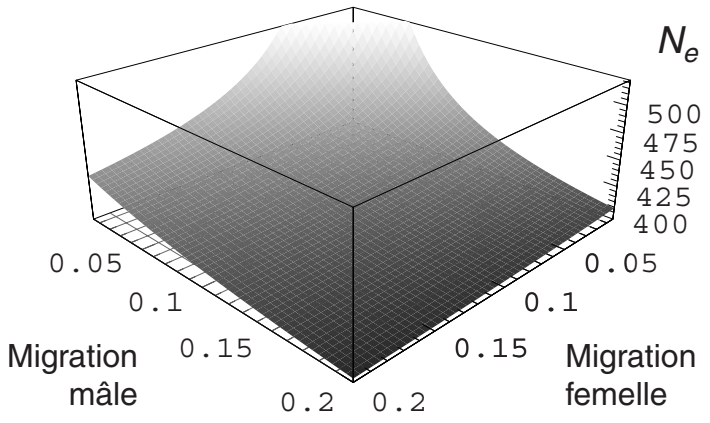


Figure 3.2: Taille efficace dans une population structurée d'une espèce dioïque. La population est subdivisée en 20 sous-populations, chacune étant composée de 100 individus. Le sexe-ratio est biaisé en faveur des femelles (95 : 5). La taille efficace asymptotique est donnée en fonction des taux de migration des mâles et des femelles.

individus. En cela, il a généralisé le modèle de Chesser *et al.* (1993) et y a apporté quelques corrections. Ces modèles font suite aux développements de Chesser (1991a,b), qui considérait une fécondité fixée pour les femelles. Tout comme les modèles adaptés aux espèces monoïques, ces modèles prédisent une augmentation de la taille efficace asymptotique avec la diminution des taux de dispersion (mais aucune dynamique démographique explicite n'est prise en compte). En revanche les biais de sexe-ratio, tout comme des différences de la distribution du nombre de descendants par sexe, peuvent conduire à des effets asymétriques de la dispersion limitée. La figure 3.2 montre une telle asymétrie. L'effet de la migration des mâles pour de forts taux de migration des femelles, y est beaucoup plus prononcé que celui de la migration des femelles pour de forts taux de migration des mâles. Le biais de sexe-ratio (tout comme la dispersion limitée) augmente l'effet de dérive à l'intérieur des sous-populations. Ici la dispersion limitée des mâles, qui sont peu nombreux, augmente l'effet de dérive à l'intérieur des sous-populations, bien plus que la dispersion limitée des femelles, qui sont plus nombreuses. Par conséquent, la dérive à l'échelle de la métapopulation est d'autant plus faible que le sexe le

moins représenté est celui qui disperse également le moins (voir également Wang 1997a, figure 4).

3.2.2 Structure temporelle

Structure temporelle dans des populations isolées

De même que la structure spatiale, la structure temporelle des populations influence l'intensité de la dérive qui s'y exerce. On peut d'ailleurs considérer la structure en âges comme un cas particulier de structure spatiale où les flux de gènes se produisent entre les différentes classes d'âge (voir, par exemple, Charlesworth 2001; Whitlock et Barton 1997). Là encore, de nombreuses définitions de la taille efficace ont été employées. Felsenstein (1971) a calculé la taille efficace de consanguinité d'une population structurée en âges, à partir des changements de probabilités d'identité par descendance de paires de gènes prises au hasard dans la population. Dans son modèle, Felsenstein (1971) obtient une taille efficace de consanguinité (égale à la taille efficace de variance) égale au nombre d'individus reproducteurs, multiplié par le temps de génération et divisé par un terme qui dépend des taux de mortalité avant reproduction. Hill (1972) a calculé la taille efficace de variance d'une population structurée en âges, à partir de la variance des fréquences alléliques par génération. Hill (1972) a obtenu une taille efficace égale à la taille efficace de variance d'une population à générations discrètes où le nombre de nouveaux individus entrant dans la population par génération serait identique, ainsi que la variance du nombre de descendants par famille sur toute la durée de vie des familles. Les approches de Felsenstein (1971) et de Hill (1972) ont été développées de façon exacte par d'autres auteurs encore (voir Choy et Weir 1978; Johnson 1977). Par ailleurs, Johnson (1977) a montré que la taille efficace de consanguinité de Felsenstein (1971) est égale à la taille efficace de variance définie par Hill (1972).

Le lien entre taille efficace et probabilités de coalescence (§ 3.1.2) est particulièrement utile pour décrire les modèles de populations structurées en classes. Dans des modèles complexes en effet, où la démographie peut être

explicite (Orive 1993), la taille efficace est définie comme l'inverse du taux moyen de coalescence entre paires de gènes ancestrales (voir Norborg 1997; Rousset 1999a,b), c'est-à-dire du taux asymptotique de coalescence. Rousset (1999a) a montré que la taille efficace ainsi définie est égale au produit de la taille efficace de consanguinité de Felsenstein (1971) et du temps de génération. Orive (1993) a défini une expression de la taille efficace en fonction des temps de coalescence de paires de gènes dans différentes classes. Ses résultats sont numériquement très proches de ceux de Felsenstein (1971). Toutefois, comme dans la définition de Felsenstein (1971), la taille efficace définie par Orive (1993) dépend du temps moyen de génération et pose donc le problème de la définition de ce temps de génération (voir Rousset 1999a, pour une discussion sur le sujet).

Dans le cas où toutes les paires de gènes dans une classe peuvent coalescer en un gène de n'importe quelle autre classe, la structure en âges a pour conséquence de diminuer la taille efficace, dès lors que tous les individus n'ont pas la même valeur reproductive[10] (Nagylaki 1980; Rousset 1999a). En revanche, lorsqu'une paire de gènes dans une classe peut descendre d'une paire de gènes d'une autre classe mais ne peut pas coalescer dans cette classe, alors la taille efficace augmente et peut compenser l'effet dû à la variance de valeur reproductive individuelle. Dans ce cas, la taille efficace d'une population peut être plus importante que le nombre total d'individus dans la population (Rousset 1999a).

Ce dernier cas de figure reflète exactement ce qui se passe dans une population de plantes annuelles dont les graines peuvent rester dormantes pendant quelques générations. Les gènes qui se trouvent dans des graines qui ont dormi plus d'un an sont des gènes qui étaient déjà dans des graines dormantes une génération auparavant, et ne peuvent donc pas avoir coalescé (puisqu'il

[10]La notion de valeur reproductive a été introduite par Fisher (1930). Il s'agit de la contribution d'une classe d'individus (âge, sexe, etc.) à la descendance future de la population. Le nombre de descendants issus de cette classe, leur survie et le temps nécessaire à leur production sont autant de termes qui entrent dans le calcul de la valeur reproductive d'une classe. Les valeurs reproductives de toutes les classes d'une population structurée sont données par le vecteur propre à gauche de la matrice qui décrit les probabilités de transition entre toutes les classes. La valeur reproductive d'un individu est égale à la valeur reproductive de la classe de cet individu, divisée par le nombre d'individus qui composent cette classe.

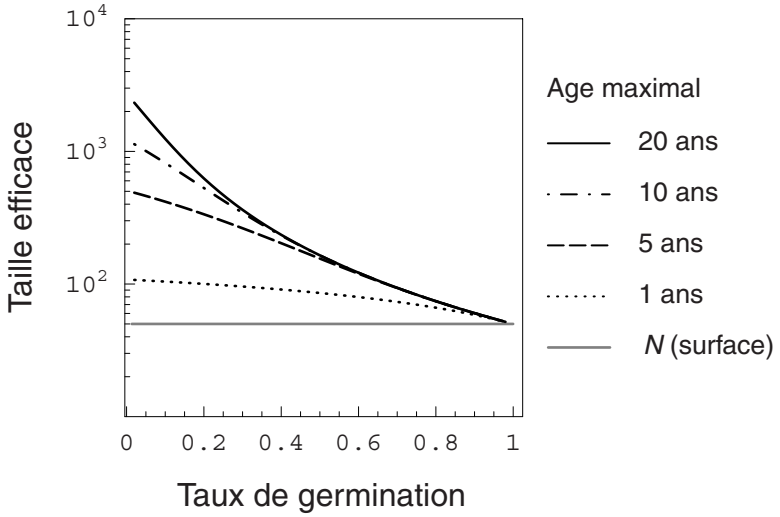


Figure 3.3: Taille efficace asymptotique d'une population de plantes avec une banque de graines. La taille efficace est représentée en fonction du taux de germination des graines, pour différentes valeurs de l'âge maximal des graines. La ligne grise indique le nombre d'individus adultes ($N = 50$) dans la population. La reproduction se fait au hasard ($s = 1/N$) et le taux de mortalité des graines dormantes est égal à 0.1.

n'y a pas eu d'événement de reproduction) à la génération précédente. Dans de telles situations (plantes annuelles avec banques de graines, ou banques d'œufs de copépodes), le taux de coalescence des gènes diminue (Kaj *et al.* 2001)[11]. Si l'on reprend le modèle présenté dans le § 2.2.2, on peut calculer la taille efficace asymptotique d'une population isolée de plantes, dont les graines ne germent pas immédiatement. Cette taille efficace a été calculée numériquement comme la plus grande valeur propre de la matrice de transition des probabilités d'identité de gènes pour toutes les classes d'âge. Sur la figure 3.3, la taille efficace est représentée en fonction du taux de germination des graines chaque génération. Pour des taux de germination inférieurs à 1,

[11] Kaj *et al.* (2001) considèrent un modèle où N individus haploïdes sont produits chaque génération par des graines produites i générations auparavant. Contrairement au modèle que j'ai présenté précédemment (§ 2.2.2), il n'y a pas de référence explicite aux paramètres biologiques (taux de survie, de germination) dont dépend la structure en âges stable de la banque de graines.

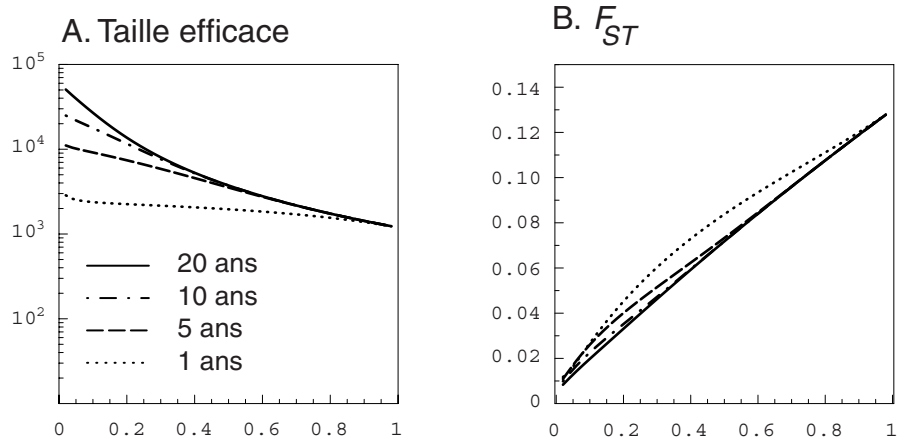


Figure 3.4: Taille efficace et différenciation génétique d'une population structurée de plantes avec une banque de graines. F_{ST} est mesuré parmi les individus en surface. La population est constituée de 20 sous-populations de $N = 50$ individus, et les taux de dispersion des graines et du pollen sont tous deux égaux à 0.02. Le taux de mortalité des graines est égal à 0.1. Voir la figure 3.3 pour la signification des tracés.

c'est-à-dire lorsqu'une proportion non nulle de graines ne germe pas chaque année, la taille efficace est supérieure au nombre total d'individus reproducteurs (en surface). L'effet est d'autant plus important que l'âge maximal des graines (et donc le nombre de classes d'âge dans la banque de graines) est élevé. Par conséquent, l'effet majeur de la banque de graines sur la distribution de la variation génétique est de diminuer l'intensité de la dérive.

Structure temporelle et structure spatiale des populations

Lorsque l'on considère non plus une mais plusieurs sous-populations connectées qui peuvent échanger du pollen et des graines, c'est-à-dire si l'on considère conjointement la structure spatiale et la structure temporelle, l'effet de la dormance est également de diminuer l'effet de la dérive génétique. La figure 3.4A montre la taille efficace d'une population subdivisée, dont les individus produisent des graines dormantes.

La taille efficace *globale* est, dans tous les cas, supérieure au nombre d'in-

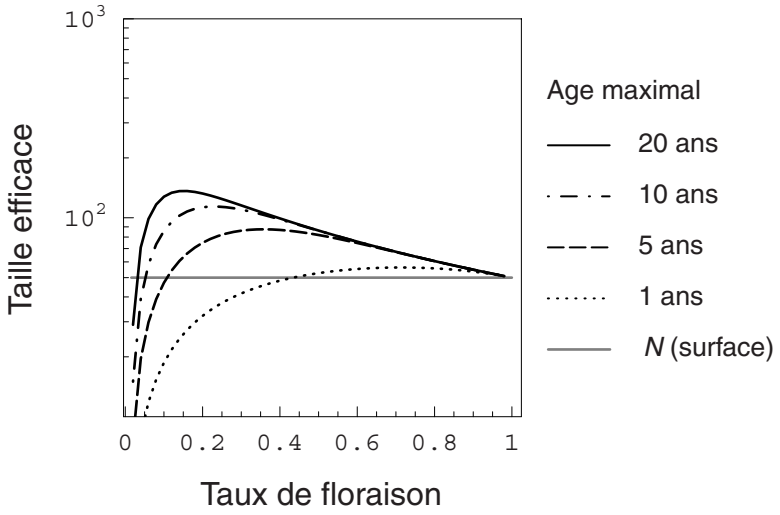


Figure 3.5: Taille efficace asymptotique d'une population de plantes monocarpiques pérennes. La population est constituée de $N = 50$ individus, au total. La reproduction se fait au hasard parmi les individus adultes et le taux de mortalité des rosettes est égal à 0.1. Voir la figure 3.3 pour d'autres légendes.

dividus en fleurs chaque année (figure 3.4A). Ceci est la conséquence, dans une moindre mesure, des flux de gènes limités entre sous-populations mais surtout de l'augmentation de taille efficace *locale* avec le taux de dormance (figure 3.3). Cette augmentation de la taille efficace locale avec le taux de dormance explique pourquoi la différenciation génétique diminue avec le taux de dormance (figure 3.4B). Par conséquent, la mesure des F-statistiques pour des populations structurées en classes d'âge ne permet pas d'inférer l'intensité des flux de gènes entre sous-populations, à moins de connaître exactement la structure en âges stable de la population.

Monocarpiques pérennes

La structure en âges qui découle de taux de germination limités n'est pas sans rappeler la structure en âges de plantes monocarpiques pérennes, c'est-à-dire de plantes dont la durée de vie est supérieure à une année, mais qui ne fleurissent qu'une fois dans leur vie. Il suffit d'imaginer que les rosettes

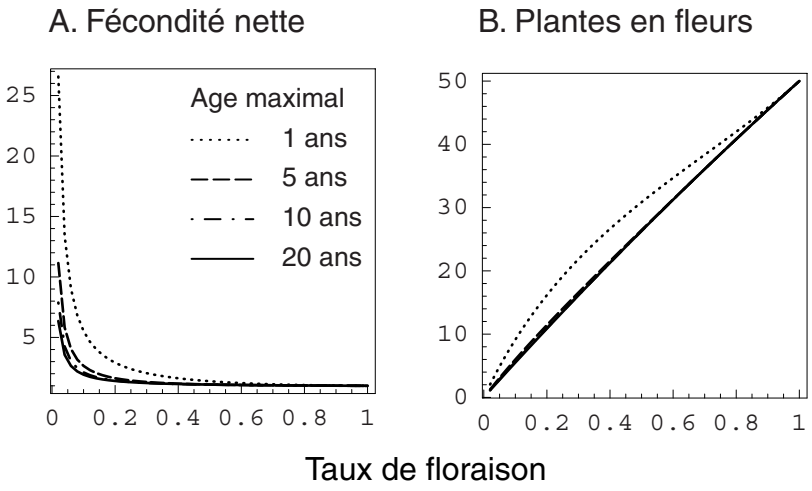


Figure 3.6: Démographie d'une population de plantes monocarpiques pérennes. A. Fécondité annuelle nette par individu. B. Nombre de plantes en fleurs. Le nombre total d'individus (plantes en fleurs et rosettes) est égal à 50, le coût à la dispersion, τ, est égal à 0.1 et l'âge maximal qu'une graine peut atteindre est indiqué sur la figure.

(les individus juvéniles avant floraison) jouent le même rôle dans nos modèles que des graines dormantes. Toutefois, si la structure en âges est la même, le type de régulation de ces populations est différent. Dans le modèle présenté jusqu'à présent (voir § 2.2.2), la régulation ne s'exerçait que pour les individus en fleur qui constituent la population de surface. Dans le cas de plantes monocarpiques pérennes, la régulation s'exerce sur l'ensemble des individus de la population, qu'ils soient aux différents stades de rosettes, ou bien au stade adulte.

La figure 3.5 montre l'évolution de la taille efficace en fonction du taux de floraison. En raison du mode de régulation, la taille efficace asymptotique peut être inférieure ou supérieure au nombre de plantes en fleurs dans la population. Pour de faibles taux de floraison, la régulation a pour effet d'augmenter la fécondité nette par individu. Très peu d'individus contribuent donc beaucoup à l'ensemble de la descendance (figure 3.6). A mesure que le taux de floraison augmente, en revanche, la variance du succès reproducteur diminue et l'effet de tampon créé par les générations chevauchantes compense

l'effet de variance de succès reproducteur. La taille efficace asymptotique est alors supérieure au nombre total d'individus dans la population (figure 3.5).

3.3 Influence de la sélection

Dans la plupart des exemples que j'ai abordés jusqu'à présent, l'hypothèse est faite que tous les individus qui constituent une population ont la même espérance de succès reproducteur, et donc que tous contribuent, en espérance, de façon égale à la génération suivante. Or s'il existe des corrélations entre les probabilités de survie ou de mortalité parmi les descendants d'une même famille, les gènes à la génération $t + 1$ ne sont plus le résultat d'un tirage aléatoire des gènes présents à la génération t, ce qui constitue l'une des propriétés fondamentales du modèle de Wright-Fisher (voir le § 3.1.1). Wright (1939) a formalisé cette idée et a donné une expression de la taille efficace en fonction de la variance du nombre de descendants par individu. Quelle que soit la nature du phénomène, la taille efficace diminue lorsque la variance du nombre de descendants par individu augmente. En particulier, toute forme de sélection qui agirait sur des caractères, héritables ou non, a donc pour effet de réduire la taille efficace (voir, pour une revue, Caballero 1994).

Dans la suite de ce chapitre, je parlerai tout d'abord des effets de la sélection agissant sur un ou plusieurs locus sur la variation neutre à des locus génétiquement liés (§ 3.3.1-3.3.2). Je conclurai ce chapitre en évoquant les limites que la liaison génétique impose à la sélection naturelle (§ 3.3.3).

3.3.1 L'effet Robertson

Qu'en est-il en particulier de l'effet de la sélection agissant sur des caractères héritables ? Robertson (1961) apporte le premier une réponse formelle à un problème posé par les sélectionneurs :

> « In a flock exposed to selection, the genetically superior individuals will tend to be most inbred. As a corollary, selection increases the approach to homozygosity, not only at loci carrying

> genes determining the character in question but at all loci. » (Morley 1954).

Robertson (1961) montre en effet que la sélection qui agit à un locus tend à augmenter l'intensité de la dérive génétique à un autre locus non lié. En effet, les gènes qui se trouvent dans un *contexte*[12] génétique favorable seront mieux représentés à la génération suivante par rapport à des gènes qui se trouvent dans un contexte génétique moins favorable. La sélection augmente ainsi la variance du succès reproducteur et donc l'effet de dérive génétique à des locus non liés au locus sélectionné (Felsenstein 1974).

Mais Robertson (1961) montre surtout que les changements de fréquences alléliques d'une génération sur l'autre ne sont pas indépendants les uns des autres dès lors que la sélection agit sur des caractères héritables : un gène qui se trouve dans un contexte génétique favorable aura tendance à augmenter en fréquence tant qu'il reste dans ce contexte, c'est-à-dire jusqu'à ce que la recombinaison casse les associations génétiques avec le gène soumis à la sélection. L'effet de la variance du succès reproducteur est donc d'autant plus fort qu'il est cumulé sur plusieurs générations (Robertson 1961; Santiago et Caballero 1995) jusqu'à la rupture complète des associations de gènes à plusieurs locus. Ce que l'on pourrait donc nommer « l'effet Robertson » résume le fait que plus un caractère sélectionné est héritable et plus l'intensité de la sélection est forte, plus les changements de fréquences alléliques sont corrélés d'une génération sur l'autre et plus l'intensité de la dérive augmente à travers tout le génome (Robertson 1961; Santiago et Caballero 1995; Woolliams *et al.* 1993). Le nombre de copies qu'un gène à un locus donné transmet à la génération suivante ne varie donc pas seulement du fait de la variation stochastique du nombre de descendants des individus qui composent la population, mais également du fait d'effets sélectifs qui se produisent à d'autres locus, non liés physiquement au locus considéré. Toutefois, le phénomène décrit par Robertson (1961) ne tient compte ni de la liaison physique des gènes entre eux (Santiago et Caballero 1998) ni de l'effet de la taille finie des populations sur les déséquilibres de liaison entre

[12] De la difficulté de traduire le terme anglais *background*, que l'on peut traduire ici par *fond* ou par *contexte*...

les locus (voir Hill et Robertson 1968; Ohta et Kimura 1969, et le § 2.3, pour une discussion sur les effets de la taille des populations sur la variance du déséquilibre de liaison).

3.3.2 Quelques exemples

A titre d'exemple, la sélection d'arrière plan[13], qui résulte de la sélection agissant contre un flux constant de mutations délétères (Charlesworth *et al.* 1993), réduit la taille efficace à des locus non-sélectionnés qui se trouvent à proximité des régions chromosomiques où la sélection agit (Charlesworth *et al.* 1997). De la même façon, la substitution de mutations avantageuses à un locus diminue également la variabilité neutre attendue aux locus liés (Barton 1995; Kaplan *et al.* 1989; Maynard Smith et Haigh 1974; Wiehe et Stephan 1993). Ce phénomène d'*auto-stop*[14] se traduit par une réduction de la variation neutre à des locus *entraînés* par la sélection agissant à d'autres locus, d'une façon tout à fait comparable à celle d'une réduction de taille efficace (Gillespie 2000). Dès lors, la notion de taille efficace prend une dimension *locale*, en ce sens que la taille efficace d'un locus en particulier, ou bien d'une portion du génome, peut être différente de la taille efficace d'un autre locus, ou bien d'une autre portion du génome. Les effets similaires des mutations délétères et des mutations avantageuses sur le polymorphisme neutre à des locus liés font qu'il est difficile, en pratique, d'écarter l'un ou l'autre de ces mécanismes pour expliquer les patrons de variation observés (Charlesworth 1994a; Gillespie 2000).

Il existe encore d'autres effets sélectifs qui peuvent influencer la distribution de la variation neutre aux locus liés. Dans des populations structurées dans l'espace, l'adaptation locale augmente non seulement la différenciation des populations aux locus où la sélection agit, mais également aux locus liés, où les valeurs de F_{ST} peuvent être fortement augmentées par rapport à l'attendu neutre (Charlesworth *et al.* 1997). En revanche, la sélection balancée agissant à un locus permet de maintenir un niveau élevé de variation

[13]Terme traduit de l'anglais, *background selection*.
[14]Terme traduit de l'anglais, *hitchhiking*.

aux marqueurs neutres se trouvant à proximité (Hudson et Kaplan 1988; Strobeck 1983).

La sélection qui agit sur un locus, qu'elle soit dirigée contre des mutations délétères ou qu'elle favorise l'augmentation en fréquence de mutations avantageuses, agit donc également sur la distribution du polymorphisme neutre à d'autres locus qui se trouvent à proximité dans le génome. Cet effet est similaire à une réduction de taille efficace. Or puisqu'une réduction de taille efficace diminue l'efficacité de la sélection, on s'attend à ce que la sélection qui agit à un locus *interfère* avec la sélection agissant à un autre locus, à proximité. Il s'agit d'un phénomène qui a été décrit pour la première fois par Hill et Robertson (1966).

3.3.3 L'effet Hill-Robertson

Hill et Robertson (1966) ont montré que la sélection à un locus diminuait l'efficacité de la sélection à d'autres locus, et ce d'autant plus que les locus étaient physiquement liés. Cet effet, que Felsenstein (1974) nomme, pour la première fois, l'« effet Hill-Robertson » peut être résumé ainsi. Un gène qui se retrouve dans un fond génétique favorable (c'est-à-dire en association avec un gène sélectionné) tend à augmenter en fréquence parmi les individus qui forment la génération suivante. Dans le cas de locus physiquement indépendants, les copies de ce gène vont, du fait de la recombinaison, se retrouver dans des fonds génétiques différents. Lorsque les locus sont liés, en revanche, les associations entre les gènes aux différents locus persistent pendant un plus grand nombre de générations. Un gène, quel qu'il soit, qui se retrouve dans un fond génétique favorable a donc tendance à être entraîné par le même fond génétique pendant plusieurs générations (Felsenstein 1987). La fréquence d'un gène parmi les individus de la génération suivante dépend bien sûr de sa propre valeur sélective. Mais la variance du succès reproducteur induite par la sélection qui agit sur son propre fond génétique, diminue l'efficacité de la sélection à ce locus (Felsenstein 1974). Dans le cas où la liaison génétique n'est pas trop forte, l'effet de la sélection à un locus se traduit par une réduction de la taille efficace aux locus qui lui sont liés.

La probabilité de fixation d'un allèle avantageux est donc plus faible lorsque la sélection agit à un autre locus, d'une manière tout à fait équivalente à celle d'une réduction de taille efficace (Barton 1995). Barton (1994, 1995, 1998) a montré récemment que cet argument était valable dans le cas général de locus faiblement liés les uns aux autres et pour différentes formes de sélection. Dans le cas de locus fortement liés, en revanche, Barton (1995) montre que l'effet de la sélection sur un locus lié ne peut être résumé par une diminution de taille efficace.

3.4 Conclusion

La notion de taille efficace est intimement liée au modèle de Wright-Fisher d'évolution des fréquences alléliques dans une population de taille finie. J'ai montré l'influence de différents types de structure (spatiale, temporelle, etc.) sur la distribution de la variation neutre. Dans la dernière partie, j'ai également souligné que la sélection qui agit à un ou plusieurs locus influence indirectement la distribution de la variation neutre, à des locus situés à proximité dans le génome. Ceci montre bien que la notion de taille efficace prend ici une dimension *locale* en ce sens que la taille efficace d'une portion du génome n'est pas nécessairement identique à celle d'une autre portion du génome.

Dans les chapitres 2 et 3, j'ai montré comment la distribution de la variation génétique dépendait d'un certain nombre de paramètres d'intérêt des modèles étudiés, tels que l'intensité de la force de dérive (intimement liée à la notion de taille efficace) ou encore les flux de gènes entre populations. Comme je l'ai évoqué à plusieurs reprises il est également possible de réaliser des inférences sur les valeurs de ces paramètres d'intérêt, à partir de mesures de la variation génétique. Dans le quatrième chapitre, je présenterai donc quelques applications pratiques qui découlent de l'analyse de ces modèles de populations structurées.

Chapitre 4

Applications

Dans ce dernier chapitre, je présenterai les principales applications des modèles que j'ai développés au cours de ma thèse. Je présenterai tout d'abord une méthode d'estimation de la taille efficace et des flux géniques à partir de mesures du déséquilibre d'identité (voir le § 2.3). Ensuite, je présenterai une méthode d'estimation des taux de migration spécifiques à un sexe. Enfin, je détaillerai une méthode de détection des marqueurs génétiques soumis à l'action, directe ou indirecte, de la sélection. Le lecteur intéressé par plus de détails sur ces différentes approches pourra se reporter (dans l'ordre des différentes sections qui composent ce dernier chapitre) aux annexes F, G, H et I.

4.1 Estimation de la taille efficace

4.1.1 Pourquoi mesurer la taille efficace et les flux de gènes ?

Comme nous venons de le voir (§ 3.2-3.3) la taille efficace d'une population est généralement différente du nombre total d'individus que l'on peut y dénombrer (voir également, pour revue, Caballero 1994). Ceci est vrai également dans le cas de populations subdivisées (voir le § 3.2, ainsi que Wang et Caballero 1999, pour une revue). Or la taille efficace a été dé-

finie comme un paramètre des modèles d'évolution des fréquences alléliques afin, justement, de décrire les propriétés de ces modèles (§ 3.1). L'efficacité de la sélection naturelle (ou artificielle) pour maintenir des mutations avantageuses ou pour favoriser l'augmentation en fréquence de nouvelles mutations favorables dépend ainsi de la taille efficace, N_e : la sélection sera d'autant moins efficace que la valeur de N_e est (localement) faible. D'autre part, de faibles valeurs de N_e limitent également la capacité de la sélection à débarrasser le génome de mutations légèrement délétères. Dans les populations de petite taille, ou plus précisément dont la taille efficace est faible, les mutations légèrement délétères ont en effet une probabilité plus forte d'être fixées par hasard, sous le simple effet de la dérive génétique. Or la fixation de ces mutations délétères diminue la valeur sélective moyenne de la population et réduit donc encore plus la taille efficace (Gabriel et Bürger 1994; Lynch et Gabriel 1990). Ce phénomène d'entraînement, qui a été décrit sous le terme de *fonte mutationnelle*[1], pourrait conduire à l'extinction des populations (Lande 1994; Lynch *et al.* 1995a,b; van Noordwijk 1994). Une belle démonstration du rôle de l'augmentation de la consanguinité sur les risques d'extinction a été apportée par une équipe de chercheurs finlandais travaillant sur le papillon damier, *Melitea cinxia* (Saccheri *et al.* 1998). Pendant dix ans, ces chercheurs ont étudié les populations de cette espèce dans les îles Åland, un ensemble de plusieurs centaines d'îles du sud-ouest de la Finlande, plus ou moins isolées les unes des autres et de tailles variables. A partir de l'étude de 42 de ces populations, dont 7 se sont éteintes entre 1995 et 1996, ils ont montré que le facteur qui expliquait le mieux la variance observée dans les taux d'extinction de ces populations était le niveau d'hétérozygotie individuelle mesurée sur des marqueurs enzymatiques. D'autre part, à travers une méta-analyse de données issues de la littérature, Frankham (1995, 1998) a montré qu'il existait une corrélation positive entre le risque d'extinction et la consanguinité dans des populations de laboratoire de souris et de drosophiles.

D'autre part, outre le fait que la dispersion dans le temps ou dans l'espace constitue le seul recours des individus pour subsister dans des habitats fragmentés et instables (Hanski et Gilpin 1997), elle constitue également

[1]Terme traduit de l'anglais, *mutational meltdown*.

un élément essentiel de l'échange d'information et de variabilité génétique (Slatkin 1987). Ces flux de gènes, ou de complexe de gènes, sont cruciaux pour déterminer la dynamique de l'adaptation locale (Maynard Smith et Hoekstra 1980).

On peut donc voir plusieurs intérêts à mesurer conjointement la taille efficace et les flux de gènes. Il peut, par exemple, s'agir d'évaluer les risques d'extinction associés à la petite taille des populations, ou bien de caractériser l'intensité de goulots d'étranglement. Mais de façon plus générale, cette évaluation conjointe des taux auxquels les gènes coalescent et sont échangés entre populations locales peut permettre de mieux appréhender les éléments favorisant ou non l'adaptation locale des populations à leur environnement (Gandon *et al.* 1996, 1998; Ronce et Kirkpatrick 2001).

La littérature compte nombre d'exemples d'estimation de flux de gènes à partir des données de polymorphisme de marqueurs neutres (Slatkin et Barton 1989). Or les approches qui reposent sur les mesures du paramètre F_{ST}[2] ou bien sur les fréquences d'allèles rares dans les populations (Slatkin 1985) permettent l'estimation du produit Nm, c'est-à-dire le produit de la taille locale des populations par le taux de migration par génération (Slatkin et Barton 1989). En ce sens, l'effet de la dérive à l'échelle des populations locales ne peut pas être distingué de celui de l'influx de gènes dans ces populations. Les méthodes d'estimation les plus récentes, qui s'appuient sur la théorie de la coalescence, permettent une estimation conjointe des produits $\mathcal{M} = 4Nm$ et $\Theta = 4N\mu$, pour chaque population dans un modèle à nombre fini d'îles (Beerli et Felsenstein 1999, 2001). Tufto *et al.* (1996) ont proposé une méthode par maximum de vraisemblance pour estimer la matrice complète de migration entre paires de populations, à partir de la distribution des fréquences alléliques. Cette méthode permet *a priori* d'estimer des taux de migration à courtes et moyennes distances, mais la taille efficace est considérée comme un paramètre de valeur connue dans le modèle. La raison pour laquelle ces méthodes ne permettent que l'estimation du produit Nm est que la vraisemblance d'un échantillon de fréquences alléliques ne

[2]Voir l'équation (2.16), p. 34 pour le lien entre la valeur d'équilibre de F_{ST} dans un modèle à nombre fini d'îles et les flux de gènes.

dépend pas d'autre chose que du produit de ces deux paramètres (Rousset 2001b). Il n'est donc pas possible, à partir des données de polymorphisme à un locus, d'estimer indépendamment les tailles locales des populations et les proportions d'individus immigrants.

Malgré l'importance des questions posées quant aux risques encourus par les populations fragmentées et/ou de petite taille, notamment dans le domaine de la biologie de la conservation, les méthodes d'estimation de la taille efficace, comme des flux de gènes, font donc face à des difficultés souvent majeures (voir, par exemple, Schwartz *et al.* 1999; Waples 1989).

4.1.2 Comment estimer la taille efficace ?

Une formulation plus juste de cette question devrait intégrer la question qui s'inscrit en filigrane (voir § 3.1) : « Quelle taille efficace estimer ? ». On trouve dans la littérature deux familles de méthodes d'estimation de la taille efficace.

La première repose sur la mesure de paramètres démographiques. En effet, puisque la différence entre la taille efficace et la taille réelle des populations dépend du régime de reproduction, de la variance de succès reproducteur, etc. (voir les § 3.2-3.3), la taille efficace peut être calculée à partir de l'évaluation directe de ces paramètres démographiques (Nunney et Elam 1994). Mais les données démographiques nécessaires à l'évaluation directe de la taille efficace sont souvent difficiles à obtenir dans les situations naturelles (Crow et Denniston 1988). De plus, l'estimation de ces paramètres lors d'une seule saison ne permet pas de prendre en compte d'éventuelles variations interannuelles de la taille des populations, ce qui revient à surestimer la taille efficace (Vucetich *et al.* 1997).

La seconde famille de méthodes d'estimation de la taille efficace rassemble les méthodes d'estimation indirecte, c'est-à-dire les méthodes qui reposent sur l'analyse du polymorphisme génétique sur des marqueurs neutres. Ces estimations indirectes reposent sur la mesure des changements temporels de fréquences alléliques (Nei et Tajima 1981; Pollak 1983; Waples 1989; Williamson et Slatkin 1999), sur la mesure des excès en hétérozygotes

dans la descendance (Luikart et Cornuet 1999; Pudovkin *et al.* 1996) ou bien encore sur la mesure de la variance du déséquilibre de liaison (Bartley *et al.* 1992; Hill 1981; Waples 1991). Toutes ces méthodes ont été construites dans un modèle de population isolée. Elles mesurent donc une taille efficace locale.

La mesure des changements temporels de fréquences alléliques (Nei et Tajima 1981; Pollak 1983; Waples 1989) permet l'estimation de la taille efficace de variance (voir § 3.1.1). Les estimateurs ont en général une variance élevée (Waples 1989). Williamson et Slatkin (1999) et Anderson *et al.* (2000) ont récemment proposé une approche du problème par la méthode du maximum de vraisemblance.

Pudovkin *et al.* (1996) ont proposé un estimateur du nombre de parents dans une population d'individus diploïdes, à sexes séparés, à partir de l'excès d'hétérozygotes par rapport à l'équilibre de Hardy-Weinberg. Dans une population de taille finie, en effet, toute différence entre les fréquences alléliques des individus mâles et femelles[3] conduit à un excès d'hétérozygotes (par rapport aux proportions attendues à l'équilibre de Hardy-Weinberg) parmi les descendants. Luikart et Cornuet (1999) ont montré que cette méthode ne donnait de résultats satisfaisants que lorsque le nombre de parents était très faible.

Enfin, la taille efficace de consanguinité, a été également estimée à partir de la mesure de la variance des corrélations des fréquences alléliques à différents locus, c'est-à-dire de la mesure de la variance du déséquilibre de liaison (Hill 1981; Waples 1991). Dans une population de taille finie, nous l'avons vu (§ 3.3.3), la dérive génétique crée des corrélations entre les fréquences des allèles à différents locus. De plus, la valeur de ces corrélations dépend de la taille efficace locale (Hill 1981; Waples 1991; Weir et Hill 1980a). La seule application de cette méthode que je connaisse est décrite dans un article de Bartley *et al.* (1992).

Il est fort probable également que ces différentes méthodes d'estimation

[3]Des différences de fréquences alléliques entre les gamètes issus des mâles et ceux issus des femelles peuvent venir d'effets d'échantillonnage dans un ensemble de taille finie de gamètes ou bien de différences de fréquences alléliques parmi les parents (Wang 1996). Voir également le § 2.2.1.

mesurent des tailles efficaces sur des échelles de temps différentes. Ainsi, les méthodes qui reposent sur la mesure des changements temporels de fréquences alléliques (Nei et Tajima 1981; Pollak 1983; Waples 1989) permettent l'estimation d'une taille efficace sur une ou quelques générations (correspondant à l'intervalle de temps séparant les mesures réalisées). En revanche, les estimations de la taille efficace qui reposent sur la mesure de la variance du déséquilibre de liaison (Hill 1981; Waples 1991) correspondent à des échelles de temps probablement plus longues (et ce d'autant plus que les locus utilisés sont liés physiquement sur le chromosome).

4.1.3 Estimer l'effectif efficace et le taux d'immigration à partir des déséquilibres d'identité

La méthode que je vais maintenant développer repose sur l'estimation des probabilités d'identité à un et deux locus. Comme nous l'avons vu précédemment, le déséquilibre d'identité dépend de la taille des populations locales et du taux d'immigration (voir la figure 4.1). Plus précisément, le mode de la fonction du déséquilibre d'identité en fonction du taux de migration dépend du produit Nm. Comme on peut le voir en effet sur la figure 4.1, les courbes qui représentent le déséquilibre d'identité en fonction du taux d'immigration ont toutes un mode pour une valeur de $Nm \approx 0.2 - 0.3$. D'autre part, la valeur du déséquilibre en ce mode dépend de la taille de la population. Par conséquent si la valeur du produit Nm peut être estimée de façon indépendante, par exemple à partir d'une mesure de F_{ST} ou d'une mesure équivalente (§ 2.1.3), alors la valeur du déséquilibre d'identité peut permettre d'estimer la taille efficace locale.

Considérons une population focale i composée de N_i individus diploïdes, qui reçoit, chaque génération une proportion m_i de gènes venant d'autres populations. En reprenant les notations adoptées précédemment (p. 26), dans les équations (2.3-2.6), on peut définir le paramètre F_i dans une population focale comme

$$F_i = \frac{Q_{1,i} - Q_2}{1 - Q_2} \tag{4.1}$$

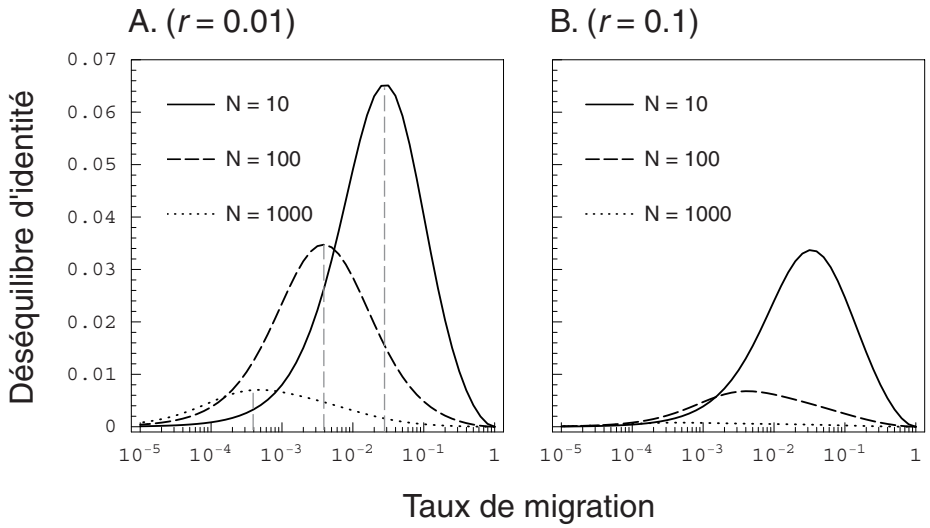


Figure 4.1: Le déséquilibre d'identité en fonction de la taille efficace et du taux d'immigration. Les valeurs attendues du déséquilibre d'identité sont données pour différentes tailles des sous-populations ($N = 10$, 100 and 1000) et pour différents taux de recombinaison. A. $r = 0.01$. B. $r = 0.1$. Les croisements se font au hasard dans chaque sous-population. Le modèle de mutation est à nombre infini d'allèles, et $\mu = 10^{-6}$. En A., Le mode de chaque courbe est matérialisé par une ligne verticale grisée (pointillée) et correspond toujours à une valeur du produit $Nm \approx 0.2 - 0.3$. Connaissant la valeur de ce produit, on peut déduire la taille efficace de la population (axes des ordonnées).

où $Q_{1,i}$ est la probabilité d'identité d'une paire de gènes dans la population focale i. On notera que la somme pondérée des F_i sur toutes les populations locales constituant une métapopulation est la corrélation intra-classe pour les probabilités d'identité par état, c'est-à-dire F_{ST} (Cockerham et Weir 1987; Rousset 1996). A l'équilibre, on peut montrer que

$$F_i = \frac{(1-\mu)^2 a_i}{(1-\mu)^2 a_i (2-(1-\mu)^2 a_i) + N_i (1-(1-\mu)^2 a_i)(2-(1-\mu)^2 a_i s)} \quad (4.2)$$

où $a_i = (1-m_i)^2$ (modèle à nombre infini d'îles). Pour de faibles taux de mutation, ce paramètre est lié au produit $N_i m_i$, par la relation $F_i \approx 1/(1+4N_i m_i)$. Dans un modèle à nombre fini K d'allèles, le paramètre F_i a

une expression identique à (4.2), avec $(1-\mu)^2$ remplacé par $(\nu K-1)/(K-1)$, où $\nu = (1-\mu)^2 + \mu^2/(K-1)$. Ce résultat suggère que, pour de faibles taux de mutation, F_i ne dépend pas du modèle de mutation (voir le tableau 4.1, p. 93).

Estimation des probabilités d'identité à un locus

Pour un allèle u, on définit la variable indicatrice x_{ijku} pour le $k$ième gène, avec $k = (1, 2)$, du $j$ième individu de la population i. $x_{ijku} = 1$ si le gène est dans l'état allélique u, et $x_{ijku} = 0$, autrement. On note P_i^u la fréquence de l'allèle u dans la population i, soit $P_i^u = \mathcal{E}\left(x_{ijku} | \mathbf{P}\right)$, où $\mathcal{E}\left(\quad | \mathbf{P}\right)$ est l'espérance conditionnelle aux fréquences alléliques $\mathbf{P}$ dans la population. Si l'on considère les moments d'ordre 2 de la variable aléatoire x_{ijku}, on a $\mathcal{E}\left(x_{ijku}^2 | \mathbf{P}\right) = P_i^u$ et $\mathcal{E}\left(x_{ijku}x_{ij'ku} | \mathbf{P}\right) = (P_i^u)^2$, où $j' \neq j$. La somme sur tous les allèles nous donne la probabilité d'identité par état d'une paire de gènes pris au hasard parmi des individus distincts

$$Q_{1,i} = \mathcal{E}\left[\sum_{u=1}^{K} (P_i^u)^2\right] \tag{4.3}$$

où $\mathcal{E}\left(\quad \right)$ est maintenant l'espérance sur toute la distribution des fréquences alléliques $\mathbf{P}$. On définit P_i^{uu} comme la fréquence des homozygotes pour l'allèle u dans la population i. Par conséquent, $\mathcal{E}\left(x_{ijku}x_{ijk'u} | \mathbf{P}\right) = (P_i^{uu})$, où $k' \neq k$. Finalement, la probabilité d'identité par état d'une paire de gènes pris au hasard dans des populations différentes est égale à

$$Q_2 = \mathcal{E}\left[\sum_{u=1}^{K} (x_{ijku}x_{i'jku})\right] \tag{4.4}$$

avec $i' \neq i$. Un estimateur non biaisé de la fréquence de l'allèle u parmi n individus nous est donné par $\hat{P}_i^u = \sum_{j=1}^{n}\sum_{k=1}^{2} x_{ijku}/(2n)$. En élevant cette expression au carré et en prenant son espérance, on a $\mathcal{E}\left[\left(\hat{P}_i^u\right)^2 | \mathbf{P}\right] =$

$[P_i^u(1+2(n-1)P_i^u)+P_i^{uu}]/(2n)$. Par conséquent,

$$\hat{Q}_{1,i}=\sum_{u=1}^{K}\left[\hat{P}_i^u(2n\hat{P}_i^u-1)-\hat{P}_i^{uu}\right]/\left[2(n-1)\right] \tag{4.5}$$

où $\hat{P}_i^{uu}=\sum_{j=1}^{n}\sum_{k'\neq k}^{2}x_{ijku}x_{ijk'u}$ est un estimateur non biaisé de la fréquence des homozygotes pour l'allèle u. Un estimateur non biaisé de la probabilité d'identité par état d'une paire de gènes pris dans des populations différentes nous est donné par

$$\hat{Q}_2=\sum_{u=1}^{K}\sum_{i'\neq i}^{d}\hat{P}_i^u\hat{P}_{i'}^u/\left[d(d-1)\right] \tag{4.6}$$

pour un échantillon de d populations. Finalement, on définit un estimateur de F_i comme

$$\hat{F}_i=\frac{\sum_u\left[\left(\hat{P}_i^u(2n\hat{P}_i^u-1)-\hat{P}_i^{uu}\right)/[2(n-1)]-\sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]\right]}{1-\sum_u\sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]} \tag{4.7}$$

Un estimateur multilocus du paramètre F_i est obtenu en prenant le rapport de la somme des numérateurs monolocus de (4.7) sur la somme des dénominateurs de (4.7) (voir Reynolds *et al.* 1983; Weir et Cockerham 1984, pour une approche similaire pour les estimateurs multilocus de F_{ST}).

Estimation des probabilités d'identité à deux locus

Pour un allèle u, on définit la variable indicatrice x_{iju} pour le $j$ième gène du $i$ième individu au premier locus, et la variable $y_{i'j'v}$ pour le j'ième gène du i'ième individu au second locus[4]. $x_{iju}=1$ si le gène i au premier locus est dans l'état allélique u et $x_{iju}=0$, autrement. De même $y_{i'j'v}=1$ si le gène i' au second locus est dans l'état allélique v et $y_{i'j'u}=0$, autrement. On note P_v^u la fréquence des haplotypes portant les allèles u et v (allèles

[4]Dans cette partie, et dans le reste du texte, la référence à la population focale est implicite. Dans un souci de clarté, tous les indices référant à la population focale ont été enlevés.

en phase[5]. $P_v^u = \mathcal{E}\left(x_{iju}y_{ijv}|\ \mathbf{P}\right)$, où $\mathcal{E}\left(\quad|\ \mathbf{P}\right)$ est l'espérance conditionnelle aux fréquences haplotypiques $\mathbf{P}$. On définit $P_{\cdot v}^{u\cdot} = \mathcal{E}\left(x_{iju}y_{ij'v}|\ \mathbf{P}\right)$ comme la fréquence des paires d'allèles u et v pris dans le même individu dans deux haplotypes distincts du même individu[6]. Comme dans le cas des estimateurs des probabilités d'identité à un locus, les probabilités d'identité à deux locus peuvent être exprimées en calculant la somme sur toutes les paires d'états alléliques u et v présents dans la population, et en prenant l'espérance sur la distribution des fréquences haplotypiques $\mathbf{P}$, soit

$$\begin{aligned} \phi &= \mathcal{E}\left[\sum_{u,v}\left(P_v^u\right)^2\right] \\ \gamma &= \mathcal{E}\left[\sum_{u,v}\left(P_v^u P_{\cdot v}^{u\cdot}\right)\right] \\ \delta &= \mathcal{E}\left[\sum_{u,v}\left(P_{\cdot v}^{u\cdot}\right)^2\right] \end{aligned} \tag{4.8}$$

Or en pratique, puisque la phase gamétique n'est en général pas connue, il n'est pas possible d'estimer P_v^u et $P_{\cdot v}^{u\cdot}$ à partir des seules données génotypiques (voir p. 2.3.1). Mais d'après l'équation (4.8), la probabilité d'identité composite Φ, définie comme la probabilité qu'à deux locus, les gènes échantillonnés dans deux individus distincts soient identiques (équation 2.30), peut s'exprimer ainsi

$$\Phi = \mathcal{E}\left[\sum_{u,v}\left(P_{\bar{v}}^{\bar{u}}\right)^2\right] \tag{4.9}$$

où $P_{\bar{v}}^{\bar{u}}$ est la moyenne des fréquences des haplotypes en *phase* et en *répulsion*, c'est-à-dire $P_{\bar{v}}^{\bar{u}} = \left(P_v^u + P_{\cdot v}^{u\cdot}\right)/2$, une quantité estimable à partir de simples données génotypiques. Un estimateur non biaisé de $P_{\bar{v}}^{\bar{u}}$ est donné par $\hat{P}_{\bar{v}}^{\bar{u}} = \sum_{i=1}^{n}\sum_{j,j'}^{2} x_{iju}y_{ij'v}/(4n)$. L'espérance du carré de cette expression,

[5] J'insiste sur le fait que la définition que j'ai adoptée du terme *phase* n'implique pas que les allèles sont sur le même chromosome, mais plutôt qu'ils sont issus du même gamète parental. Voir p. 53.

[6] De même, la définition du terme *répulsion* implique que les allèles sont issus de gamètes parentaux distincts. Voir la note précédente.

conditionnelle aux fréquences $\mathbf{P}$ des haplotypes dans la population nous est donnée par

$$\mathcal{E}\left[\left(\hat{P}_{\bar{v}}^{\bar{u}}\right)^2 \mid \mathbf{P}\right] = \left[P_{\bar{v}}^{\bar{u}}\left[1+4(n-1)P_{\bar{v}}^{\bar{u}}\right] + P_{vv}^{uu} + P_{vv}^{u\cdot} + P_{v\cdot}^{uu}\right]/(4n) \quad (4.10)$$

où P_{vv}^{uu} est la fréquence des double homozygotes pour les allèles u et v, $P_{v\cdot}^{uu}$ est la fréquence des individus qui sont homozygotes pour l'allèle u au premier locus et qui possèdent une copie de l'allèle v au second locus, et où $P_{vv}^{u\cdot}$ est la fréquence des individus qui sont homozygotes pour l'allèle v au second locus et qui possèdent une copie de l'allèle u au premier locus. Par conséquent, un estimateur non biaisé de Φ nous est donné par

$$\hat{\Phi} = \sum_{u,v}\left[\hat{P}_{\bar{v}}^{\bar{u}}\left(4n\hat{P}_{\bar{v}}^{\bar{u}} - 1\right) - \hat{P}_{vv}^{uu} - \hat{P}_{v\cdot}^{uu} - \hat{P}_{vv}^{u\cdot}\right] / \left[4(n-1)\right] \quad (4.11)$$

où $\hat{P}_{vv}^{uu}$ est la fréquence observée des double homozygotes pur les allèles u et v, $\hat{P}_{v\cdot}^{uu}$ (respectivement $\hat{P}_{vv}^{u\cdot}$) est la fréquence observée des individus homozygotes pour l'allèle u au premier locus (respectivement v au second locus) et qui possèdent une copie de l'allèle v au second locus (respectivement u au premier locus). D'après l'équation (2.31), un estimateur du déséquilibre d'identité aux locus i et j est donc

$$\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j} \quad (4.12)$$

L'espérance de cette statistique est en effet

$$\mathcal{E}\left(\hat{\eta}_{S,ij}\right) = \Phi_{ij}\left[1 - \frac{2}{n(n-1)}\right] - \frac{4(n-2)}{n(n-1)}\Gamma_{ij} - \frac{(n-2)(n-3)}{n(n-1)}\delta_{4ij} \quad (4.13)$$

où $\Gamma_{ij} = (\gamma_{3ij} + \delta_{3ij})/2$, soit pour de grandes tailles d'échantillon,

$$\mathcal{E}\left(\hat{\eta}_{S,ij}\right) \approx \Phi_{ij} - \delta_{4ij} \quad (4.14)$$

Pour des tailles d'échantillon suffisantes, $\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}$ est un estimateur non biaisé du déséquilibre d'identité aux locus i et j. De même, un

estimateur du déséquilibre d'identité ramené aux diversités génétiques totales $\eta'_{S,ij}$, dont la définition est donnée par l'équation (2.32), (p. 54) peut être défini par

$$\hat{\eta'}_{S,ij} = \frac{\hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}}{(1-\hat{Q}_{2i})(1-\hat{Q}_{2j})} \tag{4.15}$$

(voir Hedrick 1987; Ohta 1980, pour une approche semblable). Comme nous l'avons vu (§ 2.3.2, figure 2.13), cette mesure du déséquilibre d'identité η'_S ne dépend pas du modèle de mutation considéré, et ce même lorsque les croisements ne se font pas au hasard. Notons également, qu'un estimateur multilocus de $\hat{\eta'}_S$ peut être obtenu en prenant le rapport des sommes (sur les locus) des numérateurs monolocus de (2.32) sur la somme des dénominateurs de (2.32) (voir Ohta 1980, pour une approche similaire). Les estimateurs de F et de η'_S ont été évalués par la simulation d'un modèle à nombre infini d'îles (tableau 4.1)[7]. Les valeurs moyennes des estimateurs sont proches des valeurs des paramètres qu'ils estiment. De plus, elles ne dépendent pas ou très peu du modèle de mutation considéré. En revanche, l'effet de la taille de l'échantillon est plus prononcé pour les mesures du déséquilibre d'identité que pour les mesures des probabilités d'identité à un locus (tableau 4.1).

Estimation de la taille efficace

Nous venons de définir deux statistiques $\hat{F}$ et $\hat{\eta}'_S$, dont les espérances sont des fonctions des paramètres d'intérêt de notre modèle de population, à savoir la taille efficace et le taux d'immigration. De plus, leurs espérances ne dépendent pas ou très peu des paramètres de nuisance du modèle (voir la note 27 au bas de la page 56) comme le modèle de mutation (figure 2.13, tableau 4.1). Malgré tout, il faut garder à l'esprit que les valeurs des paramètres F et η'_S dépendent d'autres paramètres de notre modèle, à savoir le régime de reproduction (ici, le taux d'autofécondation) et le taux de recombinaison entre paires de locus. En toute rigueur, il s'agit de paramètres de nuisance

[7] Je rappelle que la référence à la population focale est implicite dans cette partie du texte. Tout indice référant à une population focale a été enlevé, mais il est bien clair que les estimateurs de F et de η'_S sont spécifiques à une population focale.

Tableau 4.1: Propriétés des estimateurs des probabilités d'identité à un et deux locus. La moyenne arithmétique et l'écart type (s) des estimateurs $\hat{F}$ et $\hat{\eta}'_S$ sont donnés pour différentes valeurs de N dans différents modèles de mutation. Les tailles d'échantillon (n) sont également indiquées. $Nm = 1$ et $\mu = 10^{-6}$. Le modèle de mutation est donné par la valeur de K, soit le nombre d'allèles simulés, par locus. Les propriétés des estimateurs ont été calculées à partir de 1000 simulations pour 8 locus non liés.

				$\hat{F}$			$\hat{\eta}'_S$	
K	N	n	F	$Mean$	s	η'_S	$Mean$	s
2	20	25	0.188	0.188	0.086	0.0052	0.0051	0.0075
	20	50		0.187	0.081	0.0054	0.0054	0.0056
	50	25	0.195	0.197	0.087	0.0023	0.0022	0.0060
	50	50		0.197	0.082	0.0024	0.0026	0.0043
5	20	25	0.188	0.188	0.058	0.0052	0.0052	0.0035
	20	50		0.190	0.056	0.0054	0.0054	0.0029
	50	25	0.195	0.195	0.056	0.0023	0.0024	0.0027
	50	50		0.193	0.050	0.0024	0.0024	0.0018
10	20	25	0.188	0.190	0.054	0.0052	0.0053	0.0031
	20	50		0.189	0.051	0.0054	0.0054	0.0023
	50	25	0.195	0.194	0.046	0.0023	0.0023	0.0019
	50	50		0.196	0.047	0.0024	0.0025	0.0015
∞	20	25	0.188	0.187	0.046	0.0052	0.0051	0.0022
	20	50		0.190	0.043	0.0054	0.0055	0.0020
	50	25	0.195	0.194	0.040	0.0023	0.0023	0.0014
	50	50		0.197	0.039	0.0024	0.0024	0.0011

dans notre modèle, qui doivent être estimés à partir de données indépendantes. Dans tous les résultats que je présente ici, les taux de recombinaison et le taux d'autofécondation sont supposés être connus et sont traités comme des paramètres des modèles de simulation.

La méthode d'estimation est la suivante. Le principe est de résoudre numériquement un système de deux équations à deux inconnues. Pour un ensemble de valeurs de N, on déduit une valeur de m à partir de l'équation (4.2) et de la valeur observée de F. Puis, pour chaque couple de valeurs de

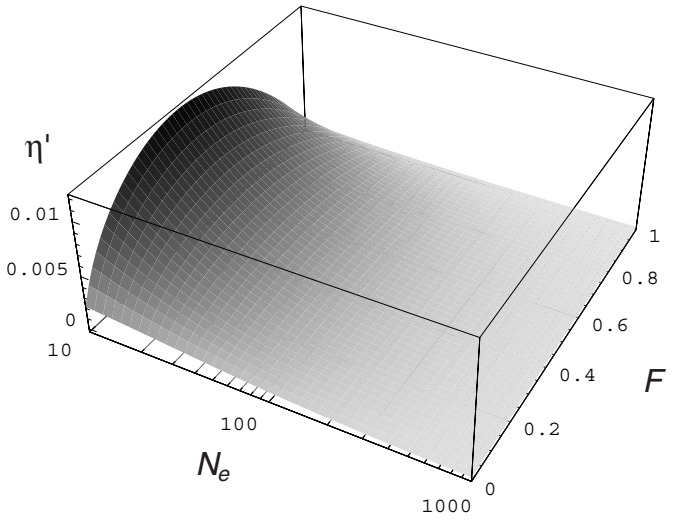


Figure 4.2: Déséquilibre d'identité η'_S attendu à l'équilibre, en fonction de la taille efficace et de F, dans un modèle à nombre infini d'allèles. Les individus se croisent au hasard dans la population. $N_e \cdot m = 1$, $\mu = 10^{-6}$. Noter l'échelle logarithmique sur l'axe des abscisses.

N et de m ainsi obtenues, la valeur d'équilibre de η'_S est calculée[8] et comparée à $\hat{\eta}'_S$. Le couple de valeurs de N et de m pour lequel on obtient la meilleure adéquation entre η'_S et $\hat{\eta}'_S$ est la solution du système d'équation considéré. Ainsi, la figure 4.2 représente le déséquilibre d'identité en fonction de la taille efficace et du taux d'immigration. A une valeur donnée de F et de N, correspond une seule valeur de η'_S.

Évaluation de la méthode d'estimation

La méthode que je viens d'évoquer a été évaluée par des simulations stochastiques. Tous les résultats que je vais maintenant présenter concernent

[8] D'après les équations (4.13) et (4.15), la valeur d'équilibre de η'_S est donnée par

$$\mathcal{E}\left(\hat{\eta}'_{S,ij}\right) = \frac{\Phi_{ij}\left[1 - \frac{2}{n(n-1)}\right] - \frac{4(n-2)}{n(n-1)}\Gamma_{ij} - \frac{(n-2)(n-3)}{n(n-1)}\delta_{4ij}}{\left(1 - Q_{2i}\right)\left(1 - Q_{2i}\right)} \tag{4.16}$$

l'estimation de la taille efficace locale, ainsi que du taux d'immigration par génération, dans une population focale, élément d'un modèle en îles. Que ce soit dans le cas d'un modèle à nombre fini ou infini d'îles[9], les résultats que je présenterai concernent donc toujours l'estimation des paramètres dans une seule population, plus ou moins isolée d'autres populations.

Les résultats des simulations dans le cadre du modèle à nombre d'allèles infini ont montré que la taille efficace est sous-estimée lorsque seulement 4 locus sont génotypés ($\overline{\widehat{N}} = 43.32$ pour $N = 50$ et $\overline{\widehat{N}} = 70.02$ pour $N = 100$). En revanche, les moyennes des valeurs estimées sont très proches des valeurs réelles (simulées), lorsque 8 locus au moins sont génotypés. La variance des estimateurs est en général élevée. L'augmentation du nombre de locus typés diminue la variance des estimateurs. Des résultats similaires ont été obtenus dans le cas du modèle à nombre fini d'allèles, bien que le nombre de locus nécessaires pour obtenir une estimation d'égale qualité soit supérieur dans le modèle de mutation à nombre fini d'allèles (pour plus de détails concernant ces résultats, voir l'annexe F).

La figure 4.3 montre les résultats obtenus dans le cas d'un modèle de mutation à nombre infini d'allèles, pour des valeurs différentes du produit Nm. Les distributions des estimations conjointes des tailles efficaces et des taux d'immigration d'une population focale sont représentées[10], pour un effort de génotypage de 8 et 16 locus non liés. Le biais et la variance diminuent avec l'augmentation de la valeur du produit Nm, c'est-à-dire avec l'augmentation de la diversité génétique totale maintenue dans la population. On remarque que les estimations conjointes de la taille efficace et du taux d'immigration sont négativement corrélées. Cela signifie que les estimations de ces deux

[9]En pratique, pour le modèle à nombre infini d'îles, une seule population focale a été simulée. Dans ce cas, la probabilité d'identité Q_2 d'un gène de la population focale avec un gène immigrant a été fixée à zéro dans le cas du modèle de mutation à nombre infini d'allèles et à $1/K$ dans le cas du modèle à nombre fini d'allèles.

[10]Les observations sont classées dans un histogramme à deux dimensions de 100×100 classes. Une distribution de probabilité est calculée en rapportant le nombre d'observations dans chaque catégorie au nombre total d'observations. Les cellules sont alors classées par ordre de probabilité décroissante et, en commençant par les cellules associées aux plus grandes probabilités, les cellules sont associées à la région de confiance de façon séquentielle jusqu'à ce que la probabilité cumulée de l'ensemble des cellules atteigne la valeur de q fixée. Ainsi définie, cette région n'est pas nécessairement continue.

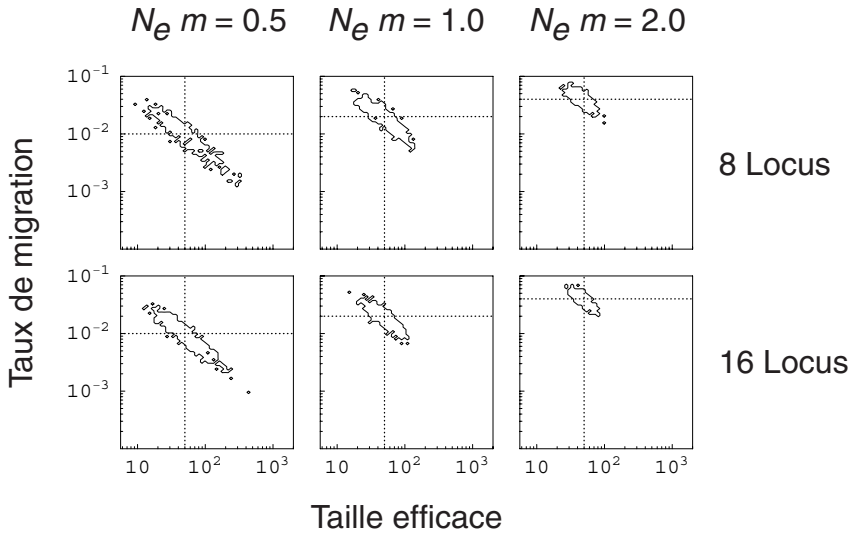


Figure 4.3: Distribution des estimateurs de N_e et m dans un modèle à nombre d'îles infini pour différentes valeurs du produit de paramètres $N \cdot m$ et pour différents nombre de locus génotypés. La taille de la population focale est fixée à $N = 50$ et les mutations se produisent selon un modèle à nombre d'allèles infini ($\mu = 10^{-6}$). Pour chaque jeu de paramètres, une enveloppe a été définie pour contenir 95% de la distribution d'échantillonnage des estimateurs de N_e et m. Les lignes pointillées indiquent les valeurs réelles (simulées) des paramètres. Noter la double échelle logarithmique.

quantités ne sont pas strictement indépendantes et qu'il existe une contrainte sur la valeur estimée du produit Nm.

Les valeurs attendues du déséquilibre d'identité et de la mesure de la divergence de la population focale ont été calculées dans un modèle à nombre infini d'îles[11]. Mais quelle est la robustesse de la méthode présentée ici au patron de migration réel des populations étudiées ? Pour répondre à cette question, un modèle à 20 îles a été simulé, toutes les populations ayant une taille égale. La figure 4.4 montre la distribution conjointe des estimations de la taille efficace et du taux d'immigration pour différentes valeurs de Nm et pour

[11]Pour calculer la valeur du déséquilibre d'identité dans un modèle à nombre fini d'îles, pas moins de 28 probabilités d'identité à deux locus seraient nécessaires, au lieu des 10 définies ici (figure 2.11).

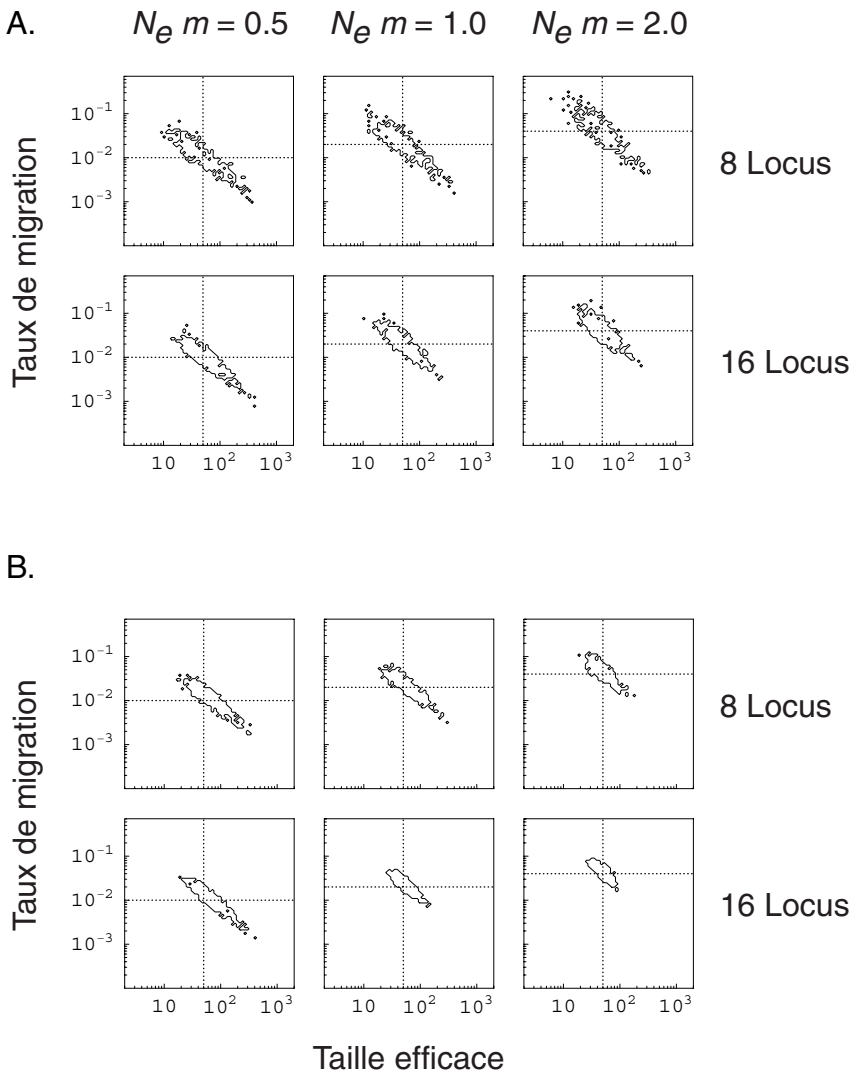


Figure 4.4: Distribution des estimateurs de N_e and m dans un modèle à 20 îles. Les sous-populations sont de taille égale ($N = 50$), échangeant des individus à un taux égal à $m = 0.02$. A. $\mu = 10^{-3}$ B. $\mu = 5 \times 10^{-3}$. Voir la figure 4.3 pour des légendes supplémentaires.

différents efforts de génotypage (en terme du nombre de locus typés). Tout comme dans le cas du modèle à nombre infini d'îles, la qualité des estimations est meilleure lorsque la valeur du produit Nm est plus élevée. Pour $\mu = 10^{-3}$ (figure 4.4A) les distributions conjointes des estimateurs sont toujours plus larges que celles obtenues pour des simulations d'un modèle à nombre infini

d'îles (comparer les figures 4.3 et 4.4A). Pour des valeurs plus élevées du taux de mutation[12] ($\mu = 5 \times 10^{-3}$), le biais et la variance des estimations sont très proches des valeurs obtenues pour des simulations d'un modèle à nombre infini d'îles (figure 4.4B). La qualité de l'estimation dépend donc, comme on pouvait s'y attendre, du niveau de variation maintenue à l'échelle de la population totale (dont la valeur dépend de la taille efficace *globale* et du taux de mutation). Pour des valeurs du produit $dN\mu$ (le produit du nombre total d'individus dans la population globale par le taux de mutation) inférieures à l'unité, le niveau de variation génétique maintenue est trop faible.

4.1.4 Limites de l'approche

La méthode d'estimation des paramètres « démographiques » que sont la taille efficace locale ainsi que l'influx de gènes, repose sur l'hypothèse d'équilibre entre les forces évolutives de mutation, de migration et de dérive. La figure 4.5 montre un ensemble de résultats obtenus dans le cas d'une population échantillonnée alors que l'équilibre stationnaire n'est pas atteint (c'est-à-dire, 10, 20, 30, 40 ou 50 générations après l'état initial). Or après seulement 50 générations, le biais et la variance des estimations conjointes de la taille efficace et du taux d'immigration sont comparables à ceux obtenus pour des populations échantillonnées 1000 générations après l'état initial. Pour 50 individus génotypés à 8 locus non liés, $\hat{N}_e = 52.33$ (s= 0.60) et $\hat{m} = 0.0216$ (s= 0.45). Les estimations du taux d'immigration sont plus biaisées, et la variance des estimateurs est plus forte pour le taux d'immigration que pour la taille efficace. Après seulement 30 ou 40 générations, les estimations de la taille efficace sont comparables à celles obtenues après 1000 générations (voir la figure 4.5). Par conséquent cette méthode permet d'estimer la taille efficace *récente* d'une population. Ceci est vrai dans le cas de locus non liés mais mérite d'être nuancé dans le cas de locus liés, pour lesquels les déséquilibres d'identités atteignent plus lentement leur valeur d'équilibre.

Il existe d'autres limites à cette approche. Tout d'abord si la taille réelle

[12]Ces deux valeurs du taux de mutation sont dans la gamme de valeurs réaliste pour des marqueurs de type microsatellite : voir Estoup *et al.* (1998).

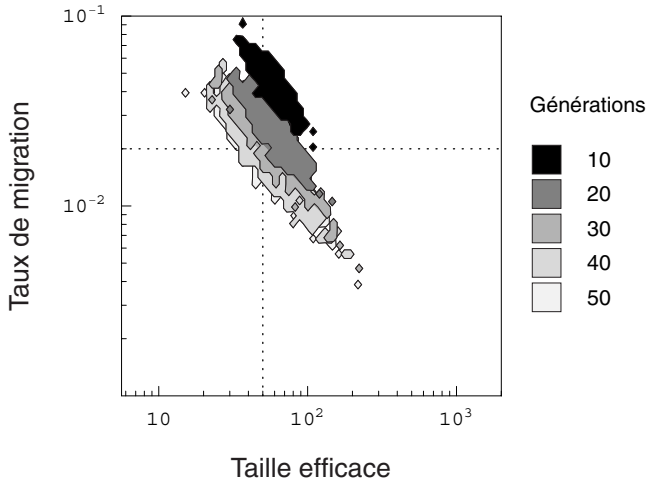


Figure 4.5: Distribution des estimateurs de N_e and m dans un modèle à nombre d'îles infini en situation de non-équilibre. La population focale comporte $N = 50$ individus et reçoit une proportion $m = 0.02$ d'individus migrant par génération. A partir d'une situation initiale où tous les gènes dans la population étaient dans des états alléliques distincts, cinq échantillons de 50 individus ont été prélevés après 10, 20, 30, 40 et 50 générations. Chaque distribution est obtenue à partir de 1000 réplicats du même processus stochastique. Les mutations se produisent selon un modèle à nombre d'allèles infini et $\mu = 10^{-6}$. Les lignes pointillées indiquent les valeurs réelles (simulées) des paramètres. Noter la double échelle logarithmique.

de la population est grande, les déséquilibres d'identité seront vraisemblablement trop faibles pour qu'une estimation correcte puisse être faite (voir, dans le cas de marqueurs non liés physiquement, la figure 4.2 p. 94). Pour estimer la taille efficace d'une population de grande taille, l'utilisation de locus liés est probablement indispensable (voir également Hill et Weir 1994). D'autre part, comme nous l'avons vu précédemment, il existe également une contrainte liée au taux d'immigration. Comme on a pu le voir sur la figure 2.15, le déséquilibre d'identité est très faible lorsque $Nm < 10^{-2}$ ou bien $Nm > 10^2$. Par conséquent, il est certain que l'approche que j'ai détaillée ici n'est valable que dans une certaine gamme de valeurs des paramètres N et m que l'on cherche à estimer.

Enfin, toute cette approche repose sur l'hypothèse que tous les locus répondent de la même manière aux forces évolutives (dérive, migration, recombinaison) que l'on a considérées dans ce modèle. Or différentes formes de sélection peuvent agir directement ou indirectement sur certains locus (voir le § 3.3. Ce point sera abordé plus en détail dans le § 4.3). D'autre part, les taux de recombinaison ne sont pas homogènes le long des génomes (voir, par exemple, Baudry *et al.* 2001). Seules des cartes génétiques précises (disponibles aujourd'hui sur des organismes modèles) pourront nous permettre de classer (de façon indépendante) les marqueurs génétiques dans des groupes de liaison adéquats.

4.2 Estimation des taux de dispersion par sexe

Chez beaucoup d'espèces, la dispersion « de naissance »[13], qui caractérise le mouvement des individus entre le lieu de leur naissance et le lieu où ils se reproduisent, dépend du sexe des individus. On dit qu'il existe chez ces espèces un *biais* de dispersion spécifique au sexe. Chez les oiseaux, ce sont les femelles qui dispersent en général plus que les mâles, tandis que chez les mammifères, ce sont en général les mâles qui dispersent plus que les femelles (Greenwood 1980).

4.2.1 Caractérisation des biais de dispersion

Pourquoi existe-t-il des biais de dispersion ?

Motro (1991, 1994) a montré qu'une stratégie de dispersion biaisée en fonction du sexe pouvait être évolutivement stable si le coût associé à la dépression de consanguinité était fort et le coût associé à la dispersion était faible (voir également Gandon 1999). D'autre part, si les coûts à la dispersion sont identiques pour les individus des deux sexes, les taux de dispersion évolutivement stables dépendent des conditions initiales. Dès que les individus d'un sexe dispersent plus que les individus de l'autre sexe, la sélection favorise des taux de dispersion encore plus forts pour les individus de ce sexe. Ce phénomène d'entraînement suggère que des contraintes phylogénétiques peuvent être déterminantes pour expliquer pourquoi chez certaines espèces ce sont les mâles qui dispersent, alors que chez d'autres espèces la situation inverse est observée (Perrin et Mazalov 1999). Mais l'évitement de la consanguinité n'est pas la seule force évolutive qui explique le maintien de stratégies de dispersion biaisée (Perrin et Goudet 2001; Perrin et Mazalov 2000). En effet, si les coûts associés à la dispersion ne sont pas identiques entre individus de sexe opposé (par exemple, parce que les individus de sexe opposé ne réagissent pas de la même manière à la compétition: voir Greenwood 1980) alors des stratégies de dispersion biaisée peuvent évoluer. C'est ce que l'on observe chez les mammifères, où la compétition pour l'accès à

[13] Terme traduit de l'anglais, *natal dispersal*.

la reproduction est perçue (avec l'évitement de la consanguinité) comme un facteur important expliquant l'évolution de la dispersion biaisée en faveur des mâles (Dobson 1982).

Or s'il existe de nombreuses espèces chez lesquelles on a observé un biais de dispersion lié au sexe, la mesure des taux de dispersion réalisés chez les individus mâles et femelles n'est pas chose aisée. Les estimations directes des taux de dispersion reposent sur des méthodes de capture-marquage-recapture d'individus ou bien sur des méthodes de suivi d'individus sur lesquels ont été installées des balises radio-émettrices. Outre la mesure des taux de dispersion moyens, ces méthodes permettent également l'estimation des distributions des distances de dispersion. En revanche, ces méthodes ne permettent pas, en général, de caractériser la dispersion efficace, c'est-à-dire le mouvement d'individus qui arrivent à se reproduire en dehors de leur lieu de naissance. De plus, la dispersion à longue distance est souvent sous-estimée, du fait des contraintes liées à la taille limitée des sites d'étude (Koenig *et al.* 1996).

Mises en évidence indirectes de biais de dispersion

La littérature est riche d'exemples où les auteurs ont tenté de détecter des biais de dispersion liés au sexe, à partir de données de polymorphisme sur des marqueurs génétiques. L'idée de comparer la distribution de la variation génétique entre des marqueurs à hérédité Mendélienne et des marqueurs à hérédité maternelle stricte vient tout naturellement à l'esprit, lorsqu'on s'intéresse aux biais de dispersion liés au sexe. Lyrholm *et al.* (1999) ont ainsi montré que l'observation de mouvements importants chez les cachalots mâles (*Physeter macrocephalus*) ainsi que l'apparente philopatrie des femelles se traduisaient par une absence de différenciation génétique sur des marqueurs microsatellites à l'échelle de l'ensemble des océans du globe, tandis que la différenciation génétique était significative sur des marqueurs mitochondriaux, à cette même échelle géographique (Lyrholm et Gyllensten 1998). Un patron de différenciation tout à fait similaire a été également observé chez le grand requin blanc *Carcharodon carcharias* (Pardini *et al.* 2001).

Chez la tortue verte *Chelonia mydas*, FitzSimmons *et al.* (1997a) ont

mesuré la différenciation génétique sur des marqueurs mitochondriaux parmi les adultes reproducteurs capturés dans plusieurs sites où les individus s'accouplent, de part et d'autre de la grande barrière de corail. Leurs résultats montrent une différenciation génétique élevée entre les différents sites ainsi qu'un patron de différenciation identique chez les mâles et les femelles. FitzSimmons *et al.* (1997a) interprètent leurs résultats comme la conséquence d'un comportement de philopatrie au site de reproduction identique chez les mâles et les femelles. En revanche, la comparaison de la variation génétique mesurée sur des marqueurs nucléaires (microsatellites et copies uniques d'ADN anonyme) et mitochondriaux, parmi des individus capturés dans les crèches (où les jeunes sont élevés) a montré une différenciation génétique plus faible sur les marqueurs nucléaires que sur les marqueurs mitochondriaux (FitzSimmons *et al.* 1997b). Les flux de gènes nucléaires sont donc suffisamment importants pour que la différenciation génétique nucléaire soit faible, en comparaison avec la différenciation sur les marqueurs mitochondriaux. Ceci peut, de prime abord, sembler contradictoire avec les résultats de FitzSimmons *et al.* (1997a) qui ont montré que les mâles reproducteurs étaient tout aussi philopatriques à leur lieu de naissance que les femelles. L'explication de ce paradoxe apparent est la suivante. Malgré l'existence de sites spécifiques pour la reproduction, les individus de différents groupes s'accouplent au cours de leurs déplacements entre leur lieu de naissance et le site où ils se reproduisent. Quel que soit le comportement de philopatrie des individus des deux sexes, ceci implique un certain brassage des gènes nucléaires (qui sont transmis par les voies mâles et femelles). En revanche, la philopatrie des femelles à leur lieu de naissance limite les flux de gènes mitochondriaux (qui, contrairement aux gènes nucléaires, ne sont transmis que par la voie femelle). Les résultats de FitzSimmons *et al.* (1997b) impliquent que les flux de gènes nucléaires sont suffisamment importants pour que la différenciation des gènes nucléaires soit faible au regard de la différenciation des gènes mitochondriaux. Dans cet exemple, les flux de gènes asymétriques entre le compartiment nucléaire et le compartiment mitochondrial ne sont pas la conséquence directe d'une différence du comportement de dispersion des mâles et des femelles (FitzSimmons *et al.* 1997b).

Toutefois, d'autres différences entre les compartiments nucléaires et mitochondriaux peuvent également être la source de discordances. D'une part, les mitochondries ont des génomes qui ne recombinent pas, ou très peu. Par conséquent, la distribution de la variation mesurée sur des marqueurs mitochondriaux peut être très fortement influencée par des effets sélectifs agissant sur le reste de la molécule (voir le § 3.3.2). D'autre part, du fait de sa transmission uniparentale, la taille efficace du génome mitochondrial est plus faible que la taille efficace du génome nucléaire.

Un des premiers exemples d'utilisation exclusive de marqueurs autosomaux que j'ai pu relever dans la littérature concerne l'étude de Kawata (1985), qui a tenté (avec un succès relatif) de mettre en évidence un biais de dispersion en faveur des mâles chez le campagnol à dos roux *Clethrionomys rufocanus bedfordiae*. Plus récemment, Favre *et al.* (1997) ont montré que le biais de dispersion en faveur des femelles chez la musaraigne musette *Crocidura russula* pouvait être détecté par une approche reposant sur l'utilisation d'indices d'assignation, développés à l'origine par Paetkau *et al.* (1995). En utilisant des marqueurs microsatellites très variables, Favre *et al.* (1997) ont montré que la distribution des probabilités d'assignation de chaque individu à la population dans laquelle il a été échantillonné[14] avait une moyenne plus faible ainsi qu'une variance plus grande chez les femelles que chez les mâles. Favre *et al.* (1997) interprètent cette plus faible probabilité de trouver des femelles nées dans le lieu de leur capture comme la conséquence de mouvements plus importants chez les femelles que chez les mâles. La même approche a été conduite avec succès par Mossman et Waser (1999) chez la souris à pattes blanches *Peromyscus leucopus* où la dispersion, cette fois-ci, est plus importante chez les mâles.

[14]Contrairement à Paetkau *et al.* (1995), Favre *et al.* (1997) ne cherchent pas à assigner chaque individu à une population, mais s'intéressent à la distribution des probabilités que chaque individu d'une population donnée appartienne à la population dans laquelle il a été échantillonné.

Différenciation génétique nucléaire spécifique au sexe

A plusieurs reprises, des études ont également tenté de caractériser des biais de flux de gènes entre individus de sexes opposés grâce aux F-statistiques (voir le § 2.1.3) (McCracken 1984; Mossman et Waser 1999; Rassmann *et al.* 1997). Chez *C. russula*, par exemple, Balloux *et al.* (1998) ont montré que la corrélation des fréquences alléliques à l'intérieur des groupes sociaux, relativement aux fréquences alléliques à l'intérieur des sous-populations (voir Chesser 1991b) était plus forte (et significative) chez les mâles que chez les femelles (chez qui la corrélation n'était d'ailleurs pas significativement différente de zéro). Très peu de ces études, à l'exception notable de Rassmann *et al.* (1997), précisent pour autant que de telles différences ne sont pas nécessairement attendues. Rassmann *et al.* (1997) soulignent en effet que si la distribution de la variation génétique varie entre mâles et femelles, cette différence ne devrait plus être observée après une génération.

Une lecture attentive des modèles développés par Chesser (1991a,b) et par Wang (1997a) montre d'ailleurs que seules des variations de succès reproducteur entre les individus de sexe opposé peuvent contribuer à des différences de structuration génétique entre mâles et femelles. En revanche, le modèle que j'ai présenté dans le § 2.2.1 souligne que des différences peuvent être observées, à la condition que les individus soient échantillonnés de part et d'autre de l'intervalle de temps qui sépare la dispersion des individus juvéniles et la reproduction des individus adultes.

4.2.2 Mesure des taux de migration mâles et femelles

Dans ce modèle, présenté en § 2.2.1, j'ai calculé les valeurs d'équilibre des F_{ST} spécifiques au sexe, chez des individus échantillonnés avant et après la dispersion. Si l'on regarde d'un peu plus près les équations (2.25) et (2.27) (p. 40), on remarque que les expressions de F_{ST} à l'équilibre sont tout à fait similaires, à un terme de migration près. Notamment, le rapport du F_{ST}^{XY}

après dispersion sur le F_{ST}^{*} avant dispersion donne l'expression suivante

$$\frac{F_{ST}^{XY}}{F_{ST}^{*}} = d_{XY} + b_{XY} \cdot F_{ST}^{XY} \qquad \text{pour tout } (X,Y) \in \{\male,\female\}^2 \tag{4.17}$$

où je rappelle que b_{XY} est la fréquence des paires d'individus tirés au hasard dans des sous-populations distinctes, et qui proviennent d'une même sous-population à la génération précédente (voir p. 38). Le dernier terme du second membre de l'équation (4.17) est négligeable devant d_{XY}. Par conséquent, F_{ST}^{*} ne diffère de F_{ST}^{XY} que par un facteur $d_{XY} = (1 - m_X\,[n/(n-1)])(1 - m_Y\,[n/(n-1)])$, où je rappelle que m_X (respectivement m_Y) est la probabilité qu'un *individu* de sexe X (respectivement de sexe Y) soit immigrant et où n est le nombre de sous-populations (voir § 2.1.2). Pour de grandes valeurs de n, $d_{XY} \approx (1 - m_X)(1 - m_Y)$. Si l'on prend le rapport du F_{ST}^{XX} spécifique au sexe des individus échantillonnés, évalué *après* la dispersion, sur le F_{ST}^{*} évalué pour des individus échantillonnés *avant* la dispersion, alors on obtient le taux de migration des individus de sexe X.

$$m_X \approx 1 - \sqrt{F_{ST}^{XX} / \; F_{ST}^{*}} \qquad \text{pour tout } X \in \{\male,\female\} \tag{4.18}$$

Par conséquent, on peut construire un estimateur du taux de migration spécifique au sexe, de la forme

$$\widehat{m}_X = 1 - \sqrt{\widehat{\theta}^{XX} / \; \widehat{\theta}^{*}} \qquad \text{pour tout } X \in \{\male,\female\} \tag{4.19}$$

où $\widehat{\theta}^{XX}$ est un estimateur multilocus de F_{ST}^{XX} évalué parmi les individus de sexe X, échantillonnés après dispersion et $\widehat{\theta}^{*}$ est un estimateur multilocus de F_{ST}^{*} évalué parmi les individus échantillonnés avant dispersion.

J'ai réalisé un modèle de simulations stochastiques, afin de tester la précision de l'estimateur du taux de migration spécifique au sexe donné en (4.19). La figure 4.6 montre deux exemples de la distribution totale des taux de migration estimés, chacun pour un jeu de paramètres. Dans la figure 4.6A, les femelles dispersent cent fois moins que les mâles. Dans seulement 0.6% des cas, le taux de migration estimé chez les femelles était supérieur au taux

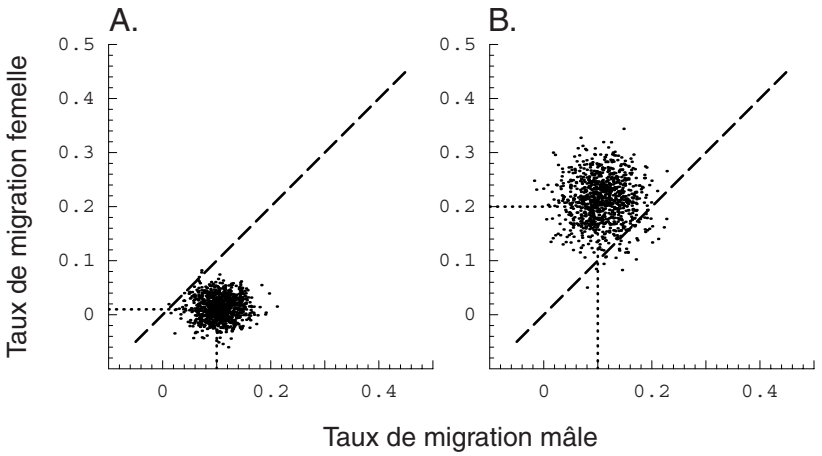


Figure 4.6: Un exemple d'estimation des taux de migration par sexe. Chaque point représente une estimation conjointe des taux de migration des mâles et des femelles. Quatre-vingt dix individus au total ont été échantillonnés au sein de dix sous-populations et génotypés à seize locus indépendants. Les tirets indiquent la première diagonale, en dessous de laquelle le taux de migration des mâles est plus faible que celui des femelles. Les lignes en pointillés donnent les valeurs réelles (simulées) des taux de migration des mâles et des femelles. $2N = 100$, $n = 20$, $m_{\male} = 0.1$ et $\mu = 10^{-3}$. A. $m_{\female} = 0.01$. B. $m_{\female} = 0.2$.

de migration estimé chez les mâles. Dans 78.6 cas sur 100 les intervalles de confiance, obtenus par la méthode du bootstrap réalisé sur les locus (voir DiCiccio et Efron 1996), étaient disjoints tout comme les intervalles de confiance moyens[15]. Dans la figure 4.6B, les femelles dispersent deux fois plus que les mâles. Dans 3.7% des cas, le taux de migration estimé des femelles était inférieur à celui des mâles. Cependant, les intervalles de confiance n'étaient disjoints que dans 44.3 cas sur 100 et les intervalles moyens étaient légèrement chevauchants.

Pour une plus grande gamme de valeurs de taux de migration des individus des deux sexes, le tableau 4.2 donne des résultats de l'estimation des taux

[15]Je définis l'*intervalle moyen* comme étant l'intervalle de confiance dont les bornes inférieures et supérieures ont pour valeur les moyennes des bornes inférieures et supérieures de tous les intervalles obtenus.

Tableau 4.2: Résultats de simulation d'un modèle à nombre fini d'îles. Les moyennes (Moy.) et les écart-types (s) des taux de migration estimés sont indiqués pour chaque sexe et pour plusieurs valeurs des paramètres $m_{\male}$ et $m_{\female}$. 20 sous-populations de 50 individus reproducteurs ont été simulées. Des échantillons sont prélevés dans 10 sous-populations parmi 50 individus avant dispersion et parmi 50 individus après la dispersion. Huit locus sont génotypés. Les valeurs moyennes des bornes des intervalles de confiance à 95% sont également indiquées.

	$\widehat{m}_{\male}$			$\widehat{m}_{\female}$	
$m_{\male}$	Moy.(s)	95% C.I.	$m_{\female}$	Moy.(s)	95% C.I.
0.01	0.01 (0.01)	$[-0.01; 0.03]$	0.01	0.01 (0.01)	$[-0.01; 0.03]$
	0.01 (0.01)	$[-0.01; 0.03]$	0.02	0.02 (0.02)	$[0.003; 0.04]$
	0.01 (0.02)	$[-0.02; 0.05]$	0.10	0.11 (0.03)	$[0.071; 0.16]$
	0.02 (0.03)	$[-0.03; 0.07]$	0.20	0.21 (0.05)	$[0.154; 0.27]$
0.02	0.02 (0.01)	$[0.001; 0.04]$	0.01	0.01 (0.01)	$[-0.01; 0.03]$
	0.02 (0.02)	$[0.000; 0.04]$	0.02	0.02 (0.02)	$[0.001; 0.04]$
	0.02 (0.02)	$[-0.01; 0.06]$	0.10	0.11 (0.03)	$[0.068; 0.15]$
	0.03 (0.03)	$[-0.02; 0.08]$	0.20	0.21 (0.04)	$[0.151; 0.28]$
0.10	0.11 (0.03)	$[0.069; 0.14]$	0.01	0.01 (0.02)	$[-0.02; 0.05]$
	0.11 (0.03)	$[0.067; 0.15]$	0.02	0.03 (0.02)	$[-0.01; 0.06]$
	0.11 (0.04)	$[0.058; 0.16]$	0.10	0.11 (0.04)	$[0.057; 0.16]$
	0.12 (0.05)	$[0.051; 0.18]$	0.20	0.21 (0.05)	$[0.141; 0.29]$
0.20	0.21 (0.04)	$[0.152; 0.27]$	0.01	0.02 (0.03)	$[-0.03; 0.07]$
	0.21 (0.05)	$[0.151; 0.28]$	0.02	0.03 (0.03)	$[-0.02; 0.08]$
	0.21 (0.05)	$[0.142; 0.29]$	0.10	0.12 (0.05)	$[0.048; 0.18]$
	0.22 (0.06)	$[0.133; 0.31]$	0.20	0.22 (0.06)	$[0.133; 0.32]$

de migration par sexe. Les taux de migration sont en général légèrement surestimés, pour l'ensemble des gammes de paramètres considérées ici. Ce léger biais provient vraisemblablement de l'approximation faite en négligeant le terme $\left(b_{XY} \cdot F_{ST}^{XY}\right)$ dans le second membre de l'équation (4.17). La précision de l'estimation augmente avec la valeur vraie (simulée) des taux de migration. Lorsque les taux de migration sont faibles (inférieurs à quelques pour-cents), les écart-types des estimations tout comme les intervalles de confiance sont

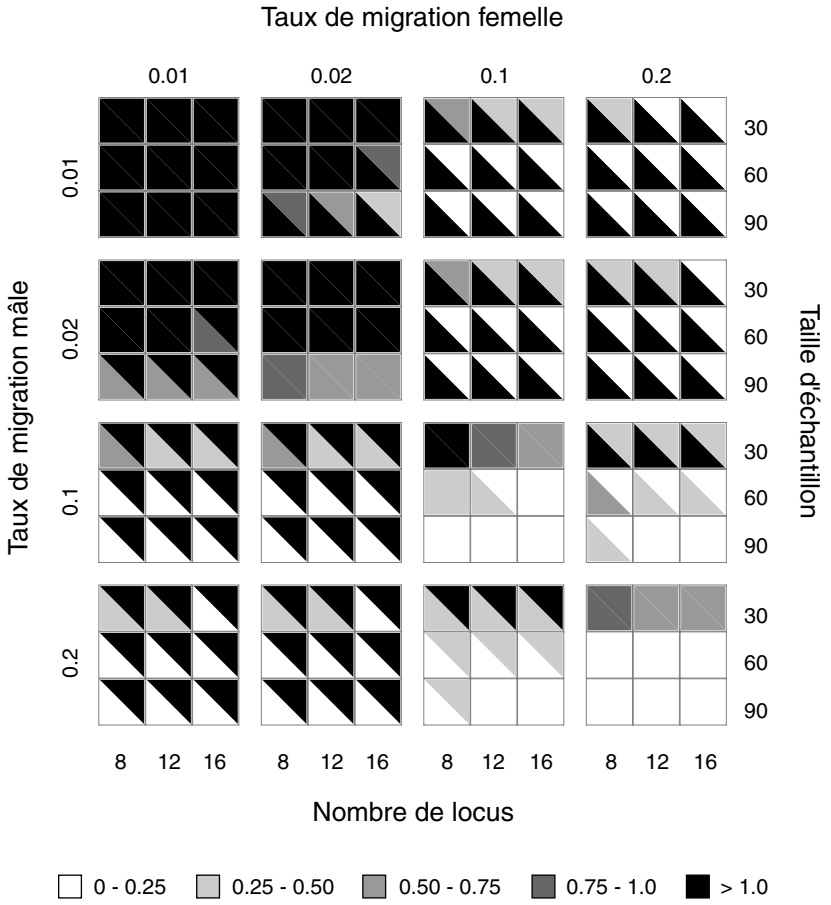


Figure 4.7: Erreur relative sur les valeurs estimées des taux de migration par sexe. Chaque carré représente une combinaison de paramètres et dans chaque carré, le triangle supérieur donne l'erreur sur les taux de migration des femelles et le triangle inférieur donne l'erreur sur les taux de migration des mâles. La taille d'échantillon indique le nombre d'individus échantillonné dans chacune des 10 sous-populations, parmi 20 sous-populations simulées. Un tiers du nombre total d'individus est échantillonné avant la dispersion, le reste après. $N = 50$, $n = 20$ et $\mu = 10^{-3}$.

importants. Le biais et la variance des estimations du taux de migration d'un sexe augmentent avec le taux de migration des individus du sexe opposé.

Pour différentes valeurs des taux de migration mâles et femelles, la figure 4.7 donne les gammes de valeur de l'erreur relative des carrés moyens

parmi les mâles (triangles des coins supérieurs) et les femelles (triangles des coins inférieurs), pour différentes stratégies d'échantillonnage (nombre d'individus échantillonnés, nombre de locus typés). L'erreur quadratique moyenne relative est calculée comme la valeur moyenne de $[(\widehat{m}_X - m_X)/m_X]^2$. Pour de faibles taux de migration ($m_{\male} \leqslant 0.02$ et $m_{\female} \leqslant 0.02$) l'erreur relative est grande ($\geqslant 0.50$). Pour des taux de migration plus importants ($m_{\male} \geqslant 0.1$ et $m_{\female} \geqslant 0.1$) les taux de migration estimés ont une précision satisfaisante ($\leqslant 0.21$ pour 90 individus échantillonnés par sous-population et 12 locus génotypés). Cependant, l'erreur standard associée aux taux de migration estimés pour le sexe le plus philopatrique est toujours plus forte que l'erreur standard associée aux taux de migration estimés pour le sexe qui disperse le plus, et ceci d'autant plus que les taux de migration sont grands (figure 4.7). La précision des estimations augmente à mesure que le nombre total d'individus échantillonnés augmente. En revanche, augmenter le nombre de locus ne semble pas améliorer la précision des estimations. Pour le même effort d'échantillonnage (c'est-à-dire le même nombre total de génotypes à un locus), il vaut toujours mieux échantillonner plus d'individus que d'augmenter le nombre de locus. D'avantage de résultats sont présentés dans l'annexe H.

Cette méthode montre que l'estimation des F_{ST} spécifiques au sexe, avant et après la dispersion des individus, permet de mesurer les taux de migration par sexe. L'équation (4.19) définit en effet un estimateur du taux de migration par sexe, dont la précision a été évaluée par des simulations stochastiques (tableau 4.2 et figures 4.6-4.7). Cette méthode peut s'appliquer à n'importe quelle espèce à sexes séparés, pourvu que l'on échantillonne les individus *avant* et *après* la dispersion. Ceci implique donc nécessairement une bonne connaissance de la biologie des espèces et de leur cycle de vie. Cette méthode peut donc s'appliquer aux nombreuses espèces dont on connaît suffisamment d'éléments du cycle de vie et pour lesquelles un biais de dispersion lié au sexe a d'ores et déjà été constaté (Greenwood et Harvey 1982; Greenwood 1980). Un échantillonnage adapté consiste, par exemple, à prélever des tissus, du sang ou toute autre source de matériel génétique, parmi les jeunes au nid (avant dispersion) et les adultes reproducteurs (après dispersion).

La plupart des méthodes développées pour mettre en évidence des biais

de dispersion à partir de données de polymorphisme sur des marqueurs génétiques ont été cantonnées à une approche qualitative visant à caractériser le mouvement relatif d'individus de sexe opposé. Récemment, Petit *et al.* (2001) ont proposé une approche quantitative, la première à ma connaissance, utilisant des marqueurs nucléaires (microsatellites) et mitochondriaux. Les flux de gènes sont généralement caractérisés par le « nombre efficace de migrants par génération », c'est-à-dire le produit de la taille efficace locale par le taux d'immigration (Beerli et Felsenstein 1999; Slatkin 1987). L'estimateur du taux de migration que j'ai proposé en (4.19) ne dépend quant à lui pas de la taille efficace locale. L'estimateur (4.19) donne la proportion d'individus qui se reproduisent avec succès dans une autre sous-population que celle où ils sont nés. Ce paramètre inclue les coûts à la dispersion.

4.2.3 Limites de l'approche

Quelles sont les limites de cette approche ? Le modèle qui a été développé fait un certain nombre d'hypothèses, dont notamment celle d'un sexe-ratio équilibré. S'il existe des biais de sexe-ratio, on s'attend à ce que la dérive soit plus forte. Dans ce cas, toute chose étant égale par ailleurs, la valeur de F_{ST}^{*} (après la reproduction) devrait donc être plus forte qu'en l'absence de biais de sexe-ratio. Mais la valeur de F_{ST}^{XY} aussi doit être plus forte et le rapport de ces paramètres ne devrait pas être très différent de (4.17). En revanche, cette approche souffre sans aucun doute de limites plus importantes liées aux hypothèse faites quant au modèle de migration. L'analyse de modèles d'isolement par la distance qui prennent en compte des biais de dispersion spécifiques à certaines classes d'individus reste une question ouverte.

4.3 Influence de la sélection sur les marqueurs

Cette section reprend l'essentiel du modèle présenté dans l'annexe I. Dans les deux sections (§ 4.1 et § 4.2) qui précèdent, je me suis attaché à présenter les résultats de deux méthodes d'estimation de paramètres « démographiques », à partir de données de polymorphisme sur des marqueurs génétiques. Ces méthodes, nous l'avons vu, reposent sur des modèles dont les hypothèses ne sont pas toujours vérifiées dans des situations naturelles (comme par exemple le modèle de dispersion des individus, ou bien les modèles de mutation des gènes). De plus, ces méthodes reposent toutes sur l'hypothèse que les marqueurs génétiques utilisés sont *neutres* vis-à-vis de la sélection naturelle.

Pourtant, il existe de nombreux cas où des facteurs sélectifs ont été mis en évidence à un ou plusieurs locus utilisés comme marqueurs génétiques dans des études de structure des populations. Ces facteurs spécifiques à un ou plusieurs locus dans un jeu de données peuvent conduire à des interprétations erronées des facteurs (migration, dérive, etc.) qui influencent la distribution de la variation génétique observée. C'est le cas par exemple dans de nombreuses études concernant le complexe d'espèces *Mytilus edulis*, où les déficits en hétérozygotes observés seraient non seulement le fait d'effets Wahlund temporels et spatiaux, mais également d'effets directs ou indirects de la sélection à certains locus (voir Raymond *et al.* 1997). D'autre part, il existe de nombreux autres exemples dans la littérature où l'étude comparative de classes distinctes de marqueurs génétiques a révélé des différences importantes dans la distribution du polymorphisme, non seulement entre classes de marqueurs mais également entre marqueurs, quelle que soit leur classe (Estoup *et al.* 1998; Fréville *et al.* 2001; Lemaire *et al.* 2000; Ross *et al.* 1999). Une question importante qui sous-tend l'ensemble des études de la distribution du polymorphisme sur des marqueurs génétiques est donc : « existe-t-il un moyen de tester la neutralité des marqueurs utilisés ? ». En d'autres termes, peut-on identifier les locus qui sont soumis à l'action de la sélection naturelle, de façon directe ou indirecte (voir § 3.3), afin de les exclure d'un jeu de données destiné à l'analyse de la structure des popula-

tions ou de leur histoire ? Cavalli-Sforza (1966) est l'un des premiers à souligner que n'importe quelle forme de sélection affectera certaines régions du génome plus que d'autres, tandis que l'histoire des populations, leur démographie, leur structure spatiale ou bien encore leur système génétique[16] affectera l'ensemble du génome de la même manière.

La formalisation mathématique de cette approche a conduit Lewontin et Krakauer (1973) à proposer deux tests de neutralité, qui reposent tous les deux sur la distribution attendue de la statistique $\widehat{F}$, « la variance standardisée des fréquences alléliques », c'est-à-dire un estimateur du paramètre F_{ST}. Le premier test de Lewontin et Krakauer (1973) est un test « goodness of fit » qui repose sur la comparaison de la distribution observée de $\widehat{F}$ à une distribution du χ^2 à $(n-1)$ degrés de liberté, où n est le nombre de populations échantillonnées. Le second test repose sur la comparaison de la variance observée de $\widehat{F}$ entre locus, notée s^2_F, avec l'attendu théorique, soit approximativement,

$$\sigma^2 = \frac{k\overline{F}^2}{n-1} \tag{4.20}$$

où $\overline{F}$ est la moyenne des valeurs de $\widehat{F}$ sur tous les locus et où k est une constante qui, d'après Lewontin et Krakauer (1973), ne doit pas dépasser 2, quelle que soit la distribution des fréquences alléliques. Le rapport s^2_F/σ^2 devrait donc être distribué comme un χ^2 avec un nombre de degrés de liberté égal au nombre de locus bi-alléliques.

Ce que n'ont pas pris en compte Lewontin et Krakauer (1973), c'est le fait que puisque les populations d'une espèce partagent le plus souvent une histoire commune et puisque les populations échangent, en des proportions variables, des gènes grâce à la dispersion des individus qui les composent, alors les valeurs de $\widehat{F}$ à différents locus sont corrélées entre elles. Si l'on imagine par exemple les conséquences de la fragmentation de l'aire de répartition d'une espèce, on s'attend à ce que les populations de cette espèce dérivent d'une population ancestrale par divergences successives. Cette structure hiérarchique

[16] Il s'agit de l'ensemble des caractères liés à la production de variabilité génétique (mutation, recombinaison, etc.) décrit par Darlington (1939) sous le terme de *genetic system*.

des populations qui composent une espèce a pour conséquence d'augmenter la variance attendue de $\widehat{F}$ (Robertson 1975a,b), bien au delà des limites imposées par Lewontin et Krakauer (1973) dans leur modèle. Que ce soit dans des modèles très simples de dérive de populations qui divergent les unes des autres (Nei et Chakravarti 1977), dans des modèles à nombre fini d'îles (Nei *et al.* 1977), ou bien encore dans des modèles de migration par pas entre populations adjacentes (Nei et Maryuyama 1975), la variance attendue de $\widehat{F}$ est beaucoup plus forte que la valeur donnée par l'équation (4.20), ce qui rend le test de Lewontin et Krakauer (1973) assez peu approprié dans la plupart des situations naturelles (Lewontin et Krakauer 1975).

Dans une étude récente du polymorphisme de marqueurs RFLP[17] nucléaires sur l'homme à l'échelle mondiale, Bowcock *et al.* (1991) ont utilisé une approche plus pertinente. La distribution observée des estimateurs bialléliques de F_{ST}, conditionnellement aux fréquences alléliques, a été comparée à une distribution théorique, obtenue par simulation d'un scénario réaliste retraçant l'histoire de la divergence des populations humaines étudiées. Parmi 100 marqueurs RFLP, un certain nombre ont montré un niveau de variabilité plus faible ou plus fort qu'attendu sous un modèle d'évolution des fréquences strictement neutre. Beaumont et Nichols (1996) ont repris l'idée de Bowcock *et al.* (1991) pour l'appliquer au cas de marqueurs multialléliques. Mais l'avancée majeure de Beaumont et Nichols (1996) est de proposer une méthode qui ne nécessite pas de connaître *a priori* l'histoire des populations. Beaumont et Nichols (1996) ont montré, en effet, que la distribution conjointe des estimateurs de F_{ST} et de l'hétérozygotie, estimée comme $(1 - \widehat{Q}_2)$ (voir § 2.1.2), était remarquablement robuste au modèle de structure des populations. Des simulations stochastiques leur ont permis de montrer que la distribution obtenue dans le cadre du modèle à nombre fini d'îles n'étaient pas très différente de la distribution obtenue pour des modèles

[17]Terme traduit de l'anglais *Restriction Fragment Length Polymorphism*. Il s'agit du polymorphisme engendré par la digestion de fragments d'ADN, amplifiés ou non, par des enzymes de restriction bactériennes, qui coupent l'ADN de façon spécifique.

plus réalistes d'extinction-recolonisation ou de migration par pas[18] (voir la note de bas de page 9, p. 67 pour une définition des modèles de migration par pas). De même, pour des valeurs de F_{ST} inférieures à 0.5, les distributions obtenues pour des modèles qui ne sont pas à l'équilibre entre les forces de migration, de mutation et de dérive, ne sont pas très différentes des distributions obtenues pour des situations à l'équilibre entre ces forces. Seule une forte hétérogénéité dans les patrons de migration entre populations conduit à des déviations importantes de la distribution observée par rapport à celle du modèle à nombre d'îles fini, dans une gamme de valeurs d'hétérozygotie comprise entre 0.1 et 0.5 (voir la figure 3d dans Beaumont et Nichols 1996). Par conséquent, la méthode de Beaumont et Nichols (1996) n'est pas très adaptée à des situations où la connectivité entre paires de populations est inégale. De la même façon, on peut raisonnablement penser que cette méthode ne sera pas non plus très adaptée lorsque l'histoire des populations peut se résumer à une série d'événements de divergence successifs. Or mises à part les situations où l'histoire des populations est connue (comme dans le cas de Bowcock *et al.* 1991), on ne peut pas à la fois estimer les patrons de migration entre les populations, ou bien leur histoire au sens généalogique, et tester l'homogénéité des marqueurs avec les mêmes données (Felsenstein 1982). Une solution à ce problème est de focaliser notre attention sur des modèles simples, mais réalistes, de divergence de *paires* de populations (Robertson 1975b; Tsakas et Krimbas 1976). Le nombre de paramètres du modèle s'en trouve ainsi grandement réduit.

[18]Toutefois, les simulations du modèle de migration par pas n'ont été réalisées que dans le cas d'un réseau de population relativement petit (tore de 24×24 sous-populations) (voir Beaumont et Nichols 1996). Or on s'attends à ce que la variance des estimateurs de F_{ST} entre paires de populations augmente avec la distance qui sépare ces populations. Des simulations stochastiques de réseaux de populations plus grands pourraient être nécessaires afin de généraliser ce résultat.

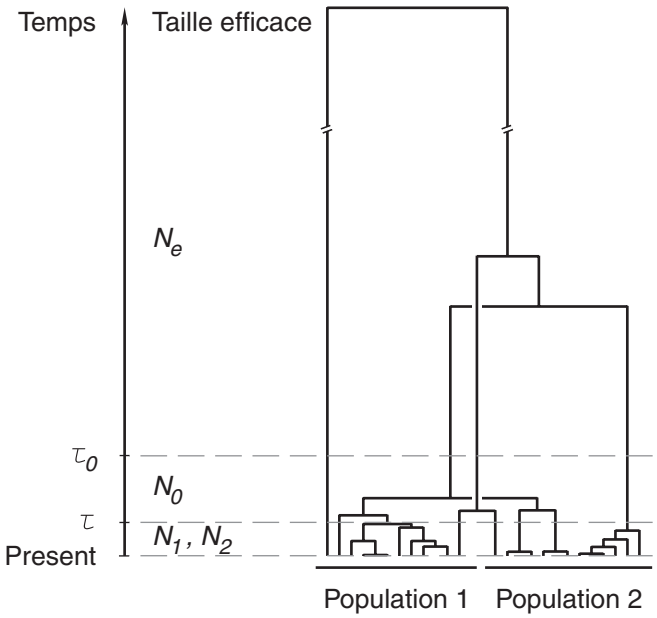


Figure 4.8: Un exemple de généalogie de $n = 10$ gènes dans un modèle de divergence de deux populations isolées. Cet exemple est le résultat d'une simulation du processus de coalescence décrit dans le texte, pour $N_1 = N_2 = 100$, $N_0 = 500$, $N_e = 1000$, $\tau = 50$, $\tau_0 = 150$ et $\mu = 10^{-3}$.

4.3.1 Un modèle de divergence d'une paire de populations

Dans cette partie, je m'efforcerai de montrer que certaines propriétés des mesures de divergence de paires de populations peuvent être utiles pour identifier des locus pour lesquels la distribution de la variation génétique s'éloigne de l'attendu théorique dans le cas neutre. Je vais donc maintenant présenter un modèle d'évolution neutre des fréquences alléliques au cours du processus de divergence de deux populations. La description de ce modèle fait largement appel à la notion de coalescence de gènes dont j'ai présenté quelques aspects importants en § 2.1.2.

Hypothèses du modèle

On considère deux populations composées de N_1 et N_2 individus haploïdes, qui ont pour origine une population unique de taille N_0 dont elles se sont séparées τ générations dans le passé. Ces deux populations ont échangé ou échangent éventuellement encore des individus, depuis le moment de leur divergence et le moment présent. Avant la divergence, la population unique n'est pas nécessairement à l'équilibre entre les forces évolutives que sont la mutation et la dérive. Au contraire, on considère que cette population peut avoir traversé un goulot d'étranglement[19], τ_0 générations avant le temps présent (où $\tau_0 > \tau$). Avant cet événement, cependant, la population ancestrale est considérée comme étant à l'équilibre entre les forces évolutives de mutation et de dérive. On considère un cycle de vie annuel strict (les générations ne se chevauchent pas) et les appariements se font au hasard. Les nouvelles mutations apparaissent à un taux μ, selon un modèle à nombre infini d'allèles. La figure 4.8 illustre ce modèle.

On note $Q_{w,i}$ la probabilité que deux gènes pris au hasard dans la population i soient identiques par descendance et Q_a, la probabilité qu'un gène échantillonné dans la population 1 soit identique à un gène échantillonné dans la population 2. De façon générique, on notera Q_h la probabilité d'identité de n'importe quelle paire de gènes, avec $h = (w, i)$ quand les deux gènes sont pris dans la population i et $h = a$ quand un gène est pris dans 1 et l'autre gène est pris dans 2. Il est possible (voir § 2.1.2) d'exprimer Q_h en fonction des temps moyens de coalescence (Slatkin 1991). En reprenant l'approximation en temps continu de Hudson (1990), on obtient

$$Q_h = \int_0^\infty \gamma^t \mathrm{c_h}(t)\mathrm{d}t \tag{4.21}$$

où $\mathrm{c_h}(t)$ est la probabilité de coalescence à t pour une paire de gènes de type h, et $\gamma = (1-\mu)^2$. Or les *temps d'attente* entre les événements de coalescence dans une population de taille N_i sont distribués selon une loi exponentielle de moyenne N_i. La probabilité d'identité pour n'importe quelle paire de gènes

[19]Ce terme, traduit de l'anglais *bottleneck*, signifie qu'une population a subi une diminution d'effectif importante au cours de son histoire.

dans la population i peut donc s'écrire

$$Q_{w,i} = \int_0^{\tau} \frac{\gamma^t}{N_i} \mathrm{e}^{-t/N_i} \mathrm{d}t + (1 - C_i) Q_0 \tag{4.22}$$

où Q_0 est la probabilité d'identité entre deux gènes dans la population ancestrale au temps τ avant la divergence, et $(1 - C_i) = \gamma^{\tau} \cdot \mathrm{e}^{-\tau/N_i}$ est la probabilité que les deux gènes n'aient ni coalescé, ni muté dans la $i$ième population, dans l'intervalle de temps $0 < t \leqslant \tau$. Le premier terme du second membre de l'équation (4.22) est la probabilité que les deux gènes aient coalescé dans l'intervalle de temps $0 < t \leqslant \tau$ et n'aient pas muté dans ce même intervalle de temps. En reprenant l'équation (4.21), la probabilité d'identité de deux gènes échantillonnés au hasard dans la population ancestrale avant la divergence au temps τ est donnée par

$$Q_0 = \int_{\tau}^{\tau_0} \frac{\gamma^{t-\tau}}{N_0} \mathrm{e}^{-(t-\tau)/N_0} \mathrm{d}t + (1 - C_0) \int_{\tau_0}^{\infty} \frac{\gamma^{t-\tau_0}}{N_e} \mathrm{e}^{-(t-\tau_0)/N_e} \mathrm{d}t \tag{4.23}$$

où $(1 - C_0) = \gamma^{\tau_0 - \tau} \cdot \mathrm{e}^{-(\tau_0 - \tau)/N_0}$ est la probabilité que deux gènes ne coalescent pas et ne mutent pas dans l'intervalle de temps $\tau < t \leqslant \tau_0$. Le premier terme du second membre de l'équation (4.23) intègre tous les événements de coalescence qui se produisent dans l'intervalle de temps $(\tau < t \leqslant \tau_0)$. Durant cette période, les temps de coalescence sont distribués selon une loi exponentielle de moyenne N_0. Le dernier terme du second membre de l'équation (4.23) intègre tous les événements de coalescence qui se produisent dans la population ancestrale, à l'équilibre entre les forces de mutation et de dérive. Ce dernier terme donne donc la probabilité d'identité de deux gènes dans une population stationnaire de taille N_e, soit $1/(1 + \theta)$, où $\theta = 2N_e\mu$. Ces intégrales peuvent être facilement résolues pour de faibles taux de mutation, c'est-à-dire dans le cas où $\gamma^t \approx \mathrm{e}^{-2\mu t}$. La solution de l'équation (4.22) est

$$Q_{w,i} \approx \frac{1}{\theta_i + 1} \left[1 - \mathrm{e}^{-T_i(\theta_i + 1)}\right] + \mathrm{e}^{-T_i(\theta_i + 1)} \cdot Q_0 \tag{4.24}$$

où $\theta_i = 2N_i\mu$ et $T_i = \tau/N_i$. La valeur de Q_0 est donnée par la solution de

l'équation (4.23), soit

$$Q_0 \approx \frac{1}{\theta_0+1}\left[1-\mathrm{e}^{-T_0(\theta_0+1)}\right]+\mathrm{e}^{-T_0(\theta_0+1)}\left(\frac{1}{\theta+1}\right) \tag{4.25}$$

où $\theta_0 = 2N_0\mu$ et $T_0 = (\tau_0 - \tau)/N_0$. Dans l'hypothèse où les flux de gènes entre les populations sont nuls après la divergence, un gène de la population 1 est identique à un gène de la population 2 uniquement s'ils sont chacun des copies d'un unique ancêtre commun avant la divergence et qu'ils n'ont pas muté, soit

$$Q_a = \gamma^\tau Q_0 \tag{4.26}$$

Comment mesurer la divergence de deux populations dans un tel modèle ? Il est possible de définir le paramètre

$$F_i = \frac{Q_{w,i} - Q_a}{1 - Q_a} \tag{4.27}$$

Dans un modèle où les deux populations en divergence ont la même taille ($N_1 = N_2 = N$), alors $(F_1 + F_2)/2$ définit la corrélation intra-classe des probabilités d'identité par descendance des paires de gènes prises à l'intérieur des populations, relativement aux paires de gènes prises entre populations (c'est-à-dire le paramètre F_{ST}). Or, comme nous l'avons vu précédemment (§ 2.1.3), les propriétés des corrélations intra-classes pour les probabilités d'identité par descendance (Cockerham et Weir 1987) peuvent être déduites des propriétés des corrélations intra-classes pour les probabilités d'identité par état, dans la limite des faibles taux de mutation (Rousset 1996, 2001b). En effet, les rapports des probabilités d'identité de la forme de l'équation (4.27) donnent la même limite, pour des faibles taux de mutation, quel que soit le modèle de mutation considéré (modèle à nombre infini ou à nombre fini d'allèles) (Rousset 1996, 1997, 2001b). Si l'on néglige les mutations (cette approximation étant raisonnable sur des temps de divergence courts), d'après les équations (4.22) et (4.26) Q_a tend vers Q_0 et $Q_{w,i} \approx C_i(1 - Q_0) + Q_0$. Par

conséquent,

$$F_i \approx 1 - \mathrm{e}^{-T_i} \tag{4.28}$$

Lorsque les deux populations ont la même taille $N_1 = N_2 = N$, il est intéressant de constater que l'équation (4.28) donne un résultat bien connu, à savoir que $F_1 = F_2 = F \approx 1 - \mathrm{e}^{-\tau/N}$ (voir, par exemple, Reynolds *et al.* 1983). On fait généralement référence au paramètre T_i comme la « longueur de branche » de la population i. Le résultat important qui justifie le modèle décrit par les équations (4.21-4.28) vient du fait que, dans la limite des faibles taux de mutation, les paramètres F_1 et F_2 ne dépendent pas des paramètres de nuisance θ ou T_0 (voir la note 27 au bas de la page 56 pour une définition de ce qu'est un paramètre de nuisance). Par conséquent, un estimateur des longueurs de branches $\widehat{T}_i$ peut être construit comme

$$\widehat{T}_i = -\ln(1 - \widehat{F}_i) \tag{4.29}$$

où $\widehat{F}_i$ est un estimateur de F_i, construit de la même manière que (2.13) soit, en fonction des fréquences alléliques p_{iu} dans la population i,

$$\widehat{F}_i = \frac{\widehat{Q}_{w,i} - \widehat{Q}_a}{1 - \widehat{Q}_a} = \frac{\sum_{u=1}^{k} \left[\widehat{p}_{iu}\left(n_i \widehat{p}_{iu} - 1\right)/(n_i - 1) - \widehat{p}_{1u}\widehat{p}_{2u}\right]}{1 - \sum_{u=1}^{k}\left(\widehat{p}_{1u}\widehat{p}_{2u}\right)} \tag{4.30}$$

Analyse du modèle

Des simulations stochastiques ont été réalisées afin de déterminer la variance d'échantillonnage des estimateurs (4.30) dans le modèle décrit dans le § 4.3.1. Un algorithme standard de simulation du processus de coalescence a été utilisé (voir, par exemple, Hudson 1990). Cet algorithme permet de construire une généalogie de $n_1 + n_2$ gènes dans le modèle de divergence de populations. Les longueurs de branches (c'est-à-dire les temps séparant deux événements de coalescence successifs) sont tirées dans une loi exponentielle dont la moyenne dépend de la taille de la population et du nombre de gènes

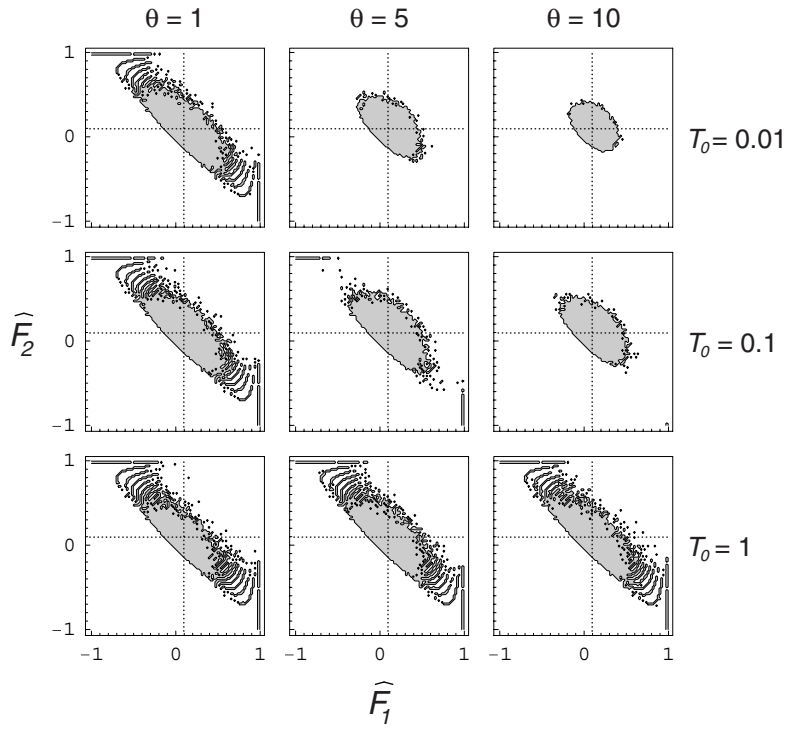


Figure 4.9: Distribution totale des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans un modèle de divergence de deux populations. Ces distributions ont été calculées pour une large gamme de valeurs des paramètres de nuisance $\theta = 2N_e\mu$ et T_0. $T_i = \tau/N_i$ est égal à 0.10 pour les deux populations en divergence (avec $\tau = 50$ et $N_1 = N_2 = 500$), soit $F_i \approx 0.0953$ (la vraie valeur de F_i est indiquée par les lignes pointillées. Pour tous les jeux de paramètres, $\mu = 10^{-4}$, $N_0 = 1000$, et $n_i = 100$. La région grisée représente 95% des valeurs simulées.

présents dans le pas de temps considéré. Enfin les mutations sont ajoutées, indépendamment, le long de la généalogie ainsi obtenue.

La figure 4.9 montre les distributions attendues de $\widehat{F}_1$ et $\widehat{F}_2$ pour différentes combinaisons de valeurs des paramètres θ et T_0, pour des longueurs de branche « réelles » (simulées) $T_1 = T_2 = 0.1$ (soit $F_1 = F_2 \approx 0.0953$). La distribution totale des estimateurs $\widehat{F}_1$ et $\widehat{F}_2$ est figurée en gris.

Sur la figure 4.10 est représentée, en gris, la distribution de $\widehat{F}_1$ et $\widehat{F}_2$ conditionnelle à la présence d'au moins 4 allèles dans l'échantillon total

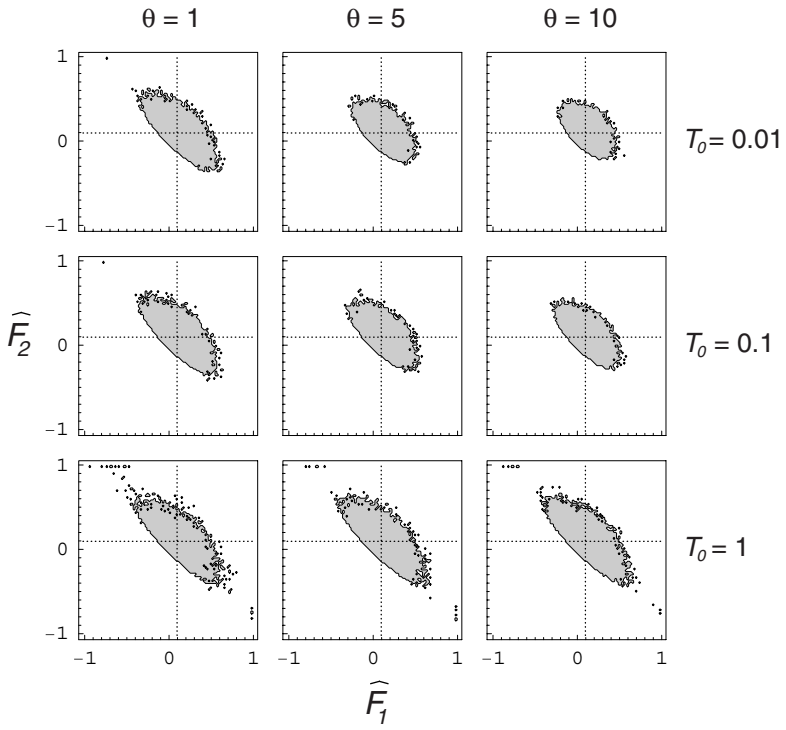


Figure 4.10: Distribution conditionnelle des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans un modèle de divergence de deux populations. Ces distributions ont été calculées pour une large gamme de valeurs des paramètres de nuisance $\theta = 2N_e\mu$ et T_0, conditionnellement à la présence de $k = 4$ allèles dans l'échantillon total. Voir la figure 4.9 pour des légendes supplémentaires.

$(n_1 + n_2)$. Les figures 4.9 et 4.10 montrent que la distribution conjointe non-conditionnelle de $\widehat{F}_1$ et $\widehat{F}_2$ dépend fortement des paramètres de nuisance du modèle (ici, θ et T_0), bien que la moyenne de cette distribution n'en dépende *a priori* pas, d'après l'équation (4.28). En effet, pour de faibles valeurs de T_0 la distribution conjointe devient plus étroite à mesure que θ augmente. Pour de grandes valeurs de θ, en revanche, la distribution devient plus étalée à mesure que T_0 augmente. Dans les deux cas, le niveau de variation qui subsiste avant la divergence des populations est le facteur important qui détermine la forme de la distribution conjointe. Pour des faibles valeurs de θ et de grandes valeurs de T_0, les lignées de gènes coalescent rapidement avant la divergence,

et le nombre d'états alléliques distincts qui peuvent être maintenus après la divergence des populations est faible. Dans ce cas, la variance des estimateurs est grande comme l'illustre la figure 4.9. En revanche, la distribution conjointe conditionnée à la présence d'au moins 4 allèles dans l'échantillon total, après la divergence (en gris foncé sur la figure 4.10) semble dépendre beaucoup moins des valeurs des paramètres de nuisance du modèle. Pour une paire de valeurs des paramètres F_1 et F_2, on peut donc construire une estimation de la distribution théorique des estimateurs de ces paramètres, conditionnée au nombre d'allèles dans l'échantillon, qui ne dépende pas (ou peu) des paramètres non estimables du modèle (tels que les taux de mutation, l'intensité d'un éventuel goulot d'étranglement, la taille efficace de la population ancestrale). Ce résultat nous permet donc de proposer une méthode pour tester l'homogénéité des patrons de différenciation génétique parmi un ensemble de marqueurs génétiques.

4.3.2 Une nouvelle méthode pour détecter des marqueurs répondant à la sélection

La méthode proposée s'appuie sur la comparaison par paires (i, j) d'échantillons provenant de populations distinctes, dont les individus ont été génotypés pour un nombre important de marqueurs. La méthode peut être résumée ainsi

(i) Pour chaque locus l, les statistiques $\widehat{F}_{il}$ et $\widehat{F}_{jl}$ sont calculées, d'après l'équation (4.30).

(ii) Les paramètres F_i et F_j sont estimés par les statistiques $\overline{\widehat{F}}_i$ et $\overline{\widehat{F}}_j$. Ces statistiques sont obtenues en calculant les moyennes sur tous les locus des estimateurs monolocus calculés en (i) pondérées respectivement par les diversités génétiques estimées à chaque locus $(1 - \widehat{Q}_{il})$ et $(1 - \widehat{Q}_{jl})$ (voir Weir et Cockerham 1984, pour un estimateur multilocus semblable de F_{ST}). Ceci revient à définir la statistique multilocus $\overline{\widehat{F}}_i$ comme le rapport de la somme sur tous les locus des valeurs du numérateur de l'équation (4.30) sur la somme des valeurs du dénominateur de cette équation.

(iii) Les longueurs de branches T_i et T_j sont estimées d'après l'équation (4.29). La distribution conjointe des statistiques $\overline{\widehat{F}}_i$ et $\overline{\widehat{F}}_j$ est alors obtenue en générant 10000 simulations indépendantes du processus de coalescence défini par les valeurs estimées $\overline{\widehat{T}}_i$ et $\overline{\widehat{T}}_j$ et pour un ensemble donné de valeurs des paramètres de nuisance du modèle. Ces 10000 simulations sont ré-itérées pour d'autres ensembles de valeurs des paramètres de nuisance (mais pour les mêmes valeurs de T_i et T_j). Ces ensembles de valeurs sont choisis pour couvrir une large gamme de valeurs des paramètres de nuisance qui permettent de générer des jeux de données simulés dans lesquels le nombre d'allèles obtenus est proche du nombre d'allèles observé dans le jeu de données initial. Les tailles d'échantillons simulées sont choisies pour être représentatives des tailles d'échantillons du jeu de données initial.

(iv) Pour chacune de ces distributions attendues de $\overline{\widehat{F}}_i$ et $\overline{\widehat{F}}_j$, toutes les distributions conditionnelles au nombre k d'états alléliques dans l'échantillon total (n_i+n_j) sont construites (pour $k = 2, 3, \ldots$). Une « zone de confiance » est alors construite, qui contient 95% des données générées par simulation.

(v) Pour chaque valeur du nombre d'allèles dans l'échantillon total, on superpose à la région de confiance construite en (iv) les points dont les coordonnées sont les valeurs estimées $\widehat{F}_1$ et $\widehat{F}_2$ calculées en (i). Si les marqueurs génétiques utilisés se comportent comme des marqueurs neutres vis-à-vis de toute forme de sélection directe ou indirecte, on s'attend à ce que 95% des points se placent à l'intérieur de la région de confiance définie en (iv).

Un exemple d'application

Cette méthode a été appliquée à un jeu de données de polymorphisme enzymatique chez l'espèce *Drosophila simulans* dans des populations africaines et européennes (Choudhary *et al.* 1992; Singh *et al.* 1987). Ce jeu de données consiste en des données de fréquences alléliques à 43 locus polymorphes (parmi 111 locus allozymiques génotypés), pour cinq populations (France,

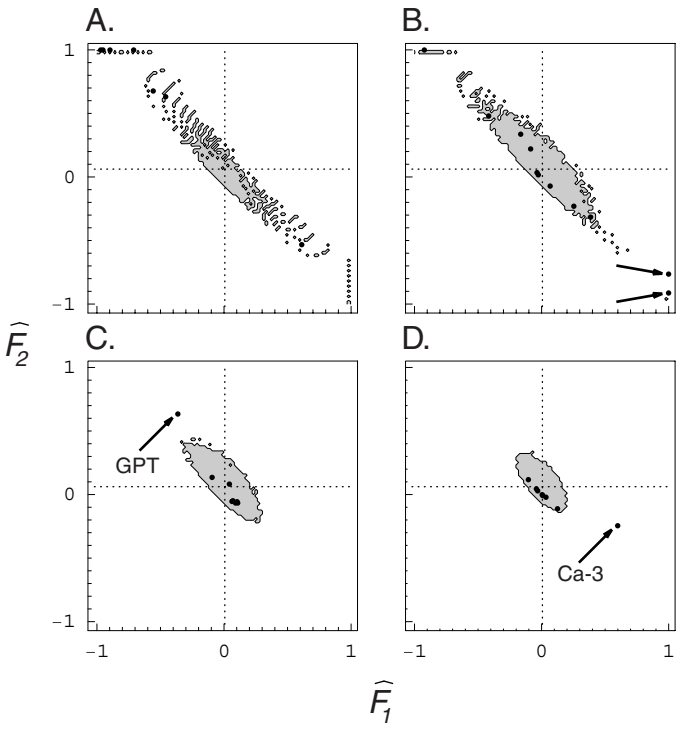


Figure 4.11: Valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ estimées pour 43 locus du jeu de données de *Drosophila simulans* dans la comparaison impliquant les populations de France ($n = 55$) et de Tunisie ($n = 52$). Les paires de valeurs par locus sont représentées chacune par un point. Les valeurs moyennes sont représentées par une ligne pointillée ($\widehat{F}_1 = 0.0064$ et $\widehat{F}_2 = 0.0617$). La région grisée représente la densité de probabilité où se trouvent 95% des données simulées pour ces valeurs de paramètres. Chacun des graphes représente une distribution conditionnée au nombre d'allèles dans l'échantillon total. A. $k = 2$. B. $k = 3$. C. $k = 4$. D. $k = 5$.

Congo, Le Cap, Seychelles, Tunisie) maintenues au laboratoire sous la forme de lignées isofemelles[20]. Les tailles d'échantillons haploïdes varient de $n = 26$ à $n = 55$.

La figure 4.11 montre le résultat de l'analyse conjointe de la popula-

[20]Les lignées iso-femelles sont construites en échantillonnant des femelles sur le terrain, puis en maintenant leur descendance au laboratoire par des croisements répétés entre frères et sœurs. Au bout de quelques générations, les lignées sont considérées comme étant homozygotes à la plupart des locus.

tion française et de la population tunisienne. Les estimateurs multilocus des paramètres F_1 (France) et F_2 (Tunisie) sont égaux à 0.0064 et 0.0617, respectivement (ces valeurs sont indiquées par les lignes pointillées). Toutes les paires de valeurs possibles de θ et T_0 ont été choisies avec $\theta = 1$, 5 ou 10 et pour $T_0 = 0.01$, 0.1 ou 1. Pour chaque ensemble de valeurs de θ et T_0, 10000 simulations du processus de coalescence ont été réalisées. Chaque distribution repose donc sur 90000 simulations indépendantes du processus de coalescence. On retrouve la plupart des points (chaque point correspondant à un locus) à l'intérieur de la limite imposée par la distribution des valeurs des estimateurs de F_1 et F_2 dans un modèle neutre de divergence de populations. Pourtant, si l'on considère la distribution conjointe des locus pour lesquels au moins trois allèles sont présents dans l'échantillon total (figure 4.11C-D), on remarque que quatre locus sont clairement à l'extérieur de la limite imposée par la distribution des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans le cas neutre[21]. On distinguera deux cas.

(i) Le premier cas concerne des locus dont l'une des valeurs mesurées $\widehat{F}_1$ ou $\widehat{F}_2$ est égale à un (et donc les points correspondant à ces locus se trouvent sur les bords des diagrammes : voir la figure 4.11B). Cette situation correspond au cas où un allèle est fixé dans l'une des populations. Or de légères variations dans les valeurs des paramètres de nuisance peuvent changer la proportion de locus pour lesquels un allèle est fixé dans une des deux populations. Il est donc possible de trouver des valeurs de paramètres telles que les locus indiqués dans la figure 4.11B se retrouvent finalement à l'intérieur de la limite imposée par la distribution des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ dans le cas neutre.

(ii) Le second cas concerne, dans notre exemple, les locus qui codent pour la glutamate pyruvate transaminase (GPT) et l'anhydrase carbonique-3 (Ca-3), c'est-à-dire des locus que l'on retrouve clairement en dehors de la limite de la « zone de confiance » à 95% définie par la distribution des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ obtenues dans un modèle strictement neutre (figures 4.11C et 4.11D).

[21] Avec 43 locus, on s'attend à ce que $0.05 \times 43 \approx 2$ locus se retrouvent, par hasard, en dehors de la distribution des valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ simulées.

Mais revenons sur le point (ii). Car si les locus qui codent pour la glutamate pyruvate transaminase ou bien pour l'anhydrase carbonique-3 sont bien à l'extérieur de la zone de confiance dans certaines comparaisons par paire impliquant la population française (comme ici dans la comparaison France-Tunisie, voir la figure 4.11), ces locus se placent au bord de la zone de confiance à 95% dans d'autres comparaisons par paire impliquant la population française. De plus, si l'on considère la région de confiance à 99% plutôt que celle à 95%, ces mêmes locus ne sont plus à l'extérieur de la région définie par la distribution des valeurs des estimateurs de F_1 et F_2 dans un modèle neutre de divergence de populations.[22]. Or l'on s'attend à ce qu'un locus, s'il a été ou est toujours sous l'effet de la sélection dans une population en particulier isolée des autres populations depuis leur divergence, soit détecté en dehors de la zone de confiance définie par le modèle neutre dans toutes, ou la plupart, des comparaisons par paire incluant cette population. Ainsi, dans l'analyse complète de ce jeu de données (non présentée ici) le locus codant pour l'isocitrate deshydrogenase-1 (détecté dans trois des quatre comparaisons par paire incluant la population des Seychelles) ainsi que 6 autres locus (détectés dans des comparaisons uniques) qui ont répondu positivement au test ne devraient pas être retenus comme des locus potentiellement soumis à l'action de la sélection.

En revanche, dans toutes les comparaisons par paire impliquant la population du Congo, deux locus codant respectivement pour la protéine larvaire 10 (Pt-10) et la phosphoglucomutase (PGM) ont été détectés comme étant à l'extérieur de la densité de probabilité où se trouvent 95% des données simulées pour les valeurs de paramètres correspondant aux temps de divergence mesurés pour la population du Congo, par rapport à toutes les autres. Dans toutes les comparaisons par paire incluant la population du Congo, en effet, les estimateurs des longueurs de branche pour ces deux locus donnent des valeurs plus grandes pour la population congolaise, par rapport à tous les autres locus (figure 4.12)[23]. Ceci suggère que la variabilité génétique est

[22]En d'autres termes, cela signifie qu'il existe des valeurs des paramètres de nuisance pour lesquelles ces locus ne se distinguent pas de locus simulés dans un modèle neutre.

[23]La seule exception notable concerne le locus codant pour la protéine larvaire 10 dans la comparaison Congo - Seychelles : voir la figure 4.12D.

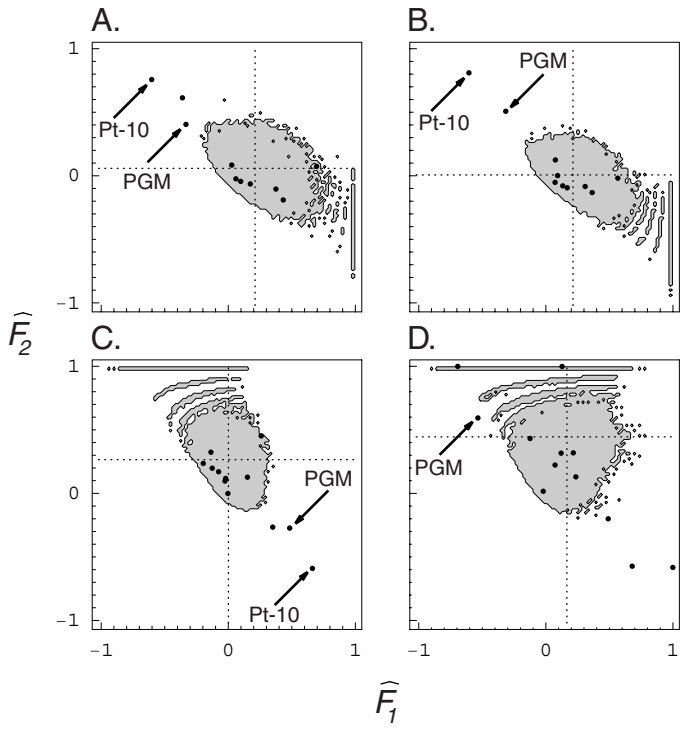


Figure 4.12: Valeurs de $\widehat{F}_1$ et $\widehat{F}_2$ estimées pour 43 locus du jeu de données de *Drosophila simulans* dans toutes les comparaisons impliquant la population du Congo ($n = 45$). Toutes les distributions sont conditionnées à $k = 4$ dans l'échantillon total. A. Comparaison entre la population française ($n = 55$) et la population congolaise. B. Tunisie ($n = 52$) et Congo. C. Congo et population du Cap (Afrique du Sud) ($n = 32$). D. Congo et Seychelles ($n = 26$).

sévèrement réduite à ces deux locus, plus qu'à tous les autres locus, et donc qu'un facteur autre que la dérive génétique agit ou a agi à ces locus dans cette population africaine. Le locus PGM a également été détecté à l'extérieur de la densité de probabilité où se trouvent 95% des données simulées pour les valeurs de paramètres correspondant aux temps de divergence mesurés pour la population des Seychelles, par rapport à toutes les autres. Ces deux locus ont été également détectés à l'extérieur de la zone de confiance à 99%, dans les deux cas. Pour ces locus, il semble donc qu'aucun scénario plausible de divergence des populations dans un modèle neutre ne puisse produire un pa-

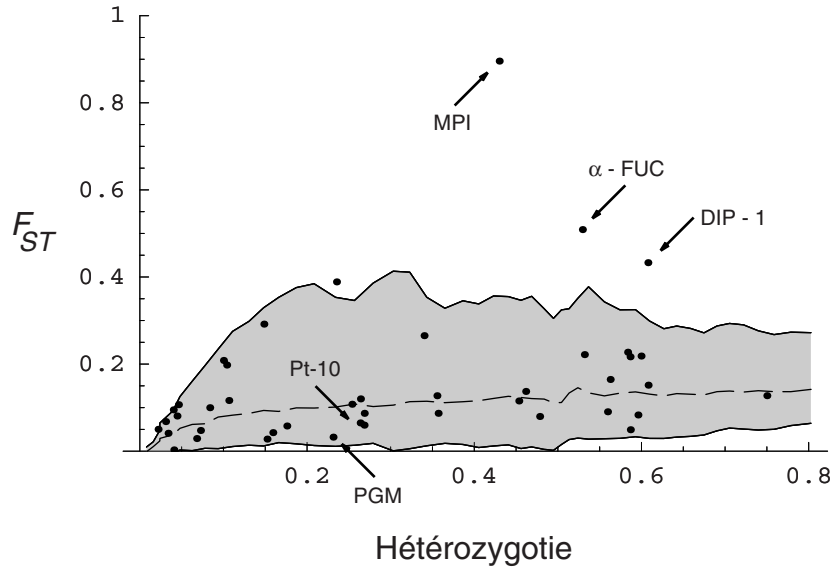


Figure 4.13: Application de la méthode de BEAUMONT et NICHOLS (1996) au jeu de données *D. simulans* décrit dans SINGH *et al.* (1987) et CHOUDHARY *et al.* (1992). Chaque point représente un locus. La distribution attendue de F_{ST} en fonction de l'hétérozygotie est figurée en gris clair, et la médiane de la distribution est indiquée par une ligne pointillée.

tron similaire à celui observé. La conclusion la plus satisfaisante semble donc impliquer l'action d'une forme de sélection sur ces locus, de façon directe ou indirecte (c'est-à-dire sur des régions se trouvant à proximité de ces locus, dans le génome : voir § 3.3).

Comparaison avec la méthode de BEAUMONT et NICHOLS (1996)

Le même jeu de données de *D. simulans* a été analysé par la méthode proposée par BEAUMONT et NICHOLS (1996)[24]. Trois locus, codant respectivement pour l'α-fucosidase (α-FUC), la dipeptidase-1 (DIP-1) et la mannose phosphatase isomerase (MPI) se trouvent clairement à l'extérieur de l'enveloppe contenant 95% de la distribution de $\widehat{F}_{ST}$ en fonction de l'hétérozygotie

[24]Le jeu de données analysé par BEAUMONT et NICHOLS dans leur article concernait l'espèce *D. melanogaster*. Voir SINGH et RHOMBERG (1987)

(figure 4.13). Or aucun de ces trois locus n'a été détecté par la méthode que j'ai présentée ici. Une interprétation possible est que la présence d'une population insulaire distante (Seychelles) pourrait mettre à mal la méthode de Beaumont et Nichols (1996). Il faut bien garder en mémoire que cette méthode repose sur des simulations d'un modèle à nombre fini d'îles. Or des populations hétérogènes quant à des paramètres tels que la taille efficace, le taux d'immigration, rendent la méthode plus sensible au risque de première espèce (Beaumont et Nichols 1996). J'ajouterai que l'isolement des populations (c'est-à-dire des taux de migration très faibles entre populations), associé à d'éventuels goulots d'étranglement démographique sont susceptibles d'augmenter également le risque de première espèce associé à l'approche de Beaumont et Nichols (1996). Dans des scénarios de divergence de populations suivie d'un isolement très fort, la fixation d'allèles privés dans certaines populations est un cas de figure relativement fréquent. Or les valeurs de F_{ST} à l'échelle globale peuvent être augmentées de façon significative à cause justement de ces locus fixés pour un allèle dans certaines populations. De tels locus risquent donc d'être identifiés comme potentiellement soumis à l'action de la sélection par la méthode de Beaumont et Nichols (1996). De telles situations (isolement de populations, goulots d'étranglement démographique) sont susceptibles d'augmenter le risque de première espèce associé à la méthode de Beaumont et Nichols (1996). En général, les locus détectés par leur approche sont des locus pour lesquels les valeurs de F_{ST} (à l'échelle de toutes les populations échantillonnées) sont beaucoup plus fortes que la valeur moyenne, sur tous les locus. Or si la sélection sur les locus n'agit qu'à une échelle géographique locale, ce qui se traduit par une réduction locale de la variation génétique à ces locus (voir le § 3.3), mais que la distribution de la variation n'est pas affectée à l'échelle globale, alors il est peu probable que la méthode de Beaumont et Nichols (1996) permette de détecter ces locus. Dans ce cas de figure, au contraire, la méthode que j'ai présentée ici apparaît plus adaptée.

Raufaste et Bonhomme (2002) ont récemment proposé une approche alternative à la méthode de Beaumont et Nichols (1996) et à la nôtre. Leur approche repose sur le comportement différent des estimateurs multi-

locus du paramètre F_{ST} de Weir et Cockerham (1984) d'une part et de Robertson et Hill (1984) d'autre part. Raufaste et Bonhomme (2000) ont en effet montré que ces deux estimateurs ont bien la même espérance dans un modèle en îles neutre, mais qu'ils ont des espérances différentes dès lors que certains allèles sont soumis à l'action de la sélection. Raufaste et Bonhomme (2002) proposent donc l'utilisation d'une statistique, la différence des deux estimateurs de F_{ST}, et une approche par simulation pour estimer la distribution attendue de cette statistique, afin de tester l'hypothèse nulle de neutralité du locus en question. Contrairement à la méthode de Beaumont et Nichols (1996) et à la nôtre, l'approche de Raufaste et Bonhomme (2002) consiste à générer la distribution attendue d'une statistique à un seul locus. En revanche, comme la méthode de Beaumont et Nichols (1996) et comme la nôtre, la prise en compte de toutes les valeurs observées des estimateurs de F_{ST} sur l'ensemble des locus d'un même jeu de données permet de mieux calibrer les paramètres des simulations stochastiques.

4.3.3 Limites de l'approche

La méthode que j'ai présentée fait l'hypothèse d'un isolement complet après l'épisode de divergence ce qui peut sembler, à juste titre, peu réaliste. Une première façon d'aborder ce problème serait d'estimer conjointement les flux de gènes et les temps de divergence à partir des données. Ceci demanderait une approche beaucoup plus lourde que celle que j'ai présentée ici (voir, par exemple, Nielsen et Slatkin 2000). Une autre façon de poser le problème est de chercher à quantifier l'effet de la migration après la séparation des populations sur le risque de première espèce associé à notre méthode. Pour cela, une version modifiée de l'algorithme présenté par Hudson (1990) a été utilisée pour simuler un processus de coalescence semblable à celui décrit précédemment mais qui intègre des flux de gènes symétriques entre les deux populations en divergence (voir, par exemple, Norborg 2001; Strobeck 1987; Takahata 1988).

Le tableau 4.3 donne une indication des performances de la méthode pour des modèles de divergence contrastés. Pour chaque jeu de valeurs des para-

Tableau 4.3: Applications de la méthode d'identification de locus soumis à la sélection. La méthode a été testée pour différents scénarios de divergence de populations. Pour chaque jeu de paramètres, 50 locus ont été simulés, pour 100 individus haploïdes au total (50 dans chaque population). La moyenne et la médiane du nombre de locus détectés comme étant potentiellement sélectionnés sont indiquées, ainsi que la valeur de p associée aux tests de Wilcoxon (tests unilatéraux ; H_0 : la distribution du nombre de locus détecté est centrée sur 2.5 ; H_1 : la distribution du nombre de locus est supérieure à 2.5).

			Locus détectés		
μ	θ	T_0	*Moyenne*	*Médiane*	p
Pas de migration : $m = 0$					
10^{-5}	1	1	1.85	2.0	0.98
10^{-5}	10	10^{-2}	1.15	1.0	1.00
10^{-3}	1	1	2.60	3.0	0.28
10^{-3}	10	10^{-2}	1.75	2.0	0.76
Faible migration : $m = 0.01$					
10^{-5}	1	1	2.30	2.5	0.79
10^{-5}	10	10^{-2}	2.25	2.0	0.77
10^{-3}	1	1	2.00	2.0	0.99
10^{-3}	10	10^{-2}	1.20	1.0	1.00
Migration modérée : $m = 0.1$					
10^{-5}	1	1	2.30	2.0	0.87
10^{-5}	10	10^{-2}	2.05	2.0	0.96
10^{-3}	1	1	2.25	2.0	0.89
10^{-3}	10	10^{-2}	1.85	2.0	0.98

mètres, 20 jeux de données artificiels ont été simulés ($n_1 = n_2 = 50$), chacun pour 50 locus indépendants. Avec 50 marqueurs, on s'attend à trouver, par hasard, $0.05 \times 50 = 2.5$ locus en dehors des limites définies par le modèle neutre. Dans chaque situation, un test de signe-et-rang de Wilcoxon (voir, par exemple, Mendenhall *et al.* 1990) permet de montrer que le nombre de locus identifiés comme étant potentiellement soumis à la sélection n'est pas significativement supérieur à 2.5 (test unilatéral). Par conséquent, cette

méthode peut être considérée comme conservative dans la mesure où, au seuil de 5%, des locus générés dans un modèle neutre de divergence ne sont pas détectés comme étant potentiellement soumis à la sélection.

Chapitre 5

Conclusion

En guise de conclusion à ces travaux, je souhaiterais tout d'abord revenir sur l'une des principales notions qui est à la source des modèles que j'ai développés. Il s'agit de la nature de la variation génétique que j'ai considérée. Je discuterai ensuite des aspects originaux des modèles que j'ai construits, mais aussi de leurs limites. Enfin, j'évoquerai quelques perspectives.

5.1 Quelle variation génétique ?

Tout au long de ce document, je me suis principalement intéressé à la variation génétique que l'on qualifie de *neutre* vis-à-vis de la sélection naturelle, c'est-à-dire de la part de la variation, héritable ou non, qui n'est influencée par aucun phénomène sélectif. Pourtant, il ne fait nul doute qu'il ne s'agit que d'un type de variation génétique parmi d'autres. Godelle *et al.* (1998) distinguent trois catégories de la diversité génétique, fondées sur des aspects mécanistes de l'apparition ou du maintien de cette diversité, à savoir la diversité génétique *neutre*, *sélectionnée transitoire* et *sélectionnée maintenue*. La diversité sélectionnée transitoire trouve sa source dans l'apparition de mutations délétères que la sélection tend à éliminer (avec plus ou moins de facilité, selon la taille de la population : voir, par exemple, Lynch *et al.* 1995a; Lynch et Gabriel 1990), ou bien dans l'apparition de mutations favorables que la sélection tend à fixer dans les populations. La dernière catégorie de

diversité génétique (la diversité sélectionnée maintenue) caractérise, quant à elle, la variation maintenue par l'effet de la sélection. Il s'agit par exemple des mécanismes fréquence-dépendants qui avantagent les génotypes rares au détriment des génotypes fréquents. Dans une revue récente des travaux publiés dans le domaine de la génétique de la conservation, Hedrick (2001) distingue également trois catégories de variation génétique. Pour autant, je qualifierais sa classification de *fonctionnelle* au regard de la classification plus *mécaniste* de Godelle *et al.* (1998). Hedrick (2001) distingue les variations *neutre*, *nuisible* et *adaptative*, opposant ainsi la variation associée aux effets délétères, à la variation associée aux effets favorables des mutations.

Godelle *et al.* (1998) et Hedrick (2001) apprécient donc différemment l'effet de la sélection (qu'elle favorise l'adaptation locale ou bien qu'elle débarrasse le génome de variants délétères) : là où les premiers privilégient l'aspect dynamique de la variation, le second s'intéresse plutôt aux effets des gènes sur la valeur sélective moyenne des populations. Malgré ces points de vue légèrement divergents, Godelle *et al.* (1998) et Hedrick (2001) qualifient de la même manière (et comme bien d'autres) le polymorphisme neutre, à savoir un polymorphisme dont la distribution ne dépend pas d'effets directs ou indirects d'une quelconque forme de sélection. Tout au long de cette thèse, j'ai montré combien la distribution de la variation neutre était influencée par toute forme de structure des populations, à savoir la structure spatiale (§ 2.1.3), la structure en classes d'individus (§ 2.2.1), la structure temporelle (§ 2.2.2). J'ai également montré combien la distribution de la variation neutre était influencée par les interactions des gènes à différents locus le long du génome (§ 2.3). Ces modèles m'ont permis, en retour, de proposer des méthodes d'inférence de paramètres d'intérêt pour le biologiste des populations, c'est-à-dire la taille efficace (§ 4.1) et les taux de dispersion (§ 4.1 et 4.2). D'un point de vue purement théorique donc, la notion de « gènes neutres » est indispensable pour comprendre l'effet des différentes forces évolutives que sont la mutation, la dérive, la migration, la recombinaison et le système de reproduction sur le niveau et la distribution de la variation génétique. Toutefois, comme je l'ai souligné dans le § 3.3 le niveau de variation génétique neutre, ainsi que sa distribution (voir Charlesworth *et al.*

1997), peuvent également être influencés par l'action d'une forme de sélection agissant à proximité dans le génome. Ce dernier point a été développé plus en détail dans le § 4.3 où j'ai présenté une méthode permettant d'apprécier l'homogénéité de réponse d'un ensemble de marqueurs génétiques à l'histoire des populations.

Le constat que la notion de diversité génétique neutre pourrait n'exister que d'un point de vue formel soulève un certain nombre de questions et souligne également les difficultés de l'approche indirecte en général (voir le § 1.1.2). D'un point de vue plus empirique, ce constat nous amène nécessairement à porter un regard plus critique sur les études de structure génétique, par exemple. Mais il offre également de nouvelles perspectives pour l'étude des populations naturelles. Cela m'amène (de façon tout à fait réductrice) à formuler trois remarques.

(i) L'approche qui a été la plus largement suivie jusqu'à aujourd'hui consiste à utiliser des marqueurs génétiques pour retracer l'histoire des populations ou bien pour interpréter leur structure (voir, par exemple, le § 1.2). Les limites de cette approche sont liées d'une part au manque de réalisme des modèles d'interprétation de la variation (lié par exemple aux hypothèses du modèle en îles: voir Bossart et Prowell 1998; Whitlock et McCauley 1999) et d'autre part à l'hétérogénéité de réponse des marqueurs moléculaires à l'histoire démographique des populations (voir Beaumont et Nichols 1996, et le § 4.3).

(ii) Une autre approche, beaucoup moins répandue, s'intéresse justement à tirer parti de l'hétérogénéité de réponse des marqueurs moléculaires à l'histoire démographique des populations. A travers une étude récente du polymorphisme génétique dans des populations morphologiquement diverses de la Littorine rugueuse (*Littorina saxatilis*), Wilding *et al.* (2001) ont identifié 15 locus AFLP[1], parmi les 306 utilisés, potentiellement soumis à l'action de la sélection (en utilisant l'approche de Beaumont et Nichols 1996). Ces quinze locus expliquent, à eux seuls, le regroupement génétique des différentes formes morphologiques de Lit-

[1]Il s'agit d'un polymorphisme de longueur de fragments amplifiés. Traduit de l'anglais *amplified fragment length polymorphism*.

torines entre elles. Tous les autres locus montrent un patron d'isolement par la distance, quelle que soit la morphologie. De telles études soulignent l'hétérogénéité des flux de gènes dans le génome, liée dans cette étude à des phénomènes de spéciation.

(iii) Enfin une dernière approche est soulignée par Hedrick (2001) qui développe l'idée selon laquelle la variation génétique neutre pourrait être, dans certains cas, un *indicateur* de la variation nuisible et adaptative. Toutefois les relations entre les différents types de diversité ne sont pas triviales (Latta 1998). Dans le même ordre d'idée, on peut supposer que la compréhension de l'histoire démographique des populations peut sous certaines conditions permettre des inférences sur le niveau de variation nuisible ou adaptative qui peut être maintenu.

Durant les vingt dernières années, l'étude du polymorphisme neutre a connu son heure de gloire avec les marqueurs enzymatiques et, plus récemment, les marqueurs microsatellites. Ces études deviennent de plus en plus aisées à mettre en œuvre de façon pratique au laboratoire, et les coûts associés ont considérablement diminué. Désormais, les outils d'interprétation de ces données font également florès mais ils montrent aussi souvent leurs limites (Rousset et Raymond 1997).

Il est probable que dans un avenir proche, les jeux de données comporteront un nombre très important de locus cartographiés dans les génomes. Ces données nous permettront sans doute de différencier des catégories de marqueurs selon leurs taux de recombinaison, ou bien selon leur réponse à un phénomène sélectif. Il sera dès lors possible de raisonner en terme d'« écologie du génome ».

5.2 Originalités et limites des approches utilisées

Les méthodes que j'ai présentées dans cette thèse s'inscrivent en droite ligne des modèles de génétique des populations classiques. Dans la plupart des cas, je n'ai fait qu'ajouter un niveau de complexité à des modèles qui

existaient déjà. Toutefois, je dégagerai trois aspects plus originaux de ces travaux.

(i) Le premier est lié aux paramètres qui ont été définis, spécifiquement à *une* population. Que ce soit pour la définition des paramètres d'identité génétique à un locus (F) ou à deux locus (η') dans la méthode d'estimation de l'effectif efficace et du taux d'immigration (§ 4.1 et annexe F) ou bien pour la définition des paramètres F_1 et F_2 dans la méthode de détection de locus potentiellement soumis à la sélection (§ 4.3 et annexe I), les paramètres ont été définis pour chaque population. Dans ces deux cas, la somme pondérée des paramètres F (pour un locus) est égale au paramètre F_{ST}. Ces paramètres spécifiques à une population ont ceci d'intéressant qu'ils n'imposent pas l'égalité des paramètres démographiques (nombre d'individus, taux d'immigration) entre toutes les populations connectées. Comme l'ont souligné Gaggiotti et Excoffier (2000), les paramètres définis à l'échelle de toutes les populations (tels que le F_{ST}), même s'ils sont corrigés par des différences entre les tailles actuelles des populations, ne permettent pas de retracer avec précision l'histoire de dérive de ces populations.

(ii) Le second aspect original concerne la mise en évidence de stratégies d'échantillonnage adéquates pour estimer des taux de migration entre populations. De la même manière que des échantillonnages répétés dans le temps peuvent permettre, dans une population isolée, de mesurer l'intensité de la dérive (Anderson *et al.* 2000; Nei et Tajima 1981; Pollak 1983; Waples 1989; Williamson et Slatkin 1999), l'échantillonnage d'individus avant et après l'épisode de dispersion permet d'en estimer l'intensité.

(iii) Enfin, la prise en compte des interactions entre locus pour estimer l'intensité des forces qui construisent ces déséquilibres d'identité (taille efficace, taux d'immigration) pourrait être utilisée dans l'avenir sur les organismes dont la carte génétique a été établie. La définition de groupes de marqueurs selon les taux de recombinaison permettrait, par exemple, d'obtenir des estimations « indépendantes » de l'intensité de

la dérive dans différentes régions du génome.

Que ces aspects soient plus ou moins originaux n'empêche pas ces approches d'être confrontées à un certain nombre de limites. Ces limites ont été signalées et développées tout au long du texte de synthèse. Toutefois, je souhaiterais revenir sur deux points qui me paraissent essentiels.

(i) Le premier point concerne le polymorphisme. Il est clair que les méthodes d'inférence que j'ai proposées nécessitent un polymorphisme suffisamment important pour donner des résultats satisfaisants, particulièrement en ce qui concerne la mesure de la taille efficace et du taux d'immigration (§ 4.1 et annexe F). Lorsque l'on s'intéresse aux espèces rares, qui sont généralement peu variables (Gitzendanner et Soltis 2000), ce problème devient majeur. Cela nous amène également à considérer le paradoxe suivant. La méthode que j'ai présentée ne permet d'estimer de façon satisfaisante que de faibles tailles efficaces. Or c'est justement dans les petites populations que le polymorphisme est le plus faible. Il est donc probable que cette méthode ne fonctionne que dans une gamme de paramètres (taille efficace et taux d'immigration) particulière, en dehors de laquelle l'estimation conjointe des paramètres ne sera pas satisfaisante. Ainsi, étant donné le faible polymorphisme mesuré chez *M. strigosa*, la mesure de taille efficace n'a pu être réalisée à partir du jeu de données obtenu chez cette espèce.

(ii) Le second point concerne le nombre de locus. La méthode d'identification des locus potentiellement soumis à l'action de la sélection (§ 4.3 et annexe I) nécessite un nombre suffisamment grand de locus. Encore une fois, le nombre de locus polymorphes généralement utilisés dans les études d'espèces rares ou menacées est à l'heure actuelle encore une limite à l'utilisation de cette méthode.

Ces limites sont naturellement communes à bon nombre d'approches utilisées aujourd'hui en génétique des populations. Mais au delà de ces limites imposées par le niveau de variation génétique tel que mesuré dans les populations naturelles, il est bien évident que les modèles que j'ai présentés mériteraient qu'un certain nombre d'hypothèses soient relâchées. Par exemple, la prise

en compte d'un modèle de dispersion plus réaliste que le modèle en îles (tel qu'un modèle d'isolement par la distance, par exemple) serait un apport indéniable à la méthode d'estimation des taux de dispersion spécifiques au sexe (§ 4.2 et annexe H).

5.3 Evolution des traits d'histoire de vie

Les traits d'histoire de vie tels que la dispersion, ou bien la dormance, dont nous avons vu qu'ils influencent fortement la distribution de la variation génétique dans les populations, sont des traits susceptibles d'évoluer. Par exemple, les traits liés à la dispersion évoluent en réponse à l'hétérogénéité temporelle et spatiale de l'environnement (McPeek et Holt 1999; Olivieri *et al.* 1995), ou bien pour éviter la dépression de consanguinité (Gandon 1999), pour échapper à la compétition entre individus apparentés (Frank 1986; Hamilton et May 1977) ou encore pour toutes ces raisons à la fois (Gandon et Michalakis 1999). De même, la dormance des graines chez les espèces végétales ou bien la diapause des œufs chez certains groupes d'animaux (insectes, crustacés) évoluent en réponse aux mêmes pressions de sélection (Olivieri 2001; Rees 1996; Venable et Brown 1988, 1993).

Les travaux récents de Rousset et Billiard (2000) ont permis de clarifier l'approche de Taylor et Frank (1996) qui définissaient la notion de « valeur sélective directe »[2] comme la valeur sélective d'un individu en fonction de son propre comportement mais aussi du comportement de tous les individus avec lesquels il peut être en compétition. En particulier, Rousset et Billiard (2000) ont montré que la définition de l'apparentement entre les individus dans ces modèles repose sur la définition de l'identité de paires de gènes à différents niveaux de hiérarchie (voir p. 27). Dans ce contexte théorique, on peut ainsi étudier l'influence des traits d'histoire de vie sur la distribution de la variation génétique (qui détermine localement les relations de parenté entre les individus) qui influence, en retour, les pressions de sélection qui s'exercent sur ce trait.

[2]Traduit de l'anglais *direct fitness*.

La prise en compte de situations plus complexes (structure en classes d'individus) et de modèles démographiques explicites, aussi bien que de systèmes génétiques plus réalistes (interactions entre gènes, recombinaison) sont actuellement, et pour quelques années encore, au cœur des développements théoriques de la discipline.

Bibliographie

Allee, W. C., O. Emerson, T. Park et K. P. Schmidt (1950) *Principles of Animal Ecology*. Saunders WB, Philadelphie.

Allsopp, A. (1952) Longevity of *Marsilea sporocarps*. *Nature*, **169** : 79–80.

Alvarez-Buylla, E. R., A. Chaos, D. Pinero et A. A. Garay (1996) Demographic genetics of a pioneer tropical tree species : patch dynamics, seed dispersal, and seed banks. *Evolution*, **50** : 1155–1166.

Anderson, E. C., E. G. Williamson et E. A. Thompson (2000) Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics*, **156** : 2109–2118.

Avery, P. J. et W. G. Hill (1979) Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. *Genetics*, **91** : 817–844.

Avise, J. C. (1994) *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Balloux, F., J. Goudet et N. Perrin (1998) Breeding system and genetic variance in the monogamous semi-social shrew, *Crocidura russula*. *Evolution*, **52** : 1230–1235.

Bartley, D., M. Bagley, G. Gall et B. Bentley (1992) Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. *Conservation Biology*, **6** : 365–375.

Barton, N. H. (1994) The reduction in fixation probability caused by substitutions at linked loci. *Genetical Research*, **64** : 199–208.

Barton, N. H. (1995) Linkage and the limits to natural selection. *Genetics*, **140** : 821–841.

Barton, N. H. (1998) The effect of hitch-hiking on neutral genealogies. *Genetical Research*, **72** : 123–133.

Barton, N. H. et M. C. Whitlock (1997) The evolution of metapopulations. In : *Metapopulation dynamics : Ecology, Genetics, and Evolution* (eds. Hanski, I. et M. Gilpin), pp. 183–210. Academic press, New York.

Baudry, E., C. Kerdelhué, H. Innan et W. Stephan (2001) Species and recombination effects on DNA variability in the tomato genus. *Genetics*, **158** : 1725–1735.

Beaumont, M. A. et R. A. Nichols (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B*, **263** : 1619–1626.

Beerli, P. et J. Felsenstein (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152** : 763–773.

Beerli, P. et J. Felsenstein (2001) Maximum-likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America*, **98** : 4563–4568.

Belovsky, G. E., C. Mellison, C. Larson et A. Van Zandt (1999) Experimental studies of extinction dynamics. *Science*, **286** : 1175–1177.

Bennett, J. H. (1954) On the theory of random mating. *Annals of Eugenics*, **18** : 311–317.

Bennett, J. H. et F. E. Binet (1956) Association between Mendelian factors with mixed selfing and random mating. *Heredity*, **10** : 51–56.

Berg, L. M., M. Lascoux et P. Pamilo (1998) The infinite island model with sex-differentiated gene flow. *Heredity*, **80** : 63–68.

Blondel, J. et J. Aronson (1999) *Biology and Wildlife of the Mediterranean Region*. Oxford University Press, Oxford.

Boileau, M. G., P. D. N. Hebert et S. S. Schwartz (1992) Non-equilibrium gene frequency divergence : persistent founder effects in natural populations. *Journal of Evolutionary Biology*, **5** : 25–39.

Bossart, J. L. et D. P. Prowell (1998) Genetics estimates of population structure and gene flow : limitations, lessons and new directions. *Trends in Ecology and Evolution*, **13** : 202–206.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd et L. L. Cavalli-Sforza (1991) Drift, admixture, and selection in human evolution : a study with DNA polymorphisms. *Genetics*, **88** : 839–843.

Caballero, A. (1994) Developments in the prediction of effective population size. *Heredity*, **73** : 657–679.

Cabin, R. J., R. J. Mitchell et D. L. Marshall (1998) Do surface plant and soil seed bank populations differ genetically ? a multipopulation study of the desert mustard *Lesquerella fendleri* (Brassicaceae). *American Journal of Botany*, **85** : 1098–1109.

Caswell, H. (1989) *Matrix Population Models*. Sinauer Associates Inc., Sunderland.

Caswell, H. (2001) *Matrix Population Models : Construction, Analysis, and Interpretation*. Sinauer Associates Inc., Sunderland, 2nd edn.

Cavalli-Sforza, L. L. (1966) Population structure and human evolution. *Proceedings of the Royal Society of London Series B*, **164** : 362–379.

Chakraborty, R. et H. Danker-Hopfe (1991) Analysis of population structure : a comparative study of different estimators of Wright's fixation

indices. In : *Statistical Methods in Biological and Medical Sciences* (eds. Rao, C. et R. Chakraborty), pp. 203–254. Elsevier Science Publishers, Amsterdam.

Channell, R. et M. V. Lomolino (2000) Dynamic biogeography and conservation of endangered species. *Nature*, **403** : 84–86.

Charlesworth, B. (1994a) The effects of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, **63** : 213–227.

Charlesworth, B. (1994b) *Evolution in Age-Structured Populations*. Cambridge University Press, Cambridge, 2nd edn.

Charlesworth, B. (2001) The effect of life-history and mode of inheritance on neutral genetic variability. *Genetical Research*, **77** : 153–166.

Charlesworth, B., M. T. Morgan et D. Charlesworth (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134** : 1289–1303.

Charlesworth, B., M. Nordborg et D. Charlesworth (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70** : 155–174.

Chesser, R. K. (1991a) Gene diversity and female philopatry. *Genetics*, **127** : 437–447.

Chesser, R. K. (1991b) Influence of gene flow and breeding tactics on gene diversity within populations. *Genetics*, **129** : 573–583.

Chesser, R. K., O. E. Rhodes, Jr, D. W. Sugg et A. F. Schnabel (1993) Effective sizes for subdivided populations. *Genetics*, **135** : 1221–1232.

Choudhary, M., M. B. Coulthart et R. S. Singh (1992) A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. VI. Patterns and processes of genic divergence between *D. melanogaster* and its sibling species, *D. simulans*. *Genetics*, **130** : 843–853.

Choy, S. C. et B. S. Weir (1978) Exact inbreeding coefficients in populations with overlapping generations. *Genetics*, **89** : 591–614.

Cockerham, C. C. (1969) Variance of gene frequencies. *Evolution*, **23** : 72–84.

Cockerham, C. C. (1973) Analyses of gene frequencies. *Genetics*, **74** : 697–700.

Cockerham, C. C. et B. S. Weir (1987) Correlations, descent measures : drift with migration and mutation. *Proceedings of the National Academy of Sciences of the United States of America*, **84** : 8512–8514.

Cockerham, C. C. et B. S. Weir (1993) Estimation of gene flow from f-statistics. *Evolution*, **47** : 855–863.

Crow, J. F. et K. Aoki (1984) Group selection for a polygenic behavioral trait : estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the United States of America*, **81** : 6073–6077.

Crow, J. F. et C. Denniston (1988) Inbreeding and variance effective population numbers. *Evolution*, **42** : 482–495.

Crow, J. F. et M. Kimura (1970) *An introduction to Population Genetics Theory*. Burgess Publishing Company, Minneapolis.

Darlington, C. D. (1939) *The Evolution of Genetic Systems*. Oliver and Boyd, Edimbourg.

De Mauro, M. M. (1993) Relationship of breeding system to rarity in the lakeside daisy (*Hymenoxys acaulis* var. *glabra*). *Conservation Biology*, **7** : 542–550.

Del Castillo, R. F. (1994) Factors influencing the genetic structure of *Phacelia dubia*, a species with a seed bank and large fluctuations in population size. *Heredity*, **72** : 446–458.

Denton, M. (1988) *Évolution : une Théorie en Crise*. Flammarion, Paris.

DiCiccio, T. et B. Efron (1996) Bootstrap confidence intervals. *Statistical Science*, **11** : 189–228.

Dobson, A. P. (1996) *Conservation and Biodiversity*. Scientific American Librairy, Freeman, New York.

Dobson, F. S. (1982) Competition for mates and predominant juvenile male dispersal in mammals. *Animal Behaviour*, **30** : 1183–1192.

Eldredge, N. (1998) *Life in the Balance - Humanity and the Biodiversity Crisis*. Princeton University Press, Princeton.

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

Estoup, A., F. Rousset, Y. Michalakis, J.-M. Cornuet, M. Adriamanga et R. Guyomard (1998) Comparative analysis of microsatellite and allozyme markers : a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology*, **7** : 339–353.

Evett, I. et B. Weir (1998) *Interpreting DNA Evidence*. Sinauer, Sunderland.

Ewens, W. J. (1969) *Population Genetics*. Methuen & Co LTD, Londres.

Ewens, W. J. (1979) *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Excoffier, L. (2001) Analysis of population subdivision. In : *Handbook of Statistical Genetics* (eds. Balding, D. J., M. Bishop et C. Cannings), pp. 271–307. John Wiley & Sons, Ltd, Chichester.

Favre, L., F. Balloux, J. Goudet et N. Perrin (1997) Female-biased dispersal in the monogamous mammal *Crocidura russula* : evidence from field data and microsatellite patterns. *Proceedings of the Royal Society of London Series B*, **264** : 127–132.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, vol. 1. John Wiley & Sons, Inc., Singapour, 3rd edn.

Felsenstein, J. (1971) Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics*, **68** : 581–597.

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

Felsenstein, J. (1982) How can we infer geography and history from gene frequencies ? *Journal of Theoretical Biology*, **96** : 9–20.

Felsenstein, J. (1987) Sex and the evolution of recombination. In : *The Evolution of Sex, An Examination of Current Ideas* (eds. Michod, R. E. et B. R. Levin), pp. 74–86. Sinauer Associates, Sunderland.

Ferrière, R. et M. Gatto (1993) Chaotic population dynamics can result from natural selection. *Proceedings of the Royal Society of London Series B*, **251** : 33–38.

Fisher, R. A. (1930) *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

FitzSimmons, N., C. J. Limpus, J. A. Normans, A. R. Goldizen et J. D. Miller (1997a) Philopatry of male marine turtles inferred from mitochondrial DNA markers. *Proceedings of the National Academy of Sciences of the United States of America*, **94** : 8912–8917.

FitzSimmons, N., C. Moritz, C. J. Limpus, L. Pope et R. Prince (1997b) Geographic structure of mitochondrial and nuclear gene polymorphims in Australian green turtle populations and male-biased gene flow. *Genetics*, **147** : 1843–1854.

Flannery, T. F. (1999) Debating extinction. *Science*, **283** : 182–183.

Frank, S. A. (1986) Dispersal polymorphism in subdivided populations. *Journal of Theoretical Biology*, **122** : 303–309.

Frankham, R. (1995) Inbreeding and extinctions : a threshold effect. *Conservation Biology*, **9** : 792–799.

Frankham, R. (1998) Inbreeding and extinctions : island populations. *Conservation Biology*, **12** : 665–675.

Fréville, H., F. Justy et I. Olivieri (2001) Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). *Molecular Ecology*, **10** : 879–889.

Gabriel, W. et R. Bürger (1994) Extinction risk by mutational meltdown : synergistic effects between population regulation and genetic drift. In : *Conservation Genetics* (eds. Loeschcke, V., J. Tomiuk et S. K. Jain), pp. 69–84. Birkhäuser Verlag, Bâle.

Gaggiotti, O. E. et L. Excoffier (2000) A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proceedings of the Royal Society of London Series B*, **267** : 81–87.

Gale, J. S. (1990) *Theoretical Population Genetics*. Unwin Hyman, Londres.

Gandon, S. (1999) Kin competition, the cost of inbreeding and the evolution of dispersal. *Journal of Theoretical Biology*, **200** : 345–364.

Gandon, S., Y. Capowiez, Y. Dubois, Y. Michalakis et I. Olivieri (1996) Local adaptation and gene-for-gene coevolution in a metapopulation model. *Proceedings of the Royal Society of London Series B*, **263** : 1003–1009.

Gandon, S., D. Ebert, I. Olivieri et Y. Michalakis (1998) Differential adaptation in spatially heterogeneous environments and host-parasite

coevolution. In : *Genetic structure and local adaptation in natural insect populations* (eds. Mopper, S. et S. Strauss), pp. 325–340. Chapman & Hall, New York.

Gandon, S. et Y. Michalakis (1999) Evolutionary stable dispersal rate in a metapopulation with extinctions and kin competition. *Journal of Theoretical Biology*, **199** : 275–290.

Gandon, S. et F. Rousset (1999) The evolution of stepping-stone dispersal rates. *Proceedings of the Royal Society of London Series B*, **266** : 2507–2513.

Gillespie, J. H. (2000) Genetic drift in an infinite population : the pseudohitchhiking model. *Genetics*, **155** : 909–919.

Gitzendanner, M. A. et P. S. Soltis (2000) Patterns of genetic variation in rare and widespread plant congeners. *American Journal of Botany*, **87** : 783–792.

Godelle, B., F. Austerlitz, S. Brachet, B. Colas, J. Cuguen, S. Gandon, P. H. Gouyon, M. Lefranc, I. Olivieri, X. Reboud et R. Vitalis (1998) Système génétique, polymorphisme neutre et sélectionné : implications en biologie de la conservation. *Genetics Selection Evolution*, **30 (Suppl. 1)** : S15–S28.

Godt, M. J. W., J. L. Hamrick et S. Bratton (1995) Genetic diversity in a threatened wetland species, *Helonias bullata* (Liliaceae). *Conservation Biology*, **9** : 596–604.

Godt, M. J. W., B. R. Johnson et J. L. Hamrick (1996) Genetic diversity and population size in four rare southern appalachian plant species. *Conservation Biology*, **10** : 796–805.

Godt, M. J. W., J. Walker et J. L. Hamrick (1997) Genetic diversity in the endangered lily *Harperocallis flava* and a close relative, *Tofieldia racemosa*. *Conservation Biology*, **11** : 361–366.

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

Goldstein, D. B. et C. Schlötterer (1999) *Microsatellites : Evolution and Applications*. Oxford University Press, Oxford.

Gomulkiewicz, R. et R. D. Holt (1995) When does evolution by natural selection prevent extinction ? *Evolution*, **49** : 201–207.

Gouyon, P. H., J. P. Henry et J. Arnould (1997) *Les Avatars du Gène. La Théorie Néodarwinienne de l'Évolution*. Belin, Paris.

Greenwood, P. et P. Harvey (1982) The natal and breeding dispersal of birds. *Annual Review of Ecology and Systematics*, **13** : 1–21.

Greenwood, P. J. (1980) Mating systems, philopatry and dispersal in birds and mammals. *Animal Behaviour*, **28** : 1140–1162.

Grosbois, V. (2001) *La dispersion : trait d'histoire de vie et paramètre démographique. Etude empirique dans une population de mouette rieuse* Larus ridibundus *L.* Thèse de doctorat, Université Montpellier II.

Hamilton, W. D. et R. M. May (1977) Dispersal in stable habitats. *Nature*, **269** : 578–581.

Hanski, I. et M. E. Gilpin (1997) *Metapopulation Biology : Ecology, Genetics, and Evolution*. Academic Press, San Diego.

Hartl, D. L. et A. G. Clark (1989) *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, 2nd edn.

Hedrick, P. W. (1987) Gametic disequilibrium : proceed with caution. *Genetics*, **117** : 331–341.

Hedrick, P. W. (2001) Conservation genetics : where are we now ? *Trends in Ecology and Evolution*, **16** : 629–636.

Hickey, R. J., M. A. Vincent et S. I. Guttman (1991) Genetic variation in running buffalo clover (*Trifolium stoloniferum*, Fabaceae). *Conservation Biology*, **5** : 309–316.

Hill, W. (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetics*, **38** : 200–216.

Hill, W. G. (1972) Effective size of populations with overlapping generations. *Theoretical population biology*, **3** : 278–289.

Hill, W. G. (1974a) Disequilibrium among several linked neutral genes in finite populations. I. mean changes in disequilibrium. *Theoretical Population Biology*, **5** : 366–392.

Hill, W. G. (1974b) Disequilibrium among several linked neutral genes in finite populations. II. variances and covariances of disequilibria. *Theoretical Population Biology*, **6** : 184–198.

Hill, W. G. et A. Robertson (1966) The effect of linkage on limits to artificial selection. *Genetical Research*, **8** : 269–294.

Hill, W. G. et A. Robertson (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38** : 226–231.

Hill, W. G. et B. S. Weir (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, **54** : 705–714.

Holland, R. F. et S. K. Jain (1981) Insular biogeography of vernal pools in the central valley of california. *The American Naturalist*, **117** : 24–37.

Holsinger, K. E., R. J. Mason-Gamer et J. Whitton (1999) Genes, demes, and plant conservation. In : *Genetics and the Extinction of Species* (eds. Landweber, L. F. et A. P. Dobson), pp. 23–46. Princeton University Press, Princeton.

Horn, R. et C. Johnson (1985) *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

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

HUDSON, R. R. et N. L. KAPLAN (1988) The coalescent process in models with selection and recombination. *Genetics*, **120** : 831–840.

JAIN, S. K. (1994) Genetics and demography of rare plants and patchily distributed colonizing species. In : *Conservation Genetics* (eds. LOESCHCKE, V., J. TOMIUK et S. K. JAIN), pp. 291–307. Birkhäuser Verlag, Bâle.

JARNE, P. et P. LAGODA (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **11** : 424–429.

JOHNSON, D. L. (1977) Inbreeding in populations with overlapping genrations. *Genetics*, **87** : 581–591.

JOHNSON, D. M. (1985) New records for longevity of *Marsilea* sporocarps. *American Fern Journal*, **75** : 30–31.

KAJ, I., S. M. KRONE et M. LASCOUX (2001) Coalescent theory for seed bank models. *Journal of Applied Probability*, **38** : 285–300.

KAPLAN, N. L., R. R. HUDSON et C. H. LANGLEY (1989) The "hitchhiking effect" revisited. *Genetics*, **123** : 887–899.

KAWATA, M. (1985) Sex differences in the spatial distribution of genotypes in the red-backed vole, *Clethrionomys rufocanus bedfordiae*. *Journal of Mammalogy*, **66** : 384–387.

KINGMAN, J. F. C. (1982a) The coalescent. *Stochastic Processes and their Applications*, **13** : 235–248.

KINGMAN, J. F. C. (1982b) On the genealogy of large populations. *Journal of Applied Probability*, **19A** : 27–43.

KOENIG, W. D., D. VAN VUREN et P. N. HOOGE (1996) Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution*, **11** : 514–517.

Kruskal, J. B. J. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, **7** : 48–50.

Lande, R. (1988) Genetics and demography in biological conservation. *Science*, **241** : 1455–1460.

Lande, R. (1994) Risk of population extinction from fixation of new deleterious mutations. *Evolution*, **48** : 1460–1469.

Latta, R. G. (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adapted traits. *The American Naturalist*, **151** : 283–292.

Latter, B. D. (1973) The island model of population differentiation : a general solution. *Genetics*, **73** : 1.

Leakey, R. et R. Lewin (1995) *The Sixth Extinction. Biodiversity and its Survival*. Phoenix - Orion Books, Londres.

Leblois, R. (2000) *Étude de l'influence de facteurs mutationnels et démographiques sur l'estimation de paramètres démographiques à partir de génotypes individuels*. Diplôme d'études approfondies, Université Montpellier II.

Lemaire, C., G. Allegrucci, M. Naciri, L. Bahri-Sfar, H. Kara et F. Bonhomme (2000) Do discrepancies between microsatellite and allozyme variation reveal differential selection between sea and lagoon in the sea bass (*Dicentrarchus labrax*) ? *Molecular Ecology*, **9** : 457–467.

Lesica, P., R. F. Leary, F. W. Allendorf et D. E. Bilderback (1988) Lack of genetic diversity within and among populations of an endangered plant, *Howellia aquatilis*. *Conservation Biology*, **2** : 275–282.

Levins, R. (1968) *Evolution in Changing Environments. Some Theoretical Explorations*. Princeton University Press, Princeton.

LEWONTIN, R. C. et J. KRAKAUER (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics*, **74** : 175–195.

LEWONTIN, R. C. et J. KRAKAUER (1975) Testing the heterogeneity of F values. *Genetics*, **80** : 397–398.

LONG, J. C. (1986) The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics*, **112** : 629–647.

LUIKART, G. et J. M. CORNUET (1999) Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, **151** : 1211–1216.

LYNCH, M., J. CONERY et R. BÜRGER (1995a) Mutation accumulation and the extinction of small populations. *The American Naturalist*, **146** : 489–518.

LYNCH, M., J. CONERY et R. BÜRGER (1995b) Mutational meltdowns in sexual populations. *Evolution*, **49** : 1067–1080.

LYNCH, M. et W. GABRIEL (1990) Mutation load and the survival of small populations. *Evolution*, **44** : 1725–1737.

LYRHOLM, T. et U. GYLLENSTEN (1998) Global matrilineal population structure in sperm whales as indicated by mitochondrial DNA sequences. *Proceedings of the Royal Society of London Series B*, **265** : 1679–1684.

LYRHOLM, T., O. LEIMAR, B. JOHANNESON et U. GYLLENSTEN (1999) Sex-biased dispersal in sperm whales : contrasting mitochondrial and nuclear genetic structure of global populations. *Proceedings of the Royal Society of London Series B*, **266** : 347–354.

MALÉCOT, G. (1948) *Les Mathématiques de l'Hérédité*. Masson, Paris.

MALÉCOT, G. (1975) Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology*, **8** : 212–241.

Maruyama, K. et H. Tachida (1992) Genetic variability and geographical structure in partially selfing populations. *Japanese Journal of Genetics*, **67** : 39–51.

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

Maynard Smith, J. et R. Hoekstra (1980) Polymorphism in a varied environment : how robust are the models ? *Genetical Research*, **35** : 45–57.

McCracken, G. (1984) Social dispersion and genetic variation in two species of emballonurid bats. *Zeitschrift für Tierpsychologie*, **66** : 55–69.

McCue, K. A. et T. P. Holtsford (1998) Seed bank influences on genetic diversity in the rare annual *Clarkia springvillensis* (Onagraceae). *American Journal of Botany*, **85** : 30–36.

McPeek, M. A. et R. D. Holt (1999) The evolution of dispersal in spatially and temporally varying environments. *The American Naturalist*, **140** : 1010–1027.

Mendenhall, W. M., D. D. Wackerly et R. L. Scheaffer (1990) *Mathematical Statistics with Applications*. PWS-KENT Publishing Company, Boston.

Miller, G. H., J. W. Magee, B. J. Johnson, M. L. Fogel, N. A. Spooner, M. T. McCulloch et L. K. Ayliffe (1999) Pleistocene extinction of *Genyornis newtoni* : human impact on australian megafauna. *Science*, **283** : 205–208.

Monod, J. (1970) *Le Hasard et la Nécessité : Essai sur la Philosophie Naturelle de la Biologie Moderne*. Seuil, Paris.

Morley, F. H. W. (1954) Selection for economic characters in australian merino sheep. IV : the effect of inbreeding. *Australian Journal of Agricultural Research*, **5** : 305–316.

MOSSMAN, C. A. et P. M. WASER (1999) Genetic detection of sex-biased dispersal. *Molecular Ecology*, **8** : 1063–1067.

MOTRO, U. (1991) Avoiding inbreeding and sibling competition : the evolution of sexual dimorphism for dispersal. *The American Naturalist*, **137** : 108–115.

MOTRO, U. (1994) Evolutionary and continuous stability in asymmetric games with continuous strategy sets : the parental investment conflict as an example. *The American Naturalist*, **144** : 229–241.

NAGYLAKI, T. (1980) The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology*, **9** : 101–114.

NAGYLAKI, T. (1983) The robustness of neutral models of geographical variation. *Theoretical Population Biology*, **24** : 268–294.

NAGYLAKI, T. (1989) Gustave Malécot and the transition from classical to modern genetics. *Genetics*, **122** : 253–268.

NAGYLAKI, T. (1995) The inbreeding effective population number in dioecious populations. *Genetics*, **139** : 473–485.

NAGYLAKI, T. (1998) Fixation indices in subdivided populations. *Genetics*, **148** : 1325–1332.

NEI, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70** : 3321–3323.

NEI, M. (1977) F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics*, **41** : 225–233.

NEI, M. (1986) Definition and estimation of fixation indices. *Evolution*, **40** : 643–645.

NEI, M. et A. CHAKRAVARTI (1977) Drift variance of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. *Theoretical Population Biology*, **11** : 307–325.

Nei, M., A. Chakravarti et Y. Tateno (1977) Mean and variance of F_{ST} in a finite number of incompletely isolated populations. *Theoretical Population Biology*, **11** : 291–306.

Nei, M. et R. K. Chesser (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, **47** : 253–259.

Nei, M. et T. Maryuyama (1975) Lewontin-Krakauer test for neutral genes. *Genetics*, **80** : 395.

Nei, M. et F. Tajima (1981) Genetic drift and estimation of effective population size. *Genetics*, **98** : 625–640.

Nielsen, R. et M. Slatkin (2000) Likelihood analysis of ongoing gene flow and historical association. *Evolution*, **54** : 44–50.

Norborg, M. (1997) Structured coalescent processes on different time scales. *Genetics*, **146** : 1501–1514.

Norborg, M. (2001) Coalescent theory. In : *Handbook of Statistical Genetics* (eds. Balding, D. J., M. Bishop et C. Cannings), pp. 179–212. John Wiley & Sons, Ltd, Chichester.

Nunney, L. et D. R. Elam (1994) Estimating effective size of conserved populations. *Conservation Biology*, **8** : 175–184.

Ohta, T. (1980) Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigen families. *Genetical Research*, **36** : 181–197.

Ohta, T. (1982a) Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **79** : 1940–1944.

Ohta, T. (1982b) Linkage disequilibrium with the island model. *Genetics*, **101** : 139–155.

Ohta, T. et M. Kimura (1969) Linkage disequilibrium due to random genetic drift. *Genetical Research*, **13** : 47–55.

Olivier, L., J. P. Galland, H. Maurin et J. P. Roux (1995) *Livre Rouge de la Flore Menacée de France.* Tome 1 : *Espèces Prioritaires*. Collection Patrimoines naturels, vol. 20. Muséum National d'Histoire Naturelle, Conservatoire Botanique National de Porquerolles & Ministère de l'Environnement, Paris.

Olivieri, I. (2001) The evolution of seed heteromorphim in a metapopulation : interactions between dispersal and dormancy. In : *Integrating Ecology and Evolution in a Spatial Context* (eds. Silvertown, J. et J. Antonovics), pp. 245–268. Blackwell Science.

Olivieri, I., Y. Michalakis et P.-H. Gouyon (1995) Metapopulation genetics and the evolution of dispersal. *The American Naturalist*, **146** : 202–228.

Orive, M. E. (1993) Effective population size in organisms with complex life-histories. *Theoretical Population Biology*, **44** : 316–340.

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

Pardini, A. T., C. S. Jones, L. R. Noble, B. Kreiser, H. Malcom, J. Bruce BD, Stevens, G. Cliff, M. C. Scholl, M. Francis, C. A. J. Duffy et A. P. Martin (2001) Sex-biased dispersal of great white sharks. *Nature*, **412** : 140–141.

Perrin, N. et J. Goudet (2001) Inbreeding, kinship, and the evolution of natal dispersal. In : *Dispersal* (eds. Clobert, J., E. Danchin, A. A. Dhondt et J. D. Nichols), pp. 123–142. Oxford University Press, Oxford.

Perrin, N. et V. Mazalov (1999) Dispersal and inbreeding avoidance. *The American Naturalist*, **154** : 282–292.

Perrin, N. et V. Mazalov (2000) Local competition, inbreeding and the evolution of sex-biased dispersal. *The American Naturalist*, **155** : 116–127.

Petit, E., F. Balloux et J. Goudet (2001) Sex-biased dispersal in a migratory bat : a characterization using sex-specific demographic parameters. *Evolution*, **55** : 635–640.

Pollak, E. (1983) A new method for estimating the effective population size from allele frequency changes. *Genetics*, **104** : 531–548.

Provine, W. B. (1971) *The Origins of Theoretical Population Genetics*. The University of Chicago Press, Chicago.

Pudovkin, A. I., D. V. Zaykin et D. Hedgecock (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144** : 383–387.

Rassmann, K., D. Tautz, F. Trillmich et C. Gliddon (1997) The microevolution of the Galápagos marine iguana *Amblyrhynchus cristatus* assessed by nuclear and mitochondrial analyses. *Molecular Ecology*, **6** : 437–452.

Raufaste, N. et F. Bonhomme (2000) Properties of bias and variance of two multiallelic estimators of F_{ST}. *Theoretical Population Biology*, **57** : 285–296.

Raufaste, N. et F. Bonhomme (2002) Single locus, multiallelic test of neutrality with two multiallelic estimators of F_{ST}. Manuscrit.

Raymond, M., R. L. Väänẗo, F. Thomas, F. Rousset, T. de Meeüs et F. Renaud (1997) Heterozygote deficiency in the mussel *Mytilus edulis* species complex revisited. *Marine Ecology Progress Series*, **156** : 225–237.

Rees, M. (1996) Evolutionary ecology of seed dormancy and seed size. *Philosophical Transactions of the Royal Society of London Series B*, **351** : 1299–1308.

Reinartz, J. A. et D. H. Les (1994) Bottleneck-induced dissolution of self-incompatibility and breeding system consequences in *Aster furcatus* (Asteraceae). *American Journal of Botany*, **81** : 446–455.

Reynolds, J., B. S. Weir et C. C. Cockerham (1983) Estimation of the coancestry coefficient : basis for a short term genetic distance. *Genetics*, **105** : 767–779.

Robert, C. (1996) *Méthodes de Monte Carlo par Chaînes de Markov*. Economica, Paris.

Robertson, A. (1961) Inbreeding in artificial selection programmes. *Genetical Research*, **2** : 189–194.

Robertson, A. (1975a) Gene frequency distribution as a test of selective neutrality. *Genetics*, **81** : 775–785.

Robertson, A. (1975b) Remarks on the Lewontin-Krakauer test. *Genetics*, **80** : 396.

Robertson, A. et W. G. Hill (1984) Deviations from Hardy-Weinberg proportions : sampling variances and use in estimation of inbreeding coefficients. *Genetics*, **107** : 703–718.

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

Ronfort, J., E. Jenczewski, T. Bataillon et F. Rousset (1998) Analysis of population structure in autotetraploid species. *Genetics*, **150** : 921–930.

Ross, K. G., D. D. Shoemaker, M. J. B. Krieger, J. DeHeer et L. Keller (1999) Assessing genetic structure with multiple classes of molecular markers : A case study involving the introduced fire ant *Solenopsis invicta*. *Molecular Biology and Evolution*, **16** : 525–543.

Rousset, F. (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, **142** : 1357–1362.

ROUSSET, F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145** : 1219–1228.

ROUSSET, F. (1999a) Genetic differentiation in populations with different class of individuals. *Theoretical Population Biology*, **55** : 297–308.

ROUSSET, F. (1999b) Genetic differentiation within and between two habitats. *Genetics*, **151** : 397–407.

ROUSSET, F. (2001a) Genetic approaches to the estimation of dispersal rates. In : *Dispersal* (eds. CLOBERT, J., E. DANCHIN, A. A. DHONDT et J. D. NICHOLS), pp. 18–28. Oxford University Press, Oxford.

ROUSSET, F. (2001b) Inferences from spatial population genetics. In : *Handbook of Statistical Genetics* (eds. BALDING, D. J., M. BISHOP et C. CANNINGS), pp. 179–212. John Wiley & Sons, Ltd, Chichester.

ROUSSET, F. et S. BILLIARD (2000) A theoretical basis for measures of kin selection in subdivided populations : finite populations and localized dispersal. *Journal of Evolutionary Biology*, **13** : 814–825.

ROUSSET, F. et M. RAYMOND (1997) Statistical analyses of population genetic data : new tools, old concepts. *Trends in Ecology and Evolution*, **12** : 313–317.

SACCHERI, M., I. KUUSSAARI, M. KANKARE, P. VIKMAN, W. FORTELIUS et I. HANSKI (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature*, **392** : 491–494.

SANTIAGO, E. et A. CABALLERO (1995) Effective size of populations under selection. *Genetics*, **139** : 1013–1030.

SANTIAGO, E. et A. CABALLERO (1998) Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics*, **149** : 2105–2117.

Schneider, S., D. Roessli et L. Excoffier (2000) *Arlequin, Version 2000 : a software for population genetics data analysis*. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Genève.

Schwartz, M. K., D. A. Tallmon et G. Luikart (1999) Using genetics to estimate the size of wild populations : many methods, much potential, uncertain utility. *Animal Conservation*, **2** : 321–323.

Sih, A., B. G. Jonsson et G. Luikart (2000) Habitat loss : ecological, evolutionary and genetic consequences. *Trends in Ecology and Evolution*, **15** : 132–133.

Singh, R. S., M. Choudhary et J. R. David (1987) Contrasting patterns of geographic variation in the cosmopolitan sibling species *Drosophila melanogaster* and *D. simulans*. *Biochemical Genetics*, **25** : 27–40.

Singh, R. S. et L. R. Rhomberg (1987) A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. II. estimates of heterozygosity and patterns of geographic differentiation. *Genetics*, **117** : 255–271.

Slatkin, M. (1975) Gene flow and selection in a two-locus system. *Genetics*, **81** : 787–802.

Slatkin, M. (1985) Rare alleles as indicators of gene flow. *Evolution*, **39** : 53–65.

Slatkin, M. (1987) Gene flow and the geographic structure of natural populations. *Science*, **236** : 787–792.

Slatkin, M. (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58** : 167–175.

Slatkin, M. (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47** : 264–279.

Slatkin, M. et N. H. Barton (1989) A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43** : 1349–1368.

Slatkin, M. et L. Voelm (1991) F_{ST} in a hierarchical island model. *Genetics*, **127** : 627–629.

Smith, F. D. M., R. M. May, R. Pellew, T. H. Johnson et K. S. Walker (1993) How much do we know about the current extinction rate ? *Trends in Ecology and Evolution*, **8** : 375–378.

Soltis, D. E. et P. S. Soltis (1986) Active enzymes from megaspores of *Marsilea* and *Regnellidium*. *American Fern Journal*, **76** : 17–20.

Soltis, P. S., D. E. Soltis, T. L. Tucker et F. A. Lang (1992) Allozyme variability is absent in the narrow endemic *Bensoniella oregona* (Saxifragaceae). *Conservation Biology*, **6** : 131–134.

Strobeck, C. (1983) Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics*, **103** : 545–555.

Strobeck, C. (1987) Average number of nucleotide differences in a sample from a single subpopulation : a test for population subdivision. *Genetics*, **117** : 149–153.

Sved, J. A. (1968) The stability of linked systems of loci with a small population size. *Genetics*, **59** : 543–563.

Swensen, S. M., G. J. Allan, M. Howe, W. J. Elisens, S. A. Junak et L. H. Rieseberg (1995) Genetic analysis of the endangered island endemic *Malacothamus fasciculatus* (Nutt.) Greene var. *nesioticus* (Rob.) Kearn. (Malvaceae). *Conservation Biology*, **9** : 404–415.

Tachida, H. et C. C. Cockerham (1986) Analysis of linkage disequilibrium in an island model. *Theoretical Population Biology*, **29** : 161–197.

Takahata, N. (1982) Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genetical Research*, **39** : 63–77.

Takahata, N. (1988) The coalescent in two partially isolated diffusion populations. *Genetical Research*, **52** : 213–222.

Taylor, P. D. et S. A. Frank (1996) How to make a kin selection model ? *Journal of Theoretical Biology*, **180** : 27–37.

Thomas, C. D., B. E. J., W. R. J., A. D. Simmons, Z. G. Davies, M. Musche et L. Conradt (2001) Ecological and evolutionary processes at expanding range margins. *Nature*, **411** : 577–581.

Tonsor, S. J., S. Kalisz, J. Fisher et T. P. Holtsford (1993) A life-history based study of population genetic structure - seed bank to adults in *Plantago lanceolata*. *Evolution*, **47** : 833–843.

Tsakas, S. et C. B. Krimbas (1976) Testing the heterogeneity of F values : A suggestion and a correction. *Genetics*, **84** : 399–401.

Tufto, J., S. Engen et K. Hindar (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics*, **144** : 1911–1921.

Tutin, T. G., V. H. Heywood, N. A. Burges, D. M. Moore, D. H. Valentine, S. M. Walters et D. A. Webb (1964) *Flora Europaea. Volume 1*. Cambridge University Press, Cambridge.

van Noordwijk, A. J. (1994) The interaction of inbreeding depression and environmental stochasticity in the risk of extinction of small populations. In : *Conservation Genetics* (eds. Loeschcke, V., J. Tomiuk et S. K. Jain), pp. 131–146. Birkhäuser Verlag, Bâle.

Venable, D. L. et J. S. Brown (1988) The selective interactions of dispersal, dormancy, and seed size as adaptations for reducing risk in variable environments. *The American Naturalist*, **360-384** : 131.

Venable, D. L. et J. S. Brown (1993) The population-dynamic functions of seed dispersal. *Vegetatio*, **108** : 31–55.

Voelker, R. A., H. E. Schaffer et T. Mukai (1980) Spontaneous allozyme mutations in *Drosophila melanogaster* : rate of occurrence and nature of the mutants. *Genetics*, **94** : 961–968.

Vucetich, J. A., T. A. Waite et L. Nunney (1997) Fluctuating population size and the ratio of effective to census population size. *Evolution*, **51** : 2017–2021.

Wahlund, S. (1928) Zuzammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, **11** : 65–106.

Wakeley, J. (2001) The coalescent in an island model of population subdivision with variation among demes. *Theoretical Population Biology*, **59** : 133–144.

Waller, D. M., D. M. O'Malley et S. C. Gawler (1987) Genetic variation in the extreme endemic *Pedicularis furbishiae* (Scrophulariaceae). *Conservation Biology*, **1** : 335–340.

Wang, J. (1996) Deviation from Hardy-Weinberg proportions in finite populations. *Genetical Research*, **68** : 249–257.

Wang, J. (1997b) Effective size and F-statistics of subdivided populations. I. Monoecious species with partial selfing. *Genetics*, **146** : 1453–1463.

Wang, J. (1997a) Effective size and F-statistics of subdivided populations. II. Dioecious species. *Genetics*, **146** : 1465–1474.

Wang, J. (1999) Effective size and F-statistics of subdivided populations for sex-linked loci. *Theoretical Population Biology*, **55** : 176–188.

Wang, J. et A. Caballero (1999) Developments in predicting the effective size of subdivided populations. *Heredity*, **82** : 212–226.

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

Waples, R. (1991) Genetic methods for estimating the effective size of cetacean populations. In : *Genetic Ecology of Whales and Dolphins* (ed. Hoelzel, A. R.), vol. Special Issue 13, pp. 279–300. International Whale Commission, Londres.

Weir, B. et W. Hill (1980a) Effect of mating structure on variation in linkage disequilibrium. *Genetics*, **93** : 477–488.

Weir, B. S. (1996) *Genetics Data Analysis II*. Sinauer Associates, Sunderland.

Weir, B. S., P. J. Avery et W. G. Hill (1980) Effect of mating structure on variation in inbreeding. *Theoretical Population Biology*, **18** : 369–429.

Weir, B. S. et C. C. Cockerham (1969) Group inbreeding with two linked loci. *Genetics*, **63** : 711–742.

Weir, B. S. et C. C. Cockerham (1973) Mixed self and random mating at two loci. *Genetical Research*, **21** : 247–262.

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

Weir, B. S. et W. G. Hill (1980b) Effect of mating structure on variation in linkage disequilibrium. *Genetics*, **95** : 477–488.

Whitlock, M. C. et N. H. Barton (1997) The effective size of a subdivided population. *Genetics*, **146** : 427–441.

Whitlock, M. C. et D. E. McCauley (1999) Indirect measures of gene flow and migration : F_{ST} doesn't equal $1/(4Nm+1)$. *Heredity*, **82** : 117–128.

Whitlock, M. C., P. C. Phillips et M. J. Wade (1993) Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. *Evolution*, **47** : 1758–1769.

Wiehe, T. H. E. et W. Stephan (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution*, **10** : 842–854.

Wilcowe, D. S., D. Rothstein, J. Dubow, A. Phillips et E. Losos (1998) Quantifying threats to imperiled species in the United States. *BioScience*, **48** : 607–615.

Wilding, C. S., R. K. Butlin et J. Grahame (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14** : 611–619.

Williamson, E. G. et M. Slatkin (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, **152** : 755–761.

Woolliams, J. A., N. R. Wray et R. Thompson (1993) Prediction of long-term contributions and inbreeding in populations undergoing mass selection. *Genetical Research*, **62** : 231–242.

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

Wright, S. (1939) Statistical genetics in relation to evolution. In : *Actualités Scientifiques et Industrielles*, Exposés de Biométrie et de la statistique biologique XIII, pp. 5–64. Hermann & Cie, Paris.

Wright, S. (1943) Isolation by distance. *Genetics*, **28** : 114–138.

Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics*, **15** : 323–354.

Wright, S. (1969) *Evolution and the Genetics of Populations. Volume II. The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Wright, S. (1977) *Evolution and the Genetics of Populations. Volume III. Experimental Results and Evolutionary Deductions*. University of Chicago Press, Chicago.

Young, A. G., H. G. Merriam et S. I. Warwick (1993) The effect of forest fragmentation on genetic variation in *Acer saccharum* marsh. (sugar maple) populations. *Heredity*, **71** : 277–289.

Deuxième Partie

Annexes

Annexe A

La biologie des extinctions

Olivieri, I. et Vitalis, R. (2001) La biologie des extinctions. *Médecine Sciences* **17** : 63-69.

---

m/s SYNTHÈSE

médecine/sciences 2001 ; 17 : 63-9

La biologie des extinctions

Isabelle Olivieri
Renaud Vitalis

Depuis l'apparition de la vie sur terre, des espèces naissent et meurent. Des crises majeures d'extinction ont ainsi ponctué l'histoire de la vie. Nous traversons aujourd'hui une nouvelle crise, la sixième, à laquelle l'homme n'est pas étranger. La dégradation des habitats, l'introduction d'espèces exotiques, la pollution sont autant de causes actuelles d'extinction. Dès lors qu'une espèce est en déclin, des processus stochastiques d'ordre démographique, génétique ou environnemental peuvent la piéger dans une spirale vers l'extinction. L'étude des espèces rares peut permettre non seulement de mieux comprendre la dynamique des processus démographiques et génétiques d'extinction, mais aussi d'apporter un éclairage nouveau sur les processus évolutifs en jeu. Car il existe également une « spirale évolutive » qui peut mener à l'extinction. Ainsi, une espèce spécialiste, très adaptée à un milieu particulier, est plus vulnérable qu'une espèce généraliste. Or la raréfaction de ce milieu favorable (liée à la dégradation des habitats) favorise dans ce type d'espèce les gènes conférant une moindre capacité à disperser, ce qui, en retour, a de bonnes chances de favoriser une plus grande adaptation locale. Un autre exemple est le cas de l'évolution des régimes de reproduction : l'auto-incompatibilité permet d'éviter l'autofécondation, mais devient un facteur limitant du potentiel reproducteur en petites populations. Des modèles suggèrent que l'existence même de locus d'auto-incompatibilité pourrait créer les conditions du maintien de ce régime de reproduction par la sélection naturelle, y compris dans les petites populations.

ADRESSES

I. Olivieri : Laboratoire génétique et environnement (CC065), Institut des sciences de l'évolution de Montpellier (ISEM), UMR Cnrs 5554, Université Montpellier II, place Eugène-Bataillon, 34095 Montpellier Cedex 05, France. R. Vitalis : ISEM et Station biologique de la Tour du Valat, 13200 Arles, France.

Depuis l'apparition de la vie sur terre, des espèces nouvelles naissent tandis que d'autres s'éteignent. Ce sont d'ailleurs justement les changements de faune et de flore dans les enregistrements stratigraphiques fossiles qui permettent de définir les périodes géologiques. L'extinction massive des espèces n'est donc pas un phénomène nouveau *(figure 1)*. L'histoire de la vie sur Terre a été ponctuée par cinq crises majeures d'extinction. Entre 65 % et 85 % des espèces animales marines se sont éteintes à l'Ordovicien (500 millions d'années – Ma), au Dévonien (345 Ma), au Trias (180 Ma) et au

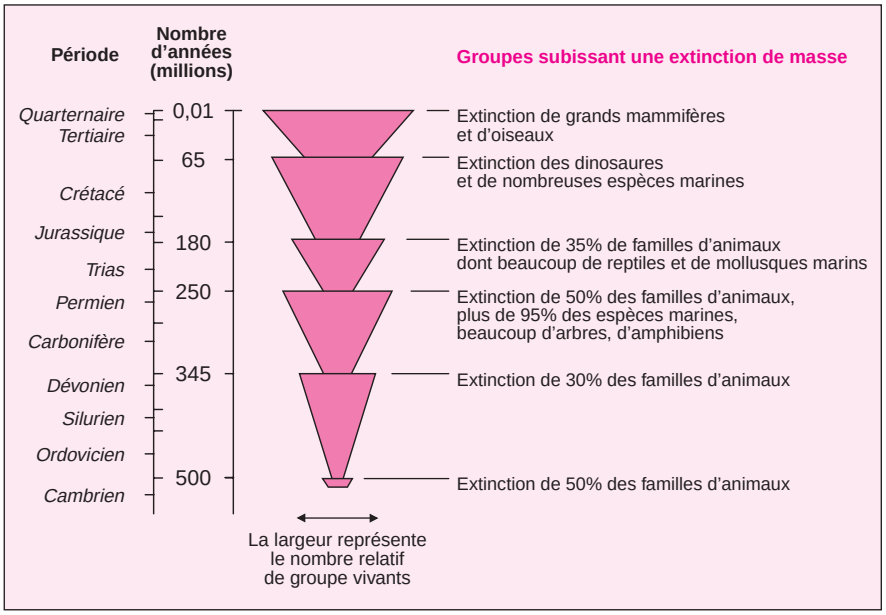


Figure 1. ***Les extinctions à l'échelle des temps géologiques.*** *La largeur de la base et celle du côté supérieur de chaque trapèze sont proportionnelles au nombre de groupes vivants à un temps géologique donné. Le nombre d'espèces croît régulièrement au fur et à mesure que le temps s'écoule. En revanche, à chacune des cinq périodes d'extinction passées, le nombre d'espèces a diminué très rapidement (instantanément, à l'échelle de ce diagramme). Ces crises majeures d'extinction ont concerné des groupes différents d'organismes. Une sixième crise, au Pléistocène, est particulièrement marquée par l'action de l'Homme, qui s'installe à cette époque sur tous les continents (d'après [30]).*

Crétacé (65 Ma). Quatre-vingt-quinze pour cent ont disparu au Permien (250 Ma). En outre, entre 90 % et 96 % des espèces se sont éteintes en dehors de ces crises majeures. Ce qui est nouveau, en revanche, c'est la rapidité du phénomène qui caractérise la phase actuelle d'extinction, amorcée il y a 100 000 ans à la fin du Pléistocène, et baptisée la « sixième extinction » par Leakey et Lewin [1] : au cours de ce siècle, une espèce d'oiseau ou de mammifère s'éteint, en moyenne, chaque année. Si l'on considère que ces deux groupes d'animaux comportent respectivement quelques 10 000 et 5 000 espèces, la « durée de vie moyenne » d'une espèce serait actuellement de 10 000 ans, soit mille fois moins que la durée de vie d'une espèce fossile, estimée à 10 millions d'années [2].

Nous étions 5 millions d'êtres humains aux débuts de l'agriculture (il y a environ 10 000 ans), nous sommes 6 milliards aujourd'hui. Une telle expansion démographique a bien sûr eu des effets sur les autres espèces animales et végétales, ne serait-ce que par l'augmentation de territoires anthropisés et la réduction concomitante d'habitats naturels qui l'ont accompagnée [3]. En raison de la nature exponentielle de la croissance démographique humaine, les effets de l'homme sur la biodiversité n'ont commencé à se faire sentir que récemment : la moitié des extinctions connues depuis 1 600 sont intervenues au XX^{e} siècle [4]. Au cours des cinquante prochaines années, la moitié des espèces actuelles pourrait être amenée à disparaître (S. Pimm, cité par Sih *et al.* [5]). Les mécanismes qui sous-tendent les cinq crises d'extinctions passées sont mal connus et sujets à controverses. En revanche, mettre en évidence la ou les causes actuelles de l'extinction des espèces ainsi que les mécanismes qui y mènent constitue l'enjeu de la compréhension et de la préservation de la biodiversité.

La cause actuelle d'extinction des espèces

Le rôle de l'homme dans la sixième extinction [1] est de mieux en mieux documenté. Par exemple, la datation précise de plus de 700 œufs fossiles de l'oiseau australien géant *Genyornis newtoni* documente sa présence continue sur le continent australien depuis plus de 100 000 ans, jusqu'à sa disparition soudaine il y a 50 000 ans, précisément lorsque les hommes arrivèrent en Australie [6]. Cette étude suggère que l'homme, plus que les faibles variations climatiques durant cette période, serait responsable de la disparition de plus de 85 % des espèces de grands vertébrés sur le continent australien [7].

Dans l'analyse récente d'une base de données comportant 245 espèces animales et végétales, Channel et Lomlino [8] montrent que, pour une espèce donnée, les populations qui persistent le plus longtemps sont plutôt situées en périphérie de l'aire de répartition de cette espèce. Or l'habitat est souvent moins favorable en limite d'aire, les populations devraient donc y être plus sujettes à l'extinction. Ces résultats suggèrent que c'est l'homme, attiré lui aussi dans les zones centrales favorables à la plupart des espèces, qui constitue la principale menace pour les autres espèces.

Une étude menée sur environ 2 000 espèces menacées d'Amérique du Nord [9] a permis de montrer que les trois principales menaces exogènes pesant sur les populations sont la destruction et la dégradation de l'habitat, l'introduction d'espèces exotiques, et la pollution *(figure 2)*. Les espèces exotiques constituent une menace particulière dans les systèmes insulaires. La pollution chimique a un très grand impact en milieu aquatique. La surexploitation agricole et cynégétique, qui a constitué la principale menace jusqu'au début du siècle, concerne encore les deux tiers des reptiles et près de la moitié des espèces de mammifères. Une telle étude reste à réaliser ailleurs qu'en Amérique du Nord, par exemple en Europe.

La destruction d'habitat s'opère à plusieurs niveaux : la réduction de la superficie totale (destruction de fragments entiers d'habitat), la fragmen-

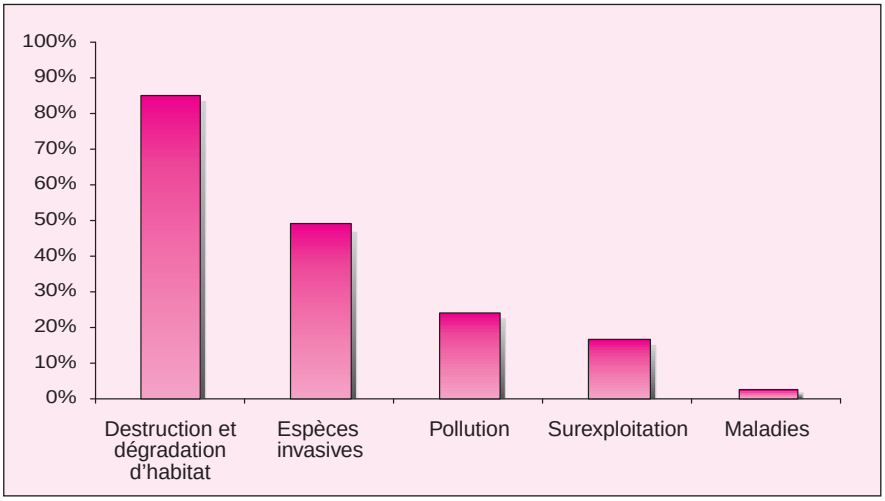


Figure 2. ***Les causes majeures d'extinction en Amérique du Nord.*** *Cette figure donne le pourcentage des espèces mises en péril par la destruction et la dégradation des habitats, les espèces invasives, la pollution, la surexploitation et les maladies. Ces catégories ne sont naturellement pas exclusives (c'est pourquoi la somme des pourcentages dépasse 100). Les espèces représentées dans ces données sont celles désignées comme étant « en péril » par l'Agence Nord-américaine de Conservation de la Nature* (Nature Conservancy)*, ainsi que toutes les espèces, sous-espèces et populations menacées, en danger, ou proposées comme telles, par la loi américaine sur les espèces en danger* (Endangered Species Act) *(d'après [9]).*

tation (qui accompagne par exemple la construction de routes et d'autoroutes), la détérioration de parcelles d'habitat (par exemple par pollution de la nappe phréatique) et la détérioration des espaces situés entre les habitats favorables (par exemple la mise en place de milieux interdisant le passage des individus migrateurs). Bien entendu, tous ces processus sont liés : la destruction de fragments d'habitat induit nécessairement une plus grande fragmentation de ce qui reste d'habitat favorable. Or les effectifs des populations naturelles augmentent avec la surface qu'elles peuvent occuper. La dégradation des habitats, avec la fragmentation qui l'accompagne nécessairement, constitue donc une réelle menace pour les espèces, en ceci qu'elle contribue à la diminution des effectifs des populations qui les composent. En outre, toutes les espèces ne sont pas égales devant ce type de risques. Chez les animaux, par exemple, la densité des populations est inversement proportionnelle à la taille ou au poids moyens des organismes : les animaux de grande taille forment des populations qui sont à la fois moins nombreuses et moins denses que celles des animaux de petite taille. Les grands mammifères qui se trouvent au sommet des pyramides trophiques ont donc besoin de plus d'espace pour que leurs populations se maintiennent à des effectifs suffisants pour prévenir les risques d'extinction. Ils constituent par conséquent les espèces les plus vulnérables à la destruction actuelle des habitats.

Une large part de la destruction d'habitat concerne la forêt tropicale. Selon Hugues *et al.* [10], le rythme actuel de destruction de la forêt tropicale de 0,8 à 2 % par an aurait pour conséquence un taux d'extinction des espèces de 0,1 à 0,3 % par an, soit 3 à 5 espèces par heure (soit 1 800 populations par heure !). Ces estimations sont fondées sur la relation établie entre la surface et le nombre d'espèces occupant cette surface par MacArthur et Wilson [11], en considérant que le nombre total d'espèces vivant sur Terre s'élève à 14 millions, dont les deux-tiers vivraient en forêt tropicale. Cette règle empirique qui relie la richesse spécifique à la surface s'appuie sur beaucoup d'études où l'on a recensé le nombre d'espèces d'une catégorie taxonomique donnée présent sur une étendue limitée dans l'espace (ces étendues, ou « îles » pouvant être de véritables îles océaniques ou bien des sommets de montagne, des lacs…). Il existe en général une relation linéaire entre le logarithme du nombre d'espèces et le logarithme de la surface occupée, de sorte qu'une réduction de la superficie de 90 % réduit la richesse spécifique de 50 %.

Les mécanismes conduisant à l'extinction : le paradigme des petites populations

Lorsque le déclin démographique d'une espèce est amorcé, des processus dus au hasard entrent en jeu, qui peuvent conduire cette espèce dans une spirale vers l'extinction. Ces effets stochastiques touchent les plus petites populations d'une espèce donnée et sont de plusieurs natures : démographique, génétique et environnementale.

Le premier effet concerne la démographie des petites populations, c'est-à-dire l'évolution et la composition de leurs effectifs. Supposons en effet que le taux de survie des individus d'une espèce soit un *paramètre* connu, de valeur 0,7, par exemple. Au fil des générations, tout se passe comme si pour chaque individu, on réalisait un « jeu » *(un tirage aléatoire)* au cours duquel cet individu survivrait avec une probabilité de 0,7 ou bien mourrait avec une probabilité de 0,3. Dans une population de très grande taille, *en moyenne* sur l'ensemble des individus, le taux de survie *réalisé* est proche de cette valeur de 0,7. Il en serait de même si le taux de survie était *mesuré* sur un très grand nombre de populations identiques, de même taille. Dans ces deux cas, en moyenne, le taux de survie est proche de la valeur du paramètre. En revanche, ce n'est plus vrai dans une population de très petite taille. Prenons le cas extrême d'une population constituée d'un seul individu : la probabilité que cet individu ne survive pas avant de se reproduire serait de 30 %. Dans notre cas particulier, c'est aussi la probabilité que cette population s'éteigne. Dans une population de 50 individus, cette probabilité serait de seulement $7{,}2.10^{-27}$!

Le deuxième effet est d'ordre génétique. Chez une espèce diploïde, comme l'homme, un enfant reçoit pour moitié des gènes de sa mère et pour moitié des gènes de son père. Par conséquent, un parent ne transmet à son descendant qu'une copie (sur deux) de ses propres gènes. A l'échelle d'une population, il arrive en outre que des individus n'aient aucun descendant et ne transmettent donc aucun gène à la génération suivante, tandis que d'autres vont transmettre plusieurs copies identiques de leurs gènes. A l'échelle d'une population, le nombre de copies différentes d'un gène (ses états alléliques) diminue donc constamment, à la mutation près. Dans une population modèle, idéale, de taille constante au cours du temps, le taux auquel la variabilité génétique est perdue est inversement proportionnel à la taille de cette population. Cette érosion de la variabilité génétique est appelée « dérive génétique ». Ce phénomène touche surtout des gènes qui n'affectent ni la survie ni la fécondité des individus (leur valeur sélective), mais aussi des gènes à effet légèrement délétère. Dans le premier cas, la dérive génétique affecte le potentiel évolutif des populations. Mais dans le second cas, si l'effet de ces gènes est trop faible pour que la sélection naturelle puisse les éliminer rapidement, certains peuvent se fixer par hasard dans la population, et diminuer la valeur sélective moyenne (c'est une des formes de ce que l'on appelle la dépression de consanguinité).

Un troisième effet est lié à la stochasticité environnementale. Ce processus est souvent divisé en effets dus aux changements cycliques mais imprévisibles du milieu ou du climat (années particulièrement sèches ou humides, intempéries) et en effets dus aux catastrophes naturelles (incendies, cyclones…).

Le rôle relatif de ces trois processus dans les mécanismes d'extinction est mal connu et encore sujet à controverses [12]. Dans une analyse de données issues de la littérature, Frankham [13, 14] a montré qu'il existait une corrélation positive entre le risque d'extinction et la consanguinité dans des populations de laboratoire de souris et de drosophiles. Un très bel exemple a été publié récemment par une équipe finlandaise qui démontre pour la première fois sur des populations naturelles que le risque d'extinction augmente avec la consanguinité [15]. Pendant dix ans, ces chercheurs ont étudié une espèce de papillon, le « damier », *Melitea cinxia*, dans les îles Aland, un ensemble de plusieurs centaines d'îles du sud-ouest de la Finlande, plus ou moins isolées les unes des autres et de tailles variables. A partir de l'étude de 42 de ces populations, dont 7 se sont éteintes entre 1995 et 1996, ils ont montré que le facteur qui expliquait le mieux la variance observée dans les taux d'extinction de ces populations était le niveau de diversité génétique mesuré sur des marqueurs enzymatiques.

Les études expérimentales sur les effets des stochasticités démographique et environnementale sont encore très rares. Dans une autre étude récente, Belovsky *et al.* [16] ont mesuré les taux d'extinction dans des populations artificielles d'un petit crustacé d'eau saumâtre, *Artemia franciscana*. Leurs résultats suggèrent, en accord avec la théorie, que la probabilité d'extinction d'une population est d'autant plus faible que l'effectif initial de la population est grand, que la capacité du milieu permet à un plus grand nombre d'individus de survivre, et que la variation environnementale est faible. Leurs résultats suggèrent également que la taille initiale de la population est un facteur moins important que ce qui est généralement supposé, tandis que la capacité maximale du milieu ainsi que la variation environnementale sont des facteurs plus importants. En outre, les oscillations engendrées par une dynamique particulière (chaotique) des effectifs des populations (liée entre autres à la surpopulation) se sont révélées être autant voire plus importantes que les phénomènes les plus souvent invoqués pour expliquer les risques d'extinction des populations naturelles. Ceci suggère qu'en l'absence de tout effet stochastique, la dynamique non linéaire d'une population, déterministe mais chaotique, peut la conduire à l'extinction.

Avant l'extinction : la rareté

Les risques d'extinction sont plus importants chez les espèces dont les populations sont de petite taille. Il est donc primordial, pour comprendre les phénomènes d'extinction, de s'intéresser à la biologie des espèces rares. Rabinowitz *et al.* [17] ont défini sept façons pour une espèce d'être rare, pour une seule façon d'être commune. Ces définitions reposent sur le croisement de plusieurs critères, tels que la taille de l'aire de répartition de l'espèce considérée, la spécificité de son habitat, la taille des populations qui la composent. Le *Tableau I* donne l'exemple de la classification de plantes britanniques et d'oiseaux d'Amérique du Nord selon ces critères. Ainsi, les principales formes de rareté diffèrent selon les types d'espèces (ainsi d'ailleurs que les zones géographiques) et donc, probablement, selon le type de menaces qui s'exercent sur les espèces rares. Une espèce présente partout mais rare parce qu'elle est partout peu abondante (rareté de type 4 ou 6) sera plus sensible à la stochasticité démographique, et moins à la stochasticité environnementale, qu'une espèce rare localement abondante mais présente en très peu de localités (rareté de type 1 ou 3).

« Évoluer » vers l'extinction

Adaptation…

La dégradation des habitats est actuellement responsable de l'extinction de trop nombreuses populations. Mais lorsque son potentiel évolutif est suffisant, une espèce peut s'adapter à son nouvel environnement partiellement détruit ou dégradé. Cette évolution, qui peut éventuellement la sauver de l'extinction, concerne notamment l'aptitude à coloniser de nouveaux sites, la tolérance à une pollution environnementale, le comportement… La capacité d'une espèce à s'adapter vite aux changements rapides du milieu est donc un facteur crucial pour sa survie [18].

Il existe aussi bien des cas où une espèce ne possède pas de capacité suffisante à s'adapter, par exemple parce qu'elle est déjà trop bien adaptée à son milieu. Prenons le cas de la dispersion. L'aire de répartition d'une espèce peut être limitée par

Tableau I. La classification selon les sept formes de rareté (d'après [17]).

	Habitat très spécifique		Habitat peu spécifique	
Taille des populations	**Espèce à grande aire de répartition**	**Espèce à petite aire de répartition**	**Espèce à grande aire de répartition**	**Espèce à petite aire de répartition**
Localement élevée	*commune*	*rare 1*	*rare 2*	*rare 3*
Plantes à fleurs	36 %	4 %	44 %	9 %
Oiseaux néotropicaux	66 %	10 %	7 %	2 %
Partout faible	*rare 4*	*rare 5*	*rare 6*	*rare 7*
Plantes à fleurs	1 %	0 %	4 %	2 %
Oiseaux néotropicaux	5 %	5 %	0 %	5 %

Application au cas des plantes à fleurs britanniques et des oiseaux néotropicaux d'Amérique du Nord [2]. Cette classification repose sur le croisement de trois facteurs : l'aire de répartition de l'espèce (l'ensemble des régions du monde où elle vit), la taille des populations qui la composent (qui peut être faible dans l'ensemble de l'aire de répartition, ou bien élevée au moins localement), ainsi que la spécificité de son habitat. Un habitat très spécifique implique une grande exigence écologique de l'espèce qui l'occupe, tandis qu'un habitat peu spécifique implique une faible exigence écologique (espèce généraliste).

l'absence de milieux favorables mais aussi par sa faible aptitude à la dispersion (et le plus souvent par les deux). Par ailleurs, l'absence de milieux favorables est elle-même liée aux exigences écologiques de l'espèce et donc à son degré de spécialisation. Or lorsqu'il y a peu de milieux favorables, les gènes conférant une moins grande aptitude à la dispersion sont sélectionnés [19, 20]. Dans ce cas, la sélection naturelle avantage également l'adaptation à un seul type de milieu, et ceci d'autant plus fortement qu'il est le seul rencontré. Et l'on peut imaginer qu'une espèce localement adaptée à son milieu ne puisse plus évoluer que dans le sens d'une spécialisation écologique toujours plus grande. Ce scénario évolutif a été suggéré pour expliquer la faible aptitude à la dispersion observée chez la « Centaurée de la Clape », *Centaurea corymbosa* Pourret (Asteraceae), une plante endémique du massif de la Clape (Hérault) [21]. Cette espèce est actuellement composée de six populations et son aire de répartition est limitée à 10 km^2.

Néanmoins, les relations entre la dispersion et la spécialisation écologique ne sont pas toujours aussi claires. En effet, bien qu'une grande aptitude à la dispersion puisse empêcher l'évolution vers une plus grande spécialisation locale et favoriser, au contraire, des individus plus généralistes, il est clair qu'elle peut également permettre l'échange d'informations génétiques nouvelles (à travers les flux de gènes) et donc favoriser l'évolution vers une plus grande spécialisation écologique locale [22, 23]. Comprendre les mécanismes conduisant à la rareté et à l'endémisme conduit donc à étudier l'évolution conjointe (la co-évolution) de la dispersion et de la spécialisation écologique.

… et mal-adaptation ?

S'il existe des cas où les espèces peuvent s'adapter rapidement aux changements du milieu et ainsi diminuer leur risque d'extinction, il arrive également que certaines s'engagent dans une spirale évolutive qui les condamne à l'extinction. Tout ne va pas forcément pour le mieux dans le meilleur des mondes… Il a été montré, par exemple, que la sélection naturelle opérant à l'échelle des individus sur certains traits d'histoire de vie (comme l'âge à la maturité, la survie adulte ou bien la probabilité de recrutement) pouvait conduire à des dynamiques non linéaires des populations [24]. Or, comme nous l'avons vu précédemment, les oscillations chaotiques des effectifs représentent un facteur important d'augmentation du risque d'extinction [16].

Un autre exemple : les régimes de reproduction chez les plantes

On trouve chez les plantes une très grande variété de systèmes de reproduction [25]. Dans certaines espèces, les individus ont des sexes séparés (certains se comportent comme des mâles, d'autres comme des femelles), alors que dans d'autres espèces les individus peuvent se comporter simultanément comme mâles et femelles (individus hermaphrodites). Dans ce cas, un individu peut se féconder lui-même (par autofécondation). On dit aussi qu'un tel individu est autocompatible. Les avantages pour l'individu qui s'autoféconde sont évidents : il n'a pas besoin de trouver un partenaire pour avoir des descendants et il leur transmet deux copies, au lieu d'une, de ses propres gènes. Nous pourrions donc être tentés de croire que de tels avantages devraient avoir favorisé l'évolution de l'autofécondation dans de nombreuses espèces. En réalité, dans une grande population, l'autofécondation est défavorisée du fait de la dépression de consanguinité. La dépression de consanguinité traduit la présence de mutations délétères, totalement ou partiellement récessives, dans les populations. Les individus génétiquement éloignés ne partagent pas (ou peu) de mutations délétères car, les mutations s'accumulant par hasard et indépendamment les unes des autres, on ne trouvera une mutation identique chez des individus différents que s'ils l'ont héritée d'un ancêtre commun. En revanche, les individus génétiquement proches ont plus de chances de partager des allèles délétères récessifs et risquent ainsi de les voir « démasqués », à l'état homozygote, dans leur descendance (car c'est en effet à l'état homozygote que ces allèles confèrent une baisse de la valeur sélective des individus qui les portent). Le passage à un régime de reproduction consanguin augmente donc la fréquence des homozygotes récessifs délétères, et diminue ainsi la valeur sélective moyenne de la popu-

lation. C'est pour cette raison qu'il est en général difficile de passer d'un régime de reproduction allogame à un régime de reproduction autogame. En revanche, dans les petites populations, la dérive génétique contribue à la fixation ou à la perte d'allèles délétères, si bien que la dépression de consanguinité tend à diminuer. On pourrait donc s'attendre à ce que les espèces rares, dont les populations sont en général de petite taille, évoluent vers un régime de reproduction autogame. Or de nombreuses espèces végétales menacées sont auto-incompatibles, c'est-à-dire que la plupart des individus de ces espèces sont incapables de s'autoféconder. C'est par exemple le cas de la « Centaurée de la Clape », évoquée ci-dessus, mais aussi du « Chou insulaire », *Brassica insularis* Moris (Brassicaceae), une endémique cyrno-sarde. Pour comprendre pourquoi l'auto-incompatibilité se maintient dans un certain nombre d'espèces rares (alors qu'elle semble être un facteur limitant du potentiel reproducteur), intéressons-nous à la dynamique de ces systèmes. Les systèmes d'auto-incompatibilité reposent sur l'existence de plusieurs types d'individus caractérisés par leur groupe d'auto-incompatibilité : deux individus d'un même groupe ne peuvent pas se croiser entre eux, tandis qu'ils peuvent se croiser avec n'importe quel individu d'un autre groupe. Les molécules impliquées dans l'appartenance à un groupe sont en général codées par un seul locus, le locus d'incompatibilité, qui détermine à lui seul les groupes d'incompatibilité : les individus appartiennent au même groupe s'ils possèdent des états alléliques identiques au locus d'incompatibilité.

Or, les populations de ces espèces végétales menacées auto-incompatibles sont en général de petite taille. Il est donc probable que l'on y trouve peu d'états alléliques distincts au locus d'auto-incompatibilité. Dans de telles conditions, où il est difficile de trouver un partenaire sexuel (*l'« effet Allee »* [26]), la sélection naturelle devrait donner l'avantage à des individus devenus autocompatibles à la suite d'une mutation des gènes responsables de l'auto-incompatiblité. Ces individus bénéficieraient alors non seulement des avantages liés à l'autofécondation, mais également de celui de ne pas « gâcher » des ovules et du pollen par des tentatives de croisements incompatibles. Or, si de tels mutants auto-compatibles existent et envahissent parfois les populations d'espèces auto-incompatibles menacées [27], cela ne semble pas être le cas général. La modélisation permet d'avancer des hypothèses explicatives [28]. Au locus d'incompatibilité, les états alléliques rares sont avantagés puisqu'ils permettent aux individus qui les portent de se reproduire avec l'ensemble de la population. Par ce mécanisme, le maintien de polymorphisme est possible et s'opère non seulement au locus d'auto-incompatibilité lui-même, mais aussi aux locus situés à proximité dans le génome… y compris aux locus impliqués dans la dépression de consanguinité. Dès lors, à ces locus, les mutations délétères se maintiennent en fréquence beaucoup plus élevée qu'en l'absence de système d'incompatibilité. On observe ainsi un excès de polymorphisme aux locus impliqués dans la dépression de consanguinité, relativement aux populations d'espèces qui ne possèdent pas de système d'incompatibilité. En outre, on peut montrer que la sélection naturelle est moins efficace pour éliminer les mutations délétères liées génétiquement au locus d'incompatibilité [28]. Or le maintien de polymorphisme aux locus impliqués dans la dépression de consanguinité avantage les croisements entre individus génétiquement éloignés, par rapport aux croisements entre individus génétiquement proches. Si les croisements entre les individus les plus éloignés génétiquement sont favorisés par la sélection, aucun mutant autocompatible (qui autoriserait le croisement d'un individu avec lui-même ou avec un individu génétiquement proche) ne pourra se répandre dans une population. La transition vers un régime de reproduction consanguin tel que l'autofécondation n'est donc pas possible dans ces conditions [28]. Les systèmes d'auto-incompatibilité créeraient donc les conditions de leur propre maintien, et ceci particulièrement dans les populations de petite taille. Dans ce cas précis encore, la sélection naturelle freine l'évolution vers un système plus avantageux.

Conclusions

L'homme est la principale menace qui pèse sur la biodiversité [29]. La dégradation des habitats réduit l'aire de répartition possible des espèces, qui, confinées en limite de leur aire, voient leurs effectifs diminuer. Dès lors, nous l'avons vu, les espèces peuvent être prises dans une spirale (démographique, génétique, évolutive) qui les mène vers l'extinction.

L'avenir de la biologie de la conservation passera sans doute par une plus grande intégration entre science et politique, entre sciences pures et sciences sociales. Il est clair que la solution à la perte généralisée de biodiversité passe par la transition démographique, et donc le développement des pays du Sud, le règlement de leur dette à l'égard des pays développés, et l'acquisition d'une « conscience écologique » dans tous les pays, développés ou non. Mais il passera aussi sûrement par une plus grande intégration des disciplines de bases que sont l'écologie et la génétique, appliquées aux espèces rares et menacées comme, d'ailleurs, aux espèces invasives. La compréhension des processus d'invasion et de spéciation constitue en effet le pendant indispensable de celle des processus d'extinction ■

Remerciements

Nous remercions Mylène Weill, Agnès Mignot, Pierre-Henri Gouyon et Robert Barbault, qui ont bien voulu relire et commenter la première version de cet article, ainsi que deux lecteurs anonymes pour leurs commentaires. Nous bénéficions d'un soutien financier du programme européen TMR « FRAGLAND » (responsable Ilkka Hanski) ainsi que du ministère de l'Aménagement du Territoire et de l'Environnement (contrat n° 98-135, attribué à IO à travers le réseau DIVERSITAS France « Populations fragmentées et biologie de la conservation »). Ceci est la publication 2000-075 de l'Institut des sciences de l'évolution de Montpellier.

RÉFÉRENCES

1. Leakey R, Lewin R. *The sixth extinction. Biodiversity and its survival.* London : Phoenix-Orion Books, 1995.

2. Dobson AP. *Conservation and biodiversity.* New York : Scientific American Library, 1995.

3. Eldredge N. *Life in the balance-humanity and the biodiversity crisis.* Princeton : Princeton University Press, 1998.

4. Smith FDM, May RM, Pellew R, Johnson TH,Walker KS. How much do we know about the current extinction rate ? *Tree* 1993 ; 8 : 375-8.

5. Sih A, Jonsson BG, Luikart G. Habitat loss : ecological, evolutionary and genetic consequences. *Trends Ecol Evol* 2000 ; 15 : 132-3.

6. Miller GH, Magee JW, Johnson BJ, *et al.* Pleistocene extinction of *Genyornis newtoni* : human impact on australian megafauna. *Science* 1999 ; 283 : 205-8.

7. Flannery TF. Debating extinction. *Science* 1999 ; 283 : 182-3.

8. Channell R, Lomolino MV. Dynamic biogeography and conservation of endangered species. *Nature* 2000 ; 403 : 84-6.

9. Wilcowe DS, Rothstein D, Dubow J, Phillips A, Losos E. Quantifying threats to imperiled species in the United States. *BioScience* 1998 ; 48 : 607-15.

10. Hughes JB, Daily GC, Ehrlich PR. Population diversity : its extent and extinction. *Science* 1997 ; 278 : 689-92.

11. MacArthur RH, Wilson EO. *The theory of island biogeography.* Princeton : Princeton University Press, 1967.

12. Lande R. Genetics and demography in biological conservation. *Science* 1988 ; 241 : 1455-60.

13. Frankham R. Inbreeding and extinctions : a threshold effect. *Cons Biol* 1995 ; 9 : 792-9.

14. Frankham R. Inbreeding and extinction: Island populations. *Cons Biol* 1998 ; 12 : 665-75.

15. Saccheri I, Kuusaari M, Kankare M, Vikman P, Fortelius W, Hanski I. Inbreeding and extinction in a butterfly metapopulation. *Nature* 1998 ; 392 : 491-4.

16. Belovsky GE, Mellison C, Larson C, Van Zandt A. Experimental studies of extinction dynamics. *Science* 1999 ; 286 : 1175-7.

17. Rabinowitz D, Cairns S, Dillon T. Seven forms of rarity and their frequency in the flora of the British Isles. In : Soulé ME, ed. *Conservation biology : the science of scarcity and diversity.* Sunderland : Sinauer Associates, 1986 : 182-204.

18. Gomulkiewicz R, Holt RD. When does evolution by natural selection prevent extinction ? *Evolution* 1995 ; 49 : 201-7.

19. Olivieri I, Michalakis Y, Gouyon PH. Metapopulation genetics and the evolution of dispersal. *Am Nat* 1995 ; 146 : 202-28.

20. Ronce O, Perret F, Olivieri I. Evolutionarily stable dispersal rates do not always increase with local extinction rates. *Am Nat* 2000 ; 155 : 485-96.

21. Colas B, Olivieri I, Riba M. *Centaurea corymbosa*, a cliff-dwelling species tottering on the brink of extinction : a demographic and genetic study. *Proc Natl Acad Sci USA* 1997 ; 94 : 3471-6.

22. Gandon S, Capowiez Y, Dubois Y, Michalakis Y, Olivieri I. Local adaptation and gene-for-gene coevolution in a metapopulation model. *Proc R Soc Lond B* 1996 ; 263 : 1003-9.

23. Gandon S, Ebert D, Olivieri I, Michalakis Y. Differential adaptation in spatially heterogeneous environments and host-parasite coevolution. In : Mopper S, Strauss S, eds. *Adaptive microevolution in natural insect populations.* New York, London : Chapman and Hall, 1998 : 325-42.

24. Ferrière R, Gatto M. Chaotic population dynamics can result from natural selection. *Proc R Soc Lond B* 1993 ; 251 : 33-8.

25. Gouyon PH, Maurice S, Reboud X, Till-Bottraud I. Le sexe, pour quoi faire ? *La Recherche* 1993 ; 250 : 70-6.

26. Allee WC, Emerson O, Park T, Schmidt KP. *Principles of animal ecology.* Philadelphia : Saunders WB, 1950.

27. Reinartz JA, Les DH. Bottleneck-induced dissolution of self-incompatibility and breeding system consequences in *Aster furcatus* (Asteraceae). *Am J Bot* 1994 ; 81 : 446-55.

28. Glémin S. Dépression de consanguinité en petites populations de plantes autoincompatibles. DEA, Université François Rabelais, Tours, 1999.

29. Biodiversité. *La Recherche* 2000 ; n° spécial : 333.

30. Barbault R. *Biodiversité : Introduction à la biologie de la conservation.* Paris : Hachette, 1997.

Summary

The biology of extinctions

Since the origin of life, new species have kept arising while others were disappearing. The new extinction crisis we are now facing, the sixth one, is mainly due to Humans. Habitat loss, introduction of exotic species, pollution are the most important current causes of extinction. Once a species has started to decline, demographical, genetical or environmental stochastic processes may trap it in a vortex towards extinction. The study of rare species gives new insights into the dynamics of these processes and the evolutionary processes involved. Indeed, there is also an « evolutionary vortex » that may lead to extinction. A species which is highly adapted to a particular habitat is more vulnerable to habitat loss than a generalist species. As patches of suitable habitats disappear, natural selection in such species favors genes that decrease dispersal propensity. This, in turn, will usually select for increased local adaptation. Natural selection might also lead to the maintenance of mating systems such as self-incompatibility, that are particularly deleterious in small populations of rare species.

TIRÉS À PART

I. Olivieri.

SOCIÉTÉ DE BIOLOGIE

Aspects moléculaires du trafic intracellulaire

Coordonnateur : Professeur Bernard Hoflack (Institut de Biologie de Lille)

Mercredi 17 janvier 2000, 16 heures

Institut des Cordeliers, Amphithéâtre Bilski-Pasquier, 15-21, rue de l'École-de-Médecine, 75006 Paris, France

Renseignements : Secrétariat de la Société de Biologie, Collège de France, 3, rue d'Ulm, 75231 Paris Cedex 05, France, Tél./Fax : 01 44 27 13 40

Annexe B

M. strigosa : une espèce menacée

Vitalis R., Colas B., Riba M. and Olivieri I. (1998) *Marsilea strigosa* Willd. : Statut génétique et démographique d'une espèce menacée. *Ecologia Mediterranea* **24** : 145-157.

---

ecologia mediterranea 24 (2), 145-157 (1998)

Marsilea strigosa Willd. : statut génétique et démographique d'une espèce menacée

Marsilea strigosa Willd. *:* genetic and demographic status of an endangered species

Renaud VITALIS[1 2], Bruno COLAS[1 2 3], Miquel RIBA[4] & Isabelle OLIVIERI[2 3]

[1] Conservatoire Botanique National Méditerranéen de Porquerolles, Castel S[te] Claire, rue S[te] Claire, 83418 Hyères Cedex, France

[2] Laboratoire Génétique et Environnement, C. C. 065, Institut des Sciences de l'Evolution de Montpellier, UMR 5554, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France
Tél. : (33) 4 67 14 47 19 ; Fax. : (33) 4 67 14 20 32 ; E-mail : vitalis@isem.univ-montp2.fr

[3] Station de Génétique et Amélioration des Plantes, Institut National de la Recherche Agronomique de Montpellier, Domaine de Melgueil, 34130 Mauguio, France

[4] Centre de Recerca Ecològica i Aplicacions Forestals, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Espagne

RESUME

Les *Marsileaceae* sont des fougères hétérosporées, aquatiques, particulièrement adaptées aux mares temporaires. Le but principal de cette étude est de déterminer le statut démographique et génétique de cette espèce, en insistant sur les effets de la structure spatiale des populations et de leur dynamique sur la diversité génétique.

Mots-clés : *Marsilea strigosa*, biologie de la conservation, génétique des populations, marqueurs enzymatiques, marqueurs RAPDs

ABSTRACT

Marsileaceae are aquatic heterosporeous ferns, well adapted to temporary ponds. The main goal of this study is to report on the genetic and demographic status of this species. We emphasize how the spatial structure of populations and temporal dynamics of the populations affect the amount of genetic variation observed.

Key-words: *Marsilea strigosa*, conservation biology, population genetics, enzymatic markers, RAPDs markers

ABRIDGED ENGLISH VERSION

Natural populations live in a temporally and spatially varying environments. They face short term threats (demographical, genetical and environmental stochasticity) and long term threats (loss of evolutionary potential), that eventually may lead to extinction. Effective population size and dispersal rates are keyparameters to understand these threats. They can be evaluated by direct (demographic surveys) and indirect (measures of genetical variability) methods. The goal of this study is to report on the demographic and genetic status of an endangered species of fern *Marsilea strigosa*, Willd. (*Marsileaceae-Pteridophyta*). This knowledge is a necessary, but not sufficient prerequiste, to propose management clues for threatened species. We sampled three populations in the south of France and seven spanish populations. The population of Roque-Haute (France) was surveyed in June 1994 and June 1995. For the Roque-Haute and Vendres populations, genetic variability was assessed at 11 enzymatic systems (17 putative loci) and 9 RAPDs loci. For all other populations, enzymatic polymorphism was assessed at 7 systems (12 loci). No polymorphism was detected, within and among sampled populations. Small population sizes and reduced colonisation ability are expected to reduce variability within population. The lack of polymorphism among populations can be interpreted in two ways : either there is still gene flow over long distances, or a single founder event was followed by successive colonisations, or both. All sampled populations could thus be issued from a single clone.

INTRODUCTION

Les populations naturelles vivent dans un environnement changeant, à la fois dans le temps et dans l'espace. Elles font face à deux catégories de menaces conduisant toutes deux éventuellement à l'extinction : une menace à court terme, liée aux variations stochastiques démographiques, génétiques et environnementales, et une menace à long terme, que représente la perte de potentiel évolutif (en l'absence de variabilité sur des caractères adaptatifs, la sélection naturelle ne peut en effet s'exercer pour promouvoir de nouvelles adaptations). Les variations continuelles du milieu menacent alors la persistance d'une espèce, incapable de faire face à de nouvelles conditions environnementales. Bien sûr, tous ces facteurs de risque sont étroitement interconnectés (Gilpin & Soulé, 1986).

Hormis les risques environnementaux (plusieurs années consécutives de sécheresse, incendies) les risques à court terme sont liés principalement aux incertitudes démographiques (dans une population de petite taille, il se peut que, par hasard, sans l'intervention de causes environnementales ou génétiques, aucun individu ne se reproduise une année) et aux incertitudes génétiques (la dérive génétique peut, parmi d'autres effets, conduire à la fixation d'allèles légèrement délétères et diminuer la valeur sélective des individus) (Shaffer, 1981 ; 1987). Dans ces deux cas, la taille efficace des populations ainsi que la capacité à coloniser de nouveaux sites favorables sont des facteurs primordiaux. La taille efficace peut être interprétée comme une mesure du taux auquel la variabilité génétique est perdue sous l'effet de la dérive génétique. De façon plus précise, on définit la taille efficace de variance comme étant la taille d'une population idéale, panmictique, qui retiendrait la même variance génétique que la population étudiée. Elle est, dans une certaine mesure, corrélée au nombre d'individus qui participent à la reproduction. La capacité à coloniser de nouveaux sites, quant à elle, sera d'autant plus forte que les individus montrent beaucoup d'aptitude à la dispersion. Elle permet de contrebalancer l'effet des extinctions locales : c'est la dynamique à l'échelle de la métapopulation (Olivieri *et al.*, 1990).

Ces paramètres clefs que sont la taille efficace et le taux de dispersion peuvent être estimés par des approches directes (suivis démographiques des populations) et indirectes (mesures de la variabilité génétique). L'approche indirecte repose sur le fait que la diversité génétique mesurée sur des marqueurs neutres vis-à-vis de la sélection naturelle est avant tout un reflet de l'histoire démographique des populations.

L'objectif de cette étude est de documenter le statut démographique et génétique de *Marsilea strigosa* Willd. (*Marsileaceae-Pteridophyta*), fougère aquatique rare, inscrite dans le Livre Rouge des Espèces Menacées de France (Olivier *et al.*, 1995), ainsi que dans la Directive Européenne "Habitats", où elle bénéficie de la mention d'espèce vulnérable. L'aire de répartition de cette espèce est disjointe. On la trouve sur le pourtour méditerranéen (France, Espagne, Baléares, Algérie, Maroc, Egypte, Italie, Sardaigne) et dans le delta de la Volga, au nord de la mer Caspienne (Tutin *et al.*, 1964).

Les *Marsileaceae* sont particulièrement adaptées aux habitats de mares temporaires. La fécondation des spores et la germination des sporophytes semblent avoir lieu au début du printemps, dans les mares inondées. Dans les mares en eau (10 à 50 cm de profondeur), les plantes adultes sont caractérisées par leurs frondes glabres, flottant à la surface de l'eau. Avec l'avancement de la saison, les mares s'assèchent et les

Marsilea développent des frondes pubescentes, plus prostrées. Sur le stolon, se forment les sporocarpes, organes contenant les spores, qui constituent ainsi un organe particulier de résistance à la sécheresse. Les spores que renferment ce sporocarpe sont sexuées. Les microspores et les macrospores vont former respectivement les gamétophytes mâles et femelles, qui contiennent les gamètes mâles (libérés dans le milieu aquatique) et femelles. Ce système de reproduction particulier des fougères dites hétérosporées, empêche l'autofécondation intra-gaméto-phytique.

Dans la perspective de proposer un plan de gestion d'une espèce menacée telle que *Marsilea strigosa*, la connaissance de son statut démographique et génétique constitue un pré-requis nécessaire, bien que non-suffisant. La combinaison des deux approches, directes et indirectes, nous informe sur la démographie de cette espèce et donc sur les risques qu'elle encourt à court terme (Ellstrand & Elam, 1993 ; Lande & Barrowclough, 1987 ; Ewens *et al.*, 1987). Dans une perspective de gestion à plus long terme, le choix de stratégies de conservation, de réintroduction de nouvelles populations, ou de renforcement des populations existantes reposera entre autres sur la connaissance de la variabilité génétique des populations qui composent l'espèce (Gilpin & Soulé, 1986).

METHODES

Echantillonnage

L'échantillonnage de *Marsilea strigosa* concerne trois populations françaises et sept populations espagnoles (Figure 1). Les seules populations françaises connues au début de l'étude se trouvaient en région Languedoc-Roussillon, dans la réserve naturelle de Roque-Haute (population connue depuis au moins 150 ans) et à Vendres (population connue depuis environ 40 ans), près de Béziers (Prelli & Boudrie, 1992). L'échantillonnage de ces deux populations a été réalisé en 1994. Une troisième population a été découverte à Saint Estève (Pyrénées-Orientales) par James Molina en juillet 1996, à la suite d'une prospection près d'un site d'où la plante avait apparemment disparu en 1982 (Amigo, 1987). Les populations espagnoles (Menorca, Mallorca, Sinarcas, Ciudad Real, Valdepeñas, Guadalajara, Huesca) ont été échantillonnées au printemps 1996. Les individus récoltés ont été maintenus en serre, à la station expérimentale de Mauguio de l'Institut National de Recherche Agronomique (I.N.R.A.) Centre de Montpellier, sous un régime de reproduction végétative obligatoire.

La population de Roque-Haute est extrêmement fragmentée. Parmi les quelques 200 mares que l'on dénombre sur le site, *Marsilea strigosa* n'en occupe qu'une quinzaine. La population de Vendres est, quant à elle, continue dans l'espace, relativement dense, et s'étale sur environ deux hectares dans les friches d'une ancienne vigne, inondée en hiver. Trois sous-populations, relativement denses et distantes de quelques dizaines de mètres composent la population de Saint Estève. Les populations espagnoles présentent également des situations contrastées : la population de Huesca couvre de façon dense et continue 7 ha d'une légère dépression dans le sol et rappelle, par sa structure spatiale continue, la population de Vendres. A Mallorca, *M. strigosa* n'a été observée que dans 2 des dizaines de petites mares creusées dans le calcaire d'un site s'étendant sur plusieurs hectares. La population de Sinarcas est constituée de deux mares distantes de quelques mètres. Dans cette population, comme dans les autres (Menorca, Ciudad Real, Valdepeñas, Guadalajara), *M. strigosa* a généralement été observée en bordure de mare, occupant peu densément l'espace.

Suivi de la population de Roque-Haute

Nous avons effectué un suivi démographique de la population de Roque-Haute, qui a consisté à déterminer la présence ou l'absence de la plante dans les différentes mares qui composent le site. Ce suivi a été réalisé de façon exhaustive sur l'ensemble des mares en juin 1994 et juin 1995.

Electrophorèse enzymatique

Quelques frondes de chaque individu ont été broyées à 4°C dans le tampon d'extraction (100 mM Tris-HCl pH=7,6 ; 4% Thioglycolate de Sodium, 2% Polyéthylène glycol 20000) et les extraits enzymatiques obtenus dans le surnageant après centrifugation ont été absorbés sur des mèches de papier filtre conservées par la suite à –80°C. Les électrophorèses enzymatiques ont été réalisées sur gels horizontaux d'amidon 12,5%. Quatre systèmes de tampon ont été utilisés : Histidine-Citrate 6,5 (Kephart, 1990), Lithium-Borate modifié de Soltis *et al.* (1983) et Kephart (1990 ; tampon de gel : 39 mM $LiOh.H_2O$, acide borique 188 mM, pH=8,3 ; tampon d'électrodes : 4

mM LiOH, 19 mM acide borique, 45 mM Tris, 7 mM acide citrique H_2O, pH=8,3), Tris-Borate-EDTA modifié de Soltis *et al.* (1983) et Wendel & Weeden (1989) ; tampon de gel : 180mM Tris, 4 mM EDTA, 100 mM acide borique, pH=8,6 ; tampon d'électrodes : 45 mM Tris, 1 mM EDTA, 25 mM acide borique, pH=8,6) et Tris-Citrate 7,2 (Soltis *et al.*, 1983). 33 systèmes enzymatiques ont été testés. 11 systèmes ont présenté une activité suffisante pour une lecture et une interprétation claires : 6-phosphogluconate déshydrogénase, aspartate aminotransférase, aldolase, estérase, β-galactosidase, malate déshydrogénase, enzyme malique, phosphatase acide, phosphoglucose isomérase, phosphoglucomutase et shikimate déshydrogénase (Tableau 1). Les systèmes de tampon utilisés pour chaque système enzymatique sont indiqués dans le tableau 1. Le génotype de 60 individus des populations de Roque-Haute et de Vendres ont été génotypés pour l'ensemble de ces systèmes enzymatiques (Tableau 2). Parmi ces 11 systèmes enzymatiques, 7 systèmes (AAT, EST, MDH, ME, PGI, PGM et SkDH) ont été résolus sur la population de Saint Estève, ainsi que sur 138 individus échantillonnés en Espagne, à Menorca, Mallorca, Sinarcas, Ciudad Real, Valdepeñas, Guadalajara et Huesca (Tableau 2).

Polymorphisme RAPD

La technique des ADNs polymorphes amplifiés aléatoirement (Randomly Amplified Polymorphic DNAs, RAPDs) a été décrite par Williams *et al.* (1990). Il s'agit de l'amplification par une réaction de polymérisation en chaîne (Polymerase Chain Reaction, PCR) de fragments « anonymes » du génome, par l'utilisation de courtes amorces oligonucléotidiques générées de façon aléatoire. Le polymorphisme observé est analysé en terme de présence ou d'absence de produit d'amplification (l'absence d'amplification pouvant être due, par exemple, à une mutation ponctuelle dans le site d'amorçage). Ce sont donc des marqueurs dominants, en général très polymorphes.

L'ADN génomique total a été extrait à partir de quelques frondes de chaque individu, selon le protocole suivant : Les frondes sont broyées dans l'azote liquide (-196°C). Le broyat est rapidement mélangé au tampon d'extraction (2% CTAB, 1% PVP, 100 mM Tris HCl, 20 mM EDTA, 1.4 M NaCl, 0.2% v/v β-mercapto-éthanol) et incubé au bain-marie à 65°C pendant 30 min. Après l'ajout d'un volume de chloroforme/alcool isoamylique (24/1) et centrifugation (13000 trs/min), la phase aqueuse est récupérée et l'ADN est précipité par l'ajout de 3/4 de volume d'isopropanol froid (-20°C). Après centrifugation et élimination du surnageant, le culot d'ADN est rincé à l'éthanol 70° et resuspendu dans 100 µL H_2O. Parmi les 30 amorces testées (Opéron), 3 ont donné un profil électrophorétique interprétable et répétable (tableau 3). Les PCRs ont été réalisées sur un thermocycleur Crocodile II (Appligene), dans un volume final de 25 µL contenant 25 ng d'ADN, 0.2 mM dNTPs, 2 mM $MgCl_2$, 1 U GoldStar DNA Polymérase (Eurogentec), 30 ng d'amorces. Après une dénaturation initiale de 4 min à 94°C, les PCRs ont consisté en 41 cycles d'1 min de dénaturation à 93°C, 2 min d'hybridation des amorces à 40°C et 2 min 30 s d'élongation à 72°C, suivi d'une élongation finale de 6 min à 72°C. Les produits d'amplification ont été analysés par électrophorèse sur gels d'agarose 1.5 %, Tris-Borate-EDTA (TBE) 0.5 X colorés au Bromure d'Ethidium (BET) et visualisés sous Ultra Violet. 43 individus des populations de Roque-Haute et Vendres ont été génotypés pour les 3 amorces RAPDs (Tableau 2).

RESULTATS

Suivi de la population de Roque-Haute

Marsilea strigosa n'occupe qu'un faible nombre de mares qui composent le site de la réserve de Roque-Haute. En 1994 la plante a été observée dans seize mares. En 1995, la plante n'a été recensée que dans onze des seize mares où elle était présente en 1994, et dans une seule autre mare (Figure 2).

Polymorphisme génétique

Onze systèmes enzymatiques ont été analysés sur les populations de Roque-Haute et Vendres (Tableau 2). Le nombre de bandes observé (Tableau 1) correspond dans presque tous les cas au nombre maximum de locus (isozymes) trouvé chez les plantes (Wendel & Weeden, 1989). Seul le système SkDH fait l'exception, puisque trois bandes ont été observées, alors que l'on n'attend en général pas plus de deux locus isozymiques chez les plantes pour ce système (Wendel & Weeden, 1989).

Système enzymatique	Acro-nyme	Numéro E.C (1)	Systèmes de tampon	Solutions de révélation	Nombre de locus
6-phosphogluconate déshydrogénase	6-PGD	1.1.1.44.	Histidine-Citrate 6.5	Wendel & Weeden, 1989	1
Aspartate aminotransférase	AAT	2.6.1.1.	Lithium-Borate modifié	Vallejos, 1983	1
Aldolase	ALD	4.1.2.13.	Tris-Citrate 7.2	Soltis *et al.*, 1983	2
Estérase	EST	3.1.1.1.	Tris-Borate EDTA modifié	Wendel & Weeden, 1989	2
β-galactosidase	GAL	3.2.1.23.	Lithium-Borate modifié	Wendel & Weeden, 1989	1
Malate déshydrogénase	MDH	1.1.1.37.	Tris-Citrate 7.2	modifié de Wendel & Weeden, 1989 (2)	2
Enzyme malique	ME	1.1.1.40.	Tris-Borate EDTA modifié	modifié de Wendel & Weeden, 1989 (2)	1
Phosphatase acide	ACP	3.1.3.2.	Lithium-Borate modifié	Soltis *et al.*, 1983	1
Phosphoglucose isomérase	PGI	5.3.1.9.	Lithium-Borate modifié	modifié de Wendel & Weeden, 1989 (2)	1
Phosphoglucomutase	PGM	2.7.5.1.	Tris-Borate EDTA modifié	modifié de Wendel & Weeden, 1989 (2)	2
Shikimate déshydrogénase	SkDH	1.1.1.25.	Tris-Citrate 7.2	modifié de Wendel & Weeden, 1989 (2)	3

(1) Enzyme Commission number
(2) c.f. Annexe

Tableau 1. Systèmes de tampon utilisés pour les électrophorèses enzymatiques et nombre de locus isozymiques résolus
Table 1. Buffer systems used for enzymes electrophoresis and number of isozymes loci observed

	Systèmes enzymatiques											Marqueurs RAPDs		
Populations	AAT	EST	MDH	ME	PGI	PGM	SkDH	6-PGD	ALD	GAL	ACP	B13	C18	P6
Roque-Haute	30	30	30	30	30	30	30	30	30	30	30	22	22	22
Vendres	30	30	30	30	30	30	30	30	30	30	30	21	21	21
St Estève	20	20	20	20	20	20	20	-	-	-	-	-	-	-
Menorca	14	14	14	14	14	14	14	-	-	-	-	-	-	-
Mallorca	6+4	6+4	6+4	6+4	6+4	6+4	6+4	-	-	-	-	-	-	-
Sinarcas	15+14	15+14	15+14	15+14	15+14	15+14	15+14	-	-	-	-	-	-	-
Ciudad Real	15	15	15	15	15	15	15	-	-	-	-	-	-	-
Valdepeñas	20	20	20	20	20	20	20	-	-	-	-	-	-	-
Guadala-jara	18	18	18	18	18	18	18	-	-	-	-	-	-	-
Huesca	32	32	32	32	32	32	32	-	-	-		-	-	-

Tableau 2. Nombre d'individus échantillonnés dans chaque population pour chaque locus
Table 2. Number of sampled individuals for each population at each locus

Amorce	Séquence (5'-3')	Nombre de locus
B13	TTCCCCCGCT	2
C18	TGAGTGGGTG	2
P6	GTGGGCTGAC	5

Tableau 3. Séquences des amorces utilisées pour les RAPDs et nombre de locus détectés
Table 3. Primers sequences used for RAPDs and number of detected loci

Il faut cependant noter que l'étude de Wendel & Weeden (1989) a essentiellement porté sur des données obtenues sur des gymnospermes ou des angiospermes, très peu d'études de ce type ayant été menées sur les fougères.Au total, 17 bandes monomorphes pour l'ensemble des individus échantillonnés ont été observées. Il n'existe donc aucune variation génétique à ces 17 locus présumés, ni à l'intérieur de ces deux populations, ni entre elles. Le polymorphisme RAPD testé sur les populations de Roque-Haute et Vendres (Tableau 2) a permis l'identification de neuf bandes amplifiées au total, pour les trois amorces utilisées (Tableau 3). Chaque bande étant assimilée à un locus, ce sont neuf locus supplémentaires qui se sont avérés monomorphes au sein de ces deux populations, comme entre elles. Les sept systèmes enzymatiques analysés pour l'ensemble des autres populations (Tableau 2) ont permis la détection de 12 locus isozymiques (Tableau 1), qui se sont avérés monomorphes au sein des populations ainsi qu'à l'échelle de l'ensemble des populations, et ont révélé la présence des mêmes allèles que ceux observés dans les populations françaises.

DISCUSSION

Marsilea strigosa fonctionne-t-elle en métapopulation à Roque-Haute ?

Cette étude est encore préliminaire et le recensement des mares occupées par *M. strigosa* doit se poursuivre pour comprendre la dynamique de la population de Roque-Haute. L'objectif de cette étude est de déterminer si cette population fonctionne en métapopulation (population subdivisée soumise à une dynamique d'extinctions et de recolonisations locales : Olivieri et *al.*, 1990) et donc de quantifier les taux d'extinction locaux ainsi que la dynamique de recolonisation. L'analyse de la relation entre la distance à la mare occupée la plus proche et la probabilité de recolonisation après une extinction locale pourrait, par exemple, nous permettre de comprendre si la recolonisation se fait par dispersion de propagules d'une mare à une autre, ou bien par refondation à partir d'un stock de sporocarpes des années antérieures. Dans ce dernier cas, la dynamique observée serait la résultante de la dynamique de mise en eau des différentes mares et refléterait la variabilité inter-annuelle des conditions hydrologiques. Il faut cependant souligner la difficulté d'estimer la densité ou le nombre de plantes présentes sur un site (forte densité d'*Isoetes* aux périodes de recensement), d'autant plus que l'existence d'un régime de reproduction clonal rend plus difficile l'interprétation du nombre d'individus observés. Bien que ces données soient encore incomplètes, elles suggèrent une capacité de dispersion réduite, étant donné le faible nombre de mares occupées, et donc une faible aptitude à la colonisation de nouveaux sites. Ce dernier point peut être appuyé par le fait que le seul organe qui puisse permettre la dispersion est le sporocarpe, qui reste accroché sur le stolon. La zoochorie pourrait donc être le seul moyen de dispersion de cette espèce. Il n'existe en revanche aucune preuve de son existence.

L'absence de polymorphisme génétique : un syndrome de la rareté ?

Mises à part les populations de Vendres et Huesca, les populations de *M. strigosa* échantillonnées sont de petite taille, en terme de nombre d'individus observés. La faible taille des populations, leur degré d'isolement (Figure 1) et une faible aptitude à la dispersion (suggérée par le suivi de la population de Roque-Haute) sont autant de facteurs qui limitent le maintien d'un polymorphisme intra-population. L'absence de polymorphisme sur l'ensemble des populations françaises et espagnoles peut s'interpréter en invoquant deux hypothèses, non mutuellement exclusives : soit il existe des flux de gènes sur de longues distances, soit un événement de fondation unique a été suivi d'événements de colonisation successifs dans les différentes populations. En considérant la faible capacité de colonisation supposée à Roque-Haute, la seconde hypothèse devrait être privilégiée. Etant donnée la capacité de reproduction végétative de cette espèce, toutes les populations échantillonnées pourraient alors être issues d'un même clone.

Les espèces rares présentent en général un polymorphisme génétique neutre réduit, bien que cela ne soit pas une règle absolue (Avise, 1994 ; Young *et al.*, 1993).

Figure 1. Répartition géographique des populations françaises et espagnoles de *Marsilea strigosa* échantillonnées entre 1994 et 1996
Figure 1. Geographical distribution of french and spanish populations of *Marsilea strigosa* sampled between 1994 and 1996

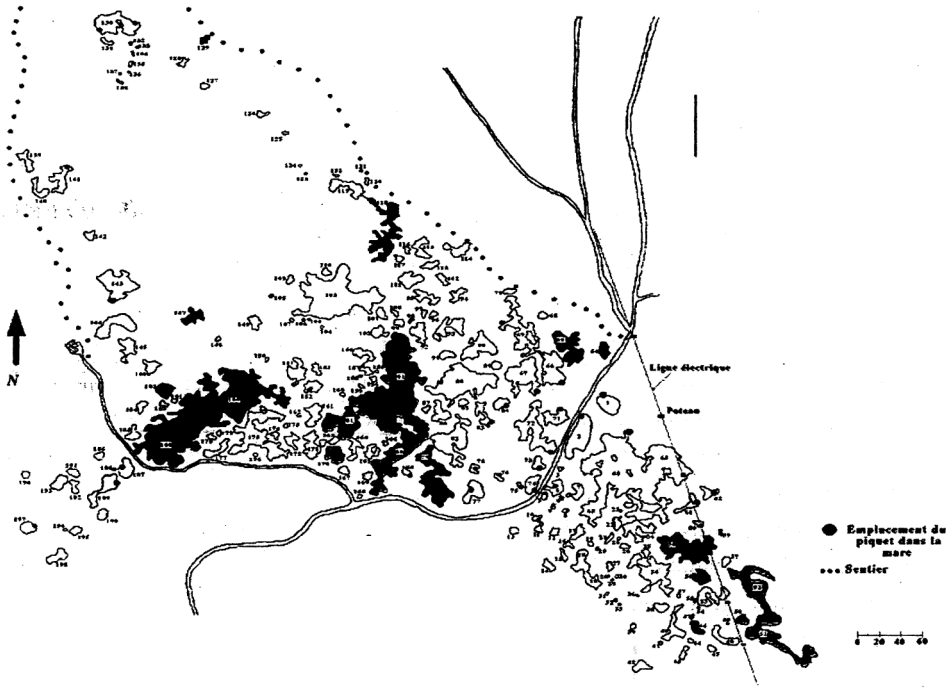


Figure 2. Suivi démographique de la population de Roque-Haute entre 1994 et 1995. Les mares sont numérotées de 1 à 198. Les mares dans lesquelles *Marsilea strigosa* a été observée en 1994 sont colorées en gris clair, les mares dans lesquelles *M. strigosa* a été observée en 1994 et en 1995 sont colorées en gris foncé et la mare 56, dans laquelle la plante n'a été observée qu'en 1995, est colorée en noir. Les lignes pointillées représentent des chemins secondaires.

Figure 2. Demographic survey of Roque-Haute population between 1994 and 1995. Ponds are numbered from 1 to 198. The ponds in which *Marsilea strigosa* was observed in 1994 are filled with light grey. The ponds in which *M. strigosa* was observed in 1994 and 1995 are filled with in dark grey and the pond 56, in which the plant was only observed in 1995, is filled with black. Dotted lines figure secondary pathways.

De nombreuses espèces végétales endémiques ou presque endémiques se sont révélées monomorphes pour un nombre comparable de locus enzymatiques. C'est le cas de *Harperocallis flava* Mc Daniel (*Liliaceae*) (Godt *et al.*, 1997), *Malacothamus fasciculatus* (Nutt.) Greene var. *nesioticus* (Rob.) Kearn. (*Malvaceae*) (Swensen *et al.*, 1995), et *Bensoniella oregona* (*Saxifragaceae*) (Soltis *et al.*, 1992). *Pedicularis furbishiae* (*Scrophulariaceae*) (Waller *et al.*, 1987), *Howellia aquaticus* (*Campanulaceae*) (Lesica *et al.*, 1988), et *Trifolium reflexum* (*Fabaceae*) (Hickey *et al.*, 1991), toutes des espèces rares ou menacées, se sont également révélées monomorphes à l'échelle intra- comme inter-populationnelle. D'autres espèces rares ou menacées ont montré un polymorphisme enzymatique très faible. C'est le cas de *Trifolium stoloniferum* (Hickey *et al.*, 1991), *Helonias bullata* (*Liliaceae*) (Godt *et al.*, 1995) ainsi que pour *Geum radiatum* (*Rosaceae*), *Carex misera* (*Cyperaceae*), *Trichophorum cespitosum* (*Cyperaceae*) et *Calamagrostis cainii* (*Poaceae*) (Godt *et al.*, 1996). Dans toutes ces études, les auteurs ont invoqué l'effet de la dérive génétique qui suit une réduction brutale de la taille de la population (goulot d'étranglement) pour expliquer le faible polymorphisme enzymatique observé. Le pourcentage de locus polymorphes ainsi que le nombre moyen d'allèles par locus (richesse allélique) sont généralement corrélés positivement à la taille des populations (Fréville *et al.*, 1998 ; Sun, 1996 ; van Treuren *et al.*, 1991 ; Raijmann *et al.*, 1994 ; Prober & Brown, 1994 ; Young *et al.*, 1996). En menant une méta-analyse sur les données de la littérature, Frankham (1996) a montré que non seulement ces derniers paramètres, mais également l'hétérozygotie, étaient corrélés positivement et de façon significative à la taille des populations. Il a également montré que la variabilité génétique neutre est significativement plus faible chez les espèces rares que chez les espèces répandues.

Quels sont les risques d'extinction associés à la génétique des petites populations ?

Des modèles théoriques montrent que l'hétérozygotie estimée à partir d'un échantillon de locus neutres est positivement corrélée à l'hétérozygotie moyenne du génome, lorsque l'hétérozygotie moyenne de la population est faible (Chakraborty, 1981 ; Mitton & Pierce, 1980). Ceci suggère que le faible polymorphisme mesuré sur des marqueurs moléculaires neutres dans les populations d'espèces rares peut refléter un faible polymorphisme à l'échelle du génome. On observe fréquemment une diminution de la valeur adaptative des individus dans les populations de petite taille. Chez *Lolium multiflorum* Lam., Polans & Allard (1989) ont mesuré une diminution de la variabilité de caractères liés à la vigueur végétative et au succès reproducteur. Une étude menée sur deux espèces menacées de poissons a montré des résultats similaires : chez *Poeciliopsis monacha* (*Poeciliidae*), la perte de variation subie après des diminutions brutales d'effectif était associée avec une diminution de la stabilité développementale, de la tolérance au stress, de l'habileté compétitrice et une augmentation de la charge parasitaire, tandis que chez *P. occidentalis*, la variabilité résiduelle mesurée sur des marqueurs enzymatiques était associée à une plus grande survie, une meilleure croissance, une plus forte fécondité et une plus grande stabilité développementale (Vrijenhoek, 1994). Une diminution de la valeur adaptative des descendants d'individus appartenant à des populations de petite taille a été observée chez *Gentiana pneumonanthe* (*Gentianaceae*) (Oostermeijer *et al.*, 1994 ; 1995). Dans dix populations naturelles d'*Ipomopsis aggreagata* (*Polemoniaceae*), Heschel & Paige (1995) ont démontré que la valeur adaptative des individus appartenant à des populations de petite taille était réduite. Bijlsma *et al.* (1994) ont observé le même phénomène chez *Scabiosa columbaria* L. (*Dipsacaceae*) et *Salvia pratensis* L. (*Labiatae*). Ces derniers auteurs nomment l'ensemble du processus « érosion génétique » (augmentation de la consanguinité dans les populations de petite taille, diminution de la valeur adaptative des individus de ces populations).

Toutes ces observations peuvent être interprétées comme étant la conséquence du « fardeau de mutation », qui désigne la réduction des performances des individus par l'accumulation de mutations délétères (Haldane, 1937). L'expression de mutations légèrement délétères récessives augmente en effet avec la consanguinité : C'est la « dépression de consanguinité ». Des modèles théoriques ont montré que la consanguinité, en augmentant l'expression des allèles délétères récessifs, et les exposant donc à la sélection, pouvait ainsi diminuer le fardeau (Barret & Charlesworth, 1991 ; Hauser *et al.*, 1994). D'autres auteurs

ont insisté sur le fait que le fardeau dépendait de la taille des populations : dans le cas des populations de petite taille, la sélection devient moins efficace, la fixation de nouvelles mutations est facilitée et la taille des populations diminue sous l'effet de la diminution de la valeur adaptative des individus (Lynch & Gabriel, 1990 ; Gabriel & Bürger, 1994). Ce processus décrit sous le terme de « fonte mutationnelle » peut finalement conduire à l'extinction (van Noodwijck, 1994 ; Lande, 1994 ; Lynch *et al.*, 1995a ; 1995b).

Des données tirées de la littérature ont permis à Frankham (1995 ; 1998) de trouver une corrélation positive entre le risque d'extinction et la consanguinité pour des populations de laboratoire de souris et de drosophiles. Mais un très bel exemple a été publié récemment par une équipe finlandaise qui démontre pour la première fois sur des populations naturelles que le risque d'extinction augmente avec la consanguinité (Saccheri *et al.*, 1998).

Quelles sont les conséquences pour la gestion à long terme des espèces menacées ?

Pour certains auteurs, identifier et maîtriser les risques d'extinction liés à la stochasticité environnementale est de première importance et doit l'emporter sur les considérations génétiques (Lande, 1988 ; Goodman, 1987 ; Shaffer, 1987 ; Menges, 1992 ; Gilligan *et al.*, 1997). A plus long terme, il s'agit également de préserver le potentiel adaptatif des espèces (Avise, 1996). Préserver la variabilité génétique des populations naturelles menacées constitue donc un enjeu majeur pour les préserver d'un risque d'extinction accru.

Malheureusement la relation entre polymorphisme neutre et potentiel adaptatif n'est pas démontrée (Lynch, 1996). D'autre part, une réduction brutale de l'effectif d'une population (goulot d'étranglement) peut avoir d'autres conséquences sur la variance additive, c'est-à-dire la part de la variance sur laquelle s'exerce la sélection naturelle. Notamment, la variance additive peut augmenter après la réduction d'effectif lorsqu'il existe des interactions entre allèles à un locus (dominance) ou entre locus (épistasie) (Carson, 1990) : malgré la diminution de viabilité qu'accompagne un goulot d'étranglement (dépression de consanguinité), on peut observer une conversion de la variance de dominance en variance additive (Wang *et al.*, 1998b). Les écarts à l'équilibre de Hardy-Weinberg qui suivent un goulot d'étranglement (Wang, 1996), ou les déséquilibres de liaison qui s'ensuivent peuvent aussi conduire à une augmentation de la variance additive (Wang *et al.*, 1998a). La réduction d'effectif permet également la conversion de certains types de variance dus aux interactions entre locus, en variance additive (Goodnight, 1988 ; Whitlock *et al.*, 1993). Ces études théoriques peuvent permettre d'interpréter les résultats obtenus en laboratoire, qui montrent une augmentation de la variance additive avec la consanguinité sur des caractères morphologiques chez la mouche domestique (Bryant *et al.*, 1986) ou sur des mesures de viabilité chez la drosophile (López-Fanjul & Villaverde, 1989). Widén & Andersson (1993) ont mesuré une plus forte variance additive et une plus grande héritabilité sur des caractères quantitatifs liés à la valeur adaptative dans les plus petites populations de *Senecio integrifolius* (Asteraceae), espèce rare et menacée en Suède. On peut donc s'attendre à des changements importants dans l'organisation de la variabilité neutre et sélectionnée dans les petites populations subissant des réductions importantes de leurs effectifs. Malheureusement, les relations entre l'une et l'autre ne sont pas triviales et dépendent fortement du déterminisme de la variabilité adaptative.

Pourquoi chercher des marqueurs neutres plus polymorphes ?

L'analyse de la variabilité génétique neutre et de sa distribution spatiale n'est donc pas directement extrapolable à la variabilité pour des caractères adaptatifs, seuls soumis par définition à l'action de la sélection naturelle. Bien qu'un effet de fondation important affectera aussi bien des locus neutres que des locus soumis à la sélection, il est possible que la mesure de la variabilité sur des caractères quantitatifs ne soit pas corrélée avec la mesure de variabilité sur des gènes neutres, en raison du déterminisme génétique de ces caractères adaptatifs (nombre de gènes impliqués, interactions entre ces gènes, type de sélection exercée sur ces gènes). En revanche, ces mesures indirectes peuvent nous permettre de faire des inférences précises sur la biologie de l'espèce, en termes de système de reproduction et de capacité de dispersion. Des marqueurs de type microsatellite (courtes séquences dispersées dans les génomes, constituées de répétitions de motifs de 1 à 5 nucléotides) sont actuellement en

cours de mise au point au laboratoire. Ces marqueurs sont parmi les plus polymorphes connus à l'heure actuelle, en raison de leur taux de mutation élevé (Jarne & Lagoda, 1996). Si ces marqueurs microsatellites permettent de détecter de la variabilité génétique entre les populations de *M. strigosa*, nous serons en mesure de quantifier les flux de gènes entre les populations échantillonnées. En terme de gestion des populations, nous pourrions ainsi savoir si une des causes de la rareté de cette espèce est sa faible capacité de dispersion, comme c'est par exemple le cas pour *Centaurea corymbosa* Pourret (*Asteraceae*) (Colas *et al.*, 1997). Une mesure de gestion pourrait alors être la création de nouvelles populations. Si en outre, ces marqueurs sont variables à l'échelle intra-populationnelle, nous pourrons caractériser le régime de reproduction (végétatif, autogamie partielle ou totale) de cette espèce.

Enfin, en guise de perspective, la présence d'échantillons de *M. strigosa* dans les herbiers de Montpellier nous offre une occasion rare d'un suivi génétique temporel de populations sur une échelle de temps assez importante. Des travaux préliminaires ont en effet montré qu'il était possible d'obtenir des germinations à partir de sporocarpes âgés de plus d'un siècle (Colas *et al.*, 1996). Le suivi de la population de Roque-Haute à l'échelle de ce siècle permettrait de tester l'hypothèse selon laquelle la diversité génétique aurait diminué au cours du temps.

Remerciements

Ces travaux ont été financés par la Région Languedoc-Roussillon et la collaboration entre les universités de Montpellier et de Barcelone a été financée par le Ministère des Affaires Etrangères (Programme PICASSO). Nous remercions l'Institut National de Recherche Agronomique, Centre de Montpellier, qui a permis la culture de spécimens en serre. Nous remercions Sylvie Muratorio qui a participé à la recherche de polymorphisme RAPD, James Molina pour son aide lors des prospections et des échantillonnages en France, Josep Lluís Gradaille Tortella et Magdalena Vicens (Jardí Botànic de Sóller) pour leur aide à Mallorca, Joël Mathez ainsi que P. A. Schäffer, pour l'accès aux Herbiers de l'Institut de Botanique de Montpellier. Merci à Hassan Souheil et Philippe March (Association de la Réserve Naturelle de Roque-Haute) pour nous avoir donné accès aux versions informatiques des cartes de la réserve. Nous remercions Frédéric Médail, Hélène Fréville ainsi qu'un lecteur anonyme pour leurs commentaires sur la version précédente de ce manuscrit.

BIBLIOGRAPHIE

Amigo, J.J., 1987. Exit la mare temporaire de San-Esteve (Saint Estève, Pyrénées-Orientales, France). *Naturalia Ruscinonensia*, 1 : 71-136.

Avise J.C., 1994. *Molecular markers, natural history and evolution*. Chapman & Hall, New York. 511 p.

Avise, J.C., 1996. Introduction : The scope of conservation genetics. *In* Avise J.C. & Hamrick J.L. (eds) *Conservation Genetics. Case histories from nature*. Chapman & Hall, New York : 1-9.

Barret S.C.H. & Charlesworth D., 1991. Effects of a change in the level of inbreeding on the genetic load. *Nature*, 352 : 522-524.

Bijlsma R., Ouborg N.J. & Treuren R. van, 1994. On genetic erosion and population extinction in plants : A case study in *Scabiosa columbaria* and *Salvia pratensis*. *In* : Loescke V., Tomiuk J. & Jain S.K. (eds), *Conservation genetics*. Birkhäuser Verlag, Basel : 255-271.

Bryant E.H., McCommas S.A. & Combs L. M., 1986. The effect of an experimental bottleneck upon the quantitative genetic variation in the housefly. *Genetics*, 114 : 1191-1211.

Carson, H.L., 1990. Increased genetic variance after a population bottleneck. *Trends Ecol. Evol.*, 5 : 228-230.

Chakraborty R., 1981. The distribution of the number of heterozygous loci in an individual in natural populations. *Genetics*, 98 : 461-466.

Colas B., Riba M. & Molina J., 1996. Statut démographique de *Centaurea corymbosa* Pourret (*Asteraceae*), *Hormatophylla pyrenaica* (Lapeyr.) Cullen & Dudley (*Brassicaceae*), et *Marsilea strigosa* Willd. (*Marsileaceae-Pteridophyta*), trois plantes rares dans le sud de la France. *Acta bot. Gallica*, 143 : 191-198.

Colas B., Olivieri I. & Riba M. 1997. *Centaurea corymbosa*, a cliff-dwelling species tottering on the brink of extinction : A demographic and genetic study. *Proc. Natl. Acad. Sci.* USA, 94 : 3471-3476.

Ellstrand N.C. & Elam D. R., 1993. Population genetic consequences of small population size : Implications for plant conservation. *Annu. Rev. Ecol. Syst.*, 24 : 217-242.

Ewens W.J., Brockwell P.J. Gani J.M. & Resnick S.I., 1987. Minimum viable population size in the presence of catastrophe. *In* : Soulé M. E. (ed.), *Viable populations for conservation*. Cambridge University Press, Cambridge : 59-68.

Frankham R., 1995. Inbreeding and extinction : A threshold effect. *Conserv. Biol.*, 9 : 792-799.

Frankham R., 1996. Relationship of genetic variation to population size in wildlife. *Conserv. Biol.*, 10 : 1500-1508.

Frankham R., 1998. Inbreeding and extinction : Island populations. *Conserv. Biol.*, 12 : 665-675.

Fréville H., Colas B., Riba M., Ronfort J. & Olivieri I., 1998. Predicting endemism from population structure of

a widespread species : Case study in *Centaurea maculosa* Lam. (*Asteraceae*). *Conserv. Biol.*, 12 : 1-10.

Gabriel W. & Bürger R., 1994. Extinction risk by mutational meltdown: synergistic effects between population regulation and genetic drift. *In* : Loeschcke V., Tomiuk J. & Jain S. K. (eds), *Conservation genetics*. Birkhäuser Verlag, Basel : 69-84.

Gilligan D.M., Woodworth L.M., Montgomery M.E., Briscoe D.A. & Frankham R., 1997. Is mutation accumulation a threat to the survival of endangered populations ?. *Conserv. Biol.*, 11 : 1235-1241.

Gilpin M.E. & Soulé M.E., 1986. Minimum viable populations : Processes of species extinction. *In* : Soulé M.E. (ed.), *Conservation biology : The science of scarcity and diversity*. Sinauer Associates, Sunderland : 19-34.

Godt M.J.W., Hamrick J.L. & Bratton S., 1995. Genetic diversity in a threatened wetland species, *Helionas bullata* (Liliaceae). *Conserv. Biol.*, 9 : 596-604.

Godt M.J.W., Johnson B.R. & Hamrick J.L., 1996. Genetic diversity and population size in four rare southern appalachian plant species. *Conserv. Biol.*, 10 : 796-805.

Godt M.J.W., Walker J. & Hamrick J.L., 1997. Genetic diversity in the endangered lily *Harpetocallis flava* and a close relative, *Tofielda racemosa*. *Conserv. Biol.*, 11 : 361-366.

Goodnight C.J., 1988. Epistasis and the effect of founder events on the additive genetic variance. *Evolution,* 42 : 441-454.

Haldane J.B.S., 1937. The effect of variation on fitness. *Am. Nat.*, 71 : 337-349.

Hauser T.P., Damgaard C. & Loeschcke V., 1994. Effects of inbreeding in small plant populations : Expectations and implications for conservation. *In* : Loeschcke V., Tomiuk J. & Jain S.K. (eds), *Conservation genetics*. Birkhäuser Verlag, Basel : 115-129.

Heschel M.S. & Paige K.N., 1995. Inbreeding depression, environmental stress, and population size variation in Scarlet Gilia (*Ipomopsis aggregata*). *Conserv. Biol.*, 9 : 126-133.

Hickey R.J., Vincent M.A. & Guttman S.I., 1991. Genetic variation in running buffalo clover (*Trifolium stoloniferum*, *Fabaceae*). *Conserv. Biol.*, 5 : 309-316.

Jarne P. & Lagoda P., 1996. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, 11 : 424-429.

Kephart S.R., 1990. Starch gel electrophoresis of plant isozymes : A comparative analysis of techniques. *Am. J. Bot.*, 77 : 693-712.

Lande R., 1988. Genetics and demography in biological conservation. *Science*, 241 : 1455-1460.

Lande R. ; 1994. Risk of population extinction from fixation of new deleterious mutations. *Evolution*, 48 : 1460-1469.

Lande R. & Barrowclough G.F., 1987. Effective population size, genetic variation, and their use in population management. *In* : Soulé M.E. (ed.), *Viable populations for conservation*. Cambridge University Press, Cambridge : 87-123.

Lesica P., ,Leary R.F., Allendorf F.W. & Bilderback D.E. 1988. Lack of genetic diversity within and among populations of an endangered plant, *Howellia aquatilis*. *Conserv. Biol.*, 2 : 275-282.

López-Fanjul C. & Villaverde A., 1989. Inbreeding increases genetic variance for viability in *Drosophila melanogaster*. *Evolution*, 43 : 1800-1804.

Lynch M., 1996. A quantitative-genetic prospective on conservation issues. *In* Avise J.C. & Hamrick J.L. (eds) *Conservation Genetics. Case histories from nature*. Chapman & Hall, New York : 471-501.

Lynch M., Conery J. & Bürger R., 1995a. Mutation accumulation and the extinction of small populations. *Am. Nat.*, 146 : 489-518.

Lynch M., Conery J. & Bürger R., 1995b. Mutational meltdowns in sexual populations. *Evolution*, 49 : 1067-1080.

Lynch M. & Gabriel W., 1990. Mutation load and the survival of small populations. *Evolution*, 44 : 1725-1737.

Menges E.S., 1992. Stochastic modeling of extinction in plant populations. *In* : Fiedler P.L. & Jain S.K. (eds), *Conservation biology : The theory and practice of nature conservation, preservation and management*. Chapman and Hall, London : 253-275.

Mitton J.B. & Pierce B.A., 1980. The distribution of individual heterozygosity in natural populations. *Genetics*, 95 : 1043-1054.

van Noordwijk A.J., 1994. The interaction of inbreeding depression and environmental stochasiticity in the risk of extinction of small populations. *In* : Loeschcke V., Tomiuk J. & Jain S. K. (eds), *Conservation genetics*. Birkhäuser Verlag, Basel : 131-146.

Olivier L., Galland J.P. & Maurin, H., 1995. *Livre rouge de la flore menacée de France. Tome 1. Espèces prioritaires*. M. N. H. N., Service du Patrimoine Naturel, Conservatoire Botanique National de Porquerolles, Ministère de l'Environnement, Direction de la Nature et des Paysages. Paris, 486 p.

Olivieri I., Couvet D. & Gouyon P.-H., 1990. The genetics of transient populations : Research at the metapopulational level. *Trends Ecol. Evol.*, 5 : 207-210.

Oostermeijer J.G.B., van Eijck M.W., van Leeuwen N.C. & den Nijs J.C.M., 1995. Analysis of the relationship between allozymic heterozygosity and fitness in the rare *Gentiana pneumonanthe* L. *J. Evol. Biol.*, 8 : 739-759.

Oostermeijer J.G.B., van Eijck M.W. & den Nijs J.C.M., 1994. Offspring fitness in relation to population size and genetic variation in the rare perennial plant species *Gentiana pneumonanthe* (Gentianaceae). *Oecologia*, 97 : 289-296.

Polans N.O. & Allard R.W., 1989. An experimental evaluation of the recovery potential of ryegrass populations from genetic stress resulting from restriction of population size. *Evolution*, 43 : 1320-1324.

Prelli R. & Boudrie M., 1992. *Atlas écologique des fougères et plantes alliées*. Editions Lechevalier, Paris.

Prober, S.M. & Brown, A.H.D., 1994. Conservation of the grassy white box woodlands : Population genetics and fragmentation of *Eucalyptus albens*. *Conserv. Biol.*, 8 : 1003-1013.

Raijmann, L.E., van Leeuwen, N.C., Kersten, R., Oostermeijer, J.G.B., den Nijs, H.C.M. & Menken, S.B.J., 1994. Genetic variation and outcrossing rate in relation to population size in *Gentiana pneumonanthe* L. *Conserv. Biol.*, 8 : 1014-1026.

Saccheri I., Kuussaari M., Kankare M., Vikman P., Fortelius W. & Hanski I., 1998. Inbreeding and extinction in a butterfly population. *Nature*, 392 : 491-494.

Shaffer M., 1981. Minimum population sizes for species conservation. *Bioscience*, 31 : 131-134.

Shaffer M., 1987. Minimum viable populations : coping with uncertainty. *In* : Soulé M.E. (ed.), *Viable populations for conservation*. Cambridge University Press, Cambridge : 69-86.

Soltis D.E., Haufler C.H., Darrow D.C. & Gastony G.J., 1983. Starch gel electrophoresis of ferns : A compilation of grinding buffers, gel and electrode buffers, and staining schedules. *Am. fern. J.*, 73 : 9-27.

Soltis P.S., Soltis D.E., Tucker T.L. & Lang F.A., 1992. Allozyme variability is absent in the narrow endemic *Bensoniella oregona* (*Saxifragaceae*). *Conserv. Biol.*, 6 : 131-134.

Sun M., 1996. Effects of population size, mating system, and evolutionary origin on genetic diversity in *Spiranthes sinensis* and *S. hongkongensis*. *Conserv. Biol.*, 10 : 785-795.

Swensen S.M., Allan G.J., Howe M., Elisens W.J., Junak S.A. & Rieseberg L.H., 1995. Genetic analysis of the endangered island endemic *Malacothamus fasciculatus* (Nutt.) Greene var. *nesioticus* (Rob.) Kearn. (*Malvaceae*). *Conserv. Biol.*, 9 : 404-415.

van Treuren, R., Bijlsma, R., van Delden, W. & Ouborg, N.J., 1991. The significance of genetic erosion in the process of extinction. I. Genetic differentiation in *Salvia pratensis* and *Scabiosa columbaria* in relation to population size. *Heredity*, 66 : 181-189.

Tutin T.G., Heywood V.H., Burges N.A., Moore D.M., Valentine D.H., Walters S.M. & Webb D.A., 1964. *Flora Europaea*. Vol. 1, Cambridge University Press, Cambridge.

Vrijenhoeck R.C., 1994. Genetic diversity and fitness in small populations. *In* : Loeschcke V., Tomiuk J. & Jain S.K. (eds), *Conservation genetics*. Birkhäuser Verlag, Basel : 37-53.

Waller D.M., O'Malley D.M.& Gawler S.C. 1987. Genetic variation in the extreme endemic *Pedicularis furbishiae* (*Scrophulariaceae*). *Conserv. Biol.*, 1 : 335-340.

Wang J., 1996. Deviation from Hardy-Weinberg proportions in finite diploid populations. *Genet. Res.* 68 : 249-257.

Wang J., Caballero A. & Hill W.G., 1998a. The effect of linkage disequilibrium and deviation from Hardy-Weinberg proportions on the changes in genetic variance with bottlenecking. *Heredity*, 81 : 174-186.

Wang J., Caballero A., Keightley, P.D. & Hill W.G., 1998b. Bottleneck effect on genetic variance : A theoretical investigation of the role of dominance. *Genetics*, 150 : 435-447.

Wendel J.F. & Weeden N.F., 1989. Visualisation and interpretation of plant isozymes. *In* : Soltis D.E. & Soltis P.S. (eds), *Isozymes in plant biology*. Chapman and Hall, London : 5-45.

Whitlock M.C. Phillips P.C. & Wade M.J., 1993. Gene interaction affects the additive genetic variance in subdivided populations with extinction and migration. *Evolution*, 47 : 1758-1769.

Widén, B. & Andersson, S., 1993. Quantitative genetics of life-history and morphology in a rare plant, *Senecio integrifolius*. *Heredity*, 70 : 503-514.

Williams J.G.K., Kubelik A.R., Livak J., Rafalski J.A. & Tingey S.V., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.*, 18 : 6531-6535.

Young A., Boyle T. & Brown T., 1996. The population genetic consequences of habitat fragmentation for plants. *Trends Ecol. Evol.* 11 : 413-418.

Young, A.G., Merriam, H.G. & Warwick, S.I., 1993. The effect of forest fragmentation on genetic variation in *Acer saccharum* Marsh. (sugar maple) populations. *Heredity*, 71 : 277-289.

ANNEXE 1

Solutions de révélation des systèmes enzymatiques, modifiées de Wendel & Weeden (1989)

Malate déshydrogénase	
Tris Hcl 0,1 M pH=9,1	5 ml
Malate de sodium	10 ml
H_2O	q.s.p. 50 ml
NAD	20 mg
PMS 0.2%	1 ml
NBT 1%	1 ml
Enzyme malique	
Malate de sodium 0,4 M pH=7,0	10 ml
Tris HCl 1M pH=8,0	5 ml
H_2O	q.s.p. 50 ml
$MgCl_2$	100 mg
NADP	10 mg
NBT 1%	1 ml
PMS 0,2%	1 ml
Glucose 6-phosphate déshydrogénase	10 U.
Phosphoglucose isomérase	
Tris HCl 1M pH=8,0	50 ml
Fructose 6-phosphate	30 mg
$MgCl_2$	100 mg
NADP	10 mg
MTT 1%	1 ml
PMS 0.2%	1 ml
Phosphoglucomutase	
Tris Hcl 0,1 M pH=8,0	50 ml
D glucose 1-phosphate	20 mg
$MgCl_2$	100 mg
NADP	10 mg
MTT 1%	1 ml
PMS 0,2%	1 ml
Glucose 6-phosphate déshydrogénase	10 U.
Shikimate déshydrogénase	
Tris Hcl 0,1 M pH=8,0	50 ml
Acide shikimique	25 mg
NADP	10 mg
MTT 1%	1 ml
PMS 0,2%	1 ml

Annexe C

Locus microsatellites chez *M. strigosa*

Vitalis R., Dubois M. -P. and Olivieri I. (2001) Characterization of microsatellite loci in the endangered species of fern Marsilea strigosa Willd. (Marsileaceae, Pteridophyta). *Molecular Ecology Notes* **1** : 64-66.

Molecular Ecology Notes (2001) **1**, 000–000

PRIMER NOTE

Characterization of microsatellite loci in the endangered species of fern *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta)

R. VITALIS,*† M.-P. DUBOIS* and I. OLIVIERI*

**Laboratoire Génétique et Environnement, C.C. 065, ISEM (UMR 5554), Université Montpellier II, 34095 Montpellier Cedex 05, France, †Station Biologique de la Tour du Valat, 13200 Arles, France*

Abstract

Absence of enzymatic polymorphism in the endangered species of fern *Marsilea strigosa* prompted us to develop microsatellite loci in this species. We used an enriched partial genomic library, from which six polymorphic microsatellite loci were obtained. From a previous attempt to develop microsatellite markers with standard techniques, a single long and polymorphic locus was obtained. This highlights the benefits from using enrichment techniques in those species where microsatellite loci may be rare in the genome.

Keywords: conservation biology, dinucleotide repeats, enriched library, *Marsilea strigosa*, microsatellites

Received 9 November 2000; revision accepted 8 December 2000

Marsilea strigosa Willd. (Marsileaceae, Pteridophytae) is a rare aquatic fern found in the Mediterranean basin and in the north of the Volga delta (Tutin *et al*. 1964). This species is listed in the French Red Book of endangered species (Olivier *et al*. 1995) and in the European 'Habitat' directive. This species completes its life-cycle in temporary ponds (Colas *et al*. 1996). Understanding the dispersal capacity and the actual reproductive system of this species may give important information for conservation and management purposes. The complete absence of enzymatic polymorphism (Vitalis *et al*. 1998) prompted us to identify microsatellite loci.

A partial genomic library was enriched with $(CT)_n$ and $(GT)_n$ microsatellite motifs following a modified protocol of Kijas *et al*. (1994). Fragments ranging from 400 to 900 bp of *Sau*3A-digested genomic DNA were ligated into dephosphorylated, *Bam*HI-digested pUC19 vector (Appligene). Single-stranded copies of inserts were obtained by asymmetric polymerase chain reaction (PCR), in 50 μL mix containing 12.5 pmoles of modified M13 Forward primer (5′-GTTTTCCCAGTCACGACGTTG-3′), 0.5 pmoles of modified M13 Reverse primer (5′-TTGTGAGCGGATAACAATTTC-3′), 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 2.5 units of *Taq* polymerase (Goldstar, Eurogentec), 75 mM Tris-HCl pH 9.0, 20 mM $(NH_4)_2SO_4$, 0.01% (w/v) Tween 20. PCR reaction was performed using a PTC-100 thermocycler (MJ Research Inc.): initial denaturation was at 94 °C for 4 min, followed by 30 cycles (94 °C for 1 min, 55 °C for 1 min and 72 °C for 1 min) and 4 min at 72 °C. Three hundred pmoles of 5′-biotinylated oligonucleotides ATAGAATAT$(CT)_{15}$ and ATAGAATAT$(GT)_{15}$ were attached separately to 60 μg of streptavidin-coated magnetic beads (MagneSphere, Promega), in 60 μL of 5× SSPE (20× SSPE: 3 M NaCl, 200 mM NaH_2PO_4, 20 mM EDTA, pH 7.4) for 15 min at room temperature. The beads were washed five times in 100 μL 5× SSPE to remove unbound oligonucleotides, and were resuspended in 50 μL of 10× SSPE. Thirty-five μL of denatured (98 °C for 10 min) single-stranded inserts were hybridized to 50 μL of beads-bounded oligonucleotides at 55 °C overnight in 100 μL of a solution of 1 M NaCl. The beads were washed four times in 100 μL of 2× SSPE at room temperature, and four times in 100 μL 2× SSPE at 55 °C to remove unbound inserts. They were resuspended in 20 μL of H_2O and bound sequences were eluted at 98 °C for 10 min. Ten μL of this enriched DNA was used as a template in PCR (see asymmetric PCR conditions above, with 12.5 pmoles of both pUC19 primers, same thermal profile). PCR products were purified (QIAquick, Qiagen) and digested with *Sau*3A. Fragments ranging from 400 to

Correspondence: R. Vitalis. Laboratoire Génétique et Environnement, C.C. 065, ISEM (UMR 5554), Université Montpellier II, 34095 Montpellier Cedex 05, France. Fax: + 33 (0) 4 67 14 36 22; E-mail: vitalis@isem.univ-montp2.fr

© 2001 Blackwell Science Ltd

Table 1 Seven microsatellite loci in *Marsilea strigosa*. For each locus, the core sequence, the size of the product (known from sequenced clones) obtained from amplification with the given forward (F) and reverse (R) primers and the GenBank accession nos are indicated. Annealing temperatures used for PCR are also given, as well as the number of alleles (N_A), the observed heterozygosity (H_O) and Nei's expected heterozygosity (Nei 1973) under random mating within subpopulations ($\bar{H}_S$)

Locus	Core sequence	Primer sequences (5′–3′)	T (°C)	Size (bp)	N_A	H_O	$\bar{H}_S$	Accession no.
IA10	$(CT)_{26}$	F: AGAGATGCTTAACAAAGTCC R: TATAGCATTCGGTTAGAGAG	55	100	11	0.03	0.16	AF317636
IE3	$(GT)_9TT(GT)_7$	F: GATCGGTATTTTATATGTG R: TATCCAGCAACTACAATAC	47	102	2	0.00	0.00	AF317637
IF7	$(AG)_{22}$	F: GCCCAAATTTGGAGCCTAAT R: AGCCTATGATTGAGTCACCACA	62	124	9	0.01	0.16	AF317638
IIIB8	$(AG)_{24}$	F: GTTGAACCACTTCACTTTTCCC R: ATAGCTTCCTAGAATCCCAGGC	56	135	8	0.04	0.18	AF317639
IIIC8	$(CT)_{21}$	F: GCGTTGTGTACTGTTCAAGACC R: TGCATGTAAATTGCCCTGAA	64	111	6	0.01	0.20	AF317640
IVH5	$(AC)_{24}AA(AC)_{10}$	F: GATCACATAGACGAGACAC R: AACATGTTGTGAAGAATTC	50	173	7	0.00	0.13	AF317641
XIIE3	$(AG)_{40}$	F: CGAGATCTCTCACATCTACACA R: AGCAATCATTTGTTGAACTCA	58	192	16	0.03	0.28	AF317642

900 bp were isolated, purified (GeneClean Kit, Bio101) and ligated into pUC19 using a 3:1 insert:vector molar ratio. The ligation was performed in 30 μL with 115 ng vector, approximately 90 ng purified PCR product and 3 units T4 DNA ligase (Biolabs). Approximately 50 ng of the ligation was transformed into 100 μL of ultracompetent cells (Epicurian Coli XL1-Blue, Stratagene). A total of 368 recombinant clones (from approximately 1500 obtained) were screened for $(CT)_n$ and $(GT)_n$ motifs, according to the PCR-based method described by Waldbieser (1995).

Thirty-one positive clones were sequenced. Among those, nine contained long $(CT)_n$ or $(GT)_n$ motifs ($n \geq 15$), 19 contained short or degenerated tandem repeats and three appeared to be false positives. Nine pairs of primers were defined. Six of them gave clear amplification products (loci IA10 to IVH5) with the PCR conditions given in Table 1. A partial genomic library was also constructed with standard protocols (see, e.g. Estoup *et al*. 1993) and screened for $(CT)_n$, $(GT)_n$, $(CAC)_n$, $(CTC)_n$, $(CTAT)_n$, $(TGTA)_n$ and $(GACA)_n$ motifs. Among 1692 recombinant clones, a single $(CT)_n$ microsatellite locus was found (locus XIIE3).

PCR reactions were performed in a 10-μL volume containing approximately 10 ng DNA, 0.3 pmoles of [γ^{33}P]-dATP labelled forward primer, 4 pmoles of (unlabelled) reverse primer, 0.2 mM dNTP, 1.5 mM $MgCl_2$, 0.5 units of *Taq* polymerase (Goldstar, Eurogentec), 75 mM Tris-HCl pH 9.0, 20 mM $(NH_4)_2SO_4$, 0.01% (w/v) Tween 20, using a PTC-100 thermocycler (MJ Research Inc.): initial denaturation was 4 min at 94 °C followed by 30 cycles (1 min at 94 °C, 1 min at annealing temperature, 1 min at 72 °C) and 4 min at 72 °C. PCR products were electrophoresized in 6% acryl-bisacrylamide and 8 M urea sequencing gels.

DNA from 138 individuals sampled in seven populations in Spain and two populations in France was extracted (CTAB 2%, Tris-HCl 100 mM, EDTA 20 mM, NaCl 1.4 M, β-mercapto-ethanol 0.2% v/v) and scored for polymorphism at the seven microsatellite loci. Highly significant differentiation was detected within populations (Fisher exact tests, $P < 0.0001$) for all loci but IE3, for which the test could not be performed (GENEPOP 3.2a; Raymond & Rousset 1995). These departures from Hardy–Weinberg equilibrium are interpreted as being the consequence of high selfing rates. Finally, it is noteworthy that enrichment techniques may be a valuable approach in those species where microsatellites are scarce in the genome.

Acknowledgements

We thank the Région Languedoc-Roussillon, the French Ministère des Affaires Etrangères (Programme PICASSO) and Fondation Sansouire for financial support. We thank A Berthomieu, J Lagnel and G Cuny for technical advice, V Vanninen and V Georgiadis for laboratory work. This is publication 2000–112 from the Institut des Sciences de l'Evolution de Montpellier.

References

Colas B, Riba M & Molina J (1996) Statut démographique de *Centaurea corymbosa* Pourret (*Asteraceae*), *Hormatophylla pyrenaica* (Lapeyr.) Cullen & Dudley (*Brassicaceae*), et *Marsilea strigosa* Willd. (*Marsileaceae-Pteridophyta*), trois plantes rares dans le sud de la France. *Acta Botanica Gallica*, **143**, 191–198.

Estoup A, Presa P, Krieg F, Vaiman D & Guyomard R (1993) $(CT)_n$ and $(GT)_n$ microsatellites: a new class of markers for *Salmo trutta* L. (brown trout). *Heredity*, **71**, 488–496.

Kijas JMH, Fowler JCS, Garbett CA & Thomas MR (1994) Enrichment

of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques*, **16**, 657–662.

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

Olivier L, Galland JP & Maurin H (1995) *Livre Rouge de la Flore menacée de France. Tome I: Espèces prioritaires*. MNHN, Service du Patrimoine Naturel, Conservatoire Botanique National de Porquerolles, Ministère de l'Environnement, Direction de la Nature et des Paysages, Paris.

Raymond M & Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenism. *Journal of Heredity*, **86**, 248–249.

Tutin TG, Heywood VH & Burges NA *et al*. (1964) *Flora Europaea*, Vol. 1. Cambridge University Press, Cambridge.

Vitalis R, Colas B, Riba M & Olivieri I (1998) *Marsilea strigosa*, Willd: Statut génétique et démographique d'une espèce menacée. *Ecologia Mediterranea*, **24**, 145–157.

Waldbieser GC (1995) PCR-based identification of AT-rich tri- and tetranucleotide repeat loci in an enriched plasmid library. *Biotechniques*, **19**, 742–744.

Annexe D

Structure multilocus chez *M. strigosa*

Vitalis R., Riba M., Colas B., Grillas, P. and Olivieri I. (2002) Multi-locus genetic structure at contrasted spatial scales of the endangered water fern *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta). *American Journal of Botany*. Accepté sous réserve de modifications mineures.

MULTILOCUS GENETIC STRUCTURE AT CONTRASTED SPATIAL SCALES OF THE ENDANGERED WATER FERN *MARSILEA STRIGOSA* WILLD. (MARSILEACEAE, PTERIDOPHYTA)[1]

RENAUD VITALIS,[2,3,6,7] MIQUEL RIBA,[4] BRUNO COLAS,[5] PATRICK GRILLAS,[3] AND ISABELLE OLIVIERI[2]

[2]Laboratoire Génétique et Environnement, C.C. 065, ISEM (UMR 5554), Université Montpellier II, 34095 Montpellier Cedex 05, France; [3]Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France; [4]Centre de Recerca Ecològica i Aplicacions Forestals, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain; and [5]Laboratoire d'Écologie (UMR 7625), C.C. 237, Université Pierre et Marie Curie (Paris VI), 7 quai St Bernard, 75252 Paris Cedex 05, France

Marsilea strigosa (Marsileaceae, Pterydophyta) is a rare water fern found in the Mediterranean basin, in temporary flooded habitats only. We analyzed the level and the distribution of genetic variation at seven microsatellite loci, both at the Mediterranean scale and at a narrower scale within a highly fragmented French metapopulation. Genetic diversity among individuals within each pond suggests that M. strigosa reproduces predominantly through selfing. The very high population differentiation at the Mediterranean scale indicates that gene flow (if any) is highly restricted. Similar differentiation is also found at the scale of a single metapopulation. The distribution of multilocus genotypes suggests that the genetic variation in this species is maintained mainly through the interplay of mutation and low recombination.

Key words: Conservation biology; Marsilea strigosa; microsatellites; multilocus analyses; population structure.

It is commonly admitted that rare, narrowly distributed or threatened species are genetically depauperate. The average number of neutral polymorphic loci and the allelic richness are generally correlated with population size (e.g., van Treuren et al., 1991; Young, Boyle, and Brown, 1996). A meta-analysis of published data sets showed that heterozygosity was also significantly correlated to population size and that genetic variation was significantly reduced in rare species, as compared to widespread ones (Frankham, 1996). Comparing patterns of neutral genetic variation between rare and widespread plant congeners, Gitzendanner and Soltis (2000) also found that genetic diversities (except the total diversity H_T) were significantly lower in rare species. Whether the lack of genetic variation is the cause or the consequence of rarity constitutes the chicken and egg problem of conservation biology. Therefore, management recommendations based only on low genetic variation assessments may be of limited significance (Holsinger, Mason-Gamer, and Whitton, 1999). Moreover, other factors may be of primary importance for assessing extinction risks (Lande, 1999). More interestingly, understanding how genetic variation is distributed within and among natural remnant populations of rare species may help to identify the evolutionary forces that shape this variation. The estimation of demographic parameters from neutral genetic variation may, among other expertise, give the clues to the management priorities.

During the 20th, Mediterranean wetlands have been dramatically reduced in size and numbers (Blondel and Aronson, 1999). These habitats constitute an important source of richness for biodiversity, and they are often associated with seasonal flooding dynamics. The Mediterranean climate creates the conditions for ephemeral flooding and harsh summer drought in ponds and shallow wetlands. Flooding usually lasts from late fall to spring, before water evaporates during summer, leaving the pool beds dry. The plant species that may be found in those habitats generally share some life-history traits especially with regards to life-cycle duration (annuality), seed dormancy, and dispersal (Blondel and Aronson,

[1]Manuscript received_____; revision accepted_____.

The authors thank the Région Languedoc-Roussillon, the French Ministère des Affaires Etrangères (Programme PICASSO) and the Fondation Sansouire for financial support; the Station de Génétique et d'Amélioration des Plantes (INRA Mauguio) for greenhouse facilities; N. Alvarez, I. Bourrié, M. – P. Dubois, V. Georgiadis, F. Justy, and V. Vanninen for technical assistance in the laboratory; L. Desnouhes, P. Gautier, J. L. Gradaille Tortella, N. Leclainche, J. Molina, L. Rhazi and M. Vicens for their help with field work; H. Souheil and B. De Clock for permitting us to work at Roque-Haute natural reserve; C. Jacob and A. Sandoz for sharing Roque-Haute GIS data; H. Fréville, S. Maurice, and F. Viard, who made insightful comments on a first draft of this manuscript; H. Schneider and K. Pryer for providing us with their manuscript in the press; K. Niklas and two reviewers for their comments and suggestions that improved this manuscript. This is publication number 2002-001 from the Institut des Sciences de l'Évolution de Montpellier.

[6]Current address: School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK.

[7]Author for reprint requests (e-mail: vitalis@isem.univ-montp2.fr).

1999). Temporarily flooded pools provide a unique situation for population biologists. In contrast with many natural situations, the habitat is generally readily demarcated and population boundaries are sharply defined (Holland and Jain, 1981). Yet, the data are still scarce concerning the population biology and genetics of such ephemeral plant populations (Elam, 1998). Those communities of species are generally characterized by small and/or fluctuating population sizes (Elam, 1998) and the pattern of species diversity resembles that predicted by insular biogeography theory (Holland and Jain, 1981). In the Great Valley (California), vernal pools are characterized by high endemism, interpreted as the consequence of low dispersal abilities of component species and high extinction risks (Holland and Jain, 1981). Human activities are the major cause of the degradation of wetlands. Because temporary ponds and marshes are continuously lost, many taxa face increased rarity and extinction risks (Jain, 1994).

Marsilea strigosa Willd. (Marsileaceae, Pteridophyta) is a rare water fern, whose leaves resemble those of a four-leaf clover. This species is listed in the French red book of endangered species (Olivier et al., 1995), in the Bern convention on the conservation of European wildlife and natural habitats (1979), and in the European "Habitats" directive (1992). It is found in the Mediterranean (France, Spain, Algeria, Morocco, Egypt, Italy and Sardinia), as well as in the North of the Volga Delta (Tutin et al., 1964). Marsilea strigosa is found in seasonally wet habitats, where it grows in shallow water and at the edges of ponds. All the members of Marsileaceae section produce spores within specialized structures called sporocarps. The establishment of new sporophytes is remarkably rapid, which makes Marsileaceae able to grow in intermittent and ephemeral habitats. Sporocarps are drought-resistant, bean-shaped reproductive structures that allow M. strigosa populations to persist over winter. The sporocarps may remain viable even after very long periods of dry storage (Allsopp, 1952; Johnson, 1985; Soltis and Soltis, 1986). Sporocarps open in flooded ponds (on hydration, a distinct gelatinous sorophore breaks through an opening in the sporocarp) to produce chains of sori, each sorus bearing series of mega- and microsporangia. Sporangia (that remain attached to the sporocarp) release their spores. The gelatinous outer layer of the spores (perine) expands, thereby decreasing the spore density. The spores drift toward the surface of the ponds (Schneider and Pryer, in press). Female and male gametophytes develop within the mega- and microspores (endospory) and the fertilization occurs at the air/water interface. In vitro observations of sporocarp germination show that the gelatinous matrix prevents sporangia from freely moving around in water and that M. strigosa is self-compatible, indicating a possible predominantly autogamous reproductive system. Within 48 h after the release of spores from the sporangia, embryos are settled on the water/soil interface (Schneider and Pryer, in press). Then, sporophytes grow through two phases: one aquatic phase with floating smooth leaves, followed by one terrestrial phase with smaller hairy leaves. During this later phase, plants propagate clonally. At the end of the growing season, the sporocarps are produced at the stem base, attached to the rhizome. As such, they do not constitute obvious dispersal units. Thus, one may ask whether what appear as metapopulations of ponds actually behave as sets of isolated populations, or as sets of populations linked by dispersal through, e.g., recolonization following local extinctions (i.e., "true" metapopulations).

If the main evolutionary forces that shape neutral genetic variation and its distribution correlate to the species' biological characteristics (e.g., self compatibility, possibly associated to selfing, long-term storage of sporocarps in the soil, low dispersal capabilities), one expects the population genetic structure to be highly pronounced. However, the geographical scale at which such processes occur is hardly known: for example, flooding may temporarily connect ponds at a very local scale, promoting gene flow between populations within a metapopulation. We thus investigated the level and distribution of genetic variation at two contrasted scales. In this paper, we first study the multilocus structure at the Mediterranean scale (France, Morocco, and Spain; see Fig. 1). We then compare the structure to that observed at a finer scale in the highly subdivided natural reserve of Roque-Haute (France). Because genetic variation was totally absent at 12 putative enzymatic loci (seven systems) among French and Spanish populations (Vitalis et al., 1998), we used seven microsatellite loci recently developed for this species (Vitalis, Dubois, and Olivieri, 2001). [Fig. 1]

MATERIALS AND METHODS

Biological materials sampled--Mediterranean Basin samples—Individuals (N = 197) were sampled in 13 localities around the Mediterranean Basin, from June 1994 to July 1999 (Fig. 1). From all localities except the Roque-Haute natural reserve (France) (population 2 in Fig. 1), samples were collected within a single pond. In the highly fragmented Roque-Haute population, 22 individuals were collected in 13 ponds out of 16 occupied by the species (Vitalis et al., 1998). All the individuals from French and Spanish populations were maintained in a vegetative state in a greenhouse before leaves were sampled for DNA

extraction. Sampled leaves from Moroccan populations were dried in silica gel. The sample sizes ranged from 3 to 30 individuals, depending on the actual density of plants and the size of ponds (Table 1).

Roque-Haute samples-- Roque-Haute is a 150-ha basaltic plateau, within which 198 ponds are temporarily flooded during the year. The average distance between two ponds is 281 m (range 5 – 845 m). The presence or absence of M. strigosa sporophytes was assessed once a year, at the end of the growing season, from 1994 to 2000 (except year 1998) in most of the 198 ponds. Indeed, in some years, some ponds were not visited, which results in an incomplete M. strigosa presence / absence dataset. In order to conduct a finer-scale study, an extensive sampling was performed in Roque-Haute natural reserve on 1 October 1999. Leaves were sampled from 308 individuals in 26 ponds. The sample sizes ranged from 5 to 30 within each pond.

[Table 1]

Microsatellite analyses-- DNA extractions were performed on frozen or dried leaves, as described in Vitalis et al. (2001). Polymorphism was assayed on each DNA sample at seven microsatellite loci (GenBank accession numbers range from GBAN-AF317636 to GBAN-AF317642; the prefix GBAN- has been added to each GenBank accession to link the online version of American Journal of Botany to GenBank but is not part of the actual accession number.) All seven loci are dinucleotide repeats, two of which are imperfect tandem repeats (IE3 and IVH5). The polymerase chain reaction (PCR) conditions have been described in Vitalis et al. (2001). Amplification products (10 µL) were mixed with 6.7 µL formamide loading dye and then electrophoresed in 6% acryl-bisacrylamide and 8 mol/L urea sequencing gels, for 3-5 h at about 1800 V. Sequencing reactions of pUC19 vector (Appligene) were also loaded adjacent to the samples, to serve as a single base ladder. Once dried, gels were exposed to X-ray films for 24 – 48 h. Loci IIIB8, IIIC8, and XIIE3 were multiplexed, as were loci IA10 and IF7. Due to the low variation of locus IE3 at the Mediterranean scale (and to the absence of variation within all populations, including Roque-Haute), Roque-Haute 1999 samples were not scored at this locus. Locus IVH5 was monomorphic within Roque-Haute (for both 1994 and 1999 samples).

Statistical analyses-- One-locus analyses of local mating system-- For each locus and within each sample, departure from random union of gametes was tested by a score test (U test), with the alternative hypothesis of heterozygote deficiency. This test has proven to be more powerful than the exact Hardy-Weinberg test (probability test) of, e.g., Guo and Thompson (1992) when the alternative hypothesis is heterozygote excess or deficiency (see Rousset and Raymond, 1995). The tests were performed using GENEPOP, version 3.3 (Raymond and Rousset, 1995). Global tests across loci or across samples were performed using the multisample score test of Rousset and Raymond (1995). This test assumes the independence of loci. Gene diversities (heterozygosities) were estimated as $(1-\hat{Q}_0)$ and $(1-\hat{Q}_1)$, where the $\hat{Q}_i$'s are the estimates of probabilities of identity in state (IIS) of pairs of genes, either within ($\hat{Q}_0$) or between ($\hat{Q}_1$) individuals within populations. One-locus Wright's *F* statistic F_{IS} (Wright, 1951) were estimated by the estimator $\hat{f}$ of Weir and Cockerham (1984). Multilocus estimates were computed in a slightly different manner than in Weir and Cockerham (1984) or in Weir (1996). Here, the multilocus statistic was computed as the ratio of the sum of one locus IIS probability estimates $(\hat{Q}_0-\hat{Q}_1)$, over the sum of one locus IIS probability estimates $(1-\hat{Q}_1)$. This gives more weight to the $\hat{Q}_i$ estimate of a more intensively typed locus, in proportion to the intensity of typing (see Rousset, 2001).

One-locus analyses of population differentiation-- Genotypic differentiation was tested by a log-likelihood based exact test (Goudet et al., 1996) over all samples, and over all sample pairs. Unbiased estimates of the associated P values were calculated using the Markov chain method computed by GENEPOP version 3.3 (Raymond and Rousset, 1995). One-locus Wright's *F* statistic F_{ST} (Wright, 1951) were estimated by the estimator $\hat{\theta}$ of Weir and Cockerham (1984). Multilocus estimates were computed as a ratio of the sum of one locus IIS probability estimates $(\hat{Q}_1-\hat{Q}_2)$, over the sum of one locus IIS probability estimates $(1-\hat{Q}_2)$ ($\hat{Q}_2$ being an estimate of the IIS probability of pairs of genes among populations). To take into account not only the probability of identity between pairs of genes but also the mutational processes at microsatellite loci, the intraclass correlations for allele size, ρ_{ST} (see Rousset, 1996), have been estimated in an analysis of variance framework (Michalakis and Excoffier, 1996). Multilocus $\hat{\rho}_{ST}$ estimates of ρ_{ST} were computed similarly to multilocus $\hat{\theta}$ estimates (ratios of the sum of one-locus estimates). The 95% confidence intervals were obtained by bootstrapping over loci as advocated by, e.g., Weir (1996). We derived the approximate bootstrap confidence intervals (ABC) using the method

described by DiCiccio and Efron (1996). This procedure is an analytic version of the bias-corrected and accelerated (BC_a) algorithm for producing confidence limits from bootstrap distributions that applies to nonparametric problems (DiCiccio and Efron, 1996). The confidence interval endpoints are analytically approximated, rendering Monte Carlo simulations unnecessary. Isolation by distance was analyzed by computing the regression of pairwise $\hat{\theta}$ (or $\hat{\rho}_{ST}$) estimates between pairs of populations to the natural logarithm of their geographical distance (Rousset, 1997). Geographic distances were calculated with reference to the World Geodetic System 1984 or WGS84 (Defense Mapping Agency, 1987). Rank correlations were tested using the Mantel permutation procedure (Mantel, 1967). For both within- and between-populations analyses, sequential Bonferroni correction was applied to test for significance, whenever multiple tests were performed (Rice, 1989).

One-locus analyses of genetic distances-- Pairwise genetic distances between pairs of populations were computed using Cavalli–Sforza and Edwards' chord distance (Cavalli-Sforza and Edwards, 1967), with the GENDIST program (PHYLIP package version 3.5c; Felsenstein, 1993). We also computed the pairwise genetic distance matrix, which takes into account the average allele size differences between pairs of populations, i.e., the $(\delta\mu)^2$ distance of Goldstein et al. (1995). Fitch-Margoliash trees were constructed from those distance matrices with the program FITCH (PHYLIP, version 3.5c), which implements the method of Fitch and Margoliash (1967). It does not assume an evolutionary clock, so it results in an unrooted tree. The robustness of each node was evaluated by bootstrapping data over loci, for 2000 replications (Hedges, 1992), using the SEQBOOT program (PHYLIP, version 3.5c). The resulting consensus tree was obtained through the program CONSENSE (PHYLIP, version 3.5c) and displayed with the program TREEVIEW (Page, 1996).

Two-locus analyses of linkage disequilibria-- Tests for the independence of genotypes across loci (genotypic linkage disequilibrium) were performed (Fisher exact tests) for all pairs of loci within each sample. Unbiased exact P value estimates were obtained by the Markov chain method computed by GENEPOP version 3.3 (Raymond and Rousset, 1995). Linkage disequilibria were estimated by the composite Burrows' Δ_{ij} estimator (Cockerham and Weir, 1977) using the LINKDOS program (Garnier-Gere and Dillmann, 1992) adapted from Black and Krafsur's (1985) program. This estimator does not require the gamete frequencies to be known, and it is not biased by departure from panmixia. Genotypic linkage disequilibria across populations were tested against the null hypothesis (linkage equilibrium) by a χ^2 statistic (see Weir, 1979). To test for significance, sequential Bonferroni tests were performed whenever multiple tests were carried out (Rice, 1989). We also calculated Ohta's variance components of linkage disequilibrium (Ohta, 1982b) with the LINKDOS program. Ohta (1982b) defined such components to account for within- and between-subpopulation effects, in analogy with Wright's (1951) *F* statistics. She proposed a test to discriminate between epistatic selection and genetic drift as possible explanations for large observed linkage disequilibria by comparison of appropriate variance components (Ohta, 1982a).

Multilocus analyses-- Assuming that the mutation pattern of the microsatellite loci employed may be approximated as a stepwise model (SMM; Kimura and Ohta, 1978), the difference in the number of repeats between any two alleles at a locus may reflect the time since the divergence of those alleles. Thus, a genetic distance between multilocus genotypes was calculated as

$$D_{ij} = \sum_{k=1}^{2} \sum_{l=1}^{L} (a_{ikl} - a_{jkl})^2 \qquad (1)$$

where a_{ikl} and a_{jkl} give the allele size of the *k*th gene at the *l*th locus of individuals *i* and *j*, respectively (the two homologous genes from one individual are ranked in the increasing order of the size of their allelic state: with $k = 1$, the pairwise squared size difference is taken between the smallest alleles of individuals *i* and *j*). The sum is taken over *L* loci. Ordering alleles according to their size makes D_{ij} the minimum distance between pairs of multilocus genotypes. Minimum spanning networks (each network embedding all minimum spanning trees for a given distance matrix: see Kruskal, 1956) were computed with the program MINSPNET, provided with the software ARLEQUIN version 2.000 (Schneider, Roessli, and Excoffier, 2000).

RESULTS

Mediterranean samples-- One-locus analyses of local mating system-- The mean number of alleles per locus ranged from 1 to 3.1. The proportion of polymorphic loci ranged from 0.0 to 71.4%, for both the 95 and the 90% criteria (Appendix 1). Multilocus gene diversities (heterozygosities) within individuals $(1-\hat{Q}_0)$, which are equivalent to Nei's H_0 (Nei, 1973), ranged from 0.000 to 0.061 (Appendix 1). Gene diversities between individuals $(1-\hat{Q}_1)$, which are equivalent to Nei's H_S, ranged from 0.000 to 0.419 (Appendix 1). Yet, the low variability observed was not due to limited sample size

(Spearman rank order correlation between gene diversity estimates and sample sizes: $r_s = 0.223$, $P = 0.464$). We tested the null hypothesis that the alleles at one locus were randomly distributed among individuals within populations (alternative hypothesis = heterozygote deficiency). Among 34 possible tests performed (out of 91 population × locus combinations), 28 were significant at the $\alpha = 0.05$ level (for both the exact test and the score test). The sequential Bonferroni test performed over the 34 exact tests (respectively score tests) showed that 16 tests (respectively 15 tests) were significant at the $\alpha = 0.05$ level. One-locus estimates $\hat{f}$'s of the parameter F_{IS} ranged from –0.125 to 1.000, over all samples (Appendix 1). Multilocus estimates ranged from 0.591 to 1.000. In all populations but Roque-Haute, positive $\hat{f}$ estimates were unlikely to be attributable to Wahlund effect (all the samples within a locality were taken within a single pond). From the equation $\hat{s} = 2\hat{f}/(1+\hat{f})$ (see, e.g., Pollak, 1987), the estimated selfing rates within all populations but Roque-Haute ranged from 0.743 to 1.000. All the multiple tests performed per locus for all populations and per population for all loci were significant at the $\alpha = 0.05$ level, indicating departure from Hardy-Weinberg equilibrium.

One-locus analyses of population differentiation-- Genotypic differentiation among populations was highly significant at each locus (log-likelihood based test; $P < 0.0001$). One-locus $\hat{\theta}$ estimates of parameter F_{ST} ranged from 0.742 to 1.000 (Table 2). The global test across loci was also highly significant (Fisher's method; $P < 0.0001$) and the multilocus estimate of F_{ST} was equal to 0.845. Intraclass correlation for allele size estimates ($\hat{\rho}_{ST}$) ranged between 0.855 and 1.000. The $\hat{\rho}_{ST}$ estimates were significantly larger than $\hat{\theta}$ estimates (one-tailed Wilcoxon's signed-rank test; $P = 0.009$). Pairwise F_{ST} estimates were significantly correlated with geographic distance (Mantel test; $P = 0.018$), as were ρ_{ST} estimates ($P = 0.003$).

[Table 2]

One-locus analyses of genetic distances-- Genetic affinities among Mediterranean populations were inferred from a Fitch-Margoliash tree constructed from the matrix of pairwise Cavalli-Sforza and Edwards' (1967) chord distance (Fig. 2). Geographically close populations (Vendres and Roque-Haute; Sinarcas ponds 1 and 2; Ben Slimane dayas 1 and 2; Balearic samples) were grouped together, respectively. This indicates that close populations resemble each other more than they do more distant populations. As might be seen from Fig. 2, the chord distances between Balearic / Moroccan samples and all other samples were larger on average (range 0.314-0.340) than all the distances between any sample and all the others (range 0.281-0.340). The bootstrap values (expressed as the percentage of the number of occurrences of a node, out of 2000 bootstrap samples) ranged from 22 to 100%. All the nodes within the subtree made of Guadalajara, Valdepeñas, Balearic, and Moroccan populations were satisfactory resolved (bootstrap values ≥ 68%). All the other nodes, except those uniting very close populations (Vendres and Roque-Haute, on the one hand, and Sinarcas ponds 1 and 2, on the other hand), were not supported (bootstrap values < 50%), indicating that the tree topology was not yet resolved for the populations from central Spain. Withdrawing Balearic and Moroccan samples did not change the topology nor the bootstrap values, of the tree part that linked the remaining samples. This ensured that long-branch attraction phenomena (Felsenstein, 1978) were not responsible for the lack of consistency among Spanish samples.

[Fig. 2]

We also computed the pairwise genetic distance matrix, which takes into account the average allele size differences between pairs of populations, i.e., the $(\delta\mu)^2$ distance of Goldstein et al. (1995). A Fitch-Margoliash tree was drawn from this pairwise distance matrix (Fig. 3). Geographically closer populations were grouped together, except populations from Moroccan dayas and the two ponds from Sinarcas, which were not significantly grouped (bootstrap value < 50%). Surprisingly, Valdepeñas population was found to be more related than Ben Slimane daya 2 to the group made of Ben Slimane daya 1 and Balearic populations. The branch lengths were far longer with Goldstein et al.'s (1995) $(\delta\mu)^2$ distances than with Cavalli-Sforza and Edwards' (1967) distances. However, some small branch lengths made difficult the identification of clear population clusters, at least as compared with the chord distance Fitch-Margoliash tree (Fig. 2). As with the chord distance Fitch-Margoliash tree, some nodes uniting Spanish and French populations were not well supported (bootstrap values < 50%).

[Fig. 3]

Two-locus analyses of linkage disequilibria-- Genotypic disequilibria were tested across all pairs of loci, within each sample. Among 50 possible tests performed (out of 273 population × locus-pair combinations), nine were significant at the $\alpha = 0.05$ level (five, when the sequential Bonferroni procedure was computed). Burrows' Δ_{ij} estimates of genotypic linkage disequilibria between pairs of loci (Cockerham and Weir, 1977) were highly significantly different from zero, even after Bonferroni correction for multiple tests ($P < 0.0001$). Table 3 gives Ohta's variance components of linkage disequilibrium (Ohta, 1982b). By analogy with F statistics, Ohta derived the variance

components of linkage disequilibrium, as the intraclass correlations of genes at different loci of gametes within a class, relative to genes in another class. The variance components of the disequilibrium within a population relative to that of the total population (noted D'^2_{IS}, the correlation of genes at two loci from one gamete relative to the total and D^2_{ST}, the correlation of genes from different gametes of one subpopulation relative to the total) are fairly large, as is the total variance of linkage disequilibrium (D^2_{IT}). Moreover, D'^2_{IS} (respectively D^2_{ST}) estimates were significantly larger than D'^2_{ST}, the variance of the total disequilibrium (respectively D^2_{IS}, the variance of the disequilibrium within a subpopulation) estimates (one-tailed Wilcoxon's signed-rank test; P < 0.0001). Thus, the relationships $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} > D^2_{IS}$ were fulfilled, which indicated that genetic drift and limited dispersal among populations may be mainly responsible for the observed linkage disequilibrium (Ohta, 1982a).

[Table 3]

Multi-locus analyses-- Among 181 complete genotypes out of 197 sampled individuals (some data were missing for 16 multilocus genotypes), 60 distinct genotypes were found, out of which 16 were heterozygote at one locus at least. Except in one case, no multilocus genotype was shared among two distinct populations (only one genotype, which was fixed in population Mallorca 51, was found as a single copy in the Menorca population). The minimum spanning network constructed from the pairwise allele size difference distance defined in Eq. 1 is shown in Fig. 4. Note that the distances separating two nodes (each node representing a single sampled genotype) are functions of the natural logarithm of the actual distances between two multilocus genotypes (Eq. 1). The striking result is that, in all cases, all the multilocus genotypes sampled in a single population are clustered together. Moreover, the distances between pairs of genotypes are always shorter within a single group than among closely related groups (although the logarithm transformation in Fig. 4 conceals such variation). The multilocus genotypes from geographically close populations are more closely related one to each other than to those from any other population (Roque-Haute and Vendres; Sinarcas 1 and 2; Balearic populations). Yet, Moroccan populations make the exception, with genotypes from Ben Slimane daya 1 more closely related to genotypes from Valdepeñas than to genotypes from Ben Slimane daya 2.

[Fig. 4]

Roque-Haute samples-- Demographic survey-- Between 1994 and 1997, M. strigosa was found in 11.5 ponds on average (range 6 – 16), out of the 198 ponds prospected. In 1999, however, a total of 26 ponds was occupied by the water fern (23 in 2000, but three ponds out of the 26 ponds occupied in 1998 were not prospected in 2000). Year 1999 was characterized by a late (May) reflooding of ponds. We could distinguish four categories of ponds: ponds 46, 59, 64, 118, and 182, in which M. strigosa was present from 1994 to 2000; Ponds 56, 58, 65, 79, 129, 147, and 180, in which M. strigosa was only absent in 1996; Ponds 50, 51, 52, 81, and 170, in which M. strigosa was present in 1994 and then in 1999-2000. The ponds from those three categories are shown in light grey in Fig. 5. Finally, M. strigosa was observed for the first time in 1999 in the ponds 1, 24, 62, 63, 66, 69, 163, and 165. Ponds from this latter category are shown in dark grey in Fig. 5. In pond 115, M. strigosa was observed in 1996, 1999, and 2000.

One-locus analyses-- The mean number of alleles per locus ranged from 1 to 2.4. The proportion of polymorphic loci ranged from 0.00 to 1.00 (95% criterion) or to 0.80 (90% criterion) (see Appendix 2, for details). Multilocus gene diversities (heterozygosities) within individuals $(1-\hat{Q}_0)$, ranged from 0.000 to 0.021 (Appendix 2). Gene diversities between individuals $(1-\hat{Q}_1)$, ranged from 0.000 to 0.367 (Appendix 2). Again, small sample sizes could not explain the low gene diversity observed (Spearman rank order correlation r_s = 0.178, P = 0.383). We tested the null hypothesis of random mating within ponds, for each locus. Among 15 possible tests performed (out of 130 pond × locus combinations), 13 were significant at the α = 0.05 level (for both the exact test and the score test). The sequential Bonferroni test performed over the 15 exact tests showed that ten tests were statistically significant at the α = 0.05 level (and so were the score tests). One-locus estimates $\hat{f}$ of the parameter F_{IS} ranged from – 0.020 to 1.000, over all samples (Appendix 2). Multilocus estimates ranged from 0.923 to 1.000. Estimated selfing rates thus ranged from 0.960 to 1.000. All the multiple tests performed per locus for all populations were highly significant. Among seven possible tests performed per pond for all loci (26 combinations), six were significant at the α = 0.05 level.

Genotypic differentiation among ponds was highly significant at each locus (log-likelihood based test; P < 0.0001). One-locus $\hat{\theta}$ estimates of parameter F_{ST} ranged from 0.592 to 0.872 (Table 4). The global test across loci was also highly significant (Fisher's method; P < 0.0001) and the

multilocus estimate of F_{ST} was equal to 0.723. Intraclass correlation for allele size estimates ($\hat{\rho}_{ST}$) ranged between 0.470 and 0.885. $\hat{\rho}_{ST}$ estimates were not significantly different than $\hat{\theta}$ estimates (two-tailed Wilcoxon's signed-rank test; P = 0.686). Pairwise $\hat{\theta}$ estimates were significantly correlated with geographic distance (10000 permutations Mantel test; P < 0.0001), as were $\hat{\rho}_{ST}$ estimates (P < 0.0001). [Table 4]

Multi-locus analyses-- Among 297 complete genotypes out of 308 sampled individuals (some data were missing for 11 multilocus genotypes), 20 distinct genotypes were found, out of which five were heterozygous (at a single locus). Figure 5 shows the spatial distribution and the frequency of multilocus genotypes within the Roque-Haute natural reserve. Two single genotypes (A and J) represent more than 70% of the total sampled genotypes. In 19 ponds (out of 26) a single genotype was found (see sample sizes in Appendix 2). The striking point is the clear-cut separation between haplotypes A and J. There was no clear relationship between the levels of polymorphism maintained within single ponds and the presence / absence pattern in those ponds (e.g., polymorphic ponds 118, 180, and 182 were among the ponds where M. strigosa was the most often observed, but polymorphic ponds 62, 81, and 163 were not). In any case, the genotypes sampled in the newly colonized ponds of year 1999 were consistent with the strong clustering of genotypes observed (see the dark-grey filled ponds in Fig. 5). [Fig. 5]

In nine ponds out of 13, multilocus genotypes were identical in years 1994 and 1999. Yet, the genotype B was found in pond 64 in 1994 and not in 1999, and the genotype O was found in pond 129 in 1994 and not in 1999. The genotype A was present in both ponds, in 1994 and 1999. The genotype D was found in pond 58 in 1994 and not in 1999 (but the genotype J was found there in 1994 and 1999). In pond 180, the genotypes C and J were found in 1994 and not in 1999, while genotypes M and O, which were found in 1999 were absent from the 1994 sample. Yet, the genotype A was found in pond 180 in 1994 and 1999. Note that the two genotypes M and O are genetically close to genotype A (see Fig. 6).

To infer the relationships between pairs of genotypes, a minimum spanning network (Kruskal, 1956) was constructed from the pairwise allele size difference distance of Eq. 1. Note that the distances separating two nodes on Fig. 6 (each node representing a single sampled genotype) are proportional to the actual distances between two multilocus genotypes (Eq. 1). Whenever a pond sample was composed of more than one genotype, those genotypes were closely related (Fig. 6: see, e.g., ponds 81, 118, 163, and 180). [Fig. 6]

DISCUSSION

Characterization of variation-- Allozyme variation at 12 putative loci was totally absent from the French and Spanish samples studied here (Vitalis et al., 1998). Microsatellite variation was low, with an average number of alleles over all Mediterranean samples equal to 1.65 (SD = 0.61; see Appendix 1). The difference between the two classes of markers is probably due to the difference in mutation rates. The mutation rates of microsatellite loci have been reported to fall in the range of 10^{-6} to 10^{-2} (Jarne and Lagoda, 1996; Goldstein and Schlötterer, 1999), i.e., up to 10^{4} times larger than that of allozymes (see, e.g., Voelker, Schaffer, and Mukai, 1980). The effective number of alleles per locus (estimated as the inverse of the IIS probability for pairs of genes within the total pooled Mediterranean sample) was positively correlated with the weighted average number of repeats (Spearman correlation r_s = 0.57; P = 0.180). The lack of significance was due to locus IVH5. With this locus removed, the correlation was higher and significant (r_s = 0.89; P = 0.019). This could indicate that replication slippage may be an important factor in generating variation at those microsatellite loci (see, e.g., Innan, Terauchi, and Miyashita, 1997). Because available data and simulation-based tests are not yet appropriate to test whether microsatellite loci mutate according to a stepwise mutation model (SMM; Kimura and Ohta, 1978), an infinite allele model (IAM; Kimura and Crow, 1964) or a mixture of both (two-phase mutation models or TPM; Slatkin, 1995), there is no real consensus about the actual mutational processes at play along microsatellite sequences (see, e.g., Jarne and Lagoda, 1996). Although the TPM has been considered the most appropriate model (Di Rienzo et al., 1998), neither the SMM nor even the IAM could be rejected (Valdes, Slatkin, and Freimer, 1993).

Marsilea strigosa reproductive system-- Within-population lack of variation (fixation) at most loci prevented the estimation of F_{IS} parameters and the application of exact tests for deviation from Hardy-Weinberg proportions in most cases (57 out of 91 for the Mediterranean data set; 115 out 130 for the Roque-Haute data set). Exact tests performed per locus and per sample were far from being all significant. In nearly all cases, however, the absence of significance was due to small sample sizes. Such low statistical power for largely inbred samples is not unexpected (see Robertson and Hill, 1984; Rousset and Raymond, 1995). Moreover, all

the multisample score tests but one (pond 59) were significant. Selfing rate estimates based on multilocus $\hat{f}$ estimates overall ranged from 0.743 to 1.000 and should be considered as underestimating the actual selfing rate. Indeed, in a finite population of size N, where the overall amount of selfing is s, $F_{IS} = \sigma / (2 - \sigma)$, with $\sigma = s - 1 / N$ (see, e.g., Rousset, 1996). Thus, what is really estimated is the actual rate at which selfed progeny are produced, minus the inverse of (unknown) population size. Although the presence of male and female spores (heterospory) prevents intragametophytic selfing (the fusion of male and female gametes derived from the same gametophyte), M. strigosa clearly reproduces mainly through intergametophytic selfing (the fusion of male and female gametes from different gametophytes derived from the same sporophyte, which is analogous to seed plant selfing). Were M. strigosa clonally reproducing only, strong heterozygote excesses would be expected. Therefore, our results imply that M. strigosa reproduces sexually through an annual life cycle and that sporocarps are the only means for M. strigosa to survive to winter harsh conditions.

These results are consistent with a recent study by Schneider and Pryer (in press) of the structure and function of spores in the Marsileaceae. Although Schneider and Pryer (in press) argued that the mechanism to float at the water/air interface for an extended period could favor the mixing of spores of different individuals (and thereby increase the potential for outbreeding), we did not find any evidence for an effective outbreeding mating system in M. strigosa. Yet, Schneider and Pryer (in press) observations that a small chamber is formed at the aperture point between the two outer layers (perine and exine) of the megaspore may give a clue to the functional aspects of M. strigosa reproductive system. Schneider and Pryer (in press) argued that this free space (referred to as a sperm lake) between the solid and gelatinous layers of the megaspore might function primarily as a trap for male gametes. Because of the large size of megaspores, a vortex-like effect would indeed occur and would force sperm cells in the local neighborhood to be caught up in the vortex and to be propelled into the sperm lake. Similar observations have been made in aquatic flowering plants with hydrophilous pollination (see, e.g., Cox, 1988).

One-locus population differentiation at contrasted geographic scales-- We observed considerable and highly significant differentiation at both Mediterranean and Roque-Haute scales (multilocus estimates $\hat{\theta}$ = 0.845 and 0.723, respectively; see Tables 2 and 4). This indicates that gene flow is highly restricted, not only between Mediterranean populations (average pairwise distance = 576 km, SD = 357 km) but also between ponds within the Roque-Haute natural reserve (average pairwise distance between sampled ponds = 267 m, SD = 157 m). Among the Mediterranean samples, $\hat{\rho}_{ST}$ estimates of the intraclass correlations for allele sizes (ρ_{ST}) were significantly larger than $\hat{\theta}$ estimates of F_{ST} while the same was not true among Roque-Haute 1999 samples (see Tables 2 and 4). This result is in agreement with the analytical predictions by Rousset (1996), who showed that, with stepwise mutations occurring, F_{ST} increases much less than ρ_{ST} in the more distant populations. Although those exact results hold in equilibrium models of population structures (Rousset, 1996), a similar lack of variation for F_{ST} measures was also obtained through simulations of nonequilibrium situations (Slatkin, 1993). A similar relationship was obtained on Bulinus truncatus, a highly selfing freshwater snail, at smaller but comparable geographic scales (Viard et al., 1996). Nonsignificant differences between estimates of F_{ST} and ρ_{ST} have also been reported (Michalakis and Veuille, 1996; Estoup et al., 1998). According to (Estoup et al., 1998), the absence of significant differences could be the result of large dispersal rates and / or recent populations divergence, two situations under which differentiation does not depend much on the mutation process (Slatkin, 1995; Rousset, 1996). Our results are thus consistent with theoretical expectations, with significantly larger $\hat{\rho}_{ST}$ values than $\hat{\theta}$ values at the Mediterranean scale and nonsignificant differences at the Roque-Haute natural reserve scale, where the average time of divergence among pools is expected to be more recent than among Mediterranean populations and where (even limited) gene flow is expected to be higher at smaller geographic scales.

There are but few studies to compare our results with. With the exception of few studies using polymorphic DNA markers (see, e.g., Schneller et al., 1998; Keiper and McConchie, 2000; Landergott et al., 2001; Pryor et al., 2001) most of the population genetic studies in pteridophytes have been carried out with allozymes. Moreover this is, to our knowledge, the first report of the population genetic structure of a heterosporous fern (heterospory being a characteristic trait of Marsileaceae and Salviniaceae: see, e.g., Pryer, 1999). Keiper and McConchie (2000) reported high F_{ST} estimates (average = 0.783) at 469 amplified fragment length polymorphisms (AFLPs) among eight natural populations of the umbrella fern Sticherus flabellatus (R. Br) St John in Australia. With the exception of Hemionitis palmata, with an overall

estimated F_{ST} of 0.698 (Ranker 1992), such a high population differentiation is not common in Pteridophytes (Keiper and McConchie, 2000). Maki and Asada (1998) showed that gene flow among populations of Polystichum otomasui (a narrow endemic fern in Japan) was high enough to impede genetic differentiation. Maki and Asada (1998) argued that, as for most homosporous ferns, aerodynamic properties of spores make them likely to be highly efficient dispersal units in this species. Rare but effective long-distance dispersal events were also invoked to account for the distribution of genetic variation in Dryopteris remota, an apomictic homosporous fern (Schneller et al., 1998). In Marsileaceae however, spores are produced within the sporocarps, which therefore must be considered as the units of long-distance dispersal (Johnson, 1985). Our results suggest that sporocarps may not be as efficient dispersing units as homosporous fern spores are likely to be.

Multilocus population structure-- At the Mediterranean scale, some pairs of loci showed significant genotypic disequilibria within samples. For each pair of loci, Burrows' Δ_{ij} estimates of genotypic linkage disequilibria across all populations (Cockerham and Weir, 1977) were highly significantly different from zero. Ohta's variance components of linkage disequilibrium (Ohta, 1982b) within subpopulations relative to that of the total populations (D'^2_{IS} and D^2_{ST}) were very large and significantly greater than the variance of the total disequilibrium (D'^2_{ST}) and that of disequilibrium within a subpopulation (D^2_{IS}), respectively (see Table 3). Because limited migration is very effective in increasing differentiation of gamete types among subpopulations, this result is expected through random genetic drift alone (Ohta, 1982a). This very strong multilocus genotypic structure across Mediterranean populations, together with several suspicions that the microsatellite loci typed here fitted with a SMM (see above), led us to consider the multilocus genotypic population structure of M. strigosa.

The minimum spanning network (MSN) constructed from the pairwise distance given by Eq. 1 between sampled multilocus genotypes was strikingly representative of the actual population clustering (Fig. 4). In a single case, one genotype was found in two populations (genotype 16 in populations 3 and 4). The relative branching of populations was congruent with the $(\delta\mu)^2$-based tree (Fig. 3), with, e.g., Ben Slimane daya 1 branching to Valdepeñas population. Although the similarity between multilocus genotypes MSN and $(\delta\mu)^2$ distance based dendrogram might not be expected in general situations, this can be interpreted here as the consequence of the quasi absence of shared multilocus genotypes between populations, as well as the low frequency of shared alleles between populations, at most loci. The clustering of multilocus genotypes reflects the quasi absence of gene flow among Mediterranean populations. The source of genetic variation within populations seems therefore to be limited to dramatically rare events of mutation and migration, followed by effective recombination (although the latter should also be considered as a rare event, given the high rate of selfing inferred from the data).

Fine-scale population structure-- At a smaller scale, within the Roque-Haute natural reserve metapopulation, we observed a very strong clustering of multilocus genotypes (Fig. 5). Except genotype N (ponds 115 and 180), all the genotypes common to two or more populations were strongly geographically clustered. A striking result comes from the relationships between the multilocus genotypes within single (polymorphic) ponds (Figs 5-6). Within all the polymorphic ponds (see, e.g., ponds 81, 118, 163, and 180) multilocus genotypes are closely linked in the MSN based on allele size differences (Eq. 1) (see Fig. 6, and the box in Fig. 5.) This is consistent with the strong one-locus differentiation. There was no clear pattern between long-term presence of M. strigosa in a given pond and the genetic diversity retained in that pond. For example, no M. strigosa sporophyte was observed in pond 81 for 5 yr, and yet the averaged heterozygosity within this pond was the highest of all ponds: $(1-\hat{Q}_1)$ = 0.367 (see Appendix 2). Long- or mid-term storage of sporocarps in the soil has the same effect on genetic variation as seed banks. Because sporocarp storage promotes the age structuring of populations, it may increase the effective population size and thus dampen the effect of genetic drift (see, e.g., Rousset, 1999). In the homosporous fern Athyrium filix-femina, spore banks were shown to retain a considerable amount of genetic variation (Schneller, 1998). Moreover, the newly colonized ponds in 1999 do not undermine the multilocus structure observed. In all cases, the multilocus genotypes present in the newly colonized ponds are compatible with short-scale dispersal events from adjacent ponds. This was particularly striking for polymorphic ponds 62 and 163. The overall pattern of genetic structure is therefore consistent with rare dispersal events occurring at a short scale (isolation by distance). Interestingly, this demonstrates that, reminiscent of the Mediterranean pattern of multilocus population structure (see Fig. 4), genetic variation may arise primarily as the result of mutation and recombination.

Implications for conservation-- Our results show that gene flow is extremely reduced between Mediterranean populations, as well as between close ponds within a single site. The low dispersal ability of M. strigosa surely jeopardizes its chance to recolonize empty sites, even if rare events (e.g., the odd climatic conditions such as the late reflooding in spring 1999) may favor the movements of dispersal units or agents from time to time. There are only a few studies that aim at documenting the level and the distribution of genetic variation of such ephemeral plant populations (Elam, 1998). It is therefore difficult to extrapolate our results to other plant species living in temporarily flooded habitats. Yet, the presence of M. strigosa in far distant populations at the Mediterranean scale (see Fig. 1) and its apparently low ability to disperse is puzzling. We cannot discriminate between a scenario of on-going habitat fragmentation (in which case the actual populations are remnant of some more common wetlands in a distant past) and a scenario of on-going loss of dispersal agents (like, e.g., water birds). For both scenarios, however, the low genetic variation maintained within populations should not be considered as the major threat to M. strigosa. Instead, some agricultural practices (e.g., draining, overpumping of ground water), along with many other threats (invasion of alien species, pollution, etc.), imperil the existing populations. In France, the third known M. strigosa population (Saint-Estève, Eastern Pyrenees) disappeared in 1982, following the management of a temporarily flooded pond into a permanent one (Amigo, 1987). Yet, a new M. strigosa population was found close to the extinct one in 1996, within an 82-ha wet land in Torremilla (J. Molina, pers. comm.). Given the low dispersal capabilities of this species, management strategies should thus first focus on the protection and the conservation of M. strigosa remnant populations and suitable habitats.

LITERATURE CITED

Allsopp, A. 1952. Longevity of Marsilea sporocarps. Nature 169: 79-80.

Amigo, J. J. 1987. Exit la mare temporaire de Saint-Estève (Saint Estève, Pyrénées-Orientales, France). Naturalia Ruscinonensia 1: 71-136.

Black, W. C., and E. S. Krafsur. 1985. A Fortran program for the calculation and analysis of two-locus linkage disequilibrium coefficients. Theoretical and Applied Genetics 70: 491-496.

Blondel, J., and J. Aronson. 1999. Biology and wildlife of the Mediterranean region. Oxford University Press, Oxford, UK.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analyses: models and estimations procedures. American Journal of Human Genetics 19: 233-257.

Cockerham, C. C., and B. S. Weir. 1977. Digenic descent measures for finite populations. Genetical Research 30: 121-147.

Cox, P. A. 1988. Hydrophilous pollination. Annual Review of Ecology and Systematics 19: 261-279.

Defense Mapping Agency. 1987. Department of Defense World Geodetic System 1984, its definition and relationships with local geodetic systems.DMA Technical report 8350.2. Washington, Washington, USA.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill, M. L. Petzl-Erler, G. K. Haines, and D. H. Barch. 1998. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148: 1269-84.

DiCiccio, T. J., and B. Efron. 1996. Bootstrap confidence intervals. Statistical Science 11: 189-228.

Elam, D. R. 1998. Population genetics of vernal pools plants: theory, data and conservation implications. In C. W. Witham, E. T. Bauder, D. Belk, W. R. Ferren Jr, and R. Ornuff [eds.], Ecology, conservation, and management of vernal pool ecosystems - Proceedings from a 1996 Conference, 180-189. California Native Plant Society, Sacramento, California, USA.

Estoup, A., F. Rousset, Y. Michalakis, J.-M. Cornuet, M. Adriamanga, and R. Guyomard. 1998. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (Salmo trutta). Molecular Ecology 7: 339-353.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27: 401-410.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, USA.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. Science 155: 279-284.

Frankham, R. 1996. Relationship of genetic variation to population size in wildlife. Conservation Biology 10: 1500-1508.

Garnier-Gere, P., and C. Dillmann. 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. Journal of Heredity 83: 239.

Gitzendanner, M. A., and P. S. Soltis. 2000. Patterns of genetic variation in rare and widespread plant congeners. American Journal of Botany 87: 783-792.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. Genetic

absolute dating based on microsatellites and the origin of modern humans. Proceedings of the National Society of Sciences, USA 92: 6723-6727.

Goldstein, D. B., and C. Schlötterer. 1999. Microsatellites: evolution and applications. Oxford University Press, Oxford, UK.

Goudet, J., M. Raymond, T. de Meeus, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics 144: 1933-1940.

Guo, S. W., and E. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportion for multiples alleles. Biometrics 48: 361-372.

Hedges, S. B. 1992. The number of replications needed for the accurate estimation of the bootstrap *P* value in phylogenetic studies. Molecular Biology and Evolution 9: 366-369.

Holland, R. F., and S. K. Jain. 1981. Insular biogeography of vernal pools in the Central Valley of California. American Naturalist 117: 24-37.

Holsinger, K. E., R. J. Mason-Gamer, and J. Whitton. 1999. Genes, demes, and plant conservation. In L. F. Landweber, and A. P. Dobson [eds.], Genetics and the extinction of species, 23-46. Princeton University Press, Princeton, New Jersey, USA.

Innan, H., R. Terauchi, and N. T. Miyashita. 1997. Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana. Genetics 146: 1441-1452.

Jain, S. K. 1994. Genetics and demography of rare plants and patchily distributed colonizing species. In V. Loeschcke, J. Tomiuk, and S. K. Jain [eds.], Conservation genetics, 291-307. Birkhäuser Verlag, Basel, Switzerland.

Jarne, P., and P. Lagoda. 1996. Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution 11: 424-429.

Johnson, D. M. 1985. New records for longevity of *Marsilea* sporocarps. American Fern Journal 75: 30-31.

Keiper, F. J., and R. McConchie. 2000. An analysis of genetic variation in natural populations of Sticherus flabellatus [R. Br. (St John)] sing amplified fragment length polymorphism (AFLP) markers. Molecular Ecology 9: 571-581.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49: 725-738.

Kimura, M., and T. Ohta. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proceedings of the National Society of Sciences, USA 75: 2868-2872.

Kruskal, J. B. J. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society 7: 48-50.

Lande, R. 1999. Extinction risks from anthropogenic, ecological, and genetic factors. In L. F. Landweber, and A. P. Dobson [eds.], Genetics and the extinction of species, 1-22. Princeton University Press, Princeton, New Jersey, USA.

Landergott, U., R. Holderegger, G. Kozlowski, and J. J. Schneller. 2001. Historical bottlenecks decrease genetic diversity in natural populations of Dryopteris cristata. Heredity 87: 344-355.

Maki, M., and Y.-J. Asada. 1998. High genetic variability revealed by allozymic loci in the narrow endemic fern Polystichum otomasui (Dryopteridaceae). Heredity 80: 604-610.

Mantel, N. 1967. The detection of disease clustering and generalized regression approach. Cancer Research 27: 209-220.

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142: 1061-1064.

Michalakis, Y., and M. Veuille. 1996. Length variation of CAG / CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. Genetics 143: 1713-1725.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences, USA 70: 3321-3323.

Ohta, T. 1982a. Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proceedings of the National Society of Sciences, USA 79: 1940-1944.

Ohta, T. 1982b. Linkage disequilibrium with the island model. Genetics 101: 139-155.

Olivier, L., J. P. Galland, H. Maurin, and J. P. Roux. 1995. Livre rouge de la flore menacée de France. Tome 1: espèces prioritaires. Collection Patrimoines naturels, vol. 20. Muséum National d'Histoire Naturelle, Conservatoire Botanique National de Porquerolles & Ministère de l'Environnement, Paris, France.

Page, R. D. M. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

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

Pryer, K. M. 1999. Phylogeny of marsileaceous ferns and relationships of the fossil Hydropteris pinnata reconsidered. International Journal of Plant Sciences 160: 931-954.

Pryor, K. V., J. E. Young, F. J. Rumsey, K. J. Edwards, M. W. Brudford, and H. J. Rogers. 2001. Diversity, genetic structure and evidence of outcrossing in British populations of the rock fern Adiantum capillus-veneris using

microsatellites. Molecular Ecology 10: 1881-1894.

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

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43: 223-225.

Robertson, A., and W. G. Hill. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics 107: 703-718.

Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142: 1357-1362.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics 145: 1219-1228.

Rousset, F. 1999. Genetic differentiation in populations with different classes of individuals. Theoretical Population Genetics 55: 297-308.

Rousset, F. 2001. Inferences from spatial population genetics. In D. Balding, M. Bishop, and C. Cannings [eds.], Handbook of statistical genetics, 239-269. John Wiley, Chichester, UK.

Rousset, F., and M. Raymond. 1995. Testing heterozygote excess and deficiency. Genetics 140: 1413-1419.

Schneider, H., and K. M. Pryer. in press. Structure and function of spores in the aquatic heterosporous fern family Marsileaceae. International Journal of Plant Sciences.

Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin, Version 2000: a software for population genetics data analysis. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva, Switzerland.

Schneller, J., R. Holderegger, F. Gugerli, K. Eichenberger, and E. Lutz. 1998. Patterns of genetic variation detected by RAPDs suggest a single origin with subsequent mutations and long-distance dispersal in the apomictic fern Dryopteris remota (Dryopteridaceae). American Journal of Botany 85: 1038-1042.

Schneller, J. J. 1998. How much genetic variation in fern populations is stored in the spore banks ? A study of Athyrium filix-femina (L.) Roth. Botanical Journal of the Linnean Society 127: 195-206.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47: 264-279.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457-462.

Soltis, D. E., and P. S. Soltis. 1986. Active enzymes from megaspores of Marsilea and Regnellidium. American Fern Journal 76: 17-20.

Tutin, T. G., V. H. Heywood, N. A. Burges, D. M. Moore, D. H. Valentine, S. M. Walters, and D. A. Webb. 1964. Flora Europaea, vol. 1. Cambridge University Press, Cambridge, UK.

Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133: 737-49.

van Treuren, R., R. Bijlsma, W. van Delden, and N. J. Ouborg. 1991. The significance of genetic erosion in the process of extinction. I. Genetic differentiation in Salvia pratensis and Scabiosa columbaria in relation to population size. Heredity 66: 181-189.

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

Vitalis, R., B. Colas, M. Riba, and I. Olivieri. 1998. Marsilea strigosa Willd.: statut génétique et démographique d'une espèce menacée. Ecologia Mediterranea 24: 145-157.

Vitalis, R., M.-P. Dubois, and I. Olivieri. 2001. Characterization of microsatellite loci in the endangered species of fern Marsilea strigosa Willd. (Marsileaceae, Pteridophyta). Molecular Ecology Notes 1: 64-66.

Voelker, R. A., H. E. Schaffer, and T. Mukai. 1980. Spontaneous allozyme mutations in Drosophila melanogaster: rate of occurrence and nature of the mutants. Genetics 94: 961-968.

Weir, B. S. 1979. Inferences about linkage disequilibrium. Biometrics 35: 235-54.

Weir, B. S. 1996. Genetics data analysis II. Sinauer, Sunderland, Massachusetts, USA.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15: 323-354.

Young, A., T. Boyle, and T. Brown. 1996. The population genetic consequences of habitat fragmentation for plants. Trends in Ecology and Evolution 11: 413-418.

APPENDIX 1. Allele frequencies, average observed heterozygosity $(1-\hat{Q}_0)$, within-population expected heterozygosity $(1-\hat{Q}_1)$, and Weir and Cockerham's (1984) estimate ($\hat{f}$) of F_{IS}, at the seven polymorphic loci in the 13 Mediterranean sampled populations of Marsilea strigosa. Mean values of these statistics are also given per population, as well as the mean number of alleles per locus (N_A) and the percentage of polymorphic loci, at 90% (P_{90}: only allele frequencies ≥ 0.90 accounted for) and 95% (P_{95}: only allele frequencies ≥ 0.95 accounted for).

Locus IA10	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 50$)	12 ($N = 54$)	13 ($N = 60$)
	68			1.000	1.000	1.000								
	78									1.000				
	82											1.000		
	90		0.932											
	92		0.068											
	98	0.071							0.375					1.000
	100								0.250					
	102	0.095					0.875	1.000	0.083		1.000			
	104	0.786												
	106	0.048					0.125		0.208					
	112								0.083					
	122												0.056	
	130												0.037	
	132												0.796	
	134												0.037	
	136												0.074	
$(1-\hat{Q}_0)$		0.143	0.045	0.000	0.000	0.000	0.000	0.000	0.083	0.000	0.000	0.000	0.185	0.000
$(1-\hat{Q}_1)$		0.381	0.132	0.000	0.000	0.000	0.250	0.000	0.803	0.000	0.000	0.000	0.365	0.000
$\hat{f}$		0.625	0.656	-	-	-	1.000	-	0.896	-	-	-	0.492	-

APPENDIX 1. Continued

Locus IE3	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 52$)	13 ($N = 54$)
	102	1.000	1.000				1.000	1.000	1.000	1.000	1.000	1.000		
	104												1.000	1.000
	106			1.000	1.000	1.000								
$(1-\hat{Q}_0)$		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\hat{f}$		-	-	-	-	-	-	-	-	-	-	-	-	-

Locus IF7	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 26$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 60$)	13 ($N = 60$)
	98			1.000	1.000	1.000							1.000	1.000
	112	0.024												
	114	0.071												
	116	0.024	0.136											
	118	0.095	0.864				0.500	0.214			1.000	0.865		
	120	0.024					0.500	0.786				0.135		
	122									1.000				
	124								1.000					
	126	0.762												
$(1-\hat{Q}_0)$		0.190	0.000	0.000	0.000	0.000	0.000	0.143	0.000	0.000	0.000	0.038	0.000	0.000
$(1-\hat{Q}_1)$		0.419	0.247	0.000	0.000	0.000	0.571	0.381	0.000	0.000	0.000	0.242	0.000	0.000
$\hat{f}$		0.546	1.000	-	-	-	1.000	0.625	-	-	-	0.841	-	-

APPENDIX 1. Continued

Locus IIIB8	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 56$)	13 ($N = 60$)
	113			0.071	1.000	1.000								
	115						1.000	1.000						
	117								0.708	0.850	1.000	1.000		
	123												1.000	0.967
	125													0.033
	129	0.095												
	131		0.136	0.929										
	133	0.119												
	135	0.738	0.864						0.292	0.150				
	137	0.048												
$(1-\hat{Q}_0)$		0.048	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.300	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.450	0.247	0.143	0.000	0.000	0.000	0.000	0.447	0.267	0.000	0.000	0.000	0.067
$\hat{f}$		0.894	1.000	1.000	-	-	-	-	0.814	-0.125	-	-	-	1.000

Locus IIIC8	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 60$)	13 ($N = 60$)
	101		0.182				1.000	0.214	0.500					
	103	1.000	0.818								1.000	0.077	1.000	
	105							0.786				0.923		0.017
	107													0.967
	109													0.017
	111								0.500	1.000				
	135			0.929	1.000	1.000								
	139			0.071										
$(1-\hat{Q}_0)$		0.000	0.000	0.000	0.000	0.000	0.000	0.143	0.000	0.000	0.000	0.000	0.000	0.067
$(1-\hat{Q}_1)$		0.000	0.312	0.143	0.000	0.000	0.000	0.381	0.545	0.000	0.000	0.148	0.000	0.066
$\hat{f}$		-	1.000	1.000	-	-	-	0.625	1.000	-	-	1.000	-	-0.009

APPENDIX 1. Continued

Locus IVH5	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 58$)	13 ($N = 58$)
	159			1.000	1.000	1.000								
	167													1.000
	173								0.583					
	177								0.083					
	181								0.333					
	185												1.000	
	211	0.048												
	213	0.952	1.000				0.750	0.429		1.000		1.000		
	215						0.250	0.571			1.000			
$(1-\hat{Q}_0)$		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.095	0.000	0.000	0.000	0.000	0.429	0.571	0.591	0.000	0.000	0.000	0.000	0.000
$\hat{f}$		1.000	-	-	-	-	1.000	1.000	1.000	-	-	-	-	-

APPENDIX 1. Continued

Locus XIIE3	Allele (bp)	1 ($N = 42$)	2 ($N = 44$)	3 ($N = 28$)	4 ($N = 10$)	5 ($N = 6$)	6 ($N = 16$)	7 ($N = 14$)	8 ($N = 24$)	9 ($N = 20$)	10 ($N = 18$)	11 ($N = 52$)	12 ($N = 60$)	13 ($N = 60$)
	150	0.048												
	156	0.095												
	160	0.833												
	162		0.182											
	164	0.024	0.773					0.071						
	170		0.045											
	176												0.867	
	178												0.133	
	184						1.000	0.929	0.042					
	186										0.056			
	188								0.667					
	190								0.083	0.700				
	192								0.208	0.300		0.019		
	198										0.944			
	200			0.929										
	204											0.038		
	206											0.904		
	210			0.071		1.000						0.038		0.867
	212				1.000									0.100
	214													0.033
$(1-\hat{Q}_0)$		0.048	0.000	0.000	0.000	0.000	0.000	0.143	0.083	0.000	0.111	0.038	0.000	0.000
$(1-\hat{Q}_1)$		0.307	0.385	0.143	0.000	0.000	0.000	0.143	0.545	0.467	0.111	0.186	0.239	0.246
$\hat{f}$		0.845	1.000	1.000	-	-	-	0.000	0.847	1.000	0.000	0.793	1.000	1.000

APPENDIX 1. Continued

Total	1	2	3	4	5	6	7	8	9	10	11	12	13
N_A	3.1	1.9	1.4	1	1	1.4	1.6	2.6	1.3	1.1	1.7	1.7	1.7
P_{95}	57.14	71.43	42.86	0.00	0.00	42.86	57.14	71.43	28.57	14.29	42.86	28.57	14.29
P_{90}	71.43	71.43	42.86	0.00	0.00	42.86	57.14	71.43	28.57	14.29	42.86	28.57	42.86
$(1-\hat{Q}_0)$	0.061	0.006	0.000	0.000	0.000	0.000	0.061	0.036	0.043	0.016	0.011	0.025	0.010
$(1-\hat{Q}_1)$	0.236	0.189	0.062	0.000	0.000	0.179	0.211	0.419	0.105	0.016	0.083	0.085	0.055
$\hat{f}$	0.741	0.966	1.000	-	-	1.000	0.710	0.915	0.591	0.000	0.866	0.706	0.824

APPENDIX 2. Allele frequencies, average observed heterozygosity $(1-\hat{Q}_0)$, within-population expected heterozygosity $(1-\hat{Q}_1)$, and Weir and Cockerham's (1984) estimate ($\hat{f}$) of F_{IS}, at the five polymorphic loci in the 26 Roque-Haute samples. Mean values of these statistics are also given per population, as well as the mean number of alleles per locus (N_A) and the percentage of polymorphic loci, at 90% (P_{90}: only allele frequencies ≥ 0.90 accounted for) and 95% (P_{95}: only allele frequencies ≥ 0.95 accounted for). The samples fixed for a single (multilocus) genotype have been grouped in the following. Group I gathers ponds 1 (N = 10), 64 (N = 28), 65 (N = 58), 66 (N = 18), 69 (N = 10), 79 (N = 10), 129 (N = 20) and 165 (N = 26). Group II gathers ponds No. 24 (N = 16), 50 (N = 20), 51 (N = 20), 52 (N = 18), 56 (N = 10), 58 (N = 28), and 63 (N = 18).

Locus IA10	Allele (bp)	Group I	Group II	46 (N = 38)	59 (N = 8)	62 (N = 10)	81 (N = 40)	115 (N = 30)	118 (N = 60)	147 (N = 20)	163 (N = 10)	170 (N = 28)	180 (N = 34)	182 (N = 22)
	86						0.100							
	88						0.150							
	90	1.000	1.000	0.974	1.000	1.000	0.750	1.000	0.967		1.000	1.000	1.000	1.000
	92			0.026					0.033	1.000				
$(1-\hat{Q}_0)$		0.000	0.000	0.053	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.053	0.000	0.000	0.426	0.000	0.066	0.000	0.000	0.000	0.000	0.000
$\hat{f}$		-	-	0.000	-	-	1.000	-	-0.02	-	-	-	-	-

APPENDIX 2. Continued

Locus IF7	Allele (bp)	Group I [a]	Group II	46 ($N = 38$)	59 ($N = 8$)	62 ($N = 10$)	81 ($N = 40$)	115 ($N = 30$)	118 ($N = 60$)	147 ($N = 20$)	163 ($N = 10$)	170 ($N = 28$)	180 ($N = 34$)	182 ($N = 20$)
	114						0.200							
	116						0.050		0.117		0.600	1.000		
	118	1.000	1.000	1.000	1.000	1.000	0.750	1.000	0.883	1.000	0.400		1.000	1.000
$(1-\hat{Q}_0)$		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.000	0.000	0.000	0.416	0.000	0.213	0.000	0.600	0.000	0.000	0.000
$\hat{f}$		-	-	-	-	-	1.000	-	0.843	-	1.000	-	-	-
Locus IIIB8	Allele (bp)	Group I [b]	Group II	46 ($N = 38$)	59 ($N = 8$)	62 ($N = 10$)	81 ($N = 40$)	115 ($N = 30$)	118 ($N = 60$)	147 ($N = 20$)	163 ($N = 8$)	170 ($N = 28$)	180 ($N = 36$)	182 ($N = 22$)
	131								0.733					
	133						0.250							
	135	1.000	1.000	1.000	1.000	1.000	0.750	1.000	0.267	1.000	1.000	1.000	1.000	1.000
$(1-\hat{Q}_0)$		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.000	0.000	0.000	0.395	0.000	0.405	0.000	0.000	0.000	0.000	0.000
$\hat{f}$		-	-	-	-	-	1.000	-	1.000	-	-	-	-	-
Locus IIIC8	Allele (bp)	Group I [c]	Group II	46 ($N = 38$)	59 ($N = 8$)	62 ($N = 10$)	81 ($N = 38$)	115 ($N = 30$)	118 ($N = 60$)	147 ($N = 20$)	163 ($N = 8$)	170 ($N = 28$)	180 ($N = 32$)	182 ($N = 22$)
	101		1.000	1.000	0.750	0.600								
	103	1.000			0.250	0.400	0.737	1.000	0.900	1.000	1.000	1.000	1.000	1.000
	105						0.263		0.100					
$(1-\hat{Q}_0)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$	0.000	0.000	0.000	0.000	0.500	0.600	0.409	0.000	0.186	0.000	0.000	0.000	0.000	0.000
$\hat{f}$	-	-	-	-	1.000	1.000	1.000	-	1.000	-	-	-	-	-

APPENDIX 2. Continued

Locus XIIE3	Allele (bp)	Group I	Group II	46 ($N = 38$)	59 ($N = 8$)	62 ($N = 10$)	81 ($N = 40$)	115 ($N = 30$)	118 ($N = 60$)	147 ($N = 20$)	163 ($N = 10$)	170 ($N = 28$)	180 ($N = 34$)	182 ($N = 22$)
	162								0.700					
	164	1.000	1.000	0.974	1.000	1.000	0.900		0.300	1.000	1.000	1.000	0.588	0.909
	166			0.026									0.176	
	168						0.100	1.000						0.091
	170												0.235	
$(1-\hat{Q}_0)$		0.000	0.000	0.053	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.053	0.000	0.000	0.189	0.000	0.434	0.000	0.000	0.000	0.603	0.182
$\hat{f}$		-	-	0.000	-	-	1.000	-	1.000	-	-	-	1.000	1.000
Total		Group I	Group II	46	59	62	81	115	118	147	163	170	180	182
N_A		1	1	1.2	1.2	1.2	2.4	1	2	1	1	1	1.2	1.2
P_{95}		0.00	0.00	0.00	0.20	0.20	1.00	0.00	0.80	0.00	0.00	0.00	0.20	0.20
P_{90}		0.00	0.00	0.00	0.20	0.20	0.80	0.00	0.60	0.00	0.00	0.00	0.20	0.00
$(1-\hat{Q}_0)$		0.000	.000	0.021	0.000	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000
$(1-\hat{Q}_1)$		0.000	0.000	0.021	0.100	0.120	0.367	0.000	0.261	0.000	0.130	0.000	0.121	0.037
$\hat{f}$		-	-	0.000	1.000	1.000	1.000	-	0.923	-	1.000	-	1.000	1.000

[a] Sample size for pond 65: $N = 56$
[b] Sample size for pond 165: $N = 26$
[c] Sample size for pond 79: $N = 8$

TABLE 1. Description of sampled populations in the Mediterranean. Longitudes and latitudes refer to international meridian (Greenwich) and to the World Geodesic System 1984 or WGS84 (Defense Mapping Agency, 1987). For each population, the pond size and some characteristics of Marsilea strigosa covering area within or around the pond are given , as well as the sampling date and the sample size (number of distinct sampled individuals).

			Geographic coordinates				Sampling date	Sample size
Country	Population		Longitude	Latitude	Pond size (*m*)	Characteristics		
France	1	Vendres	+3°14′17′′2	43°16′47′′0	1000 × 800	continuous water	10 Jun 1994	21
	2	Roque-haute	+3°21′59′′4	43°18′17′′0	(see Fig. 5)	highly fragmented	9 Jun 1994	22
Spain	3	Menorca (*Illa d'en Colom*)	+4°16′41′′7	39°55′13′′8	47 × 30	continuous water	17 May 1996	14
	4	Mallorca – 4 (*Cap Blanc*)	+2°45′56′′0	39°23′26′′5	5 × 3	continuous water	19 May 1996	5
	5	Mallorca – 51 (*Cap Blanc*)	+2°48′30′′6	39°23′05′′8	8 × 2	continuous water	19 May 1996	3
	6	Sinarcas – 1 (*Casilla del Médico*)	-1°14′23′′1	39°45′14′′8	60 × 40	edges of the pond	20 May 1996	8
	7	Sinarcas – 2 (-)	-1°14′32′′1	39°47′17′′0	28 × 20	edges of the pond	20 May 1996	7
	8	Valdepeñas (*Cañada Blanquilla*)	-3°27′45′′7	38°47′09′′4	100 × 100	edges of the pond	22 May 1996	12
	9	Guadalajara (*Laguna Grande*)	-3°15′25′′2	40°53′07′′8	725 × 500	edges of the pond	22 May 1996	10
	10	Ciudad Real (*Laguna de la Perdiguera*)	-4°15′10′′0	38°54′05′′9	600 × 400	edges of the pond	21 May 1996	9
	11	Huesca (*El Basal*)	+0°08′26′′4	41°37′02′′5	360 × 270	continuous water	26 May 1996	26
Morocco	12	Ben Slimane *daya 1*	-7°05′46′′2	33°38′14′′8	200 × 150	continuous water	3 Jul 1999	30
	13	Ben Slimane *daya 2*	-7°02′25′′2	33°33′52′′7	150 × 150	continuous water	30 Mar 1999	30

TABLE 2. Genetic differentiation among Mediterranean populations. Intraclass correlations of the IIS probability ($\hat{\theta}$) and of allele size differences ($\hat{\rho}_{ST}$), for pairs of genes within subpopulations relative to the whole population. Population differentiation was highly significant at all loci ($<10^{-4}$). The 95% confidence interval (95% CI) is given for each multilocus estimate. Confidence intervals were obtained by deriving the approximate bootstrap confidence intervals (ABC) method described by DiCiccio and Effron (1996).

Locus	$\hat{\theta}$	$\hat{\rho}_{ST}$
IA10	0.825	0.993
IE3	1.000	1.000
IF7	0.822	0.983
IIIB8	0.839	0.855
IIIC8	0.855	0.983
IVH5	0.900	0.998
XIIE3	0.742	0.990
All	0.845	0.987
95% CI	[0.799 – 0.923]	[0.963 – 0.995]

TABLE 3. Ohta's (1982a,b) components of linkage disequilibrium at the Mediterranean basin scale.

Pairs of loci			D^2_{IS}	D^2_{ST}	D'^2_{IS}	D'^2_{ST}	D^2_{IT}
IA10	-	IE3	0.000	1.369	5.770	0.223	5.993
IA10	-	IF7	0.025	1.000	4.359	0.210	4.569
IA10	-	IIIB8	0.023	1.025	4.226	0.235	4.461
IA10	-	IIIC8	0.010	1.167	5.069	0.248	5.317
IA10	-	IVH5	0.009	1.031	4.599	0.275	4.874
IA10	-	XIIE3	0.028	0.759	2.519	0.232	2.751
IE3	-	IF7	0.000	0.814	2.630	0.473	3.102
IE3	-	IIIB8	0.000	0.809	2.636	0.414	3.050
IE3	-	IIIC8	0.000	0.804	2.886	0.180	3.066
IE3	-	IVH5	0.000	0.758	2.593	0.414	3.007
IE3	-	XIIE3	0.000	0.776	2.693	0.177	2.870
IF7	-	IIIB8	0.021	0.888	3.135	0.309	3.444
IF7	-	IIIC8	0.011	0.998	4.028	0.144	4.171
IF7	-	IVH5	0.008	0.870	3.491	0.312	3.804
IF7	-	XIIE3	0.016	0.706	2.449	0.191	2.640
IIIB8	-	IIIC8	0.005	0.951	3.592	0.181	3.773
IIIB8	-	IVH5	0.002	1.012	4.076	0.262	4.338
IIIB8	-	XIIE3	0.050	0.716	2.604	0.226	2.830
IIIC8	-	IVH5	0.010	1.097	4.970	0.219	5.189
IIIC8	-	XIIE3	0.005	0.711	2.441	0.204	2.645
IVH5	-	XIIE3	0.015	0.746	2.642	0.216	2.858

TABLE 4. Genetic differentiation among Roque-Haute pools. Intraclass correlations of the IIS probability ($\hat{\theta}$) and of allele size differences ($\hat{\rho}_{ST}$), for pairs of genes within subpopulations relative to the whole population. Population differentiation was highly significant at all loci ($<10^{-4}$). The 95% confidence interval (95% CI) is given for each multilocus estimate. Confidence intervals were obtained by deriving the approximate bootstrap confidence intervals (ABC) method described by DiCiccio and Effron (1996).

Locus	$\hat{\theta}$	$\hat{\rho}_{ST}$
IA10	0.637	0.530
IF7	0.632	0.470
IIIB8	0.592	0.663
IIIC8	0.872	0.885
XIIE3	0.638	0.646
All	0.723	0.678
95% CI	[0.684 – 1.000]	[0.604 – 0.817]

CAPTIONS FOR FIGURES

Fig. 1. Location map of 13 M. strigosa sampled populations, in France, Morocco, and Spain. Geographic coordinates (longitudes and latitudes) refer to the international meridian (Greenwich), and to the World Geodetic System 1984 or WGS84 (Defence Mapping Agency 1987).

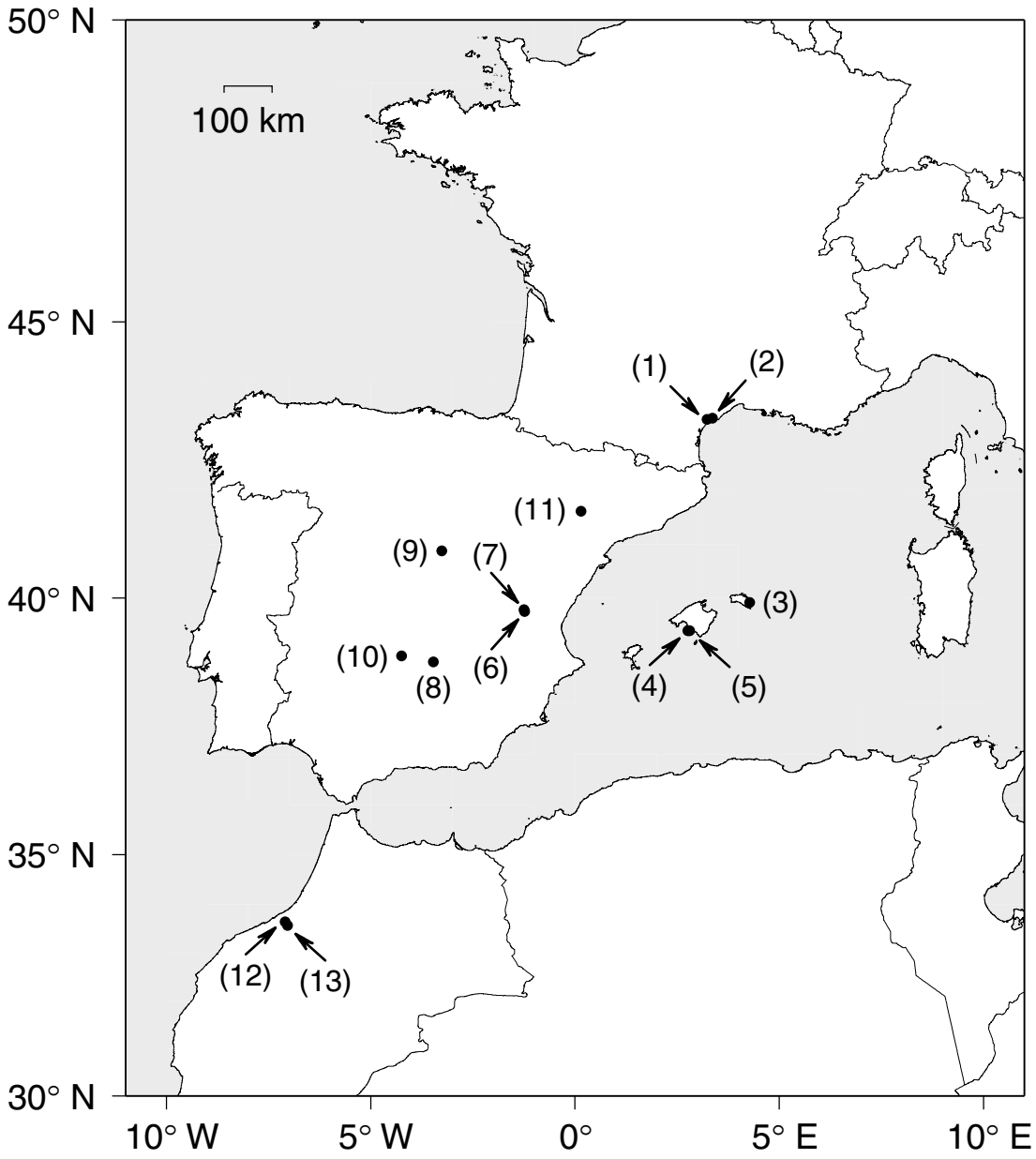
Fig. 2. Fitch-Margoliash tree inferred from pairwise Cavalli-Sforza and Edwards' (1967) chord distance between 13 Mediterranean populations. Population numbers refer to Table 1. Bootstrap values higher than 50% (calculated from 2000 bootstrap samples) are given for each node.

Fig. 3. Fitch-Margoliash tree inferred from pairwise $(\delta\mu)^2$ distance (Goldstein et al., 1995), which takes into account the average allele size differences between population. Population numbers refer to Table 1. Bootstrap values higher than 50% (calculated from 2000 bootstrap samples) are given for each node.

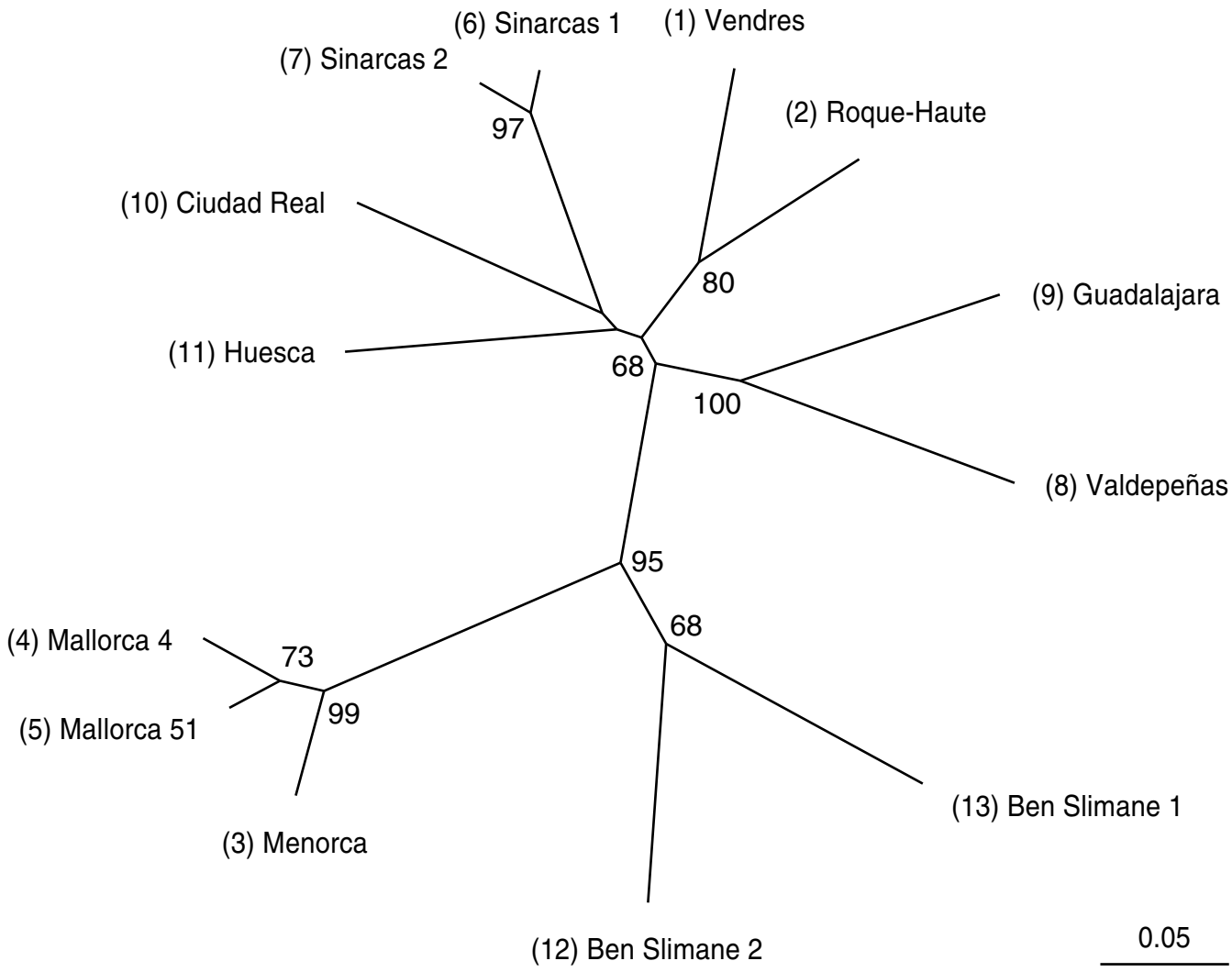
Fig. 4. Minimum Spanning Network (MSN) inferred from the pairwise distance of mean allele size differences between multilocus genotypes (eqn 1), among all Mediterranean sampled populations. Genotypes are numbered as a function of their appearance in the data set. Black circled nodes are heterozygous genotypes. Genotypes are linked one to each other by lines sizing the logarithm of distance separating these genotypes. Ellipses embed all the genotypes sampled in a single population.

Fig. 5. Roque-Haute natural reserve map. The 26 sampled ponds (out of 198) are filled in grey. The ponds filled in dark grey are the newly colonized ones in year 1999 (bold italicised numbers). All the other sampled ponds are filled in light grey (plain numbers). For each pond, the genotypic composition is indicated by a pie diagram (the smaller charts are for monomorphic sampled ponds and the larger charts are for polymorphic sampled ponds). Numbers refer to the pond numbers. Each colour characterizes a single multilocus genotype. The miniature Minimum Spanning Network (MSN) from Fig. 6 is given, to illustrate the relationships between multilocus genotypes. Note that the colour gradation tentatively reflects the genetic distances between multilocus genotypes.

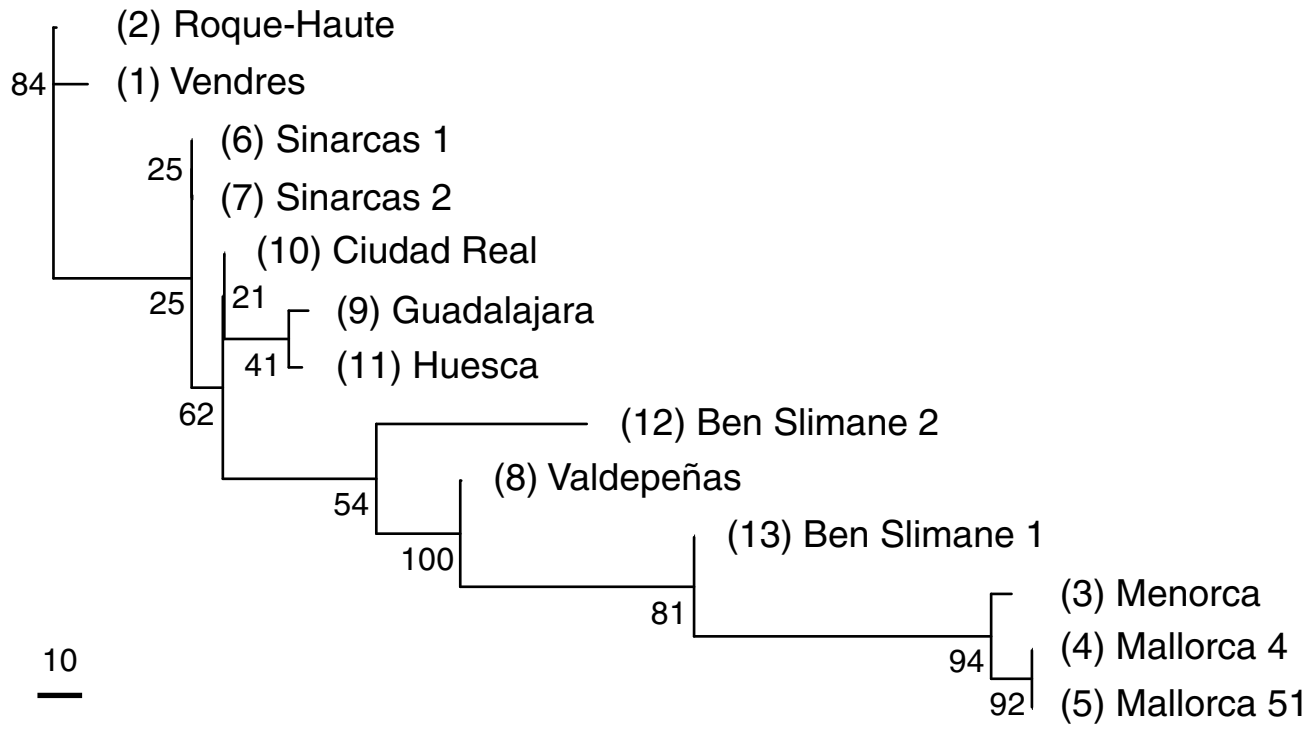
Fig. 6. Minimum Spanning Network (MSN) inferred from the pairwise distance of mean allele size differences between multilocus genotypes (Eq. 1), within the Roque-Haute natural reserve. The relative size of each node (each representing a single sampled genotype) is proportional to the natural logarithm of the genotype frequency, among all sampled ponds. The distance separating two nodes is proportional to the actual distance between the two multilocus genotypes represented by the nodes. Black circled nodes are heterozygous genotypes. Dashed lines embed all the multilocus genotypes (except the frequent A and J, for clarity) sampled in a single polymorphic population.



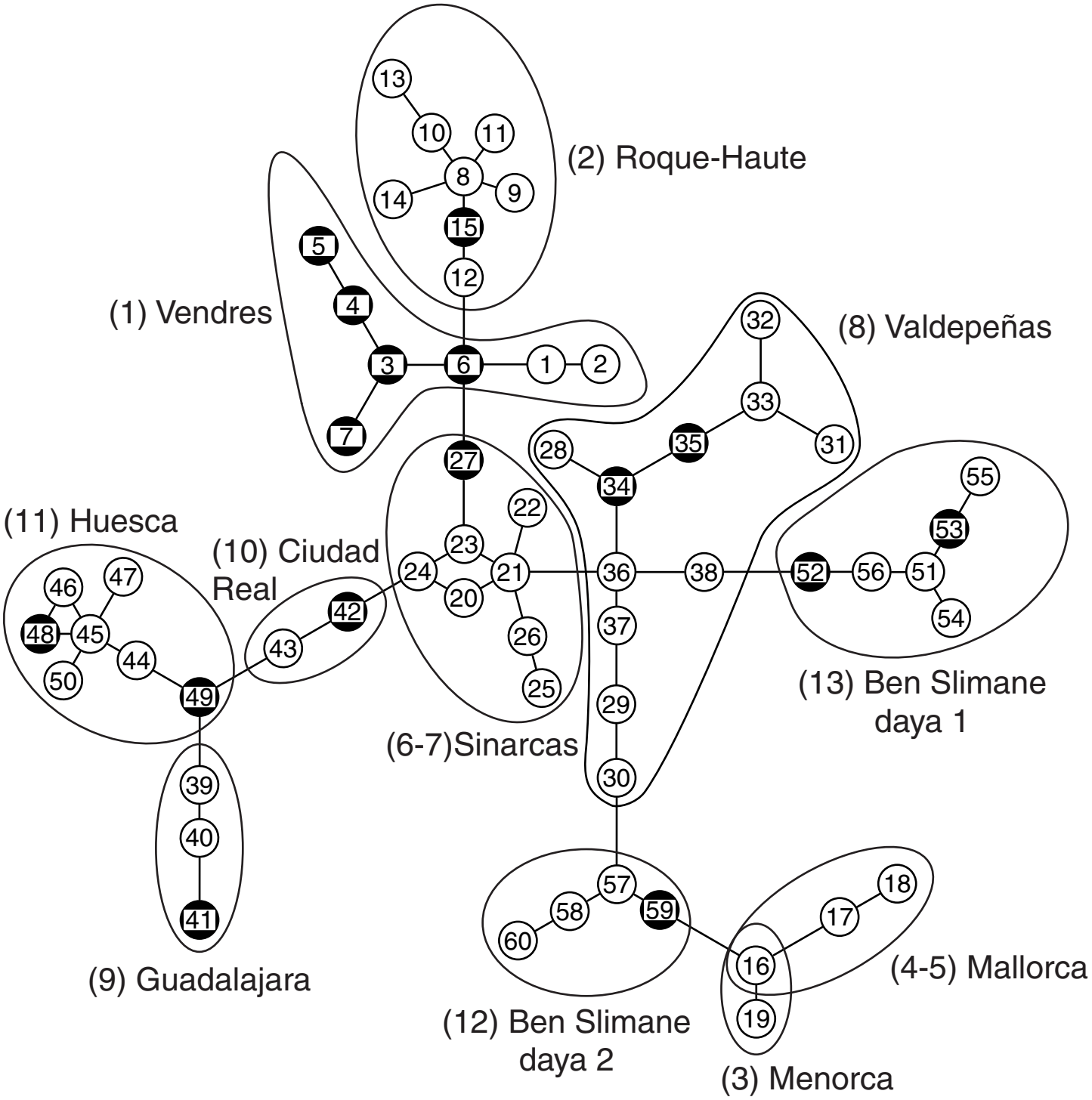
Vitalis *et al.* Figure 1



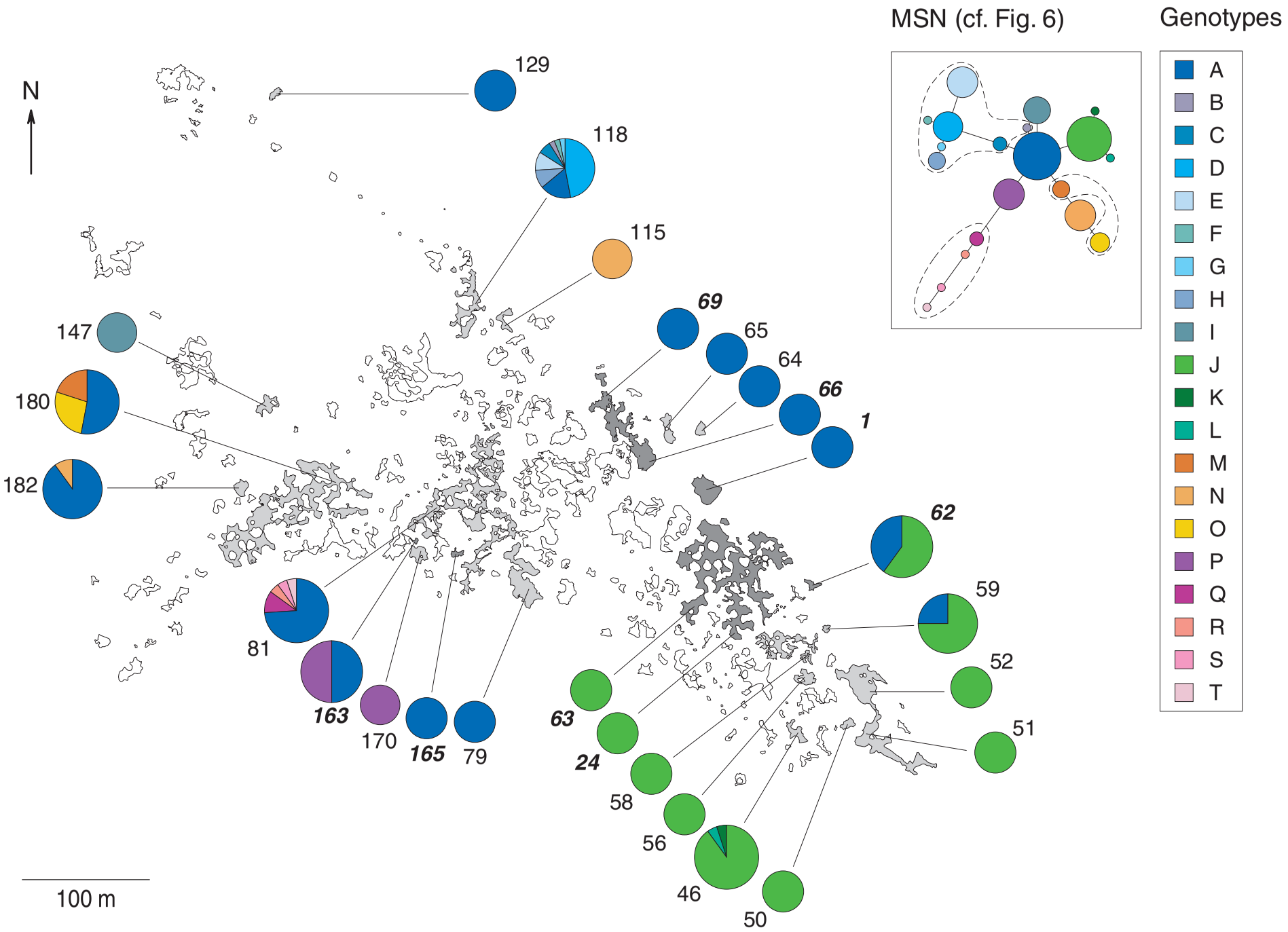
Vitalis *et al.* Figure 2



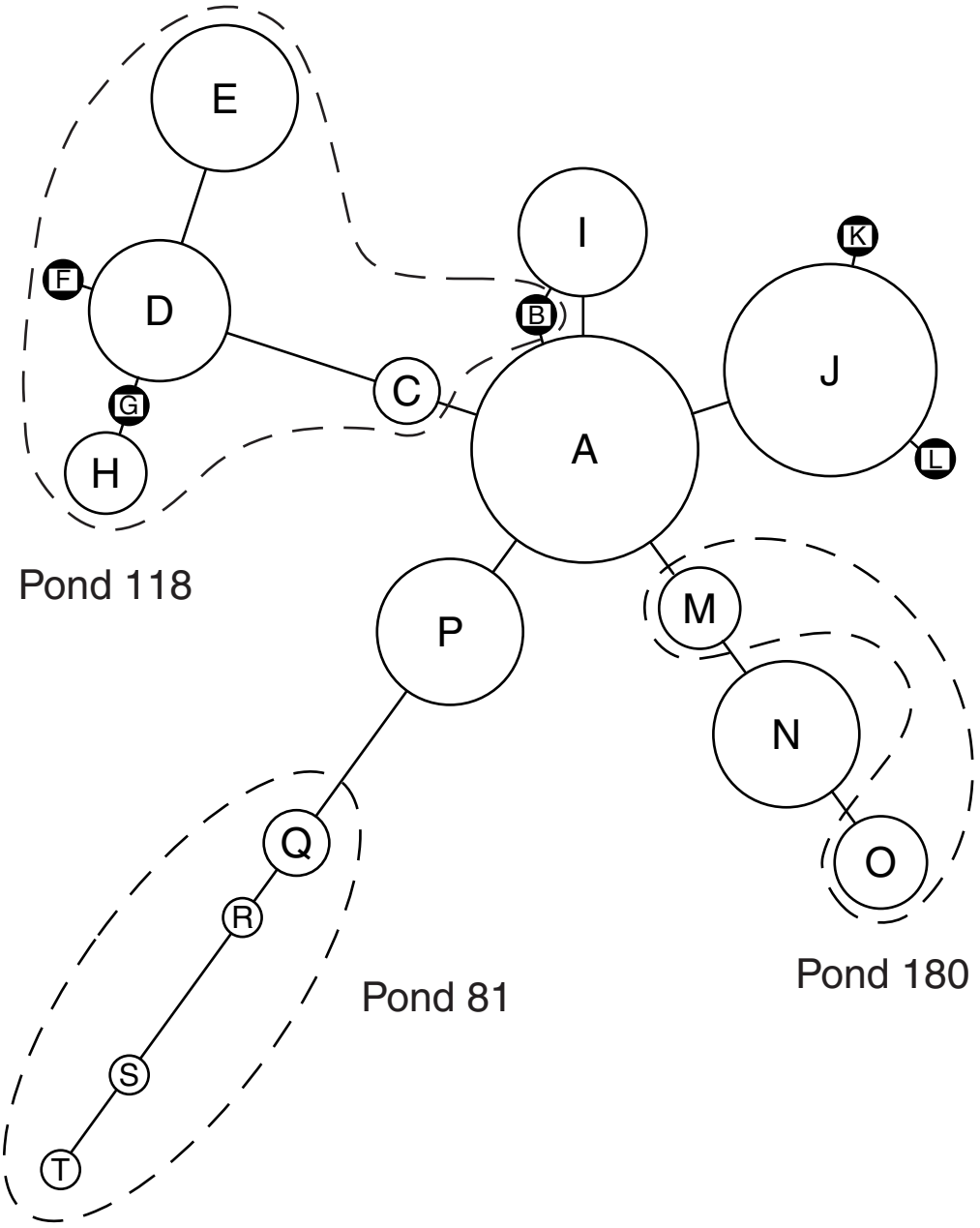
Vitalis *et al.* Figure 3



Vitalis *et al.* Figure 4



Vitalis *et al.* Figure 5



Vitalis *et al.* Figure 6

Annexe E

Le déséquilibre d'identité

Vitalis R. and Couvet D. (2001) Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genetical Research* **77** : 67-81.

Genet. Res., Camb. (2001), **77**, *pp.* 67–81. *With* 6 *figures.* Printed in the United Kingdom © 2001 Cambridge University Press

Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population

RENAUD VITALIS[1,2]* AND DENIS COUVET[3]

[1]*Laboratoire de Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France*
[2]*Station Biologique dé la Tour du Valat, Arles, France*
[3]*CRBPO – Museum National d'Histoire Naturelle, Paris, France*

(*Received 8 May 2000 and in revised form 28 August 2000*)

Summary

Measures of association of genes at different loci (linkage disequilibrium) are widely used to determine whether the structure of natural populations is clonal or not, to map genes from population data, or to test for the homogeneity of response of molecular markers to background selection, for example. However, the usual definitions of parameters for gametic associations may not be suitable for all these purposes. In this paper, we derive the recursion equations for one- and two-locus identity probabilities in an infinite island model. We study the role of drift, gene flow, partial selfing and mutation model on the expected association of genes across loci. We define the 'within-subpopulation identity disequilibrium' as the difference between the joint two-locus probability of identity in state and the expected product of one-locus identity probabilities. We evaluate this parameter as a function of recombination rate, effective size, gene flow and selfing rate. Within-subpopulation identity disequilibrium attains maximum values for intermediate immigration rates, whatever the selfing rate. Moreover, identity disequilibrium may be very small, even for high selfing rates. We discuss the implications of these findings for the analysis of data from natural populations.

1. Introduction

It is commonly admitted that the extent to which genes are associated across loci may be useful for the analysis of population structure or for the detection of selection pressures. Therefore, it is important to determine the amount of gametic association for neutral genes expected in a population. There is, however, no real consensus on the relevance of different measures of gametic associations.

(i) *The mean linkage disequilibrium*

The pairwise gametic disequilibrium parameter between two loci is defined as the excess of coupling gametes over that expected under random association. Let P_v^u be the frequency of gametes carrying allele u at the first locus and allele v at the second locus. Let P^u and P_v be the frequencies of alleles u and v at the first and the second locus, respectively. Then, a general expression for linkage disequilibrium is given by $D_v^u = P_v^u - P^u \cdot P_v$. This parameter may also be viewed as a covariance since it is expressed as the difference between the joint frequency and the product of single frequencies of genes at two loci.

Extending Geiringer's (1944) theory, Bennett (1954) demonstrated that, in an infinite population, the linkage disequilibrium among any set of neutral loci always tends to zero. However, the rate of approach to equilibrium may be very slow if genes are tightly linked. Hill & Robertson (1966) provided the equations for the change in linkage disequilibrium between a pair of loci in a finite random mating population. They showed that, if none of the genes has an effect on fitness, the expected linkage disequilibrium is zero. This is true whatever the number of loci (Hill, 1974*a*).

* Corresponding author. Tel: +33 4 67 14 32 50. Fax: +33 4 67 14 36 22. e-mail: vitalis@isem.univ-montp2.fr

In an infinite partially selfing population, there is no statistical association between allele frequencies at linked loci at mutation–drift equilibrium (Bennett & Binet, 1956). While the value of linkage disequilibrium is therefore expected to be zero, it has been shown that selfing slows its decay in a similar way to linkage (Weir & Cockerham, 1973). However, Hastings (1984) showed a complex interaction of linkage and partial selfing for maintaining high linkage disequilibrium in the presence of selection.

In subdivided populations without selection the mean linkage disequilibrium eventually vanishes (Nei & Li, 1973: Ohta, 1982*a*, *b*). In such situations, with restricted gene flow among groups, linkage disequilibrium can only be transient (Nei & Li, 1973; Slatkin, 1975). Thus, in the absence of selection whatever the number of loci considered, in isolated or subdivided populations, whatever the mating system and the population size, the mean linkage disequilibrium is zero at mutation–drift equilibrium.

(ii) *The variance of linkage disequilibrium*

While the mean linkage disequilibrium asymptotically converges to zero, it has been shown that there may be a large variance among segregating lines, due to finite population size (Hill & Robertson, 1968; Sved, 1968; Ohta & Kimura, 1969; Hill, 1974*b*; Weir & Hill, 1980). Since its expectation is zero, the variance of linkage disequilibrium equals the mean squared disequilibrium. Sved (1971) further studied the correlation of genotype frequencies at two linked loci by comparison of one approach based on disequilibrium parameters and a second approach based on identity-by-descent probabilities. From his results, he suggested that the measures of linkage disequilibrium in finite populations might be used to give information about effective population size (Hill, 1981).

Ohta (1982*a*, *b*) investigated the consequence of drift and gene flow on the variance of linkage disequilibrium in a finite subdivided population. She defined various statistics for linkage disequilibrium, to account for within- and between-subpopulation effects, by analogy with *F*-statistics. She also determined the variance of these statistics as a function of moments of gametes and genes frequencies in a symmetric two-allele model (Ohta, 1982*b*) and in an infinite-allele model (Ohta, 1982*a*). Finally, she proposed a test to discriminate between epistatic selection and limited gene flow as explanations for a large observed linkage disequilibrium by comparison of appropriate variance components (Ohta, 1982*a*). Tachida & Cockerham (1986) further developed the analysis and attempted to clarify the definition of linkage disequilibrium parameters. They determined the variance–covariance structure of some estimators of these parameters as functions of probabilities of identity by descent (Weir & Cockerham, 1969).

(iii) *Identity disequilibrium within individuals*

Bennett & Binet (1956) showed that, in an infinite partially selfing population, even though genes are associated at random across loci at equilibrium, there is a positive association between the genotypic states at different loci, even if they are not linked. There is an excess of double homozygotes for two independent loci compared with the product of single homozygotes at these two loci (Bennett & Binet, 1956). This has also been proved to exist in a finite population as a consequence of the variation of inbreeding among individuals (Weir & Cockerham, 1973). This quantity, referred to as 'identity disequilibrium', 'genotypic association' or even 'equilibrium constant', is a decreasing function of recombination rate in an infinite mixed selfing and random mating population (Weir & Cockerham, 1973). Furthermore, for any particular recombination rate, there is an amount of selfing which maximizes this parameter. The variation of inbreeding among individuals within a population (identity disequilibrium within individuals) has been described as a function divided in two parts, one being viewed as a correlation among united gametes and the other being the effect of linkage (Weir & Cockerham, 1969). More generally, in partially inbreeding (finite) populations there is an association between the homozygosity of different loci, even if they are not linked (Charlesworth, 1991).

It is worth noticing that this feature has been recognized as a possible explanation for the heterozygosity–fitness correlations through associative overdominance in finite, inbred populations (Strobeck, 1979; Charlesworth, 1991). More recently, David (1999) proposed an extension of this theory for studying the association between phenotypic variance and heterozygosity at marker loci. Correlations between heterozygosity and both mean and variance of phenotypes could indeed be parsimoniously explained by inbreeding.

Other measures of association have also been defined (Hill, 1975; Ohta, 1980; Takahata, 1982). Indeed, describing the two-locus population structure in terms of identity probabilities naturally leads to the definition of 'within-subpopulation identity disequilibrium', as the excess of two-locus identity probabilities over the product of single-locus probabilities. We evaluate this parameter in a structured, partially selfing population. We show that this measure has interesting properties that should be useful in the context of inferring demographic parameters. We discuss the implications of these results for the analysis of data from natural populations.

2. The model

We consider an infinite island model of population structure (Wright, 1931) with gametic migration. This model assumes that the population is subdivided into an infinitely large number of diploid monoecious subpopulations. Each generation, individuals reproduce in each subpopulation and some offspring are produced by selfing. All individuals contribute with equal probability to the next generation. Recombination occurs between loci at rate r. Thereafter, the subpopulations exchange migrant gametes at a rate m. The proportion of pairs of genes that come from a single subpopulation in the previous generation is thus $(1-m)^2$. Since we consider gametic migration, we shall define the selfing rate s as the conditional probability that two homologous sets of genes of one individual are derived from the same parent, given they are both copies of genes from one subpopulation before migration. The overall amount of selfing is thus $(1-m)^2 s$. Locally, in a completely random mating subpopulation of N hermaphroditic individuals, $s = 1/N$. Genes are sampled after dispersal. We consider two mutation models. The infinite-allele model (IAM) provides expected values of the probabilities of identity by descent (IBD). Alternatively, the symmetric K-allele model (KAM) gives the expectations for the probabilities of identity in state (IIS).

(i) *Definition of one-locus identity probabilities*

Let us define Q_0 as the probability that two homologous genes taken at one locus in a single individual are identical in state. Q_1 and Q_2 are the IIS probabilities of two genes at one locus, taken in distinct individuals, respectively in the same (Q_1) or different (Q_2) subpopulations. With local panmixia, $Q_0 = Q_1$. As we consider an infinite island model, the probability of coalescence for genes taken in distinct subpopulations is zero at equilibrium. Therefore, the equilibrium IIS probability for two genes taken in distinct subpopulations depends on the mutation model. In the IAM the IBD probability Q_2 is zero at equilibrium. In the KAM, the equilibrium IIS probability Q_2 depends on the number of allelic states: two genes can be IIS (and not IBD) with probability $= 1/K$, where K is the number of allelic states (Cockerham, 1984).

(ii) *Definition of two-locus identity probabilities with local random mating*

With local random mating, one needs to define three two-locus IIS probabilities (Whitlock *et al.*, 1993). In the following, a haplotype is a set of two genes sampled in one individual at two distinct loci which comes from a single gamete in the previous generation. The two loci need not be physically linked on a single chromosome. All two-locus probabilities are defined as IIS probabilities. However, the IAM provides expectations for IBD probabilities.

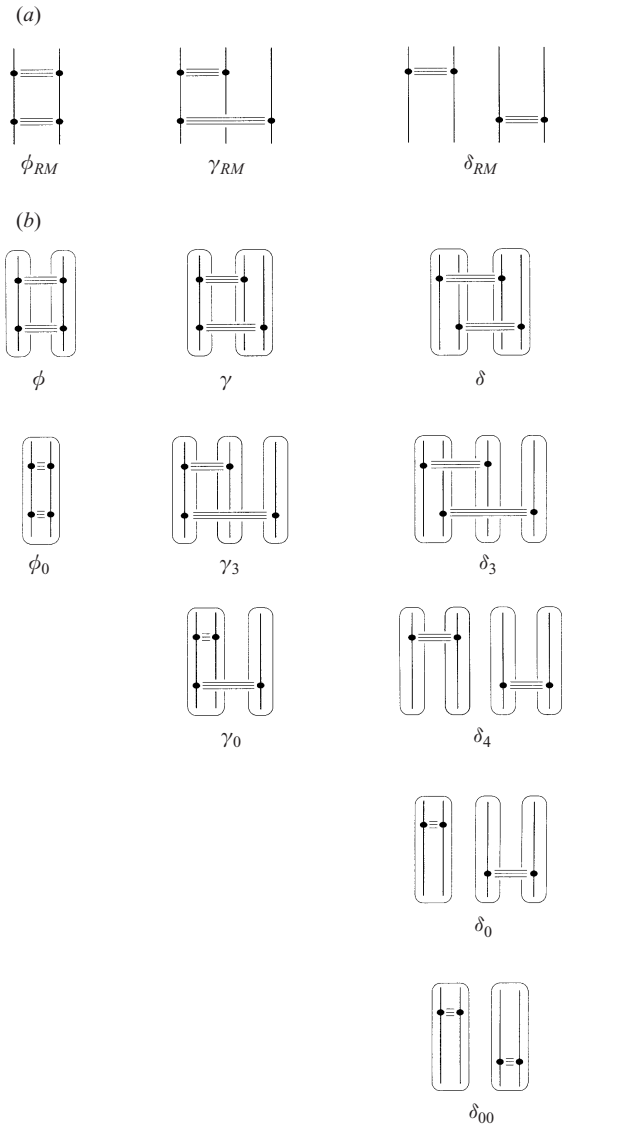


Fig. 1. Definition of two-locus probabilities for the probability of identity in state (IIS). (*a*) Random mating. Vertical lines represent sampled haplotypes, on which upper and lower positions of filled circles represent two loci. The symbol $\equiv$ among pairs of homologous genes stands for identity in state. In the infinite allele model, these coefficients define the corresponding probabilities for the identity by descent (IBD). (*b*) Non-random mating. The IIS probabilities defined above may have different values according to the number of individuals from which the haplotypes are sampled. Other probabilities are thus defined, with each diploid individual represented as a box. Only the sampled haplotypes are shown. See the text for details.

ϕ_{RM} is the probability that two randomly sampled haplotypes are IIS at both loci. γ_{RM} is the probability that, among three sampled haplotypes, one pair is IIS at one locus and a distinct pair (with a single common haplotype) is IIS at the second locus. δ_{RM} is the

probability that, when four haplotypes are sampled, two are IIS at one locus and the other two are IIS at a second locus. These identity probability parameters are schematically defined in Fig. 1*a*.

As in the one-locus case, we can derive the recursion equations for two-locus IIS probabilities. Let us take the recursion for ϕ_{RM}, for example. With probability $(1-m)^2(1-1/N)$, two randomly sampled haplotypes come from different parents within the subpopulation in the previous generation. Now, these two sampled haplotypes may be two recombinant gametes, one recombinant and one parental gamete, or two parental gametes. In that case, the IIS probability involves respectively four, three or two haplotypes in the previous generation. Therefore, they are IIS with probability $[(1-r)^2\phi_{RM}+2r(1-r)\gamma_{RM}+r^2\delta_{RM}]$. With probability $(1-m)^2/N$, the two sampled haplotypes come from a single parent in the same subpopulation in the previous generation. With probability $[(1-r)^2+r^2]$ the haplotypes are both recombinant or non-recombinant gametes. The gametes may be copies of the same or of distinct haplotypes, and therefore are IIS with probability $(\phi_{RM}+1)/2$. With probability $[2r(1-r)]$ one haplotype is recombinant, and the other is non-recombinant. They are copies of distinct gametes and are IIS with probability Q_{RM} (with Q_{RM} defined as the one-locus IIS probability for a pair of genes taken at random within one subpopulation). Although rather more complicated, since they involve triplets and quadruplets of haplotypes, recursion equations for γ_{RM} and δ_{RM} can be derived similarly.

(iii) *Definition of two-locus identity probabilities with selfing*

With non-random mating, eg. partial selfing or dioecy, more IIS probabilities need to be defined, since IIS probabilities defined for pairs, triplets or quadruplets of haplotypes (respectively ϕ_{RM}, γ_{RM} and δ_{RM} in the random mating case) may now have different values whether these haplotypes are found in two, three or four individuals (Weir *et al*., 1980; Weir & Cockerham, 1969; Weir & Hill, 1980).

Now, ϕ is the probability that two haplotypes randomly sampled in two individuals are IIS at both loci. ϕ_0 is the corresponding probability when the two haplotypes are taken within a single individual.

γ is the probability that, when a first haplotype is chosen from one individual and two others are chosen from a distinct individual, the first haplotype is IIS to the second one at the first locus and IIS to the third haplotype at the second locus. γ_3 is the probability that, among three sampled haplotypes, each taken in a distinct individual, one pair is IIS at one locus, and a distinct pair (with a single common haplotype) is IIS at the second locus. γ_0 is the probability that, when two individuals are sampled, one is homozygous at the first locus and both carry IIS genes at the second locus.

δ is the probability that, when two pairs of haplotypes are taken among two individuals, one pair is IIS at the first locus and the other pair is IIS at the second locus. δ_3 is the probability that, when four haplotypes are taken among three individuals, one pair of haplotypes taken among a first pair of individuals is IIS at the first locus and another pair of haplotypes, taken among a second pair of individuals, is IIS at the second locus. δ_4 is the probability that, when four haplotypes are sampled among four distinct individuals, two are IIS at one locus, and the other two are IIS at a second locus. δ_0 is the probability that, when three distinct individuals are sampled, one individual is homozygous at the first locus, and two sampled haplotypes among the other two individuals are IIS at the second locus. δ_{00} is the probability that, for two randomly sampled individuals, one is homozygous at the first locus and the other is homozygous at the second locus. All these identity probability parameters are schematically defined in Fig. 1*b*.

Overall, we have defined 10 two-locus identity probabilities. Along with one-locus identity probabilities Q_0 and Q_1 for each locus, all these 14 parameters are necessary and sufficient to describe the two-locus genetic structure of a single, partially selfing population. As in the one-locus case, the probability of coalescence for two-locus pairs of genes in distinct subpopulations is zero at equilibrium, in an infinite island model. Therefore the equilibrium IIS probabilities for two-locus pairs of genes taken in distinct subpopulations depend also on the mutation model. In the IAM, these probabilities are zero (IBD probabilities). In the KAM, IIS two-locus probabilities are simple products of one-locus probabilities, i.e. products of $1/K$'s over pairs of loci.

We derived the recursion equations for each of these 14 parameters, in the IAM and the KAM. Some details are given in the Appendix. With panmixia, $\phi = \phi_0$, $\gamma = \gamma_3 = \gamma_0$ and $\delta = \delta_3 = \delta_4 = \delta_0 = \delta_{00}$, and therefore the system reduces to the three two-locus IIS probabilities defined for the random mating case above.

It may not be possible to know from genotypic data, when pairs of genes are taken at two loci, whether the genes within each individual belong to the same haplotype ('coupling genes', or 'genes in phase') or not ('genes in repulsion'). Therefore, we define the composite identity probability Φ as the probability that a pair of genes taken at two distinct loci within an individual is IIS to a pair of genes taken at homologous loci in a second, distinct, individual:

$$\Phi = \frac{\phi+2\gamma+\delta}{4}. \quad (1)$$

(iv) *Definition of identity disequilibrium parameter*

Ohta (1980) defined a measure of association among amino acid sites as the excess probability of simultaneous identity over that expected from random combination of the identity at the two loci, a quantity equivalent to the covariance of non-identity (heterozygosity) at two loci within populations (Avery & Hill, 1979; Hedrick, 1987).

We define an analogous parameter, hereafter referred to as 'within-subpopulation identity disequilibrium', η_S, as the difference between the joint two-locus identity probability among two randomly chosen individuals in a population, and the expected product of one-locus identity probabilities:

$$\eta_{S,ij} = \Phi_{ij} - \delta_{4ij}. \quad (2)$$

There are some advantages to considering this parameter. First, its definition is a straightforward function of identity probabilities. Second, as further developed in a companion paper (Vitalis & Couvet, in press), statistics can easily be derived to estimate this quantity.

3. Results

(i) *The within-subpopulation identity disequilibrium as a measure of association*

Although the within-subpopulation identity disequilibrium may be viewed as the covariance for the probability of identity across a pair of loci, it is better defined as a function of probabilities of identity. This parameter depends on some demographic parameters of the population model (effective size, migration rate), but also on the mutation model considered. Indeed, it is quite sensitive to the mutation rate, and especially to the number of allelic states that can arise at each locus. A 'standardized' parameter can be defined as

$$\eta'_{S,ij} = \frac{\Phi_{ij} - \delta_{4ij}}{(1-Q_{2i})(1-Q_{2j})}. \quad (3)$$

Similar standardized parameters have already been defined (Ohta, 1980; Takahata, 1982; Hedrick, 1987).

Fig. 2 shows η_S and η'_S for various models of mutation (IAM and KAM). For large Nr, both η_S and η'_S tend to zero, indicating that two-locus identity probabilities converge to the expected product of one-locus identities. But the within-subpopulation identity disequilibrium remains significantly greater than zero for increasing values of Nr, as the mutation model approaches the IAM. Indeed, the within-subpopulation identity disequilibrium attains its maximum value when the potential number of allelic states at both loci is infinite. More generally, the within-subpopulation identity disequilibrium increases as the number of allelic states becomes large, especially when linkage is tight. Conversely, the looser the linkage, the smaller the discrepancy between IAM and KAM. In other words, drift creates associations more easily when the number of allelic states is high.

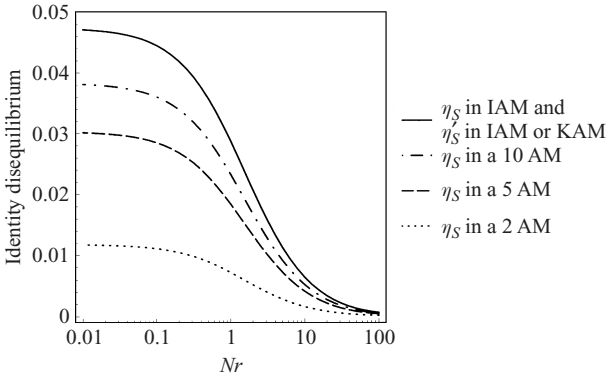


Fig. 2. Expected within-subpopulation identity disequilibrium as a function of Nr for various mutation models, with $\mu = 10^{-6}$. The other population parameters are $N = 200$ and $Nm = 1$. Random mating is assumed ($Ns = 1$). Note the logarithmic scale on the abscissa.

Fig. 2 further shows that η'_S does not depend on the mutation model, whatever the range of other population parameters in the model (proportion of selfing, recombination rate, effective population size, migration rate). Ohta (1980) and Takahata (1982) found an analogous result for a single, random mating population. Therefore the standardized within-subpopulation identity disequilibrium is closely related to linkage and nearly independent of the properties of the mutation model.

(ii) *The effect of selfing*

Golding & Strobrek (1980) derived the expression for the squared linkage disequilibrium in a single, finite, partially selfing population. They found that, for $N\mu < 1$, squared linkage disequilibrium may increase with selfing if $Nr > 1$, or decrease if $Nr < 1$. For $N\mu < 1$, squared linkage disequilibrium increases with selfing, whatever the range of Nr. Considering the effect of migration, we found in turn that the same holds as long as $Nm < N\mu < 1$. Only when $Nm > N\mu$ does the squared linkage disequilibrium increase with the selfing rate, whatever the range of Nr.

Fig. 3 depicts the expected value of within-subpopulation identity disequilibrium as a function of Nr, for a range of s values. It is shown that within-subpopulation identity disequilibrium increases with decreasing values of Nr. Moreover, increasing the proportion of selfing always increases the expected within-subpopulation identity disequilibrium. Contrary to a single population model (Golding &

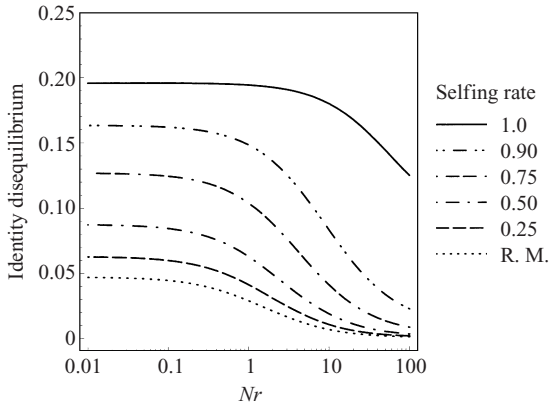


Fig. 3. Expected within-subpopulation identity disequilibrium as a function of Nr for various selfing rates and random mating (R. M.). The other population parameters are $N = 200$ and $Nm = 1$. The mutation model is an IAM with $\mu = 10^{-6}$. Note the logarithmic scale on the abscissa.

Strobeck, 1980), the measure of association among loci depends on the recombination rate when $s = 1$. One would expect that, for completely selfing organisms, the equilibrium value of a measure of association of genes among loci is not influenced by the proportion of recombination. But as noted before, the overall amount of selfing is $(1-m)^2 s$. Consequently, when $m \neq 0$, the effective rate of selfing is less than one. This is the reason why, even for $s = 1$, the within-subpopulation identity disequilibrium slightly decreases with increasingly larger values of Nr.

(iii) *The effect of migration*

Within-subpopulation identity disequilibrium is a unimodal function of the migration rate (Fig. 4). Within-subpopulation identity disequilibrium attains a maximum value for intermediate migration rates. For small migration rates, since migration introduces new gametes in fairly monomorphic subpopulations, it inflates the disequilibrium among two-locus IIS probabilities. However, as the migration rate increases and converges towards unity, the whole population approaches panmixia and the within-subpopulation identity disequilibrium tends to a small value for a large (infinite) population. Similar results were obtained for the variance of the gametic component of linkage disequilibrium in an island model (Tachida & Cockerham, 1986). While the mode of the curve depends on Nm rather than m, the maximum value of within-subpopulation identity disequilibrium decreases with N. Furthermore, the effect of migration is more pronounced for tightly linked loci.

Interestingly, this leads to the observation that, for a given value of Nm, the within-subpopulation identity disequilibrium could be used to infer the effective population size and migration rate. This point will be further developed in a companion paper (Vitalis & Couvet, in press).

(iv) *Joint effect of selfing and migration*

Fig. 5 shows the expected within-subpopulation identity disequilibrium as a function of Nr and Nm in random mating and complete selfing situations. Within-subpopulation identity disequilibrium increases with selfing, whatever the range of Nr or Nm. There are values of Nm for which within-subpopulation identity disequilibrium is always expected to be very low. One can deduce that when $Nm < 10^{-2}$ or $Nm > 10^2$, no within-subpopulation identity disequilibrium shall be found. With random mating, significant within-subpopulation identity disequilibrium is expected when Nr is smaller than one. With

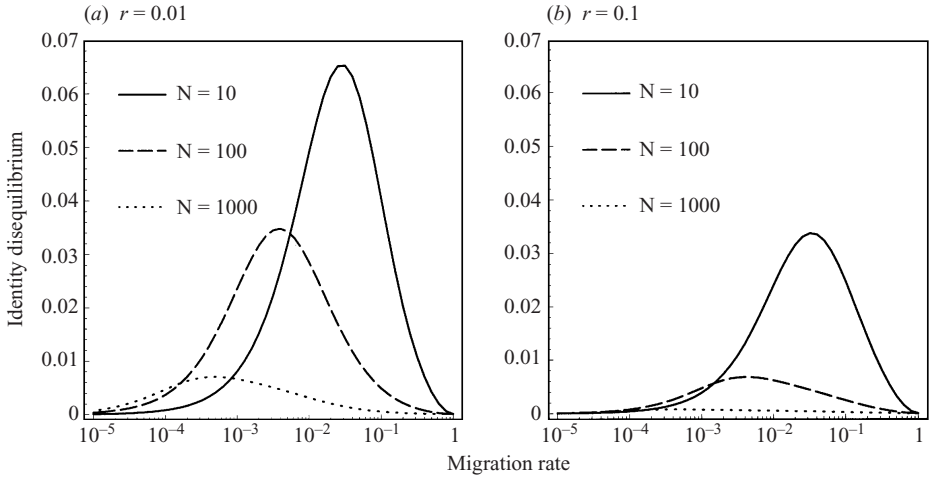


Fig. 4. Expected within-subpopulation identity disequilibrium as a function of migration rate, for various population sizes ($N = 10$, 100 and 1000) and recombination rates: (*a*) $r = 0{\cdot}01$ and (*b*) $r = 0{\cdot}1$. Random mating is assumed. The mutation model is an IAM, with $\mu = 10^{-6}$. Note the logarithmic scale on the abscissa.

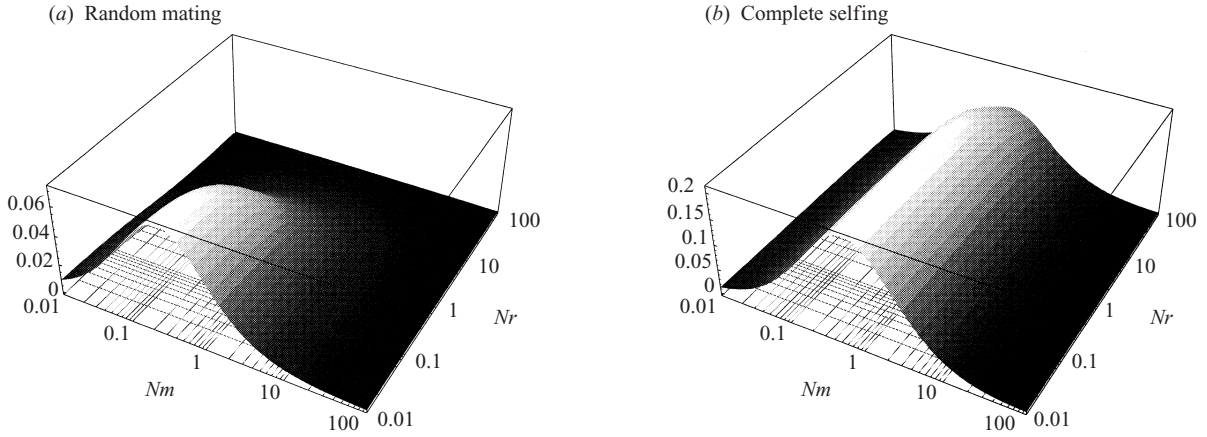


Fig. 5. Expected within-subpopulation identity disequilibrium as a function of Nm and Nr in two contrasted situations: (*a*) random mating ($Ns = 1$) and (*b*) complete selfing ($s = 1$). The population size is $N = 200$. The mutation model is an IAM, with $\mu = 10^{-6}$. Note the logarithmic scale on the x- and y-axes.

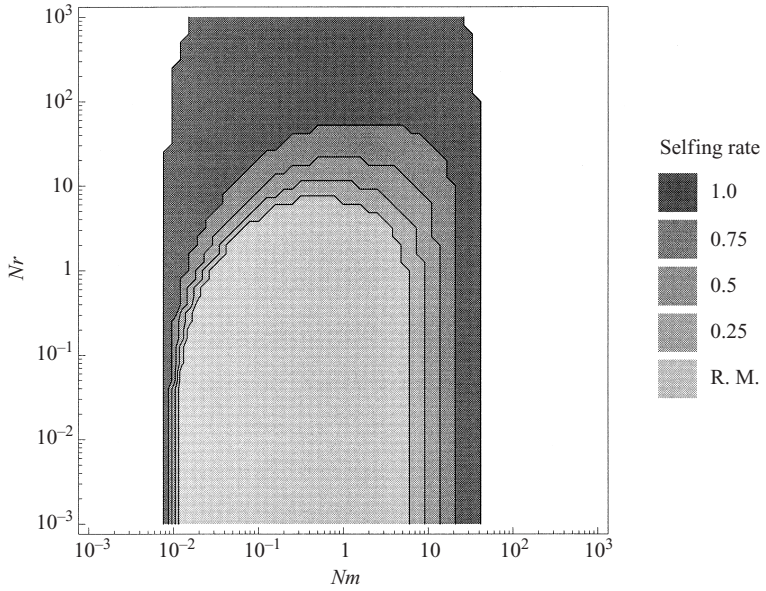


Fig. 6. Diagrammatic representation of the range of Nr and Nm necessary and sufficient to maintain significant ($> 0{\cdot}01$) within-subpopulation identity disequilibrium, for various selfing rates and random mating (R.M.). Each (continuous) region draws the parameter space in which the within-subpopulation identity disequilibrium is greater than 1%. Darker areas always include lighter ones. Note the logarithmic scale on the x- and y-axes.

complete selfing, within-subpopulation identity disequilibrium is asymptotically always maintained for $10^{-2} < Nm < 10^2$. Consequently, within the appropriate range of Nm, within-subpopulation identity disequilibrium is maintained when $Nr(1-s) < 1$, as for squared linkage disequilibrium in the single population case (Golding & Strobeck, 1980).

The effect of increased selfing is summarized in Fig. 6. Grey areas draw the range of Nr and Nm values for which $\eta'_S > 0{\cdot}01$. Darker areas always include lighter ones. Higher selfing rates maintain identity disequilibrium in a larger range of conditions. In the special case of complete selfing, with $10^{-2} < Nm < 10^2$, η'_S is asymptotically always greater than 1%.

4. Discussion

The two-locus genetic structure of a partially selfing metapopulation has been characterized by means of appropriate definitions of probabilities of identity of genes, for one and two loci (Fig. 1). A simple measure of association of genes across loci has been defined, as the difference between the joint probability of identity at two loci and the expected product of one-locus probabilities at these loci. This quantity has been referred to as the 'within-subpopulation identity disequilibrium' and is equivalent to the 'identity excess' of Ohta (1980) and to the covariance of non-identity defined by Avery & Hill (1979) (see also

Hedrick, 1987). This parameter, denoted η_S, could also be viewed as the covariance for the probability of identity of genes, across pairs of loci. Finally, we derived the expected value of this parameter in an infinite island model of population structure, with some emphasis on the consequence of mutation model, reproductive system, dispersal, population size and recombination.

(i) *The effect of the mutation model*

Within-subpopulation identity disequilibrium depends on the mutation model (Fig. 2). However, for low mutation rates, the 'standardized' definition of within-subpopulation identity disequilibrium, η'_S, has been shown to be independent of the mutation model (IAM vs KAM) and of the number of allelic states in the KAM (Fig. 2). A similar property has already been discussed for *F*-statistics by Crow & Aoki (1984) and Rousset (1996). This result is particularly important in the perspective of comparing estimates across independent pairs of loci. Although the question of estimation will be addressed elsewhere (Vitalis & Couvet, in press), this result also suggests that, since they have the same expectation, estimates could be pooled over pairs of loci.

Analogous results were obtained for the analysis of other measures of gametic associations. For example, since the squared linkage disequilibrium has been shown to depend on allele frequencies, some authors defined another measure of disequilibrium as the ratio of squared linkage disequilibrium over the product of variance in gene frequencies at the two loci. This quantity has been referred to as 'correlation' (Hill & Robertson, 1968) or 'standardized linkage disequilibrium' (Ohta, 1982*a*). Various definitions have been proposed, depending on whether ratios of expectations or expectations of ratios were taken (Ohta, 1989; Takahata, 1982; Hedrick, 1987). However, the difference between the parameters and the statistics that estimate those parameters has not always been clearly defined. This led to some controversy on the dependence of some measures of association on the underlying gene frequencies (see Hedrick, 1987; Lewontin, 1988). Here, our purpose was to define a parameter which, in a given model of population structure, does not depend strongly on any 'nuisance' parameter such as the mutation rate, or the number of (possible) allelic states.

(ii) *The conditions for the maintenance of identity disequilibrium*

Identity disequilibrium was found to be a unimodal function of migration rate. (Figs. 4, 5). All else being equal, there is a value of Nm which maximizes the identity disequilibrium. However, the value of identity disequilibrium increases with smaller effective sizes. Consequently, it is suggested that both drift within subpopulations and gene flow among subpopulations shape the distribution of identity disequilibria. Taking advantage of this result, we propose in a companion paper (Vitalis & Couvet, in press) a method-of-moments framework to infer both effective population size and migration rates from one- and two-locus identity functions.

Selfing always increases identity disequilibrium (Fig. 3). Furthermore, we showed that, within the appropriate range of Nm, identity disequilibrium is maintained when $Nr(1-s) < 1$ (Figs. 5, 6). Golding & Strobeck (1980) obtained the same result for the squared linkage disequilibrium in a partially selfing finite population.

(iii) *Implications for data analysis*

There is still much debate in the literature on whether selfing maintains linkage disequilibrium or not. However, linkage disequilibrium can only be transient, and depends on other ecological parameters, such as the dispersal pattern. Indeed, a recent AFLP study in the selfing species *Arabidopsis thaliana* revealed a low proportion of significant pairwise linkage disequilibrium among markers (Miyashita *et al*., 1999).

With the advance in molecular techniques, assessing the genetic structure of microbial populations has also received growing interest (Lenski, 1993). It has been argued that many micro-organisms may have a clonal population structure (Tibayrenc *et al*., 1990). Linkage disequilibrium, within or among classes of molecular markers, is taken as evidence of clonal, rather than sexual, reproduction. For the agent of human malaria, *Plasmodium falciparum*, Tibayrenc *et al*. (1991) reported contradictory results. Indeed, linkage disequilibrium analyses failed to prove whether the genetic population structure deviated from panmixia (Tibayrenc *et al*., 1991). The interpretation of linkage disequilibrium data may depend on how one analyses the data and on the sampling scale (Lenski, 1993; Maynard Smith *et al*., 1993). Improvement of standard methods to estimate inbreeding coefficients revealed significant departure from Hardy–Weinberg equilibrium among zygotes of the parasite in Tanzania (Hill *et al*., 1995). In Papua New Guinea, high levels of heterozygote deficits (as compared with Hardy–Weinberg proportions) were also found among *P. falciparum* zygotes, in addition with the lack of any linkage disequilibrium across loci (Paul *et al*., 1995). Tibayrenc & Lal (1996) found these two results to be incompatible. They argued that low levels of outcrossing could not prevent non-random gametic

associations. However, as we have shown in this paper, high inbreeding and gametic equilibrium may jointly occur, since the extent of identity disequilibrium depends also on gene flow. Finally, within a highly polymorphic region of chromosome 5 in *P. falciparum*, linkage disequilibrium declines with increasing distance map (Conway *et al*., 1999). Furthermore, Anderson *et al*. (2000) recently suggested that highly significant deficits of heterozygous parasite oocysts within mosquitoes could be explained by the presence of null alleles and the subsequent mis-scoring of genotypes. Therefore, there may be enough outcrossing to prevent gametic disequilibrium across unlinked markers. Genetic studies of *P. falciparum* natural populations should be greatly improved with the development of genome-wide linkage maps (Su *et al*., 1999).

However, gene mapping from population-based studies of linkage disequilibrium may be very difficult since the relationship between the recombination fraction and the physical distance is obscured by other population factors, such as genetic drift, reproductive system or population structure (Hill & Weir, 1994; Devlin & Rish, 1995). Instead, if the true map is established from genetic crosses, measures of within-subpopulation identity disequilibria conditioned on recombination rates may be used to infer population parameters of interest, such as the effective size or the immigration rate (Vitalis & Couvet, in press).

The analytical theory we presented here predicts the expected value of η'_S under the assumptions of our model. That expectation is unconditional on any locus being polymorphic. Yet Slatkin (1994) found, through simulations of stationary populations, that there may be a large number of statistically significant linkage disequilibria (Fisher's exact test performed only on pairs of polymorphic loci) even though the population parameters are such that the expected (unconditional) squared linkage disequilibrium is low. Therefore, our results may underestimate the within-subpopulation identity disequilibrium which is likely to be estimated in natural situations from only those loci that are polymorphic. This question deserves further attention.

Finally, we should indicate that other identity disequilibria could be considered. Indeed, within-individual identity disequilibrium can be defined as the difference between the joint homozygosity at two loci and the product of single probabilities for each locus. The analysis of the expected joint distribution in a neutral model of these different disequilibrium parameters could be used, for example, to build neutrality tests for molecular markers.

We thank K. Dawson, B. Godelle and I. Olivieri for helpful comments on a previous draft of this manuscript. We thank W. G. Hill and two anonymous referees for constructive comments. R.V. acknowledges financial support from the Training and Mobility of Researchers programme 'FRAGLAND' of the European Commission, coordinated by I. Hanski, and from the Fondation Sansouire. This is contribution number 2000-079 of the Institut des Sciences de l'Évolution de Montpellier.

Appendix

(i) *Expected values of one- and two-locus identity in the IAM*

Recursion equations for the probabilities of identity by descent (IBD) can generally be written in the form

$$\mathbf{Q}^{t+1} = \mathbf{A}\mathbf{Q}^{t} + \mathbf{B},$$

where $\mathbf{Q}$ is a vector of identity probabilities, $\mathbf{A}$ is a transition matrix, and $\mathbf{B}$ is a vector of coalescent terms. Since two-locus identity probabilities are partial functions of one-locus identity probabilities, we will consider $\mathbf{Q}$, $\mathbf{A}$ and $\mathbf{B}$ as partitioned vectors and matrices. Let

$$\mathbf{Q} = \left(\frac{\mathbf{Q_1}}{\mathbf{Q_2}}\right),$$

with $\mathbf{Q_1} = (Q_{0i}, Q_{0j}, Q_{1i}, Q_{1j})^{\mathrm{T}}$ being the sub-vector of one-locus identity probabilities, for loci i and j, and $\mathbf{Q_2} = (\phi_0, \phi, \gamma, \gamma_3, \gamma_0, \delta, \delta_3, \delta_4, \delta_0, \delta_{00})^{\mathrm{T}}$ being the sub-vector of two-locus identity probabilities, among loci i and j, as defined in the main text. Accordingly, the transition matrix $\mathbf{A}$ is composed of sub-matrices $\mathbf{A_1}$, $\mathbf{A_2}$ and $\mathbf{A_3}$

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{A_1} & \mathbf{0} \\ \hline \mathbf{A_2} & \mathbf{A_3} \end{array}\right)$$

and so follows the vector $\mathbf{B}$

$$\mathbf{B} = \left(\frac{\mathbf{B_1}}{\mathbf{B_2}}\right).$$

Considering the recursion equations for one-locus identity probabilities, the mean change over one generation is given by

$$\mathbf{A_1} = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D_1 M_1} \quad \text{and} \quad \mathbf{B_1} = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D_1 C_1}$$

where $\nu_i = (1-\mu_i)^2$ is the probability that, at locus i, neither gene in a pair has mutated; $\mathbf{D_1} = (1-m)^2\,\mathbf{I}$ is a 4×4 matrix containing the probabilities that both genes in a random pair are resident ($\mathbf{I}$ is the identity matrix); $\mathbf{C_1} = (\frac{s}{2}, \frac{s}{2}, \frac{1}{2N}, \frac{1}{2N})^{\mathbf{T}}$ is a vector containing the conditional probabilities that they have coalesced in the previous generation given that they are resident; $\mathbf{M_1}$ is a 4×4 matrix containing the conditional probabilities that two genes in a pair were alike in the previous generation given that they are resident:

$$\mathbf{M_1} = \begin{Bmatrix} \frac{s}{2} & 0 & 1-s & 0 \\ 0 & \frac{s}{2} & 0 & 1-s \\ \frac{1}{2N} & 0 & 1-\frac{1}{N} & 0 \\ 0 & \frac{1}{2N} & 0 & 1-\frac{1}{N} \end{Bmatrix}.$$

The recursion equations for the two-locus identity probabilities as functions of one- and two-locus probabilities are given by

$\mathbf{A_2} = \nu_i \nu_j \mathbf{D_2 M_2}$

$\mathbf{A_3} = \nu_i \nu_j \mathbf{D_2 M_3}$,

with the probabilities that joint pairs of genes have coalesced in the previous generation given by

$\mathbf{B_2} = \nu_i \nu_j \mathbf{D_2 C_2}$,

where $\mathbf{D_2}$ is a 10×10 diagonal matrix containing the probabilities that both pairs of genes were resident in the previous generation:

$\mathbf{D_2} = \mathrm{diag}[(1-m)^2, (1-m)^2, (1-m)^3, (1-m)^3, (1-m)^4, \ldots, (1-m)^4]$.

$\mathbf{M_2}$ is a 4×10 matrix for the conditional one-locus probabilities that, in one of the two pairs, genes were alike (the dots in matrix $\mathbf{M_2}$ signify that the first and second columns are identical, as are the third and fourth columns); $\mathbf{C_2}$ is a vector containing the conditional probabilities that joint pairs of genes have coalesced in the previous generation, given that they were in the same population; and $\mathbf{M_3}$ is a 10×10 matrix for the conditional two-locus probabilities that genes in both pairs were alike in the previous generation, given that they have not migrated:

$$\mathbf{M_2} = \begin{Bmatrix} sr(1-r) & \cdots & 0 & \cdots \\ \frac{r(1-r)}{N} & \cdots & 0 & \cdots \\ \frac{s}{4N} & \cdots & \frac{1-s}{2N} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{s}{4N} & \cdots & \frac{\alpha}{4} & \cdots \\ \frac{(1-s)^2}{4N(N-1)} + \frac{s^2}{4N} & \cdots & \frac{s(1-s)}{N} + (1-s)^2\frac{N-2}{2N(N-1)} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{1}{4N^2} & \cdots & \frac{N-1}{2N^2} & \cdots \\ \frac{s}{4N} & \cdots & \frac{\alpha}{4} & \cdots \\ \frac{s^2}{4} & \cdots & \frac{s(1-s)}{2} & \cdots \end{Bmatrix}, \quad \mathbf{C_2} = \begin{Bmatrix} \frac{s}{2}[(1-r)^2+r^2] \\ \frac{(1-r)^2+r^2}{2N} \\ \frac{s}{4N} \\ \frac{1}{4N^2} \\ \frac{s}{4N} \\ \frac{(1-s)^2}{4N(N-1)} + \frac{s^2}{4N} \\ \frac{1}{4N^2} \\ \frac{1}{4N^2} \\ \frac{s}{4N} \\ \frac{s^2}{4} \end{Bmatrix}$$

and

$$\mathbf{M_3} = \left\{ \begin{array}{ccccc}
\frac{s}{2}[(1-r)^2+r^2] & (1-s)(1-r)^2 & 2r(1-r)(1-s) & 0 & 0 \\
\frac{(1-r)^2+r^2}{2N} & (1-r)^2\frac{N-1}{N} & 2r(1-r)\frac{N-1}{N} & 0 & 0 \\
\frac{s}{4N} & \frac{s}{2}(1-r)\left(\frac{N-1}{N}\right) & \frac{s}{2}\left(\frac{N-1}{N}\right) & (1-s)(1-r)\frac{N-2}{N} & \frac{1-s}{N} \\
\frac{1}{4N^2} & (1-r)\frac{N-1}{2N^2} & \frac{N-1}{2N^2} & (1-r)\frac{(N-1)(N-2)}{N^2} & \frac{N-1}{N^2} \\
\frac{s}{4N} & (1-r)\frac{1-s}{2N} & \frac{1-s}{2N} & (1-r)\frac{(1-s)(N-2)}{N} & \frac{\alpha}{2} \\
\frac{s^2}{4N} & \frac{\beta}{4(N-1)} & \frac{\beta}{2(N-1)} & (1-s)\frac{N-2}{N-1}\alpha & \frac{2s(1-s)}{N} \\
\frac{s}{4N^2} & \frac{\alpha}{4N} & \frac{\alpha}{2N} & \frac{N-2}{2N^2}[N\alpha+2(1-s)] & \frac{\alpha}{N} \\
\frac{1}{4N^3} & \frac{N-1}{2N^3} & \frac{N-1}{N^3} & \frac{2(N-1)(N-2)}{N^3} & \frac{2(N-1)}{N^3} \\
\frac{s}{4N^2} & \frac{(1-s)}{2N^2} & \frac{(1-s)}{N^2} & (1-s)\frac{2(N-2)}{N^2} & \frac{\alpha}{N} \\
\frac{s^2}{4N} & \frac{(1-s)^2}{2N(N-1)} & \frac{(1-s)^2}{N(N-1)} & (1-s)^2\frac{2(N-2)}{N(N-1)} & \frac{2s(1-s)}{N}
\end{array} \right.$$

$$\left. \begin{array}{ccccc}
(1-s)r^2 & 0 & 0 & 0 & 0 \\
r^2\frac{N-1}{N} & 0 & 0 & 0 & 0 \\
\frac{s}{2}r\frac{N-1}{N} & (1-s)r\frac{N-2}{N} & 0 & 0 & 0 \\
r\frac{N-1}{2N^2} & r\frac{(N-1)(N-2)}{N^2} & 0 & 0 & 0 \\
r\frac{1-s}{2N} & r\frac{(1-s)(N-2)}{N} & 0 & 0 & 0 \\
\frac{\beta}{4(N-1)} & (1-s)\frac{N-2}{N-1}\alpha & (1-s)^2\frac{(N-2)(N-3)}{N(N-1)} & (1-s)^2\frac{N-2}{N(N-1)} & \frac{(1-s)^2}{4N(N-1)} \\
\frac{\alpha}{4N} & \frac{N-2}{2N^2}[N\alpha+2(1-s)] & (1-s)\frac{(N-2)(N-3)}{N^2} & (1-s)\frac{N-2}{N^2} & \frac{1-s}{4N^2} \\
\frac{N-1}{2N^3} & \frac{2(N-1)(N-2)}{N^3} & \frac{(N-1)(N-2)(N-3)}{N^3} & \frac{(N-1)(N-2)}{N^3} & \frac{N-1}{4N^3} \\
\frac{(1-s)}{2N^2} & (1-s)\frac{2(N-2)}{N^2} & (1-s)\frac{(N-2)(N-3)}{N^2} & \frac{N-2}{2N}\alpha & \frac{s(N-1)}{4N^2} \\
\frac{(1-s)^2}{2N(N-1)} & (1-s)^2\frac{2(N-2)}{N(N-1)} & (1-s)^2\frac{(N-2)(N-3)}{N(N-1)} & 2s(1-s)\frac{N-2}{2N} & s^2\frac{N-1}{4N}
\end{array} \right\},$$

with α and β defined such as

$N\alpha = (1-s)+s(N-1)$

$N\beta = (1-s)^2+s^2(N-1)^2.$

(ii) *Expected values of one- and two-locus identity in the KAM*

The *K*-allele model provides the probabilities of identity in state (IIS). It is assumed that all genes at locus *i* have the same probability μ_i of mutating to any of the *K* allelic states. Recursion equations for the probability of identity in state in a *K*-allele model can also be written in the form

$\mathbf{Q}^{t+1} = \mathbf{M}'\mathbf{Q}^t + \mathbf{C}'$,

with $\mathbf{M}'$ and $\mathbf{C}'$ defined, as for the IAM, as

$$\mathbf{M}' = \left(\begin{array}{c|c} \mathbf{M}'_1 & \mathbf{0} \\ \hline \mathbf{M}'_2 & \mathbf{M}'_3 \end{array}\right)$$

$$\mathbf{C}' = \left(\frac{\mathbf{C}'_1}{\mathbf{C}'_2}\right).$$

Note that $\mathbf{C}$ no longer contains only terms of coalescence, but also the probabilities that some genes that were different in state in the preceding generation become IIS after mutation. As there is a finite number of allelic states K_i at the ith locus, and as the unconditional probability of mutation is μ_i, each allele can mutate to another state with probability $\mu_i/(K_i-1)$. Thus, genes which were identical in state in the previous generation are still identical in state with probability $\nu'_i = (1-\mu_i)^2+\mu^2/(K_i-1)$. Alternatively, genes which were in different allelic states in the previous generation can become identical in state with probability $\omega_i = (1-\nu'_i)/(K_i-1)$. For example, all terms of the form

$$Q_{1i}^{t+1} = \nu_i \frac{1+Q_{0i}^t}{2}$$

in the IAM become, in the KAM,

$$Q_{1i}^{t+1} = \frac{\nu'_i + Q'^t_{0i}}{2},$$

where Q'_{0i} and Q'_{1i} are defined as the conditional IIS probabilities for pairs of genes, after mutation, given the IIS probabilities Q_{0i} and Q_{1i} before mutation. For pairs of genes taken h steps apart in the hierarchy ($h = 0$ for genes taken within individuals or $h = 1$ for genes taken among individuals)

$$\begin{aligned} Q'_{hi} &= \nu'_i Q_{hi} + \omega_i(1-Q_{hi}) \\ &= \rho_i Q_{hi} + \omega_i, \end{aligned}$$

with $\rho_i = \nu'_i - \omega_i$. More generally, we need to rewrite recursion equations by multiplying each term of IAM recursions by ρ_i instead of ν_i, and replacing Q_{hi}'s by Q'_{hi}'s. Consequently, the recursion equations for the one-locus IIS probabilities are given by

$$\mathbf{M}'_1 = \mathbf{D}_1 \mathbf{M}_1 \begin{Bmatrix} \rho_i & 0 & 0 & 0 \\ 0 & \rho_j & 0 & 0 \\ 0 & 0 & \rho_i & 0 \\ 0 & 0 & 0 & \rho_j \end{Bmatrix}$$

and

$$\mathbf{C}'_1 = \begin{Bmatrix} \nu_i & 0 & 0 & 0 \\ 0 & \nu_j & 0 & 0 \\ 0 & 0 & \nu_i & 0 \\ 0 & 0 & 0 & \nu_j \end{Bmatrix} \mathbf{D}_1 \mathbf{C}_1 + \mathbf{D}_1 \mathbf{M}_1 \begin{Bmatrix} \omega_i \\ \omega_j \\ \omega_i \\ \omega_j \end{Bmatrix} + (\mathbf{I}-\mathbf{D}_1) \begin{Bmatrix} \frac{1}{K_i} \\ \frac{1}{K_j} \\ \frac{1}{K_i} \\ \frac{1}{K_j} \end{Bmatrix}.$$

Note that the last term in the last equation represents the fact that two genes at locus i which were in different subpopulations in the previous generation have a probability $1/K_i$ to be identical in state, since the number of subpopulations is assumed to be infinitely large.

The natural extension for two-locus IIS identity probabilities is as follows. If we now denote by Q_i some one-locus identity probability, and Q_{ij} some two-locus identity probability, all terms of the form $Q_{ij}^{t+1} = \nu_i \nu_j (1+Q_i^t+Q_j^t+Q_{ij}^t)/4$ in the IAM, become, in the KAM, $Q_{ij}^{t+1} = (\nu'_i \nu'_j + Q'^t_i + Q'^t_j + Q'^t_{ij})/4$, where Q'_i and Q'_j are defined as in the previous section dedicated to the one-locus case, and Q'_{ij} represents some conditional IIS probability of a pair of genes, after mutation, given the IIS probability Q_{ij} before mutation.

$$\begin{aligned} Q'_{ij} &= \nu'_i \nu'_j Q_{ij} + \omega_i \omega_j (1-Q_{ij}) \\ &= (\nu'_i \nu'_j - \omega_i \omega_j) Q_{ij} + \omega_i \omega_j. \end{aligned}$$

Again, we need to rewrite recursion equations by multiplying each term of IAM recursions by appropriate factors. Consequently, the recursion equations for the two-locus probabilities of identity in state are now given by

$$\mathbf{M}_2' = \mathbf{D}_2\mathbf{M}_2 \begin{Bmatrix} \nu_j\rho_i & 0 & 0 & 0 \\ 0 & \nu_i\rho_j & 0 & 0 \\ 0 & 0 & \nu_j\rho_i & 0 \\ 0 & 0 & 0 & \nu_i\rho_j \end{Bmatrix} + \mathbf{D}_2'\mathbf{M}_2'' \begin{Bmatrix} \frac{\rho_i}{K_j} & 0 & 0 & 0 \\ 0 & \frac{\rho_j}{K_i} & 0 & 0 \\ 0 & 0 & \frac{\rho_i}{K_j} & 0 \\ 0 & 0 & 0 & \frac{\rho_j}{K_i} \end{Bmatrix}$$

$$\mathbf{M}_3' = (\nu_i\nu_j - \omega_i\omega_j)\,\mathbf{D}_2\mathbf{M}_3$$

$$\mathbf{C}_2' = \mathbf{D}_2\mathbf{M}_2 \begin{Bmatrix} \nu_j\omega_i \\ \nu_i\omega_j \\ \nu_j\omega_i \\ \nu_i\omega_j \end{Bmatrix} + \mathbf{D}_2\mathbf{M}_3 \begin{Bmatrix} \omega_i\omega_j \\ \vdots \\ \omega_i\omega_j \end{Bmatrix} + \nu_i\nu_j\mathbf{D}_2\mathbf{C}_2 + \mathbf{D}_2'\mathbf{M}_2'' \begin{Bmatrix} \frac{\omega_i}{K_j}+\frac{\omega_j}{K_i} \\ \vdots \\ \frac{\omega_i}{K_j}+\frac{\omega_j}{K_i} \end{Bmatrix} + \left(\frac{\nu_i}{K_j}+\frac{\nu_j}{K_i}\right)\mathbf{D}_2'\mathbf{C}_2''$$

$$+ (\mathbf{I} - \mathbf{D}_2 - \mathbf{D}_2') \begin{Bmatrix} \frac{1}{K_iK_j} \\ \vdots \\ \frac{1}{K_iK_j} \end{Bmatrix}$$

with the additional vectors and matrices $\mathbf{D}_2'$, $\mathbf{M}_2''$ and $\mathbf{C}_2''$ defined as

$$\mathbf{D}_2' = \mathrm{diag}[0, 0, 2m(1-m)^2, 2m(1-m)^2, 2m(1-m)^2(2-m), \ldots, 2m(1-m)^2(2-m)]$$

$$\mathbf{M}_2'' = \begin{Bmatrix} 0 & \cdots & 0 & \cdots \\ 0 & \cdots & 0 & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{Ns+1}{8N} & \cdots & \frac{(N-1)+N(1-s)}{4N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{1}{4N} & \cdots & \frac{N-1}{2N} & \cdots \\ \frac{Ns+1}{8N} & \cdots & \frac{(N-1)+N(1-s)}{4N} & \cdots \\ \frac{s}{4} & \cdots & \frac{1-s}{2} & \cdots \end{Bmatrix} \quad \text{and} \quad \mathbf{C}_2'' = \begin{Bmatrix} 0 \\ 0 \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{Ns+1}{8N} \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{1}{4N} \\ \frac{Ns+1}{8N} \\ \frac{s}{4} \end{Bmatrix}.$$

References

Anderson, T. J. C., Paul, R. E. L., Donelly, C. A. & Day, K. P. (2000). Do malaria parasites mate non-randomly in the mosquito midgut? *Genetical Research* **75**, 285–296.

Avery, P. J. & Hill, W. G. (1979). Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. *Genetics* **91**, 817–844.

Bennett, J. H. (1954). On the theory of random mating. *Annals of Eugenics* **18**, 311–317.

Bennett, J. H. & Binet, F. E. (1956). Association between Mendelian factors with mixed selfing and random mating. *Heredity* **10**, 51–56.

Charlesworth, D. (1991). The apparent selection on neutral marker loci in partially inbreeding populations. *Genetical Research* **57**, 159–175.

Cockerham, C. C. (1984). Drift and mutation with a finite number of allelic states. *Proceedings of the National Academy of Sciences of the USA* **81**, 530–534.

Conway, D. J., Roper, C., Oduola, A. M. J., Arnot, D. E., Krelsner, P. G., Grobusch, M. P., Curtis, C. F. & Greenwood, B. M. (1999). High recombination rate in natural populations of *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the USA* **96**, 4506–4511.

Crow, J. F. & Aoki, K. (1984). Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the USA* **81**, 6073–6077.

David, P. (1999). A quantitative model of the relationship between phenotypic variance and heterozygosity at marker loci under partial selfing. *Genetics* **153**, 1463–1474.

Devlin, B. & Rish, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics* **15**, 25–57.

Golding, G. B. & Strobeck, C. (1980). Linkage disequilibrium in a finite population that is partially selfing. *Genetics* **94**, 777–789.

Hastings, A. (1984). Maintenance of high disequilibrium in the presence of partial selfing. *Proceedings of the National Academy of Sciences of the USA* **81**, 4596–4598.

Hedrick, P. W. (1987). Gametic disequilibrium: proceed with caution. *Genetics* **117**, 331–341.

Hill, W. G. (1974*a*). Disequilibrium among several linked neutral genes in finite populations. I. Mean changes in disequilibrium. *Theoretical Population Biology* **5**, 366–392.

Hill, W. G. (1974*b*). Disequilibrium among several linked neutral genes in finite populations. II. Variances and covariances of disequilibria. *Theoretical Population Biology* **6**, 184–198.

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics* **38**, 209–216.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Hill, W. G. & Weir, B. S. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics* **54**, 705–714.

Hill, W. G., Babiker, H. A., Ranford-Cartwright, L. C. & Walliker, D. (1995). Estimation of inbreeding coefficients from genotypic data on multiple alleles, and their application to estimation of clonality in malaria parasites. *Genetical Research* **65**, 53–61.

Lenski, R. E. (1993). Assessing the genetic structure of microbial populations. *Proceedings of the National Academy of Sciences of the USA* **90**, 4334–4336.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics* **120**, 849–852.

Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences of the USA* **90**, 4384–4388.

Miyashita, N. T., Kawabe, A. & Innan, H. (1999). DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* **152**, 1723–1731.

Nei, M. & Li, W.-H. (1973). Linkage disequilibrium with the island model. *Genetics* **101**, 139–155.

Ohta, T. (1980). Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigene families. *Genetical Research* **36**, 181–197.

Ohta, T. (1982*a*). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **79**, 1940–1944.

Ohta, T. (1982*b*). Linkage disequilibrium with the island model. *Genetics* **101**, 139–155.

Ohta, T. & Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

Paul, R. E. L., Packer, M. J., Walmsley, M., Lagog, M., Ranford-Cartwright, L. C., Paru, R. & Day, K. P. (1995). Mating patterns in malaria parasite populations of Papua New Guinea. *Science* **269**, 1709–1711.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation process. *Genetics* **142**, 1357–1362.

Slatkin, M. (1975). Gene flow and selection in a two-locus system. *Genetics* **81**, 787–802.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.

Strobeck, C. (1979). Partial selfing and linkage: the effect of a heterotic locus on a neutral locus. *Genetics* **92**, 305–315.

Su, X.-Z., Ferdig, M. T., Huang, Y., Huynh, C. Q., Liu, A., You, J., Wooton, J. C. & Wellems, T. E. (1999). A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286**, 1351–1353.

Sved, J. A. (1968). The stability of linked systems of loci with a small population size. *Genetics* **59**, 543–563.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.

Tachida, H. & Cockerham, C. C. (1986). Analysis of linkage disequilibrium in an island model. *Theoretical Population Biology* **29**, 161–197.

Takahata, N. (1982). Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genetical Research* **39**, 63–77.

Tibayrenc, M. & Lal, A. (1996). Self-fertilization, linkage disequilibrium, and strain in *Plasmodium falciparum*. *Science* **271**, 1300–1301.

Tibayrenc, M., Kjellberg, F. & Ayala, F. J. (1990). A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proceedings of the National Academy of Sciences of the USA* **87**, 2414–2418.

Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Brenière, S. F., Dardé, M.-L. & Ayala, F. J. (1991). Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proceedings of the National Academy of Sciences of the USA* **88**, 5129–5133.

Vitalis, R. & Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* (in press).

Weir, B. S. & Cockerham, C. C. (1969). Group inbreeding with two linked loci. *Genetics* **63**, 711–742.

Weir, B. S. & Cockerham, C. C. (1973). Mixed self and random mating at two loci. *Genetical Research* **21**, 247–262.

Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.

Weir, B. S., Avery, P. J. & Hill, W. G. (1980). Effect of mating structure on variation in inbreeding. *Theoretical Population Biology* **18**, 369–429.

Whitlock, M. C., Phillips, P. C. & Wade, M. J. (1993). Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. *Evolution* **47**, 1758–1769.

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

Annexe F

Estimation de la taille efficace

Vitalis R. and Couvet D. (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157** : 911-925.

Copyright © 2001 by the Genetics Society of America

Estimation of Effective Population Size and Migration Rate From One- and Two-Locus Identity Measures

Renaud Vitalis*,†,‡ and Denis Couvet§

**Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, 34095 Montpellier Cedex 05, France, †Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France, ‡Laboratoire Génome, Populations et Interactions, Université Montpellier II, 34095 Montpellier Cedex 05, France and §CRBPO-Museum National d'Histoire Naturelle, 75005 Paris, France*

Manuscript received May 10, 2000
Accepted for publication November 9, 2000

ABSTRACT

Standard methods for inferring demographic parameters from genetic data are based mainly on one-locus theory. However, the association of genes at different loci (*e.g.*, two-locus identity disequilibrium) may also contain some information about demographic parameters of populations. In this article, we define one- and two-locus parameters of population structure as functions of one- and two-locus probabilities for the identity in state of genes. Since these parameters are known functions of demographic parameters in an infinite island model, we develop moment-based estimators of effective population size and immigration rate from one- and two-locus parameters. We evaluate this method through simulation. Although variance and bias may be quite large, increasing the number of loci on which the estimates are derived improves the method. We simulate an infinite allele model and a *K* allele model of mutation. Bias and variance are smaller with increasing numbers of alleles per locus. This is, to our knowledge, the first attempt of a joint estimation of local effective population size and immigration rate.

IN finite populations, genes undergo a random sampling process, known as genetic drift. As a consequence of this process, genetic variation is continuously lost. A powerful way to illustrate this point is that if we could trace the genealogy of genes, going backward in time, we would observe a continuous decrease in the number of "ancestor" genes. For neutral genes, the dynamic of this process depends mostly on the effective population size. Indeed, the expected time elapsed since two genes diverged from their common ancestor (the coalescence time) increases with the effective population size. The effective population size, noted N_e, has been defined as the size of an ideal population, in which a given genetic parameter (*e.g.*, the rate of inbreeding) takes the same expected value as in the population under scrutiny (Wright 1931). This suggests that there may be many different effective sizes depending on which parameter is chosen. Indeed, the inbreeding effective size has been defined from the expected rate of increase in homozygosity, and the variance effective size is derived from the expected rate of increase of variance in allele frequency per generation (Crow and Denniston 1988). Ewens (1979) also defined the eigenvalue effective size as the first nonunit eigenvalue of the transition matrix for allelic changes in a population. In the particular case of the Wright-Fisher model, inbreeding and variance effective sizes take the same value (Crow and Kimura 1970, chap. 8) as does the eigenvalue effective size (Ewens 1982). This model assumes a finite monoecious population of constant size, nonoverlapping generations, random mating, and equal expected contribution of individuals to the next generation (Fisher 1930; Wright 1931).

The conditions under which the different definitions of N_e coincide can be understood by considering the relationship between the rate of coalescence and the effective population size. The asymptotic rate at which two genes coalesce (the probability of coalescence) in a single ideal monoecious diploid population of N individuals is $1/(2N)$. But consider a structured population with different classes of individuals. Those classes may represent groups of individuals differing by their sex, age, stage, social rank, or geographical position. In contrast with unstructured populations, lines of descent may now coalesce at different rates within classes and among classes.

There is a strict relationship between probabilities of identity by descent and coalescence times (Malécot 1975; Slatkin 1991). Indeed, the leading eigenvalue of the transition matrix describing the increase in identity by descent for pairs of genes gives the long-term rate of coalescence of lineages. It is also equal to the leading nonunit eigenvalue of the transition matrix that gives the change in the distribution of allelic states (Whitlock and Barton 1997). This is true for an arbitrarily structured population. Therefore, the long-term rate of coalescence defines an asymptotic effective size that gives the asymptotic rate of increase in inbreeding,

Corresponding author: Renaud Vitalis, Laboratoire de Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université de Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France.
E-mail: vitalis@isem.univ-montp2.fr

gene frequency variance, and the expected rate of change in the distribution of allelic states (Whitlock and Barton 1997).

Mating systems, variance in the reproductive success of individuals, changes in population size through time, skewed sex ratios, and overlapping generations are some factors expected to make the effective size different from the census size (Caballero 1994). The effective population size also affects the efficiency of natural selection in maintaining advantageous mutations or promoting the spread of new ones. A small effective size also reduces the ability of natural selection to purge slightly deleterious alleles. Slightly deleterious mutations are more likely to be fixed, decreasing the mean fitness of the population and, thus, further reducing the effective population size (Lynch and Gabriel 1990; Gabriel and Bürger 1994). This runaway process, described as the mutational meltdown, eventually leads to extinction (Lande 1994; van Noordwijk 1994; Lynch *et al.* 1995a,b). But despite its importance in predicting population health or inferring past demography, estimation of effective size in natural populations is still a difficult task (Waples 1989; Schwartz *et al.* 1999).

Since the discrepancy between the census and the effective size of a population depends mainly on the parameters of the mating system and relative reproductive success of individuals, effective population size can be calculated from the direct evaluation of these parameters (Nunney and Elam 1994). However, the demographic data needed to calculate the effective population size are difficult to obtain practically in natural situations (Crow and Denniston 1988). Moreover, single-season assessments of these parameters are known to overestimate long-term effective population size, since interannual fluctuations in population size are not taken into account (Vucetich *et al.* 1997). As an alternative to direct evaluation of effective population size from ecological data, indirect estimates of effective size from genetical data have attracted much attention. So far, indirect estimates have been based on the temporal change in allele frequencies (Nei and Tajima 1981; Pollak 1983; Waples 1989; Williamson and Slatkin 1999), the excess of heterozygotes in progeny (Pudovkin *et al.* 1996; Luikart and Cornuet 1999), and linkage disequilibrium (Hill 1981; Waples 1991; Bartley *et al.* 1992).

It is worth noting that all of these approaches assume a single isolated population. However, in natural situations, it is more likely that populations of the same species are connected, to some extent, by gene flow (Slatkin 1987). Indeed, from an evolutionary and ecological perspective, the only way a species can persist in fragmented and changing habitats is to disperse (Hanski and Gilpin 1997). Moreover, patterns of dispersal are of primary importance in preventing or favoring local adaptation (Maynard Smith and Hoekstra 1980). There have been several attempts to estimate gene flow from genetic data (Slatkin and Barton 1989). However, all the approaches based on F_{ST} or the rare allele method of Slatkin (1985) cannot untangle the effect of drift from the migration pattern (Slatkin and Barton 1989). Indeed, gene flow is estimated as $N_e m$, the product of effective population size and immigration rate, and is referred to as the "effective number of migrants" per generation. Even more powerful methods, based on coalescent theory, have the same drawback (Slatkin and Maddison 1989; Beerli and Felsenstein 1999). Tufto *et al.* (1996) developed a maximum-likelihood method to estimate patterns of migration from allele frequency distribution. Their method provides estimates for short and long range migration rates, but with the effective population size treated as a known parameter.

Here, we propose a method-of-moments approach to infer effective population size on the one hand and immigration rate on the other hand, based on estimates of functions of one- and two-locus identity probabilities. One-locus parameters are functions of first and second moments of allele frequencies. Two-locus identity probabilities are functions of higher-order moments, up to the fourth (Ohta 1982a,b). One-locus identity probabilities are influenced primarily by genetic drift and local immigration. Two-locus identity probabilities are also influenced by drift and migration, through the formation of gametic disequilibria (Weir and Cockerham 1969; Cockerham and Weir 1977; Ohta 1982a,b; Tachida and Cockerham 1986). Thus, the joint analysis of one- and two-locus identity probabilities provides more information on the parameters of interest than the analysis of only one-locus parameters.

In a companion article, we derived the expected values for two-locus probabilities for identity in state in an island model with partial selfing (Vitalis and Couvet 2001). We defined a two-locus parameter, which we call the "within-subpopulation identity disequilibrium," as the excess of two-locus identity probabilities over the product of single-locus probabilities among individuals within subpopulations. Here, we first recall the definitions of one- and two-locus identity probabilities and derive the expectation for one-locus parameters. We then define appropriate statistics for estimating one- and two-locus parameters, whose expectations depend on the population parameters of interest (effective size N_e, migration rate m) and not on some nuisance parameters, such as the mutation rate or the model of mutation. Then, we present a simple method-of-moments joint estimation of effective size and migration rate. We further explore the reliability of N_e and m estimators by means of stochastic simulations.

THEORY

One-locus population structure: The expectation of any descriptive statistic of population genetic structure

can ultimately be expressed as a function of some probabilities of identity in state for pairs of genes taken at different hierarchical levels (within individuals, among individuals within a population, among populations).

Let us consider an infinite island model of population structure (Wright 1940). This model assumes that the whole population (or the species) is subdivided into an infinite number of subpopulations that exchange genes. We may relax the usual assumption of equal subpopulation sizes and immigration rates and then define as many sets of parameters as there are subpopulations. A focal subpopulation i has N_{ei} diploid individuals and receives m_i migrant genes per generation. We define $Q_{0,i}$ as the probability that two homologous genes randomly sampled in one individual from subpopulation i are identical in state (IIS), $Q_{1,i}$ as the IIS probability for two genes randomly sampled in subpopulation i, and Q_2 may also be defined as the IIS probability for two genes randomly sampled in the pool of immigrants.

Let us first assume an infinite allele model (IAM). In this model, mutation always creates a new allelic state in the population. Therefore, genes that are IIS are also identical by descent (IBD), *i.e.*, two exact copies (without mutation) of the same ancestral gene (Malécot 1948). Let $a_i = (1 - m_i)^2$ be the frequency of pairs of genes that come from the same subpopulation in the previous generation. Each generation, some offspring are produced by selfing. We define s as the conditional probability that two homologous genes of one individual were produced by the same individual, given that they are copies of genes from one subpopulation in the previous generation. Mutations arise at rate μ, and $\nu = (1 - \mu)^2$. The recursion equations for IBD probabilities in the IAM are given by

$$\begin{aligned} Q_{0,i}^{t+1} &= \nu\left[a_i\left(s\frac{1 + Q_{0,i}^t}{2} + (1 - s)Q_{1,i}^t\right) + (1 - a_i)Q_2^t\right] \\ Q_{1,i}^{t+1} &= \nu\left[a_i\left(\frac{1}{N_{ei}}\frac{1 + Q_{0,i}^t}{2} + \left(1 - \frac{1}{N_{ei}}\right)Q_{1,i}^t\right) + (1 - a_i)Q_2^t\right] \\ Q_2^{t+1} &= \nu Q_2^t. \end{aligned} \quad (1)$$

A useful parameter to consider is

$$F_i = \frac{Q_{1,i} - Q_2}{1 - Q_2}. \quad (2)$$

The weighted sum of F_i over subpopulations is the intraclass correlation for the probability of identity in state of genes within subpopulations relative to the whole population, which is F_{ST} (Cockerham and Weir 1987; Rousset 1996). At equilibrium,

$$F_i = \frac{\nu a_i}{\nu a_i(2 - \nu a_i) + N_{ei}(1 - \nu a_i)(2 - \nu a_i s)}. \quad (3)$$

For small mutation rates, this parameter is inversely related to the effective number of migrants per generation, $N_{ei}m_i$,

$$F_i \approx \frac{1}{1 + 4N_{ei}m_i}. \quad (4)$$

A measure of this parameter can therefore be used to make some inferences about the product of the local effective population size and immigration rate.

We also derived the recursion equations for IIS probabilities in the K allele model (KAM). In this model, there are a finite number (K) of possible allelic states at a locus. Since the unconditional probability of mutation is μ, each gene mutates toward a particular allele with probability $\mu/(K - 1)$. Thus, genes that were IIS in the previous generation are still IIS in the current generation with probability $\nu' = (1 - \mu)^2 + \mu^2/(K - 1)$. Also, genes that were different in state in the previous generation can become IIS in the current generation with probability $\omega = (1 - \nu')/(K - 1)$. Therefore, recursion equations for the IIS probabilities in the KAM are given by

$$\begin{aligned} Q_{0,i}^{t+1} &= a_i\left(s\frac{\nu' + Q_{0,i}'^t}{2} + (1 - s)Q_{1,i}'^t\right) + (1 - a_i)Q_2'^t \\ Q_{1,i}^{t+1} &= a_i\left(\frac{1}{N_{ei}}\frac{\nu' + Q_{0,i}'^t}{2} + \left(1 - \frac{1}{N_{ei}}\right)Q_{1,i}'^t\right) + (1 - a_i)Q_2'^t \\ Q_2^{t+1} &= Q_2'^t, \end{aligned} \quad (5)$$

with $Q_h'^t = \nu' Q_h^t + \omega(1 - Q_h^t)$ defined as the conditional IIS probabilities for pairs of genes taken in the hth state of the hierarchy ($h = 0, 1, 2$), after mutation, given the IIS probabilities Q_h^t before mutation (see Crow and Aoki 1984; Rousset 1996, for similar developments). The range of $Q_{1,i}$ is strongly dependent on the total number of alleles at a locus. In particular, when there are K possible allelic states at a locus, the theoretical range of $Q_{1,i}$ is bounded below by $1/K$ (see Appendix 10 in Crow and Kimura 1970, p. 515). On the other hand, at equilibrium,

$$F_i = \frac{\alpha a_i}{\alpha a_i(2 - \alpha a_i) + N_{ei}(1 - \alpha a_i)(2 - \alpha a_i s)}, \quad (6)$$

with $\alpha = (\nu' K - 1)/(K - 1)$. Equation 6 is of the same form as Equation 3, with α replacing ν. Therefore, for $\mu \ll m$, F_i has the same expectation in the IAM and the KAM. This result suggests that F_i is an appropriate function of IIS probabilities, since this parameter is nearly independent of the number of allelic states (Table 1).

Two-locus population structure: The one-locus theory for IIS probabilities can be extended to the two-locus case (Cockerham 1984; Goodnight 1987, 1988). In a random mating population, three two-locus IIS probabilities need to be defined (see Whitlock *et al.* 1993) for pairs, triplets, and quadruplets of sampled haplotypes (we call a haplotype a set of two genes taken at two distinct loci, which is inherited from a single parent, after recombination). However, in any case of departure

TABLE 1

Expected one-locus identity probabilities and two-locus identity disequilibrium, with special reference to the effect of the mutation model

μ	K	Q_1	F	η_S	η'_S
10^{-6}	2	0.597586	0.195172	0.000599	0.002398
	5	0.356142	0.195177	0.001536	0.002399
	10	0.275661	0.195179	0.001944	0.002400
10^{-4}	∞	0.194397	0.194397	0.002390	0.002390
10^{-5}		0.195108	0.195108	0.002399	0.002399
10^{-6}		0.195179	0.195179	0.002400	0.002400

Parameters of the model are $N_e = 50$, $N_e m = 1$ for a sample size of $n = 50$ diploid individuals. Random mating is assumed in the population.

from random mating, more coefficients must be defined, since IIS probabilities defined for pairs, triplets, or quadruplets of haplotypes may have different expectations whether these haplotypes are sampled in two, three, or four individuals. For this purpose, we defined a total of 10 IIS two-locus probabilities (see Vitalis and Couvet 2001, for further details). Figure 1 depicts the definition of two-locus IIS probabilities.

For example, for two-individual IIS probabilities, ϕ is defined as the probability that two haplotypes sampled in two distinct individuals are IIS at the two loci considered. γ is defined as the probability that, when a single haplotype of one individual and the two haplotypes of a second individual are sampled, both pairs of homologous genes taken among the two individuals are IIS. And δ is the probability that, when both haplotypes of each individual are sampled, one pair (across individuals) is IIS at the first locus, and the other (distinct) pair is IIS at the second locus. See Figure 1 and Vitalis and Couvet (2001) for the definition of other IIS probabilities.

In a diploid population, the gametic phase is usually not known. Therefore, we define the compound IIS two-locus probability for two pairs of genes, each taken at two distinct loci, among individuals in a population as

$$\Phi = \frac{\phi + 2\gamma + \delta}{4}. \quad (7)$$

The expected gametic disequilibrium can be expressed by the within-subpopulation identity disequilibrium ($\eta_{S,ij}$), across loci i and j. Identity disequilibrium is equivalent to the covariance for a pair of one-locus identity probabilities in a random pair of individuals,

$$\eta_{S,ij} = \Phi_{ij} - \delta_{4ij}, \quad (8)$$

where δ_{4ij} is the two-locus probability of identity by descent among loci i and j, when all genes are sampled from distinct haplotypes. η_S is equivalent to the excess

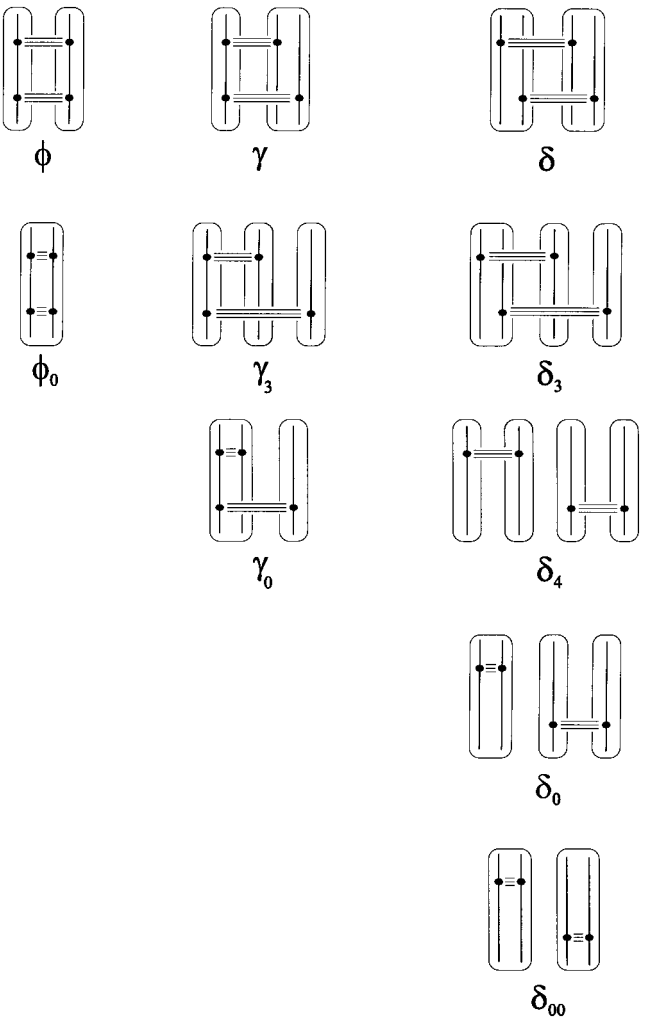


Figure 1.—Definition of two-locus probabilities for the probability of identity in state (IIS). Vertical lines show sampled haplotypes, on which upper and lower positions of solid circles represent two loci. Each diploid individual is represented by a box. Horizontal lines (≡) between pairs of homologous genes stand for identity in state. In the infinite allele model, these coefficients define the corresponding probabilities for identity by descent (IBD). Only the sampled haplotypes are shown.

probability of simultaneous identity over that expected from random combination of the identity at two loci (Ohta 1980). It is also equivalent to the covariance of nonidentity at two loci within populations (Avery and Hill 1979). We derived the recursion equations for all these two-locus probabilities in the IAM (IBD probabilities) and in the KAM (IIS probabilities; Vitalis and Couvet 2001). The parameter η_S is a monotonic decreasing function of N_e, for a given value of F (Figure 2). Therefore, there is a single pair of N_e and m values that provides a given pair of η_S and F values.

It follows also from this graph that, besides any consideration about the variance of the estimates, large effective sizes would be difficult to estimate, since the amplitude of the variation of η_S with effective population size

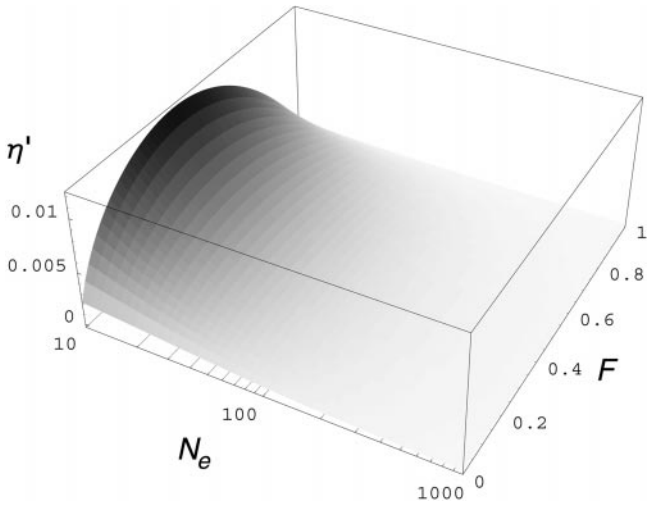


FIGURE 2.—Expected identity disequilibrium η'_S as a function of effective population size and one-locus identity probability F, in an infinite allele model with $\mu = 10^{-6}$. The effective number of immigrant individuals per generation ($N_e m$) was fixed to 1 and random mating was assumed in the population. Note the logarithmic scale on the x-axis.

decreases as the effective size increases. Moreover, the amplitude of variation depends on F, suggesting that the efficiency of estimation should vary with the population genetic diversity. The joint analysis of one- and two-locus identity probabilities for unlinked loci should thus permit the estimation of effective population size, provided this parameter is not of too high an order. Obtaining reliable estimates for large population sizes would require tightly linked loci (see also HILL and WEIR 1994).

ESTIMATION

One-locus statistics: For any given allele u, we define the indicator variable x_{ijku} to describe the state of the kth gene, with $k = (1, 2)$, of the jth individual in the ith subpopulation. $x_{ijku} = 1$ if the allelic state is u, $x_{ijku} = 0$ otherwise. Let P_i^u be the frequency of allele u in the subpopulation i. Then, $P_i^u = \mathscr{E}(x_{ijku}|\mathbf{P})$, with $\mathscr{E}(\ |\mathbf{P})$ denoting the expectation, conditional on the array $\mathbf{P}$ of all the allele frequencies. Considering the second moments of the random variable x_{ijku}, it follows that $\mathscr{E}(x_{ijku}^2|\mathbf{P}) = P_i^u$ and $\mathscr{E}(x_{ijku}x_{ij'k'u}|\mathbf{P}) = (P_i^u)^2$, with $j' \neq j$ and $k' \neq k$. Summing over all alleles gives the probability that two genes randomly taken from distinct individuals are IIS,

$$Q_{1,i} = \mathscr{E}\left[\sum_{u=1}^{K}(P_i^u)^2\right], \tag{9}$$

where $\mathscr{E}$ denotes now the expectation over the distribution of allele frequencies $\mathbf{P}$. We also define P_i^{uu} as the frequency of homozygotes for allele u in the subpopulation i. Then, $\mathscr{E}(x_{ijku}x_{ijk'u}|\mathbf{P}) = (P_i^{uu})^2$. The IIS probability for two genes randomly taken in the whole population can be defined as

$$Q_2 = \mathscr{E}\left[\sum_{u=1}^{K}(x_{ijku}x_{i'j'k'u})\right] \tag{10}$$

with $i' \neq i$. An unbiased estimator of the frequency of allele u among n sampled individuals is given by $\hat{P}_i^u = \sum_{j=1}^{n}\sum_{k=1}^{2}x_{ijku}/(2n)$. Expanding the square of this expression and taking expectation gives $\mathscr{E}[(\hat{P}_i^u)^2|\mathbf{P}] = [P_i^u(1 + 2(n-1)P_i^u) + P_i^{uu}]/(2n)$. Therefore,

$$\hat{Q}_{1,i} = \sum_{u=1}^{K}[\hat{P}_i^u(2n\hat{P}_i^u - 1) - \hat{P}_i^{uu}]/[2(n-1)], \tag{11}$$

where $\hat{P}_i^{uu} = \sum_{j=1}^{n}\sum_{k'\neq k}^{2}x_{ijku}x_{ijk'u}$ is an unbiased estimator of the frequency of homozygotes for allele u. An estimator for the IIS probability for genes taken in different subpopulations is given by

$$\hat{Q}_2 = \sum_{u=1}^{K}\sum_{i'\neq i}^{d}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)] \tag{12}$$

for d sampled subpopulations. Finally, approximating the expectation of a ratio by the ratio of expectations, an estimator of F_i can be given as

$$\hat{F}_i = \sum_u[(\hat{P}_i^u(2n\hat{P}_i^u - 1) - \hat{P}_i^{uu})/[2(n-1)] - \sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]]/(1 - \sum_u\sum_{i'\neq i}\hat{P}_i^u\hat{P}_{i'}^u/[d(d-1)]). \tag{13}$$

To combine the information over loci, we define a multilocus estimator as the ratio of the sum of locus-specific numerators over the sum of locus-specific denominators (REYNOLDS *et al.* 1983; WEIR and COCKERHAM 1984).

Two-locus statistics: Now, x_{iju} is the indicator variable that describes the state of gene j in individual i at a first locus and y_{ijv} is the indicator variable that describes the state of gene j in individual i at a second, distinct, locus. For the sake of clarity, all subpopulation indices are dropped in this section. $x_{iju} = 1$ if the allelic state at the first locus is u, $x_{iju} = 0$ otherwise; and $y_{ijv} = 1$ if the allelic state at the second locus is v, $y_{ijv} = 0$ otherwise. Let P_v^u be the frequency of two-locus haplotypes bearing alleles u and v (alleles in phase). Then $P_v^u = \mathscr{E}(x_{iju}y_{ijv}|\mathbf{P})$, with $\mathscr{E}(\ |\mathbf{P})$ denoting the expectation, conditional on the array $\mathbf{P}$ of all the haplotype frequencies. We also define $P_{\cdot v}^{u\cdot} = \mathscr{E}(x_{iju}y_{ij'v}|\mathbf{P})$ as the frequency of pairs of alleles u and v taken from the same individual but from different haplotypes (alleles in repulsion). As in the one-locus case, summing over all possible pairs of alleles and then taking expectations over the distribution of haplotype frequencies $\mathbf{P}$ give the following IIS probabilities:

$$\begin{aligned}\phi &= \mathscr{E}\left[\sum_{u,v}(P_v^u)^2\right]\\ \gamma &= \mathscr{E}\left[\sum_{u,v}(P_v^uP_{\cdot v}^{u\cdot})\right]\\ \delta &= \mathscr{E}\left[\sum_{u,v}(P_{\cdot v}^{u\cdot}P_{\cdot v}^{u\cdot})^2\right].\end{aligned} \tag{14}$$

However, since the gametic phase is usually unknown, P_v^u and $P_{\cdot v}^{u\cdot}$ are not measurable in practice from genotypic

data. We rather evaluate their mean, noted $P_{\bar{v}}^{\bar{u}}$, as $P_{\bar{v}}^{\bar{u}} = (P_v^u + P_{\cdot v}^{u\cdot})/2$. Therefore, the IIS probability for two haplotypes among individuals within a population is defined as

$$\Phi = \mathscr{E}\left[\sum_{u,v}(P_{\bar{v}}^{\bar{u}})^2\right]. \tag{15}$$

An unbiased estimate of $P_{\bar{v}}^{\bar{u}}$ is given by $\hat{P}_{\bar{v}}^{\bar{u}} = \sum_{i=1}^{n}\sum_{j,j'}^{2} \times x_{iju}y_{ij'v}/(4n)$. Expanding the square of this expression and then taking expectation, conditional on the array **P** of all the haplotype frequencies, gives

$$\mathscr{E}\left[(\hat{P}_{\bar{v}}^{\bar{u}})^2|\mathbf{P}\right] = [P_{\bar{v}}^{\bar{u}}[1 + 4(n-1)P_{\bar{v}}^{\bar{u}}] + P_{vv}^{uu} + P_{vv}^{u\cdot} + P_{v\cdot}^{uu}]/(4n), \tag{16}$$

where P_{vv}^{uu} is the frequency of double homozygotes for alleles u and v, $P_{v\cdot}^{uu}$ is the frequency of individuals that are homozygotes for allele u at the first locus and that carry one copy of allele v at the second locus, and where $P_{vv}^{u\cdot}$ is the frequency of individuals that are homozygotes for allele v at the second locus and that carry one copy of allele u at the first locus. Therefore, an unbiased estimator for Φ is

$$\hat{\Phi} = \sum_{u,v}[\hat{P}_{\bar{v}}^{\bar{u}}(4n\hat{P}_{\bar{v}}^{\bar{u}} - 1) - \hat{P}_{vv}^{uu} - \hat{P}_{v\cdot}^{uu} - \hat{P}_{vv}^{u\cdot}]/[4(n-1)], \tag{17}$$

where $\hat{P}_{vv}^{uu}$ is the observed frequency of double homozygotes for alleles u and v, and where $\hat{P}_{v\cdot}^{uu}$ (respectively $\hat{P}_{vv}^{u\cdot}$) is the observed frequency of homozygotes for allele u (respectively v) that carry one copy of allele v (respectively u) at the second locus. An estimator of the identity disequilibrium among loci i and j is given by

$$\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}. \tag{18}$$

Indeed, the expectation of this statistic is

$$\mathscr{E}(\hat{\eta}_{S,ij}) = \Phi_{ij}\left[1 - \frac{2}{n(n-1)}\right] - \frac{4(n-2)}{n(n-1)}\Gamma_{ij} - \frac{(n-2)(n-3)}{n(n-1)}\delta_{4ij} \tag{19}$$

with $\Gamma_{ij} = (\gamma_{3ij} + \delta_{3ij})/2$. For large samples sizes,

$$\mathscr{E}(\hat{\eta}_{S,ij}) \approx \Phi_{ij} - \delta_{4ij}. \tag{20}$$

Thus, provided the sample size is not too small, $\hat{\eta}_{S,ij} = \hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}$ is an unbiased estimator of the identity disequilibrium among loci i and j.

However, η_S depends on the underlying mutation model (Table 1). There has been some debate in the literature about the dependence on some measures of gametic disequilibrium of the underlying allellic frequencies (Hedrick 1987; Lewontin 1988). At the origin of this dispute is the fact that Hedrick (1987) considered allelic frequencies as fixed parameters. An alternative approach is to consider allelic frequencies as random variables, whose distribution depends on some parameters of the population model. In this sense, we expect that a good measure of disequilibrium shall satisfy the following criteria: (i) The parameter depends on the population parameters of interest, and not on any other parameter; (ii) the expectation of an estimator of this quantity, taken over replicates of the stochastic process of drift, depends only on the parameters of interest (unbiased statistic); and (iii) the distribution of this unbiased estimator depends on the parameters of interest, and not on any other parameter. This allows the measures to be compared across loci or populations, as well as to be pooled over loci.

Various parameters of gametic disequilibrium among loci i and j have been standardized with the product of the Hardy-Weinberg heterozygosities or nonidentities (Hedrick 1987; Ohta 1980),

$$\eta'_{S,ij} = \frac{\Phi_{ij} - \delta_{4ij}}{(1 - Q_{2i})(1 - Q_{2j})}. \tag{21}$$

An estimator for the standardized identity disequilibrium $\eta'_{S,ij}$ among loci i and j is

$$\hat{\eta}'_{S,ij} = \frac{\hat{\Phi}_{ij} - \hat{Q}_{1i}\hat{Q}_{1j}}{(1 - \hat{Q}_{2i})(1 - \hat{Q}_{2j})}. \tag{22}$$

Whereas the absolute measure of identity disequilibrium η_S depends on the mutation rate and the model of mutation, the standardized measure η'_S is insensitive to the mutational process (Table 1), even if the mating system departs from panmixia (not shown). The combination of identity disequilibrium measures over pairs of loci is achieved by taking the ratio of averaged numerators and denominators over pairs of loci and over samples (Ohta 1980).

Estimators for F and η'_S have been evaluated through the simulation of a population in an infinite island model. Several mutation models (from the two-allele model to the infinite allele model) were compared. For five alleles and more, the mean values for the parameters are close to their expectations and do not depend on the mutation model (Table 2). This agrees with the criteria given above. The effect of sample size is more pronounced for identity disequilibrium measures than for one-locus parameter measures. Indeed, increasing the sample size decreases the variance among η'_S estimates.

Estimation of N_e by the method-of-moments: For a focal population within an infinite island model, we have two statistics $\hat{F}$ and $\hat{\eta}'_S$, whose expectations are known functions of the population parameters of interest, namely the effective population size N_e and the immigration rate m. Their expectations do not depend on any nuisance parameters such as the mutation model, but do depend on some other parameters, such as the reproductive system or the extent of (physical) linkage of markers. However, these latter parameters may be estimated from independent data. Therefore, we assume in the following that the selfing rate, as well as the recombination rate among loci, is known.

We solve numerically a system of two simultaneous

TABLE 2

Estimated properties for one- and two-locus parameters

				$\hat{F}$			$\hat{\eta}'_S$	
K	N_e	n	F	Mean	σ	η'_S	Mean	σ
2	20	25	0.187917	0.187614	0.086208	0.005229	0.005133	0.007491
	20	50		0.186695	0.080968	0.005362	0.005379	0.005641
	50	25	0.195172	0.196953	0.087050	0.002338	0.002209	0.006024
	50	50		0.196924	0.081832	0.002398	0.002552	0.004265
5	20	25	0.187919	0.187542	0.058231	0.005230	0.005152	0.003507
	20	50		0.190018	0.055673	0.005363	0.005352	0.002881
	50	25	0.195177	0.194750	0.056116	0.002339	0.002394	0.002734
	50	50		0.192533	0.050428	0.002399	0.002437	0.001846
10	20	25	0.187919	0.190234	0.054219	0.005230	0.005332	0.003074
	20	50		0.189385	0.050953	0.005363	0.005355	0.002304
	50	25	0.195179	0.193779	0.046168	0.002339	0.002337	0.001872
	50	50		0.196065	0.047252	0.002400	0.002505	0.001540
∞	20	25	0.187920	0.186700	0.046215	0.005230	0.005111	0.002249
	20	50		0.189787	0.043126	0.005363	0.005515	0.002018
	50	25	0.195179	0.194349	0.040393	0.002339	0.002265	0.001428
	50	50		0.197306	0.039202	0.002400	0.002423	0.001133

Arithmetic means as well as standard deviations (σ) of $\hat{F}$ and $\hat{\eta}'_S$ are given for various simulated effective population sizes and mutation models. Estimates are based on 1000 measures (see text for details) from a simulated population of size N_e, receiving a proportion $1/N_e$ of migrant haplotypes each generation. A total of 50 individuals were sampled without replacement. Random mating was assumed in the population. Eight loci were simulated. The mutation rate was 10^{-6}. K indicates the number of alleles simulated per locus.

equations, with two unknowns, as follows. For a wide range of N_e values starting from $N_e = 2$, Equation 6 is solved for m, with $F = \hat{F}$. Then, for each pair of N_e and m values, the expected value of $\hat{\eta}'_S$ is calculated as

$$\mathcal{E}(\hat{\eta}'_{S,ij}) = [\Phi_{ij}[1 - 2/[n(n-1)]] - \Gamma_{ij}[4(n-2)]/[n(n-1)] - \delta_{4ij}[(n-2)(n-3)]/[n(n-1)]]/[(1-Q_{2i})(1-Q_{2i})] \quad (23)$$

(see Equations 19 and 22) from the recursive equations given in Vitalis and Couvet (2001). The solutions for N_e and m are then obtained for the best fit between this expected value and $\hat{\eta}'_S$ over a wide range of N_e values (up to 10 times the true N_e).

ASSESSING THE METHOD-OF-MOMENTS ESTIMATOR

Simulation procedure in an infinite island model: We evaluated the method of inference through simulations. We focused our analysis on a focal population receiving migrants from an infinitely large number of populations. In practice, a single local population consisting of N_e diploid individuals was simulated. Gametic dispersal occurred prior to reproduction. Under the IAM, at every locus, immigrant individuals carried alleles that were absent from the local population. Therefore, the probability to draw IIS genes from two different populations (Q_2) was considered to be zero in the IAM and $Q_2 = 1/K$ in the KAM. In both cases, Q_2 was taken as a fixed parameter when calculating statistics for the population parameters. Local random mating was assumed in all the results presented thereafter. For each set of simulations, the initial population was formed with $2N_e$ different alleles at each locus. For each set of parameters, the results from 10 independent simulations were pooled. In each simulation replicate, the first measure was realized after 1000 generations, and then every 100 generations, to avoid temporal correlation across samples, for 100 times. Fifty individuals were sampled for each measure. Indeed, since the identity disequilibrium may be very small in many circumstances, reasonably large sample sizes are required. However, this sample size is representative of those found in the literature (see Williamson and Slatkin 1999, and references therein).

Estimation efficiency was also assessed by examining the percentage of successful inferences as compared with unsuccessful ones. Indeed, among the 1000 identity measures obtained for a set of simulations, there were some cases in which reliable N_e estimates could not be inferred. Overall, we distinguished three cases.

Case 1: For a given F value, the $\hat{\eta}'_S$ estimate overrides its possible range. This is interpreted as a very small effective population size ($N_e \leq 2$).

Case 2: The estimate for $\hat{\eta}'_S$ is negative. This is interpreted as an infinite effective population size. We also excluded estimates that were >10 times the true N_e in this latter category.

Case 3: The estimation of N_e is reliable [lying in the interval $(2, 10 \times N_e)$].

Simulation results: Concerning the estimation of effective population size, a first source of bias is expected

TABLE 3

Results from simulations for an infinite allele model

		Estimated $\hat{N}_e$				Estimated $\hat{m}$			Inference efficiency		
L	N_e	H	c.v.	95% C.I.	m	Mean	c.v.	95% C.I.	1	2	3
4	20	17.75	0.63	8.10–51.88	0.05	0.0553	0.51	0.0194–0.1070	0.0	0.2	99.8
	50	43.32	0.88	19.21–192.77	0.02	0.0232	0.62	0.0052–0.0487	0.0	5.3	94.7
	100	70.02	1.07	26.91–465.71	0.01	0.0141	0.70	0.0025–0.0337	0.0	17.3	82.3
8	20	18.75	0.45	10.18–39.06	0.05	0.0537	0.40	0.0235–0.0933	0.0	0.0	100.0
	50	47.74	0.61	24.92–131.15	0.02	0.0212	0.46	0.0071–0.0384	0.0	0.3	99.7
	100	91.70	0.94	41.95–426.41	0.01	0.0110	0.60	0.0023–0.0229	0.0	5.5	94.5
12	20	18.86	0.37	9.94–36.64	0.05	0.0530	0.39	0.0251–0.0910	0.0	0.0	100.0
	50	48.45	0.48	26.06–107.73	0.02	0.0209	0.42	0.0085–0.0366	0.0	0.0	100.0
	100	98.84	0.80	47.53–332.14	0.01	0.0102	0.54	0.0028–0.0205	0.0	2.9	97.1
16	20	18.81	0.36	10.73–34.68	0.05	0.0533	0.38	0.0260–0.0888	0.0	0.0	100.0
	50	48.64	0.46	26.86–105.44	0.02	0.0210	0.41	0.0088–0.0366	0.0	0.0	100.0
	100	97.49	0.76	50.83–292.93	0.01	0.0105	0.49	0.0033–0.0205	0.0	0.8	99.2

Results are based on 1000 measures (see text) from a simulated population of size N_e receiving a proportion m of migrant individuals per generation. n is the number of diploid individuals that were sampled without replacement. Harmonic means (H) for N_e and arithmetic means for m are given for various simulated effective population sizes and various numbers of loci. Coefficients of variation (c.v.), the ratio of standard deviation over the mean, are also indicated. A 95% confidence interval (C.I.) is defined from the 5th and 95th percentiles. An assessment of inference efficiency is given with special reference to three distinct cases in the process of inference. Case 1 is the percentage of too large η_S values for a fit to be reliable, which is expected for a very small effective population size. Case 2 is the percentage of either negative η_S estimates, which give an infinite effective size, or $\hat{N}_e$ estimates >10 times the true N_e. Case 3 is the percentage of successful inferences from which results are given.

to result from the following. One- and two-locus identity parameters are inversely related to effective population size. As a consequence, even if estimates of parameters are unbiased, effective population size estimates may be biased, since the expectation of an inverse function is not the inverse of the expectation of the function, as discussed in Cockerham and Weir (1993). This is why we present harmonic rather than arithmetic means of $\hat{N}_e$ estimates.

In the IAM, harmonic means for $\hat{N}_e$ may be underestimated (43.32 for $N_e = 50$ and 70.02 for $N_e = 100$) when 4 loci are pooled (Table 3). However, for a greater number of loci, estimates are in very close agreement with true values (Figure 3, Table 3). Indeed, increasing the number of loci reduces the confidence interval, as given by the interval between the 5th and 95th percentiles of the overall distribution of estimates. Increasing the number of loci also increases the percentage of successful inferences (Table 3). Estimations of immigration rates were also in close agreement with the truth. Estimates generally have a high variance, as indicated by the coefficients of variation. However, the variance was substantially reduced when at least 8 loci were sampled. The estimation was very efficient in most cases. In all cases, we note that estimating effective size <2 never occurred (Case 1). Conversely, for 8 loci and less, estimating infinite effective size (Case 2) occurred only when the true effective size was large. For 12 loci and more, the percentage of successful estimation (Case 3) was >95% (Table 3).

In the KAM, harmonic means for N_e are biased downward for 8 loci and less, in a more significant way than in the infinite allele model (Figure 3, Table 4). This point is also indicated by a lower percentage of successful inferences (Table 4). The variance was smaller in a 10-allele rather than in a 5-allele model. Again, increasing the number of loci reduces both the variance and the confidence interval for the estimates. The estimation was efficient (>85% of successful estimates) for 12 loci and more, or for 8 loci with at least 10 alleles. As in the IAM, unsuccessful estimates (Case 1) occurred only when the true effective size was small, and too large estimates (Case 2) occurred only when the true effective size was large. Again, estimates of immigration rates were close to the truth, although they were slightly overestimated in all cases.

We also evaluated the method for various migration rates in an infinite allele model with 8 and 16 independent loci (Table 5, Figure 4). Bias and variance were reduced with a higher effective number of migrants per generation, $N_e m$. Indeed, the dispersion around the mean was large for low migration rate. In this model, lower immigration rates imply also lower levels of gene diversity within subpopulations, and therefore a poorer estimation efficiency.

We obtained a confidence region from the joint distri-

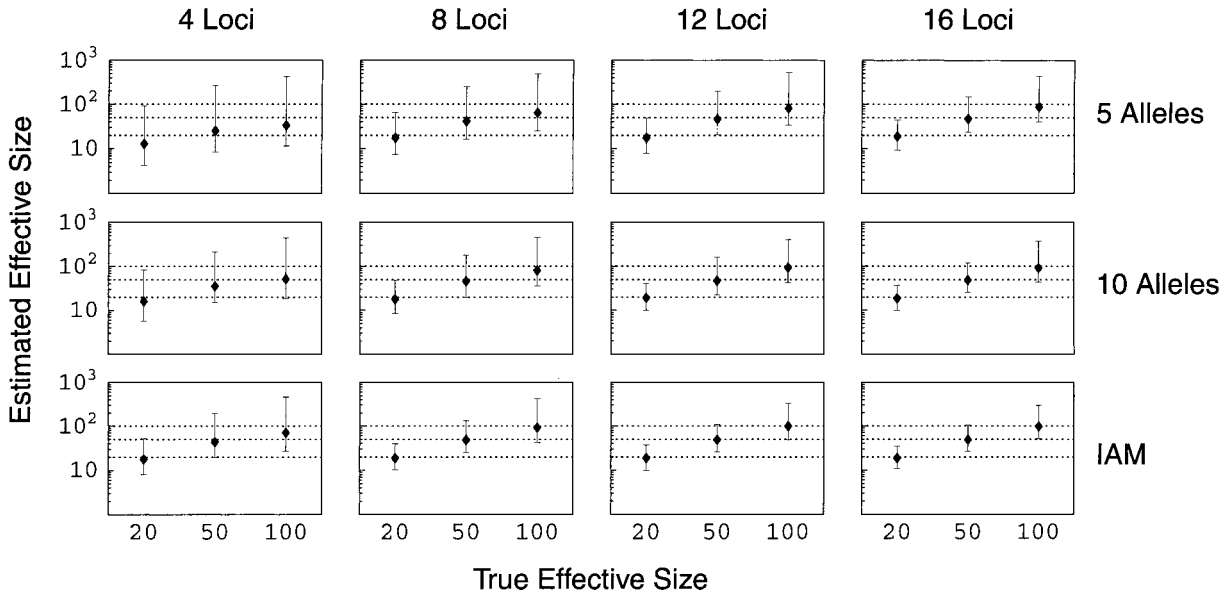


FIGURE 3.—Harmonic means and distribution of N_e estimates for a range of parameters. Three true effective sizes were simulated (20, 50, and 100, as indicated by horizontal dashed lines) with $N_e m = 1$. Simulations were performed for various numbers of unlinked loci, as well as for various mutation models, ranging from 5- and 10-allele models to the infinite allele model, with $\mu = 10^{-6}$. For each set of parameters, 1000 measures of one- and two-locus parameters were performed. The confidence limit gives the 5th and 95th percentiles of the distribution of all realized inferences. Note the logarithmic scale on the *y*-axis. See Tables 3 and 4 for a summary of these sets of simulations.

bution of N_e and m estimates in the following way: All simulations were run to generate a large number of observations, each of which consisted of a pair of N_e and m estimates. These observations were binned to a two-dimensional array of size 100 × 100. The bins did not overlap, had the same width, and were evenly distributed in both dimensions. All bin counts were standardized by the total number of counts in the two-dimensional array, and the discrete probability distribution was derived. Then, the cells were sorted in order of decreasing probability. Finally, starting from the cells with the highest associated probabilities, cells were sequentially added to the confidence region until the cumulative probability of the whole set of cells obtained was less than or equal to the preliminary fixed q-value. From this procedure, we obtained for each simulation a region within which a fraction q of the data lay. This confidence region was not constrained to be continuous.

ROBUSTNESS TO MODEL ASSUMPTIONS

Finite number of demes: The reliance on the infinite island model might be seen as a serious drawback of our method. However, developing the theory in a finite island model would necessitate as many as 28 distinct two-locus identity coefficients (instead of 10 here) and would be thus far more complicated. In contrast, we assumed that the number of demes had a small effect on our estimators. Therefore, we chose to test our method on data sets generated from a more realistic population model. For this purpose, we simulated a finite island model. Twenty subpopulations were used, each containing the same number of reproducing individuals and receiving the same number of immigrant haplotypes per generation. As in the infinite island model, a single subpopulation was repeatedly sampled.

In contrast to the infinite island model, migration *per se* cannot maintain polymorphism within a subpopulation in a finite island model. Therefore, in our simulations, mutations will arise at a sufficient rate to maintain polymorphism. We chose the mutation rate so that the product $dN_e\mu$ is at least equal to 1. With smaller $dN_e\mu$ values, not enough variation was maintained at equilibrium. In particular, we used two mutation rates, $\mu = 10^{-3}$ and $\mu = 5 \times 10^{-3}$, giving $dN_e\mu = 1$ and $dN_e\mu = 5$ (with $N_e = 50$ and $d = 20$). These two mutation rates remain in the range of realistic values, as given by ESTOUP and ANGERS (1998) for microsatellite markers.

As one might expect, the results depend on the level of genetic variation (Table 5). With $\mu = 10^{-3}$, the bias and variance of the N_e and m estimates are always larger than those obtained with the infinite island model. With $\mu = 5 \times 10^{-3}$, we obtained results that were in very close agreement with those obtained with the infinite island model. With only eight loci, bias and coefficient of variation can even be smaller with the simulations based on the finite island model. In all cases, however, the bias for N_e estimates was positive, although it was always negative when an infinite island model was simulated.

Figure 5 shows the joint distributions of N_e and m estimates over different numbers of loci and different

TABLE 4

Results from simulations for a *K* allele model

			Estimated $\hat{N}_e$				Estimated $\hat{m}$			Inference efficiency		
L	*K*	N_e	*H*	c.v.	95% C.I.	*m*	Mean	c.v.	95% C.I.	1	2	3
4	5	20	12.92	1.02	4.23–91.29	0.05	0.0743	0.81	0.0117–0.1899	1.2	9.5	89.3
		50	25.18	1.23	8.37–268.07	0.02	0.0384	0.93	0.0048–0.1081	0.1	26.2	73.7
		100	33.44	1.27	11.58–423.45	0.01	0.0294	1.06	0.0026–0.0835	0.0	36.1	63.9
	10	20	16.03	0.92	5.63–81.84	0.05	0.0603	0.72	0.0120–0.1404	0.9	2.7	96.4
		50	35.21	1.04	14.70–210.67	0.02	0.0286	0.78	0.0048–0.0657	0.0	16.5	83.5
		100	51.26	1.23	18.31–440.86	0.01	0.0191	0.82	0.0024–0.0488	0.0	28.2	71.8
8	5	20	17.49	0.77	7.41–65.3	0.05	0.0577	0.63	0.0160–0.1294	0.2	0.6	99.2
		50	41.36	1.03	16.27–248.39	0.02	0.0248	0.74	0.0042–0.0610	0.0	9.7	90.3
		100	64.02	1.14	25.25–489.45	0.01	0.0158	0.75	0.0023–0.0393	0.0	22.3	77.7
	10	20	17.80	0.57	8.34–47.65	0.05	0.0553	0.51	0.0197–0.1133	0.0	0.0	100.0
		50	45.57	0.80	19.60–179.6	0.02	0.0222	0.59	0.0056–0.0466	0.0	2.3	97.7
		100	80.23	1.01	36.02–460.98	0.01	0.0126	0.64	0.0022–0.0276	0.0	14.3	85.7
12	5	20	17.53	0.63	7.94–49.46	0.05	0.0570	0.56	0.0197–0.1173	0.0	0.0	100.0
		50	46.00	0.89	20.59–200.37	0.02	0.0218	0.62	0.0048–0.0449	0.0	3.1	96.9
		100	81.25	1.05	34.29–525.36	0.01	0.0127	0.72	0.0021–0.0304	0.0	14.5	85.5
	10	20	19.22	0.44	9.89–41.09	0.05	0.0538	0.45	0.0239–0.1001	0.0	0.0	100.0
		50	47.08	0.79	22.74–164.06	0.02	0.0214	0.52	0.0061–0.0424	0.0	0.9	99.1
		100	92.90	0.90	42.03–400.13	0.01	0.0108	0.62	0.0025–0.0235	0.0	7.8	92.2
16	5	20	18.87	0.48	9.26–44.62	0.05	0.0534	0.47	0.0239–0.1004	0.0	0.0	100.0
		50	46.83	0.75	23.65–148.73	0.02	0.0221	0.53	0.0067–0.0457	0.0	1.5	98.5
		100	87.19	1.00	39.74–430.81	0.01	0.0117	0.65	0.0021–0.0262	0.0	11.0	89.0
	10	20	18.62	0.41	10.09–37.61	0.05	0.0546	0.42	0.0248–0.0967	0.0	0.0	100.0
		50	48.60	0.52	25.83–117.28	0.02	0.0213	0.47	0.0077–0.0392	0.0	0.0	100.0
		100	91.16	0.88	44.05–370.74	0.01	0.0110	0.56	0.0026–0.0221	0.0	4.2	95.8

See Table 3 legend for details. *K* is the number of alleles simulated per locus.

TABLE 5

Results from simulations for various population models

	Estimated $\hat{N}_e$				Estimated $\hat{m}$			Inference efficiency		
$N_e m$	*H*	c.v.	95% C.I.	*m*	Mean	c.v.	95% C.I.	1	2	3
Simulations based on an infinite island model: $\mu = 10^{-6}$										
0.5	42.70	0.95	16.59–222.61	0.01	0.0117	0.68	0.0022–0.0273	0.0	5.4	94.6
1.0	47.79	0.68	24.16–133.76	0.02	0.0211	0.48	0.0069–0.0392	0.0	0.3	99.7
2.0	48.45	0.31	30.66–82.54	0.04	0.0420	0.32	0.0226–0.0669	0.0	0.0	100.0
Simulations based on a finite island model: $\mu = 10^{-3}$										
0.5	38.87	0.98	14.62–251.29	0.01	0.0166	1.21	0.0023–0.0446	0.2	13.7	86.1
1.0	38.95	1.01	14.87–226.16	0.02	0.0334	1.25	0.0044–0.0934	0.4	9.3	90.3
2.0	36.95	0.99	15.33–228.84	0.04	0.0633	1.12	0.0081–0.1913	3.0	10.0	87.0
Simulations based on a finite island model: $\mu = 5 \times 10^{-3}$										
0.5	56.72	0.77	27.00–219.13	0.01	0.0141	0.59	0.0032–0.0296	0.0	5.3	94.7
1.0	53.49	0.64	27.91–152.62	0.02	0.0258	0.51	0.0082–0.0500	0.0	0.7	99.3
2.0	50.32	0.44	30.43–97.78	0.04	0.0502	0.48	0.0219–0.0906	0.0	0.2	99.8

See Table 3 legend for details. For all sets of parameters, eight loci were scored among 50 sampled individuals. An infinite island model was assumed.

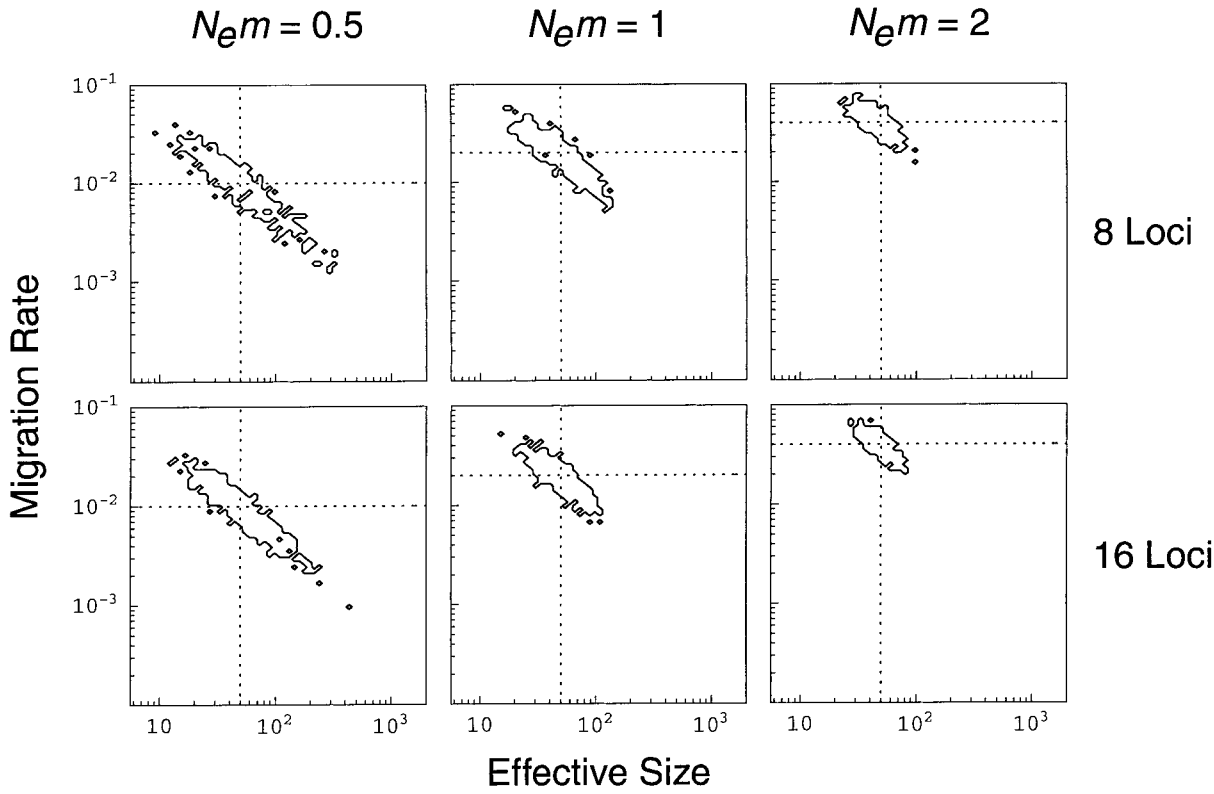


FIGURE 4.—Joint distribution of N_e and m estimates for various effective numbers of migrant individuals per generation and various numbers of loci. In all cases, the true local population size was fixed to 50. Local random mating was assumed, and an infinite allele model of mutation was used ($\mu = 10^{-6}$). For each set of parameters, 1000 estimates of one- and two-locus parameters were obtained (see text for details). The regions plotted are the 95% confidence regions for the sampling distribution of the estimators and were obtained directly from two-dimensional histograms of N_e and m estimates, as explained in the text. Dotted lines show the true values for the parameters. Note the logarithmic scale in both dimensions. See Table 5 for a summary of this set of simulations.

effective numbers of migrants per generation, for two different mutation rates. As in the infinite island model, we obtained better estimates with larger numbers of migrants per generation (N_em). With $\mu = 10^{-3}$, even with 16 loci, the distributions of pairwise N_e and m estimates are always broader than the distributions obtained with data generated from infinite island model simulations (to compare, see Figure 4). With $\mu = 5 \times 10^{-3}$ (Table 5, Figure 5B), the bias and variance were very close to the values obtained in the infinite island model. Bias and variance were also greatly reduced when 16 loci were scored (Figure 5). Increasing the sample size up to $n = 100$ also increases the estimation efficiency: more successful inferences are made, all of which give slightly less biased estimates with less variance.

Departure from equilibrium: Among the methods that aim at inferring population parameters, many rely on the hypothesis of equilibrium between mutation, migration, and drift. Our approach is no exception. To test whether our method was sensitive to recent departures from migration drift equilibrium, we conducted the same simulations as before with an infinite island model of population structure. But this time, individuals were sampled 10, 20, 30, 40, or 50 generations after the initial state. This process was repeated 1000 times.

Surprisingly, after only 50 generations, bias and variance of joint N_e and m estimates were as small as after 1000 generations (Figure 6). With eight loci sampled among 50 individuals, $\hat{N}_e = 52.33$ [coefficient of variation (c.v.) = 0.60] and $\hat{m} = 0.0216$ (c.v. = 0.45). Estimates were more biased and the variance of the distribution was larger for m than for N_e estimates. Moreover, after 30–40 generations, N_e estimates showed bias and variance as low as after 1000 generations. In all cases, the joint inference was successful in >99% of cases.

DISCUSSION

Effective population size is directly related to the asymptotic rate of coalescence for neutral genes. Thus, this population parameter determines the rate at which neutral genetic variation is lost from the population. Therefore, any attempt to provide a reliable estimate of effective population size should deserve careful evaluation. Our method is, to our knowledge, the first attempt for a joint estimation of (local) effective population size and immigration rate.

An advantage of using this method-of-moments to estimate N_e is that it requires only a single sample. Our estimates of effective population size are in general

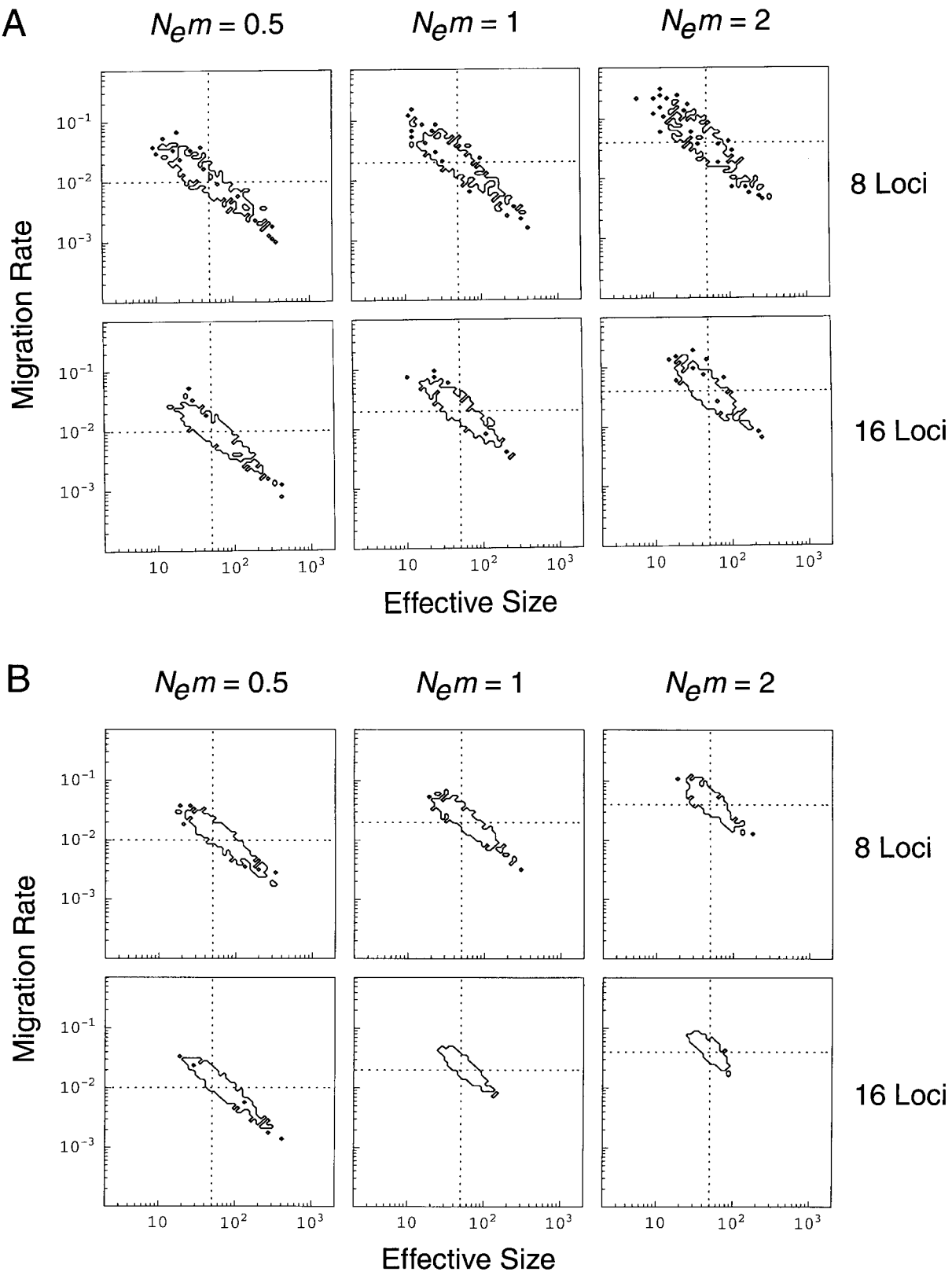


FIGURE 5.—Joint distribution of N_e and m estimates for different values of N_em and different numbers of loci. A finite island model was simulated with 20 demes of equal effective population sizes $N_e = 50$. The migration rate was set to $m = 0.02$. Each time, 50 diploid individuals within a single subpopulation were sampled. See Figure 4 for additional details. (A) $\mu = 10^{-3}$. (B) $\mu = 5 \times 10^{-3}$. Note the logarithmic scale on *x*- and *y*-axes. See Table 5 for a summary of this set of simulations.

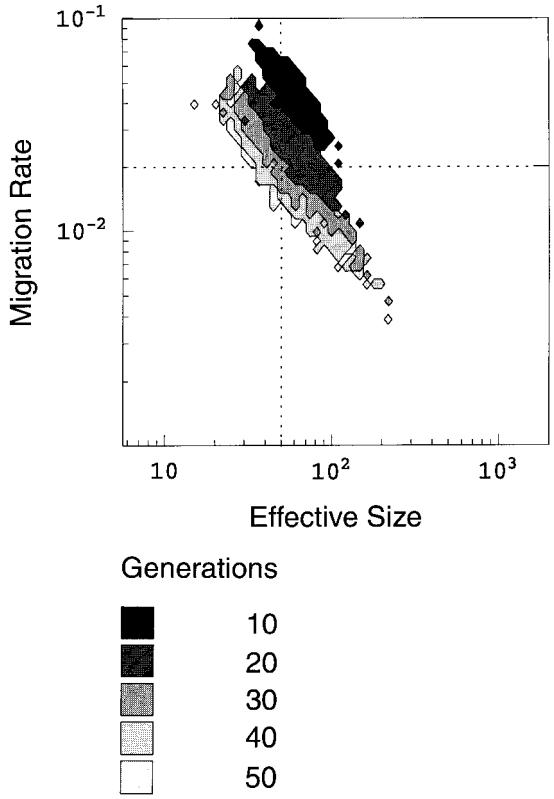


FIGURE 6.—Joint distribution of N_e and m estimates in non-equilibrium situations. An infinite island model was simulated with $N_e = 50$ and $m = 0.02$. Five data sets were obtained as follows. Starting from an initial state where all genes in the population are distinct, 50 individuals were sampled after 10, 20, 30, 40, or 50 generations. In each case, the process was repeated 1000 times. The 95% confidence regions for each data set are shown. An infinite allele model of mutation was used ($\mu = 10^{-6}$). Dotted lines show the true values for the parameters. Note the logarithmic scale on both axes.

slightly biased (Figure 3). The dispersion around the mean, as shown by the coefficient of variation as well as by the 95% confidence interval, is large and increases with the true population size (Figure 3, Table 3). Bias and variance are also decreased when the local immigration rate is increased (Figure 4). In this latter situation, the genetic diversity within the population is increased, as well as the gametic disequilibrium, making the estimation procedure more efficient.

However, we show that the bias and variance of effective population size estimates can be substantially reduced when $\hat{F}$ and $\hat{\eta}'_S$ are estimated over 8 loci or more in the infinite allele model, or 12 loci or more in K allele model (Figure 3, Table 4). Increasing the number of allelic states has also been shown to improve the estimation. We thus recommend using a large number of highly polymorphic loci for this method to be reliable. With the advances of molecular techniques in the last decade, this recommendation (using at least 8 highly polymorphic loci) is not unrealistic. Indeed, it is now common practice to work with at least 8–10 highly polymorphic loci, such as microsatellite markers. In all the results that we presented, the sample size was representative of those used in empirical studies found in the literature (WILLIAMSON and SLATKIN 1999). Of course, the joint estimates may also be improved with larger sample sizes. As we have shown, the reliance of our method on the infinite island model of population structure is not tremendous (Table 5, Figure 5). Although the mutation rates we used may seem to be high, they fall in the range of realistic values for microsatellite markers (ESTOUP and ANGERS 1998). When using unlinked loci, our method seems to give reasonably robust estimates even when recent changes in population structure or size have arisen (Figure 6). This may not be true, however, if one intends to use closely linked loci (to estimate larger effective sizes, for example).

We did not directly compare the results obtained with this method to other tentative estimations. The main reason is that alternative methods do not always estimate the same quantities. Another reason is that the number of required samples may be different. Variance of effective population size has been tentatively estimated from temporal changes in allele frequency (NEI and TAJIMA 1981; POLLAK 1983; WAPLES 1989). These estimates also exhibited high variance (WAPLES 1989). WILLIAMSON and SLATKIN (1999) recently proposed a maximum-likelihood method to estimate effective population size from temporal change in allele frequencies. This improvement of the method, although having smaller bias and lower variance, still exhibited a large variance. However, WILLIAMSON and SLATKIN (1999) stated that the maximum-likelihood estimate of population size from temporal change in allele frequency should be difficult to compute exactly when there are more than two alleles per locus. Maximum-likelihood methods for highly polymorphic loci could be implemented using Monte Carlo Markov chain computations.

PUDOVKIN *et al.* (1996) provided an estimate of the effective number of breeders in a population from the observed excess of heterozygotes in the progeny relative to Hardy-Weinberg expected proportions in the base population. Any heterozygote excess relative to the expected distribution for random mating populations results from a difference in allele frequencies among different gamete pools. Indeed, in finite diploid populations, a difference in the allele frequencies between male and female uniting gametes is expected from the sampling error due to the finite size of male and female gamete pools and from the difference in allele frequencies among male and female parents (WANG 1996). LUIKART and CORNUET (1999) further evaluated the accuracy and precision of this method. This procedure provides reliable estimates for very small numbers of breeders, but is sensitive to the mating system. Moreover, this method does not hold for consanguineous reproductive systems.

Effective population size has also been evaluated from

the measure of the variance of the correlation of allele frequencies between pairs of loci (Hill 1981; Waples 1991; Bartley *et al.* 1992). In an infinite isolated random mating population at mutation-drift equilibrium, the correlation of (neutral) allele frequencies at a pair of loci is zero. In a finite population, however, genetic drift causes the allele frequencies at independent loci to be correlated. Moreover, the magnitude of this correlation depends on the effective population size (Weir and Hill 1980; Hill 1981; Waples 1991). This method gave discouraging results and has not been evaluated through simulations.

Finally, none of these methods allow for migration or population subdivision. They assume single isolated populations that do not receive any migrant individuals from elsewhere. If this hypothesis were false, estimates of effective population size would be biased upward. Moreover, all previous attempts to estimate gene flow could not untangle the effect of local drift from immigration.

In conclusion, we address the issue of obtaining a confidence interval in practical cases. Resampling methods have their drawbacks. For example, the bootstrap does not seem to be adapted to correctly estimating second moments statistics. Estimates of one- and two-locus identity probabilities, which are second and fourth moments of allelic frequency, are obtained through the comparison of pairs of haplotypes in a sample. Random sampling with replacement within a sample increases the probability of sampling the same individuals twice. Therefore, bootstrapping over individuals indubitably overestimates the measures of identity among pairs of individuals. So, we suggest obtaining a confidence interval by means of stochastic simulations. Since we suppose the population to be at migration-drift equilibrium, we can simulate the stochastic process of dispersal and reproduction with estimated values of N_e and m used as input parameters. The distribution of new estimates could then be used to build a confidence interval.

We thank K. Dawson, G. Luikart, O. Hardy, I. Olivieri, and C. Garza for helpful comments on a previous draft of this manuscript. We thank M. Slatkin and two anonymous reviewers for constructive criticism of this manuscript. R.V. acknowledges financial support from the Training and Mobility of Researchers programme "FRAGLAND" of the European Commission, coordinated by I. Hanski, from contract number BIO4-CT96-1189 of the Commission of the European Communities (DG XII), and from the Fondation Sansouire. This is contribution 2000-099 of the Institut des Sciences de l'Évolution de Montpellier.

LITERATURE CITED

Avery, P. J., and W. G. Hill, 1979 Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. Genetics **91:** 817–844.

Bartley, D., M. Bagley, G. Gall and B. Bentley, 1992 Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. Conserv. Biol. **6:** 365–375.

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Caballero, A., 1994 Developments in the prediction of effective population size. Heredity **73:** 657–679.

Cockerham, C. C., 1984 Additive by additive variance with inbreeding and linkage. Genetics **108:** 487–500.

Cockerham, C. C., and B. S. Weir, 1977 Digenic descent measures for finite populations. Genet. Res. **30:** 121–147.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

Cockerham, C. C., and B. S. Weir, 1993 Estimation of gene flow from *F*-statistics. Evolution **47:** 855–863.

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA **81:** 6073–6077.

Crow, J. F., and C. Denniston, 1988 Inbreeding and variance effective population numbers. Evolution **42:** 482–495.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Burgess Publishing, Minneapolis, MN.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology*, edited by G. R. Carvalho. IOS Press, Amsterdam.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W. J., 1982 On the concept of effective population size. Theor. Popul. Biol. **21:** 373–378.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, London.

Gabriel, W., and R. Bürger, 1994 Extinction risk by mutational meltdown: synergistic effects between population regulation and genetic drift, pp. 69–84 in *Conservation Genetics*, edited by V. Loeschcke, J. Tomiuk and S. K. Jain. Birkhäuser Verlag, Basel.

Goodnight, C. J., 1987 On the effect of founder events on epistatic genetic variance. Evolution **41:** 80–91.

Goodnight, C. J., 1988 Epistasis and the effect of founder events on the additive genetic variance. Evolution **42:** 441–454.

Hanski, I., and M. E. Gilpin, 1997 *Metapopulation Biology: Ecology, Genetics, and Evolution*. Academic Press, San Diego.

Hedrick, P. W., 1987 Gametic disequilibrium: proceed with caution. Genetics **117:** 331–341.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genetics **38:** 200–216.

Hill, W. G., and B. S. Weir, 1994 Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. **54:** 705–714.

Lande, R., and B. S. Weir, 1994 Risk of population extinction from fixation of new deleterious mutations. Evolution **48:** 1460–1469.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Luikart, G., and J. M. Cornuet, 1999 Estimating the effective number of breeders from heterozygote excess in progeny. Genetics **151:** 1211–1216.

Lynch, M., and W. Gabriel, 1990 Mutation load and the survival of small populations. Evolution **44:** 1725–1737.

Lynch, M., J. Conery and R. Bürger, 1995a Mutation accumulation and the extinction of small populations. Am. Nat. **146:** 489–518.

Lynch, M., J. Conery and R. Bürger, 1995b Mutational meltdowns in sexual populations. Evolution **49:** 1067–1080.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité*. Masson, Paris.

Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **24:** 268–294.

Maynard Smith, J., and R. Hoekstra, 1980 Polymorphism in a varied environment: How robust are the models? Genet. Res. **35:** 45–57.

Nei, M., and F. Tajima, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

Nunney, L., and D. R. Elam, 1994 Estimating effective size of conserved populations. Conserv. Biol. **8:** 175–184.

Ohta, T., 1980 Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigene families. Genet. Res. **36:** 181–197.

Ohta, T., 1982a Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA **79:** 1940–1944.

Ohta, T., 1982b Linkage disequilibrium with the island model. Genetics **101:** 139–155.

Pollak, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

Pudovkin, A. I., D. V. Zaykin and D. Hedgecock, 1996 On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics **144:** 383–387.

Reynolds, J., B. S. Weir and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics **105:** 767–779.

Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357–1362.

Schwartz, M. K., D. A. Tallmon and G. Luikart, 1999 Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. Anim. Conserv. **2:** 321–323.

Slatkin, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

Slatkin, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–792.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

Slatkin, M., and N. H. Barton, 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

Slatkin, M., and W. P. Maddison, 1989 A cladistic measure of gene flow inferred from the phylogeny of alleles. Genetics **123:** 603–613.

Tachida, H., and C. C. Cockerham, 1986 Analysis of linkage disequilibrium in an island model. Theor. Popul. Biol. **29:** 161–197.

Tufto, J., S. Engen and K. Hindar, 1996 Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics **144:** 1911–1921.

van Noordwijk, A. J., 1994 The interaction of inbreeding depression and environmental stochasticity in the risk of extinction of small populations, pp. 131–146 in *Conservation Genetics*, edited by V. Loeschcke, J. Tomiuk and S. K. Jain. Birkhäuser Verlag, Basel.

Vitalis, R., and D. Couvet, 2001 Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. Genet. Res. (in press).

Vucetich, J. A., T. A. Waite and L. Nunney, 1997 Fluctuating population size and the ratio of effective to census population size. Evolution **51:** 2017–2021.

Wang, J., 1996 Deviation from Hardy-Weinberg proportions in finite populations. Genet. Res. **68:** 249–257.

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

Waples, R. S., 1991 Genetic methods for estimating the effective size of cetacean populations, pp. 279–300 in *Genetic Ecology of Whales and Dolphins*, Special Issue 13, edited by A. R. Hoelzel. International Whale Commission, London.

Weir, B. S., and C. C. Cockerham, 1969 Group inbreeding with two linked loci. Genetics **63:** 711–742.

Weir, B. S., and C. C. Cockerham, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Weir, B. S., and W. G. Hill, 1980 Effect of mating structure on variation in linkage disequilibrium. Genetics **93:** 477–488.

Whitlock, M. C., and N. H. Barton, 1997 The effective size of a subdivided population. Genetics **146:** 427–441.

Whitlock, M. C., P. C. Phillips and M. J. Wade, 1993 Gene interaction affects the additive genetic variance in subdivided populations with migration and extinction. Evolution **47:** 1758–1769.

Williamson, E. G., and M. Slatkin, 1999 Using maximum likelihood to estimate size from temporal changes in allele frequencies. Genetics **152:** 755–761.

Wright, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

Wright, S., 1940 Breeding structure of population in relation to speciation. Am. Nat. **74:** 232–248.

Communicating editor: M. Slatkin

Annexe G

Estim 1.0

Vitalis, R. and Couvet, D. (2001) ESTIM 1.0 : a computer program to infer population parameters from one- and two-locus gene identity probabilities. *Molecular Ecology Notes* **1** : 354-356.

Molecular Ecology Notes (2001) **1**, 354–356

PROGRAM NOTE

ESTIM 1.0: a computer program to infer population parameters from one- and two-locus gene identity probabilities

R. VITALIS*† and D. COUVET‡

*Laboratoire Génétique et Environnement, C.C. 065, ISEM (UMR 5554), Université Montpellier II, 34095 Montpellier Cedex 05, France, †Station Biologique de la Tour du Valat, 13200 Arles, France, ‡CRBPO — Museum National d'Histoire Naturelle, 75005 Paris, France

Abstract

Estimating effective population size is an important issue in population and conservation genetics. Recently, we proposed a new method to infer effective size and migration rate from one- and two-locus identity probability measures. We now announce the release of a user-friendly Microsoft® Windows program that uses this method to provide joint estimates of local effective population size and immigration rate for each subpopulation in a population genetics data set.

Keywords: effective size, identity disequilibrium, migration rate

Received 6 February 2001; revision accepted 4 April 2001

The distribution of neutral genetic variation in natural populations is shaped by history and by demography, as well as by the degree to which populations are subdivided and isolated. In particular, the effective population size of local populations, as well as the proportion of novel variants that successfully immigrate in a population, give the long-term rate at which pairs of neutral genes coalesce. Estimating these parameters is therefore a major, but still difficult, task in population and conservation genetics (Schwartz *et al*. 1999).

Previous attempts have been made to infer effective population size from temporal changes in allele frequency (Pollak 1983; Waples 1989; Williamson & Slatkin 1999), from heterozygote excess in progeny (Pudovkin *et al*. 1996; Luikart & Cornuet 1999) and from linkage disequilibrium (Hill 1981; Waples 1991; Bartley *et al*. 1992). All but the latter method rely on one locus theory. In other words, all but the latter method take measures of genetic variation at distinct loci as independent and identically distributed. However, in small-sized populations, random genetic drift creates the conditions for significant associations of gene copies to occur across loci (Hill & Robertson 1968). In a previous contribution, we defined an appropriate parameter, the 'within-subpopulation identity disequilibrium' (noted η_s), which describes the extent to which gene identities are correlated across loci (Vitalis & Couvet 2001b). We also provided the expected value of this parameter in a partially selfing subdivided population, as a function of (local) effective population size, migration rate, selfing proportion and recombination fraction (Vitalis & Couvet 2001b). From this theory, we further proposed a new method to infer (local) effective population size and migration rate from one- and two-locus genetic data (Vitalis & Couvet 2001a). The rationale is a method-of-moments approach that provides the joint estimates of N_e and m from estimates of F (the one-locus averaged correlation of genes within the focal subpopulation, relative to the whole population) and η_s (the two-locus averaged identity disequilibrium within the focal subpopulation, relative to the whole population). We showed that this method performed well in estimating local effective population size and immigration rate for a variety of sampling conditions (see Vitalis & Couvet 2001a for extensive simulation results). Here, we provide a user-friendly program that calculates the multilocus estimates of the parameters F and η_s, as discussed in Vitalis & Couvet (2001a), for each subpopulation in a multilocus data set. This interface has been named ESTIM, an acronym standing for 'effective size from two-locus identity measures'.

Correspondence: R. Vitalis. Laboratoire Génétique et Environnement, C.C. 065, ISEM (UMR 5554), Université Montpellier II, 34095 Montpellier Cedex 05, France. Fax: + 33 (0) 4 67 14 36 22; E-mail: vitalis@isem.univ-montp2.fr

© 2001 Blackwell Science Ltd

Table 1 Evaluation of the reliability of nonparametric confidence intervals at 95%, provided for each point estimate. See the main text for the details concerning the stochastic simulations. Estimated N_e is the harmonic mean of all reliable point estimates. For each set of parameters, the lower (resp. upper) bound of the mean confidence interval is the harmonic mean of all reliable lower (resp. upper) bounds of point confidence intervals. The inference efficiency is assessed with special reference to three distinct cases. Case 1 gives the percentage of negative η_s estimates, which gave infinite effective population sizes. Case 2 gives the percentage of cases where the true effective size (N_e) lies outside the confidence interval given for each point estimate. Case 3 gives the percentage of reliable estimates (estimates are averaged over Case 2 and Case 3)

Number of loci	True N_e	Estimated N_e	Mean Confidence Interval		Inference efficiency		
			Lower bound	Upper bound	1	2	3
8	20	19.20	9.04	37.94	0.4	19.9	79.7
	50	46.17	19.35	92.23	6	10.7	83.3
	100	76.50	30.26	143.33	18.6	5.8	75.6
12	20	19.64	11.93	32.55	0	27.2	72.8
	50	48.35	24.77	94.62	2.5	14.8	82.7
	100	87.63	40.21	161.88	11	8.8	80.2
16	20	20.25	13.95	29.93	0	35.7	64.3
	50	51.10	30.33	91.72	0.6	23.4	76.0
	100	97.16	47.89	188.72	4.6	8.4	87.0

This program computes a confidence interval for each estimate, at any α-level chosen by the user. Confidence intervals are obtained by deriving the approximate bootstrap confidence intervals (ABC) method described by DiCiccio & Efron (1996). This procedure is an analytic version of the bias-corrected and accelerated (BC_a) algorithm for producing confidence limits from bootstrap distributions, that applies to nonparametric problems (DiCiccio & Efron 1996). The confidence interval endpoints are analytically approximated, rendering Monte Carlo simulations unnecessary. This method, as do all re-sampling approaches, requires a careful examination and understanding of the nature of the observed data. Weir (1996) advocated the bootstrapping to be carried out over loci. This implies that each locus gives an independent realization of the process of genetic sampling, which depends on some population parameters of interest (effective size, migration rate, number of demes, …). In our case, since we are interested in evaluating a measure of association of genes across loci, the latter hypothesis is clearly unrealistic. In contrast, each pair of loci gives the proper unit to be re-sampled. Because the multilocus estimate of within-subpopulation identity disequilibrium is defined as the sum of all pairwise identity disequilibria over the sum of all pairwise heterozygosity products (Vitalis & Couvet 2001a), all possible pairwise estimates of the parameters of interest shall be taken as the appropriate units of re-sampling. Concerning the one-locus parameters F, the observed data thus consist of pairwise averages, over pairs of loci, for all possible pairs. Concerning the within-subpopulation identity disequilibria (η_s), the observed data consist of two-locus estimates for all possible pairs of loci. This procedure amounts to bootstrapping pairs of loci for the whole data set. From F and η_s confidence intervals endpoints, N_e and m confidence intervals endpoints are deduced following the same inference procedure as described in Vitalis & Couvet (2001a) for point estimates.

We performed stochastic simulations to test whether this procedure was efficient in determining confidence intervals. We simulated a finite island model with $d = 20$ subpopulations of N_e individuals, each receiving $N_e \cdot m = 1$ effective migrant per generation. Random mating was assumed and the loci were assumed to be (physically) unlinked. Mutations occurred at rate μ, with $d \cdot N_e \cdot \mu = 1$, and followed the infinite allele model. For each set of parameters, after 1000 initial generations, 50 individuals per population were sampled in 10 randomly chosen subpopulations. This was repeated 100 times, every 100 generations. The results (averaged over replicate subpopulations) are summarized in Table 1. We would expect that, for a given set of parameters, the true (simulated) N_e value is contained in the estimated confidence interval in 95 cases out of 100. We observed the confidence interval to contain the true value in 64.3–87.0% of cases (Table 1). This may arise from the fact that, although the confidence intervals for F and η_s contain the true values of these two parameters in 95% of the time, linear transforms of F and η_s confidence interval endpoints are not expected to contain the true values of N_e and m in the same proportion of cases. However, the results were found to be satisfactory in more than two-thirds of simulations.

As already discussed elsewhere (Vitalis & Couvet 2001a), the dispersion around the mean of effective population size estimates decreases when the 'true' population size increases. Besides any consideration about the variance of the estimates, it should be kept in mind that large effective sizes are difficult to estimate because the amplitude of variation of η_s with N_e decreases as N_e increases

(Vitalis & Couvet 2001a,b). Moreover, the efficiency of estimation varies with population genetic diversity (Vitalis & Couvet 2001a). For unlinked loci, our moment-based approach should thus provide reliable estimates of effective population size, provided the 'true value' of this parameter is not of too high an order. Tightly linked loci are required to obtain reliable estimates for large population sizes (see also Hill & Weir 1994).

The software opens input files in GENEPOP format (Raymond & Rousset 1995). Some loci and/or some subpopulations may be excluded from the analysis (at least two loci and two subpopulations are required). By default, all subpopulations and all loci are accounted for. The values of the model parameters (mating system, recombination fraction, mutation rate) may be fixed by the user. Output files are in text format. The number of loci is limited to 100 and the number of populations to 100. For each population, the sample size can not exceed 1000 individuals. The software can be freely downloaded by anonymous FTP at ftp://isem.isem.univ-montp2.fr/pub/pc/estim. This software was programed using Borland® Delphi Version 4.0 (Inprise Corporation) on a PC computer running with operating system Microsoft® Windows NT 4.0 (Build 1381: Service Pack 6, 130 420 Kb RAM available).

Acknowledgements

We thank K. Belkhir and N. Raufaste for programing tips, as well as I. Olivieri for supporting this project, M. Hochberg for commenting this manuscript and C. Petit for testing the software. R.V. acknowledges financial support from the European Training and Mobility of Researchers program 'FRAGLAND' allocated to I. Hanski, and from the Fondation Sansouire. This is publication 2001–037 from the Institut des Sciences de l'Evolution de Montpellier.

References

Bartley D, Bagley M, Gall G, Bentley B (1992) Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. *Conservation Biology*, **6**, 365–375.

DiCiccio TJ & Efron B (1996) Bootstrap confidence intervals. *Statistical Science*, **11**, 189–228.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetics*, **38**, 200–216.

Hill WG & Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Hill WG & Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, **54**, 705–714.

Luikart G & Cornuet J-M (1999) Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, **151**, 1211–1216.

Pollak E (1983) A new method for estimating the effective population size from allele frequency changes. *Genetics*, **104**, 531–548.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144**, 383–387.

Raymond M & Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenism. *Journal of Heredity*, **86**, 248–249.

Schwartz MK, Tallmon DA, Luikart G (1999) Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. *Animal Conservation*, **2**, 321–323.

Vitalis R & Couvet D (2001a) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, **157**, 911–925.

Vitalis R & Couvet D (2001b) Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genetical Research*, **77**, 67–81.

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

Waples RS (1991) Genetic methods for estimating the effective size of cetacean populations. In: *Genetic Ecology of Whales and Dolphins* (ed. Hoelzel AR) Special Issue 13, pp. 279–300. International Whale Commission, London.

Weir BS (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland.

Williamson EG & Slatkin M (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, **152**, 755–761.

Annexe H

Différenciation sexe-spécifique

Vitalis R. (2002) Sex-specific genetic differentiation and coalescence times : Estimating sex-biased dispersal rates. *Molecular Ecology* **11** : 125-138.

Molecular Ecology (2002) **11**, 125–138

Sex-specific genetic differentiation and coalescence times: estimating sex-biased dispersal rates

RENAUD VITALIS*†
Laboratoire Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier (UMR 5554), Université Montpellier II, F-34095 Montpellier Cedex 05, France, †Station Biologique de la Tour du Valat, Le Sambuc, F-13200 Arles, France

Abstract

I derive the equilibrium values of sex-specific F_{ST} parameters, in an island model for a dioecious species with sex-biased dispersal and binomial distribution of family size before dispersal (as assumed in a Wright–Fisher population). I show that F_{ST} may take different values among males and among females whenever dispersal is a trait conditioned on gender. This has not always been recognized, because some models assumed that genes are sampled before dispersal. In particular, the ratios of sex-specific F_{ST} parameters evaluated after dispersal over F_{ST} evaluated before dispersal are simple functions of sex-specific dispersal rates. Therefore, a simple moment-based estimator of sex-specific dispersal rate is proposed. This method is based on the comparison of F_{ST} estimated before and after dispersal and assumes equilibrium between migration and drift. I evaluate this method through stochastic simulations for a range of sex-specific dispersal rates and sampling effort (sample size, number of loci scored).

Keywords: coalescence times, *F*-statistics, island model, population genetics, sex-biased dispersal

Received 1 March 2001; revision received 30 August 2001; accepted 13 September 2001

Introduction

It has long been recognized that natal dispersal, the movement of individuals from their birth place to the place where they aim to reproduce, may be sex-biased. Females generally disperse more than males among bird species, while the reverse is usually true among mammals (Greenwood 1980). There may be several reasons for this. Motro (1991, 1994) showed that an unbiased dispersal strategy is not evolutionary stable as long as the cost of inbreeding is high and the cost of dispersal is low. In such conditions, sex-biased dispersal strategies are selected for (Gandon 1999). If the dispersal costs are the same for both sexes, the equilibrium sex-specific dispersal rates are driven by initial conditions and genetic drift. Once one sex disperses more than the other, a runaway process is engaged and selection favours the former to disperse still more. This suggests that phylogenetic constraints may determine which sex actually disperses (Perrin & Mazalov 1999). Yet, if males and females do not equally suffer from local competition, they may pay different costs (or gain different benefits) for dispersal (Greenwood 1980) and sex-biased dispersal may evolve. This is what arises with some breeding systems, as observed among mammals (Dobson 1982), in which competition for mates (along with inbreeding avoidance) is thought to be the most important factor responsible for the evolution of sex-biased dispersal.

Evaluating the evolutionary outcomes of sex-biased dispersal requires to test these predictions across species and therefore to assess the existence and the magnitude of sex-biased dispersal. This addresses the question of the detection of sex-biased dispersal in the field. Direct estimates of dispersal distributions (through mark–recapture or radio-tracking) are often difficult to obtain in natural situations. Whatever the technique employed, long distance dispersal may be underestimated because of the area limitation of the study sites (Koenig *et al*. 1996). Moreover, all of these methods aim at measuring the actual distributions of individual movements, rather than to measure the effective rates at which individuals successfully reproduce outside their birth place. With the advance in genetic techniques, there may be new opportunities to obtain indirect estimates of sex-specific dispersal. Indeed, the level and the distribution

Correspondence: Renaud VITALIS. Laboratoire Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier (UMR CNRS 5554), Université Montpellier II, Place Eugène Bataillon, F-34095 Montpellier Cedex 05, France. Fax: +33 4 67 14 36 22; E-mail: vitalis@isem.univ-montp2.fr

© 2002 Blackwell Science Ltd

of dispersal plays a major role in determining the genetic structure of natural populations.

Kawata (1985) attempted to detect sex differences in the distribution of genetic variation at allozyme loci in the red-backed vole, *Clethrionomys rufocanus bedfordiae*. He found no significant difference in allele frequencies among sexes, but showed a tendency for allele distributions to be more heterogeneous in females than in males, a result consistent with a male-biased dispersal pattern. Favre *et al*. (1997) used highly polymorphic microsatellite markers to detect sex-biased dispersal patterns in the white-toothed shrew *Crocidura russula*, in which dispersal is female-biased. In their approach, an assignment index (Paetkau *et al*. 1995) was calculated for each individual's genotype as its expected frequency in the population in which it was trapped, assuming local random mating and linkage equilibrium. Favre *et al*. (1997) found that the distribution of female assignment indices had lower mean and larger variance than the male distribution. Thus, females had a lower probability of being born locally, a result expected with female-biased dispersal. This approach was also used with the white-footed mouse *Peromyscus leucopus*, in which dispersal is male-biased (Mossman & Waser 1999). These authors found that males had significantly lower assignment indices than females.

There has been several attempts to detect sex-biased gene flow from *F*-statistics (McCracken 1984; Rassmann *et al*. 1997; Mossman & Waser 1999). In the white-toothed shrew *C. russula*, Balloux *et al*. (1998) found that the 'coancestry' (the correlation of gene frequencies within breeding groups relative to gene frequencies within subpopulations, see Chesser 1991b) was higher among males than among females. The coancestry was significantly different from zero for males whereas it was not for females.

Sex-biased dispersal may also cause the distribution of genetic variation to differ between maternally inherited mitochondrial DNA (mtDNA) markers and biparentally inherited nuclear markers. In the widely distributed sperm whale *Physeter macrocephalus*, Lyrholm *et al*. (1999) showed the absence of genetic differentiation for microsatellite markers between oceans, while the genetic differenti-ation for mtDNA markers was significant (Lyrholm & Gyllensten 1998). This was interpreted as a direct consequence of greater movements of males across oceans, consistent with the observations that females are more philopatric than males in this species (see also Pardini *et al*. 2001; for a similar differentiation pattern in the great white shark *Carcharodon carcharias*). In contrast, FitzSimmons *et al*. (1997) found that mtDNA haplotype frequencies were not different between males and females of the marine turtle *Chelonia mydas* and that F_{ST} estimates among the study sites were large and similar for both sexes. This was understood as the consequence of male and female philopatry.

On closer examination, however, few studies acknowledge the fact that gender-biased dispersal detection may be unexpected in certain situations. For example, Rassmann *et al*. (1997) noticed that if the distribution of genetic variation differed among males and females because of the presence of male immigrants, this difference should not be noticeable in the next generation. So, in which condition may one expect sex-biased population differentiation? I will argue that the ability to detect contrasted population differentiation among sexes depends on whether genes are sampled before or after dispersal.

In this paper, I define sex-specific probabilities of identity for pairs of genes taken at different hierarchical levels and derive the recursion equations that are used to obtain the equilibrium values of those probabilities. I also define sex-specific *F*-statistics parameters as appropriate ratios of identity probabilities (Cockerham & Weir 1987; Rousset 1996). I show that F_{ST} estimates may differ among males and females, if dispersal rates are sex-specific and if genes are sampled after dispersal. I also show that the ratios of sex-specific F_{ST}'s estimated after dispersal over F_{ST} estimated before dispersal are simple functions of sex-specific dispersal rates. Therefore, a new method is proposed to infer sex-specific dispersal rates from allele frequency data collected before and after dispersal. This method is evaluated through stochastic simulations.

Equilibrium values of sex-specific *F*-statistics

F-statistics are defined as intraclass correlations for the probability of identity in state (IIS correlations) (Cockerham & Weir 1987; Rousset 1996). Yet, the infinite allele model (IAM) provides the value of the probability of identity by descent (IBD probabilities) and in the low mutation rate limit, for two given classes, IIS and IBD correlations converge to the same value (Slatkin 1991; Rousset 1996). Thus, the properties of *F*-statistics can be deduced from the properties of intraclass correlations for IBD probabilities (Rousset 1996).

Let Q_0 be the probability that the two copies of a gene within an individual after dispersal are identical by descent, whatever its sex. Let Q_1^{XY} be the IBD probability of two gene copies sampled within the same subpopulation from two individuals of sex X and Y after dispersal, whatever $X \in \{\male, \female\}$ and $Y \in \{\male, \female\}$ (so that two individuals of sex X and Y may be two males, two females, or one male and one female). Let Q_2^{XY} be the corresponding probability for two gene copies sampled among distinct subpopulations, after dispersal. Let me now define the corresponding IBD probabilities for pairs of genes sampled before dispersal. Since the IBD probabilities of pairs of genes sampled before dispersal take the same values among males and among females (Wang 1997), there is no need to consider distinct coefficients for different pairs of individuals of the same or opposite sex. Let Q_0^* be the IBD probability of two copies of a gene sampled before

dispersal within an individual, whatever its sex. Let Q_1^* be the corresponding probability for two gene copies sampled within the same subpopulation from two individuals before dispersal, and Q_2^* the IBD probability of two gene copies sampled among subpopulations before dispersal.

I consider a finite island model of population structure (Wright 1951). Dispersal is achieved by juvenile organisms, before reproduction. A migrant individual in a subpopulation is equally likely to come from any of the $(n-1)$ other subpopulations. Generations do not overlap. Let m_X denote the probability that an individual of sex X has immigrated. Each generation, after migration, the frequency of pairs of individuals taken at random in one subpopulation that come from a single subpopulation before migration is $a_{XY} = (1-m_X)(1-m_Y) + m_X \cdot m_Y/(n-1)$, for pairs of individuals of sex X and Y, with $X \in \{\sigma, \varphi\}$ and $Y \in \{\sigma, \varphi\}$. Conversely, the frequency of pairs of individuals taken at random from two subpopulations after migration that originate from the same subpopulation before migration is $b_{XY} = (1-a_{XY})/(n-1)$. All subpopulations have the same size N. The sex ratio is unbiased, so that each generation, in each subpopulation, $N/2$ males are equally likely to mate with $N/2$ females. In each subpopulation, the number of offspring produced is assumed to be very large (infinite) and the random number L_{XY} of surviving offspring of sex X descending from a parent of sex Y is binomially distributed as $L_{XY} = \mathrm{b}\left(\frac{N}{2}, \frac{2}{N}\right)$. Therefore, an offspring is equally likely to descend from any of the $N/2$ parents of sex X (Nagylaki 1995). Mutations (IAM) arise during meiosis at rate μ and $\gamma = (1-\mu)^2$ is the probability that two genes have not undergone mutation in one generation.

As dispersal is achieved through the movement of individuals, and as genes are sampled before local reproduction, the IBD probability of a pair of genes taken within an individual after dispersal equals that probability before dispersal, i.e.

$$Q_0(t+1) = Q_0^*(t+1) \tag{1}$$

Before dispersal, the IBD probability of a pair of genes taken within an individual equals, baring mutation, the IBD probability of a pair of genes taken among its parents. Therefore,

$$Q_0^*(t+1) = \gamma Q_1^{\sigma\varphi}(t) \tag{2}$$

The genes sampled among individuals in a subpopulation at generation $(t+1)$ come from the same subpopulation before dispersal with probability a_{XY} and from different subpopulations with probability $(1-a_{XY})$. These genes are IBD with probability $Q_1^*(t+1)$ and $Q_2^*(t+1)$, respectively. Thus,

$$Q_1^{XY}(t+1) = a_{XY}Q_1^*(t+1) + (1-a_{XY})Q_2^*(t+1) \tag{3}$$

Now, the within-subpopulation IBD probability $Q_1^*(t+1)$ for two genes taken at random before dispersal from two offspring depends on whether these genes come from male or female gametes. With probability 1/4, the two genes are issued from male (*resp.* female) gametes, and descend from the same father (*resp.* mother) in the previous generation with probability $2/N$. In this case, the two gene copies are IBD with probability $\gamma[Q_0(t)+1]/2$. These two gametes can also come from two different fathers (*resp.* mothers) with probability $(1-2/N)$ and then are IBD with probability $\gamma Q_1^{\sigma\sigma}(t)$[*resp.* $\gamma Q_1^{\varphi\varphi}(t)$]. With probability 1/2, one of the sampled genes comes from a male gamete while the other comes from a female gamete. Therefore, these two genes cannot descend from the same parent, and are IBD with probability $\gamma Q_1^{\sigma\varphi}(t)$. Thus,

$$Q_1^*(t+1) = \gamma\left[\frac{Q_0(t)+1}{2N} + \left(1-\frac{2}{N}\right)\frac{Q_1^{\sigma\sigma}(t) + Q_1^{\varphi\varphi}(t)}{4} + \frac{Q_1^{\sigma\varphi}(t)}{2}\right] \tag{4}$$

The genes sampled among subpopulations come from the same subpopulation before dispersal with probability b_{XY} and from different subpopulations with probability $(1-b_{XY})$. Again, these genes are IBD with probability $Q_1^*(t+1)$ and $Q_2^*(t+1)$, respectively. Thus,

$$Q_2^{XY}(t+1) = b_{XY}Q_1^*(t+1) + (1-b_{XY})Q_2^*(t+1) \tag{5}$$

Now, a pair of genes sampled before dispersal among subpopulations comes from two male gametes, two female gametes or one male and one female gametes in proportion 1/4, 1/4 and 1/2, respectively. These genes are IBD with probability $\gamma Q_2^{\sigma\sigma}(t)$, $\gamma Q_2^{\varphi\varphi}(t)$ and $\gamma Q_2^{\sigma\varphi}(t)$, respectively. Therefore,

$$Q_2^*(t+1) = \gamma\left[\frac{Q_2^{\sigma\sigma}(t) + 2Q_2^{\sigma\varphi}(t) + Q_2^{\varphi\varphi}(t)}{4}\right] \tag{6}$$

Putting eqns 1–6 together, we can obtain the recurrence equations that give the IBD probabilities of pairs of genes sampled after dispersal. In Appendix I, I rewrite eqns 1–6 in a matrix form, and give the expressions for equilibrium solutions of IBD probability vectors. Below, I define sex-specific *F*-statistics as appropriate ratios of sex-specific IBD probabilities (see Cockerham & Weir 1987; Rousset 1996)

$$\begin{aligned} F_{\mathrm{IS}}^{XY} &= \frac{Q_0 - Q_1^{XY}}{1 - Q_1^{XY}} \\ F_{\mathrm{ST}}^{XY} &= \frac{Q_1^{XY} - Q_2^{XY}}{1 - Q_2^{XY}} \\ F_{\mathrm{IT}}^{XY} &= \frac{Q_0 - Q_2^{XY}}{1 - Q_2^{XY}} \end{aligned} \tag{7}$$

From the above definitions, we have the following relationship

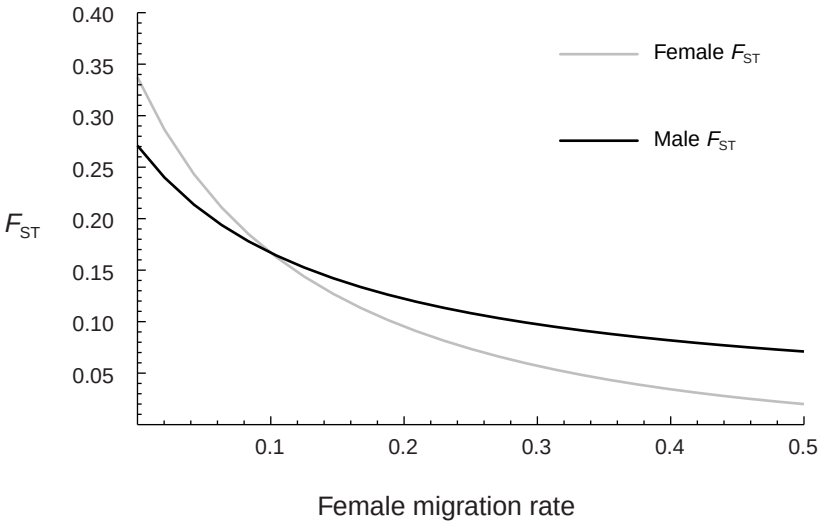


Fig. 1 Equilibrium values of sex-specific *F*-statistics. $F_{ST}^{♂♂}$ and $F_{ST}^{♀♀}$ are plotted against female dispersal rate and the male dispersal rate is constant. $2N = 20$, $n = 20$, $Nm_{♂} = 1$ and $\mu = 10^{-6}$.

$$(1 - F_{IT}^{XY}) = (1 - F_{IS}^{XY})(1 - F_{ST}^{XY}) \text{ for all } (X, Y) \in \{♂, ♀\}^2 \quad (8)$$

The equilibrium values of IBD probabilities derived from eqn A1.3 give the following equilibrium values of IBD correlations defined in eqn 7

$$F_{IS}^{XY} = \frac{-\gamma\left[a_{XY} - \gamma(a_{♂♀} - \bar{b})\right]}{2N(1 - \gamma\bar{d}) + \gamma\left[a_{X'Y'} - \gamma(a_{♂♀} - \bar{b})\right]} \quad (9)$$

and

$$F_{ST}^{XY} = \frac{\gamma d_{XY}}{2N(1 - \gamma\bar{d}) + \gamma(a_{♂♂} + a_{♀♀} - b_{XY}) - \gamma^2(a_{♂♀} - \bar{b})} \quad (10)$$

with $\bar{a} = (a_{♂♂} + 2a_{♂♀} + a_{♀♀})/4$, $\bar{b} = (b_{♂♂} + 2b_{♂♀} + b_{♀♀})/4$ and $\bar{d} = \bar{a} - \bar{b}$. We also note with a prime (e.g. in X') the alternate sex of X, i.e. $X' = ♂$ if $X = ♀$ and $X' = ♀$ if $X = ♂$. With dispersal independent of sex, i.e. $m_{♂} = m_{♀} = m$, eqns 9 and 10 reduce to

$$F_{IS} = \frac{-\gamma(a - \gamma d)}{2N(1 - \gamma d) + \gamma(a - \gamma d)} \quad (11)$$

$$F_{ST} = \frac{\gamma d}{2N(1 - \gamma d) + \gamma\left[b + (2 - \gamma)d\right]} \quad (12)$$

where $a = (1 - m)^2 + m^2/(n - 1)$, $b = (1 - a)/(n - 1)$ and $d = a - b$. This is a slightly modified version of the model considered by Rousset (1996). Figure 1 depicts the equilibrium male and female population differentiation, as given by the values of $F_{ST}^{♂♂}$ and $F_{ST}^{♀♀}$. In this figure, male migration rate is fixed ($m_{♂} = 0.1$) and sex-specific F_{ST}'s are plotted against female migration rate. The two curves cross at $m_{♀} = m_{♂} = 0.1$. With females dispersing less than males, $F_{ST}^{♀♀} > F_{ST}^{♂♂}$. The inverse is true whenever females disperse more than males.

Distribution of coalescence time

Although the relationship of sex-specific F_{ST}'s to sex-specific migration rates (see Fig. 1) is obviously sound, the further examination of the properties of sex-specific coalescence times may emphasize this result. There is a strict relationship between IBD probabilities and coalescence times (Slatkin 1991; Rousset 1996). The probability of identity of any pair of genes is the probability that neither gene has undergone mutation since they coalesced, i.e. since their most recent common ancestor (Hudson 1990). In Appendix II, I derived the expressions for sex-specific coalescence probabilities (i.e. the coalescence probabilities conditional upon the gender of sampled individuals).

I computed eqn A2.3 over time to obtain the distribution of coalescence probabilities $C_h^{XY}(t)$ (see Appendix II) for pairs of genes sampled within or among individuals of the same or different sex, within or among subpopulations (Fig. 2). Note that the abscissas give the number of generations going backward in time (i.e. the number of generations *elapsed* before present). In Fig. 2, the females migrate at a rate a hundred times superior to that of the males. The probabilities of coalescence for genes sampled within subpopulations (within or among individuals) coincide whatever the sex of the sampled individuals. Yet, the probabilities of coalescence for genes sampled among subpopulations differ among males and females in the early generations ($t < 50$ generations). More precisely, the probability that a pair of genes taken in different subpopulations coalesce in a female is higher than the probability that such a pair of genes coalesce in a male. Therefore, the mean coalescence time of genes taken in females from different subpopulations $\left[T_2^{♀♀} = \sum_{t=1}^{\infty} t \cdot C_2^{♀♀}(t)\right]$ is smaller than the mean coalescence time of genes taken in males from different subpopulations $\left[T_2^{♂♂} = \sum_{t=1}^{\infty} t \cdot C_2^{♂♂}(t)\right]$. Also, in the low mutation limit, $F_{ST}^{XY} \approx (T_2^{XY} - T_1^{XY})/T_2^{XY}$, where $T_h^{XY} = \sum_{t=1}^{\infty} t \cdot C_h^{XY}(t)$ is the mean coalescence time of genes taken h steps apart in the hierarchy (Appendix II; see Slatkin 1991 and Rousset

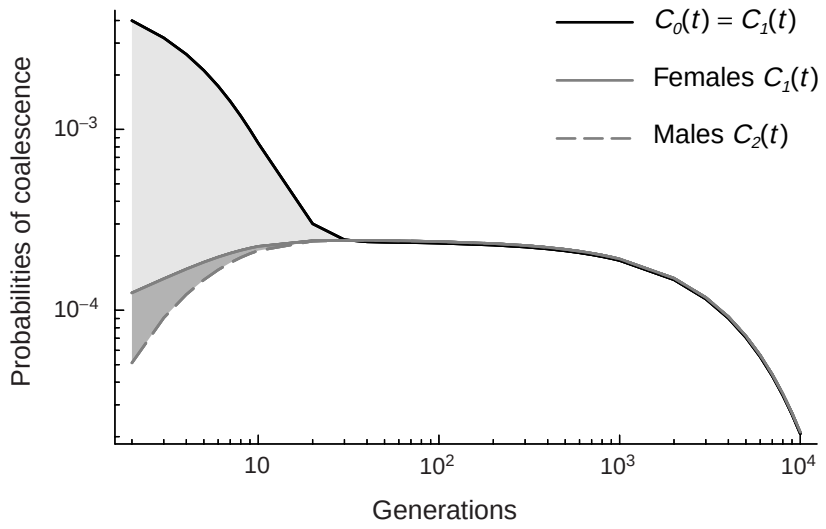


Fig. 2 Distribution of coalescence probabilities. The probabilities $C_h(t)$ of coalescence at time t are given for the different pairs of genes. Note the logarithmic scales on x- and y-axis. $2N = 100$, $n = 20$, and $m_{\varphi} = 0.2 = 100 \times m_{\sigma}$. The area between $C_1(t)$ and $C_2^{\varphi\varphi}(t)$ (female F_{ST}) is in light grey. The area between $C_1(t)$ and $C_2^{\sigma\sigma}(t)$ (male F_{ST}) is the union of the light grey and the dark grey areas.

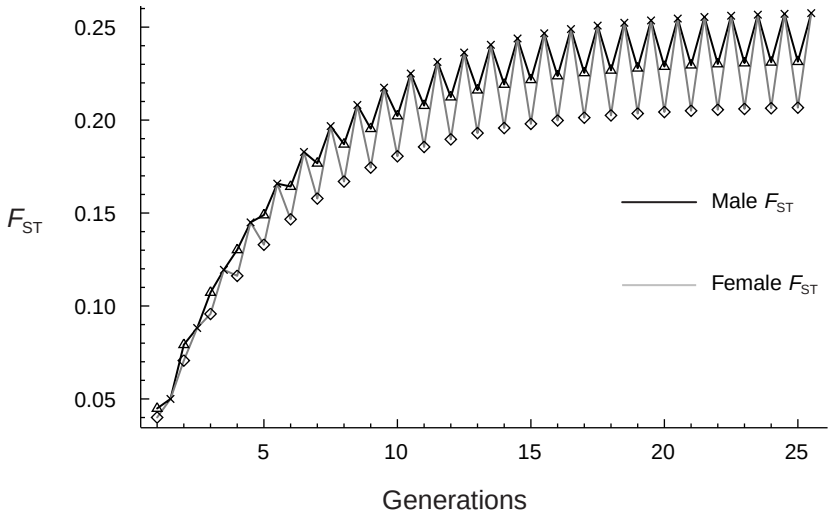


Fig. 3 Within-generation variation of sex-specific F_{ST}'s. Black lines show the evolution of male F_{ST} through time, for genes sampled before (×) or after (△) dispersal. Grey lines show the evolution of female F_{ST} through time, for genes sampled before (×) or after (◇) dispersal. $2N = 20$, $n = 20$, $Nm_{\sigma} = 0.5$, $Nm_{\varphi} = 1$ and $\mu = 10^{-6}$.

1996; for the general case with unbiased dispersal). With non sex-biased dispersal, Rousset (1996, 2001) has shown that, in the low mutation limit, the 'initial area' between the probability distribution of $C_1(t)$ and the probability distribution of $C_2(t)$ is the value of F_{ST} (see the light grey area in Fig. 2). From Fig. 2, it is clear that the 'initial area' between the probability distribution of $C_1^{XX}(t)$ and the probability distribution of $C_2^{XX}(t)$ is smaller for females than for males. This is so because a pair of genes taken in different subpopulations has a higher chance to coalesce in a female than in a male. Therefore, the difference between male and female F_{ST} values is only due to the difference between the male and female probabilities of coalescence for pairs of genes taken in different subpopulations, in a recent past.

Within-generation variation of *F*-statistics

If genes are sampled immediately after reproduction and if the distribution of the number of offspring (before dispersal) that descend from a single parent is binomial, then there is no difference between male and female *F*-statistics (Chesser 1991a,b; Wang 1997). At equilibrium, the values of (sex-unspecific) *F*-statistics for genes sampled before dispersal are found to be

$$F_{IS}^{*} = \frac{-\gamma\left[1-\gamma(a_{\sigma\varphi} - \bar{b})\right]}{2N(1-\gamma\bar{d}) + \gamma\left[(a_{\sigma\sigma} + a_{\varphi\varphi} - 1) - \gamma(a_{\sigma\varphi} - \bar{b})\right]} \quad (13)$$

and

$$F_{ST}^{*} = \frac{\gamma}{2N(1-\gamma\bar{d}) + \gamma(a_{\sigma\sigma} + a_{\varphi\varphi}) - \gamma^2(a_{\sigma\varphi} - \bar{b})} \quad (14)$$

Figure 3 shows the evolution of sex-specific F_{ST} for genes sampled before or after dispersal. For both sexes, F_{ST} takes different values when evaluated before or after dispersal. With females dispersing more than males, $F_{ST}^{\varphi\varphi} < F_{ST}^{\sigma\sigma}$. Moreover, the difference between F_{ST} among juveniles (after dispersal) and F_{ST} among their parents (before dispersal) is always larger among females than among males.

From a glance at eqns 10 and 14, it is noteworthy that

$$\frac{F_{ST}^{XY}}{F_{ST}^{*}} = d_{XY} + b_{XY} \cdot F_{ST}^{XY} \quad \text{for all } (X, Y) \in \left\{♂, ♀\right\}^2 \tag{15}$$

The last term in the right-hand side of eqn 15 is negligible compared to d_{XY}. Thus, F_{ST}^{*} differs from F_{ST}^{XY} by a factor $d_{XY} = (1 - m_{XY}[n/(n-1)])^2 \approx (1 - m_{XY})^2$, for large n. Therefore, taking the ratio of sex-specific F_{ST}^{XX} evaluated after juvenile dispersal over F_{ST}^{*} evaluated before dispersal give the sex-specific migration rate

$$m_X \approx 1 - \sqrt{F_{ST}^{XX}/F_{ST}^{*}} \quad \text{for all } X \in \left\{♂, ♀\right\} \tag{16}$$

Application to simulated data

Methodological outlines

I simulated a finite island model with $n = 20$ subpopulations, each of which was made of male and female diploid individuals. Each generation, adults reproduce within each subpopulation and their progeny disperse at a rate conditioned upon their sex (with $m_♂$ and $m_♀$ being the male and female immigration rates, respectively). After density-dependent regulation, N juveniles survive within each subpopulation. All the adults die. The sex of each juvenile is randomly chosen, so that the expected sex-ratio is one-half. A juvenile can disperse to any of the other $(n-1)$ subpopulations. The marker loci are unlinked and mutation occurs at rate $\mu = 0.001$ (IAM). For each set of parameters, 1000 generations were run before the first measures were taken. Then, 100 measures were performed every 100 generations, to avoid temporal correlations across samples. For each measure, 10 subpopulations (among 20) were sampled. In each of those subpopulations, some individuals were sampled without replacement before dispersal and some individuals were sampled after juvenile dispersal. Among the individuals sampled after dispersal, some were immigrant while some were philopatric. In both cases, individuals were sampled without regard to their sex. Note also that the number of sampled individuals in each subpopulation might in some cases exceed the number of reproducing individuals (this corresponds, e.g. to situations where sampling occurs before density-dependent regulation). Ten independent replicates of this process were simulated. Thus, one simulation run consisted in 1000 measures.

The parameters $F_{ST}^{♂♂}$ and $F_{ST}^{♀♀}$ were estimated by Weir & Cockerham's (1984) estimator for males and females in the sample, respectively. Therefore, for each measure, the sample size for males might slightly differ from the sample size for females. Multilocus estimates were obtained as the sum of one-locus estimates weighted by the heterozygosity at each locus (Weir & Cockerham 1984). Within each simulation run (1000 measures), F_{ST} estimates were significantly autocorrelated in ~8–9 cases out of 144 tests performed for all the sets of parameter values from Figs 5 and 6 (Durbin–Watson autocorrelation test; see, e.g. Sokal & Rohlf 1995, pp. 393–396). Sequential Bonferroni tests performed over the 144 tests showed that no test was significant at the $\alpha = 0.05$ level (see Rice 1989). Thus, for all the set of parameter values considered here, the samples within a simulation run could be considered as independent. The sex-specific migration rates were estimated as

$$\hat{m}_X = 1 - \sqrt{\hat{\theta}^{XX}/\hat{\theta}^{*}} \quad \text{for all } X \in \left\{♂, ♀\right\} \tag{17}$$

where $\hat{\theta}^{XX}$ is the multilocus estimator of F_{ST}^{XX} among individuals of sex X sampled after dispersal and $\hat{\theta}^{*}$ is the multilocus estimator of F_{ST}^{*} among individuals sampled before dispersal. Only the cases where $\hat{\theta}^{XX}/\hat{\theta}^{*} < 0$ (in which case $\hat{m}_X$ has no real value) were discarded from the calculation of means and standard deviations of multilocus estimates. Yet, such undefined estimates where found for only four sets of parameter values out of 144, and the proportion of discarded estimates in these cases was comprised between 0.1 and 1.1%.

For each (multilocus) $\hat{m}_X$ estimate, a 95% confidence interval was computed. The confidence intervals were obtained by bootstrapping samples over loci as advocated by, e.g. Weir (1996). I used the approximate bootstrap confidence intervals (ABC) method described by DiCiccio & Efron (1996). This procedure is an analytic version of the bias-corrected and accelerated (BC_a) algorithm for producing confidence limits from bootstrap distributions, which applies to nonparametric problems (DiCiccio & Efron 1996). The confidence interval endpoints are analytically approximated, rendering Monte Carlo simulations unnecessary and thus reducing the computational burden by an enormous factor.

For each sample, a large-sample Wilcoxon's signed-rank test (see, e.g. Mendenhall *et al*. 1990) was performed before dispersal to determine if the distributions of one-locus male and one-locus female $\hat{\theta}$ estimates differ in location (null hypothesis H_0 = 'no difference'; two-tailed test). Signed-rank tests were also performed after dispersal, but the alternative hypothesis H_1 depended on the actual sex-bias in dispersal. With $m_♂ = m_♀$, the alternative hypothesis was that the relative frequency distributions for male and female $\hat{\theta}$'s differed in location (two-tailed test). With $m_♂ < m_♀$ (*resp.* $m_♂ > m_♀$), the alternative hypothesis was that the distribution for male $\hat{\theta}$'s was shifted to the right (resp. left) of that for female $\hat{\theta}$'s (one-tailed tests). Accounting for different alternative hypotheses makes sense here because some knowledge about the species' biology will often, if not always, provide the meaningful alternative hypothesis to be tested against. For each set of parameters, the proportion of significant tests at the $\alpha = 0.05$ level was calculated over all the tests performed

before dispersal on the one hand, and after dispersal on the other hand. For the tests performed before dispersal, the distributions of male and female one-locus $\hat{\theta}$ estimates do not differ in location and the null hypothesis H_0 is true. Thus, the proportion of significant tests gives the probability to reject H_0 when it is true, i.e. the type I error. For the tests performed after dispersal, H_0 is true when dispersal is not sex-biased. Thus, the proportion of significant tests also gives the type I error. With sex-biased dispersal, however, H_0 is false and the proportion of significant tests provides the probability to reject H_0 when the alternative hypothesis H_1 is true, which is the power of the test (i.e. the complement of type II error).

Results

Figure 4 shows an example of the full distribution of some simulation results for two sets of parameter values. In Fig. 4(A), the females dispersed 10 times less than males. There was only 0.6% of cases where the migration rate estimate of females was superior to that of males. In 78.6 cases out of 100 the confidence intervals were not overlapping, and the average confidence intervals were not overlapping either. In Fig. 4(B), females disperse two times more than males. There was 3.7% of cases where the migration rate estimate of females was inferior to that of males. Yet, the confidence intervals were not overlapping in 44.3 cases out of 100, and were slightly overlapping on average. In the following, I first discuss the accuracy of sex-specific migration rate estimates and then, I evaluate the conditions under which sex-biased dispersal might be detected.

Precision of estimations

Table 1 gives some estimates of sex-specific migration rates from simulated data as well as some indications about the precision of the method. Sex-specific migration rates are slightly overestimated, for all the parameter values considered here. This may be the consequence of neglecting the second right-hand term in eqn 15. The precision of the estimates increases with the true value of the migration rates. With very low migration rates (less than a few per cent) estimates show particularly high standard deviations and wide confidence intervals. The bias and the variance of estimates for one sex increase as the other sex disperses more.

Figure 5 shows the relative mean square error (MSE) of male (bottom-left triangles) and female (upper-right triangles) migration rate estimates for the same range of population parameter values as in Table 1 and a wide range of sampling efforts (sample size and number of loci scored). The relative MSE is the average of the squared relative errors (among 1000 measures), for a given set of parameters. With low migration rates ($m_{\sigma} \leq 0.02$ and $m_{\varphi} \leq 0.02$), the MSE is quite large (MSE ≥ 0.50). With higher migration

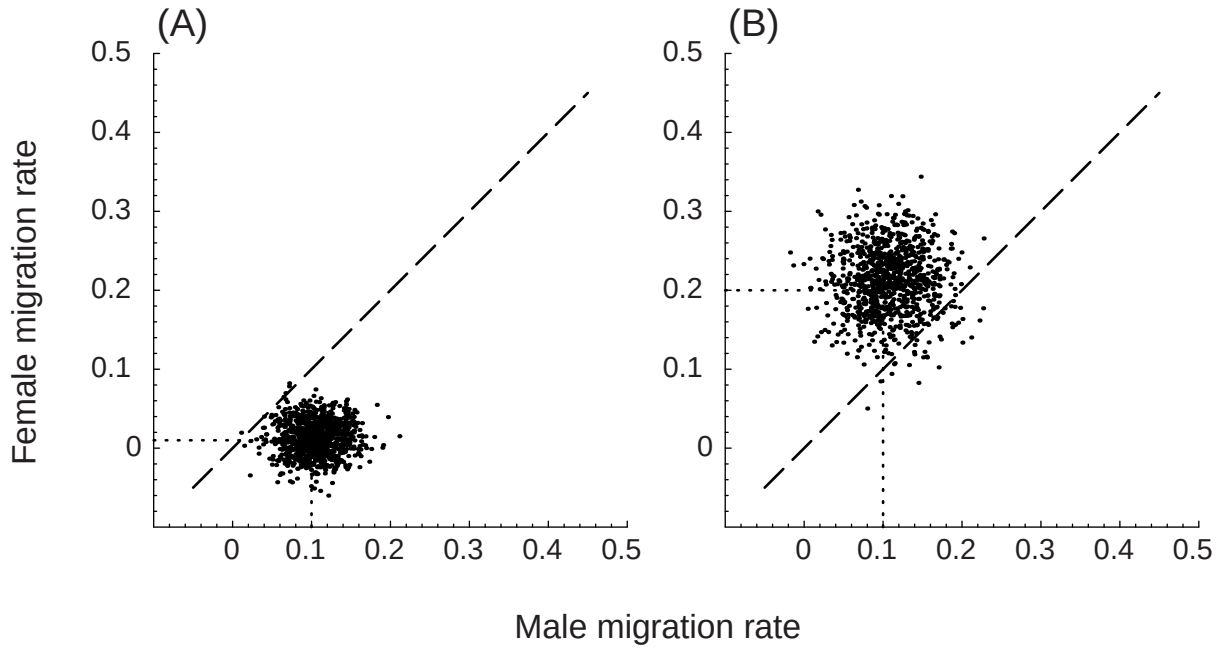


Fig. 4 Scatter plots of some simulation results. Each dot represents a joint estimate of male and female migration rates. A total of 90 individuals were sampled within each subpopulation and scored at 16 loci. The dashed line corresponds to the first diagonal, below which the male migration rate is inferior to the female migration rate. Dotted lines indicate the true values of the male and female migration rates. $2N = 100$, $n = 20$, $m_{\varphi} = 0.1$ and $\mu = 10^{-3}$. (A) $m_{\varphi} = 0.01$; 99.4% of the points are below the diagonal. (B) $m_{\varphi} = 0.2$; 96.3% of the points are above the diagonal.

Table 1 Results from simulation of a finite island model

	$\hat{m}_{\delta}$ estimates			$\hat{m}_{\varphi}$ estimates		Proportion of significant tests		
m_{δ}	mean (s.d)	95% C.I.	m_{φ}	mean (s.d)	95% C.I.	Before	After	H_1
0.01	0.0106 (0.0119)	[−0.005; 0.0261]	0.01	0.0106 (0.0123)	[−0.005; 0.0262]	$P = 0.09$	$P = 0.10$	(≠)
	0.0109 (0.0136)	[−0.007; 0.0294]	0.02	0.0218 (0.0155)	[0.0029; 0.0405]	$P = 0.10$	**$P = 0.18$**	(<)
	0.0121 (0.0221)	[−0.022; 0.0461]	0.10	0.1079 (0.0312)	[0.0706; 0.1555]	$P = 0.08$	**$P = 0.71$**	(<)
	0.0195 (0.0304)	[−0.029; 0.0690]	0.20	0.2133 (0.0451)	[0.1535; 0.2744]	$P = 0.06$	**$P = 0.91$**	(<)
0.02	0.0202 (0.0152)	[0.0012; 0.0392]	0.01	0.0121 (0.0136)	[−0.006; 0.0306]	$P = 0.08$	**$P = 0.15$**	(>)
	0.0212 (0.0164)	[0.0003; 0.0421]	0.02	0.0221 (0.0162)	[0.0008; 0.0432]	$P = 0.09$	$P = 0.10$	(≠)
	0.0240 (0.0249)	[−0.012; 0.0605]	0.10	0.1070 (0.0326)	[0.0680; 0.1466]	$P = 0.08$	**$P = 0.60$**	(<)
	0.0288 (0.0336)	[−0.022; 0.0804]	0.20	0.2125 (0.0448)	[0.1511; 0.2751]	$P = 0.06$	**$P = 0.87$**	(<)
0.10	0.1065 (0.0312)	[0.0688; 0.1446]	0.01	0.0136 (0.0222)	[−0.021; 0.0478]	$P = 0.08$	**$P = 0.69$**	(>)
	0.1065 (0.0318)	[0.0666; 0.1463]	0.02	0.0250 (0.0247)	[−0.011; 0.0612]	$P = 0.08$	**$P = 0.57$**	(>)
	0.1093 (0.0366)	[0.0577; 0.1620]	0.10	0.1096 (0.0372)	[0.0569; 0.1630]	$P = 0.08$	$P = 0.10$	(≠)
	0.1159 (0.0469)	[0.0511; 0.1828]	0.20	0.2145 (0.0522)	[0.1414; 0.2893]	$P = 0.07$	**$P = 0.41$**	(<)
0.20	0.2124 (0.0436)	[0.1523; 0.2735]	0.01	0.0171 (0.0300)	[−0.031; 0.0657]	$P = 0.06$	**$P = 0.92$**	(>)
	0.2127 (0.0455)	[0.1511; 0.2752]	0.02	0.0298 (0.0315)	[−0.021; 0.0812]	$P = 0.07$	**$P = 0.87$**	(>)
	0.2138 (0.0543)	[0.1416; 0.2880]	0.10	0.1150 (0.0454)	[0.0480; 0.1830]	$P = 0.08$	**$P = 0.42$**	(>)
	0.2208 (0.0570)	[0.1328; 0.3144]	0.20	0.2223 (0.0611)	[0.1329; 0.3170]	$P = 0.08$	$P = 0.08$	(≠)

Means and standard deviations (s.d) of sex-specific migration rate estimates are given for different values of the parameters. In each subpopulation made of $N = 50$ reproducing individuals, 50 individuals are sampled before dispersal and 50 individuals are sampled after dispersal, all being typed at 8 independent loci. The average endpoints of bootstrap 95% confidence intervals (C.I) are given for sex-specific migration rate estimates. The proportion of significant Wilcoxon's signed-rank tests are given. For the tests performed before dispersal (two-tailed tests), this proportion gives the type I error. For the tests performed after dispersal, this proportion also gives the type I error when males and females disperse at the same rate. Yet, when males and females disperse at different rates, the proportion of significant tests provides the power of the method (bold cells). For all cases, the alternative hypothesis H_1 for the tests performed after dispersal is provided. $H_1 \equiv (\neq)$ when males and females disperse at the same rate (two-tailed tests), $H_1 \equiv (>)$ when males disperse more than females (one-tailed tests) and $H_1 \equiv (<)$ when males disperse less than females (one-tailed tests).

rates ($m_{\delta} \geq 0.1$ and $m_{\varphi} \geq 0.1$), sex-specific migration rate estimates have good precision (e.g. MSE ≤ 0.21 for 90 sampled individuals per subpopulation and 12 loci scored). Yet, sex-specific migration rate estimates of the more philopatric sex have higher MSE than that of the more dispersing sex. This is particularly true when the dispersal bias is large (see Fig. 5). The precision of estimates (low MSE) increases as the total number of individuals sampled increases. Increasing the number of loci scored does not seem to greatly improve the accuracy of estimates (Fig. 5). For the same sampling effort (i.e. the same number of one-locus genotypes), it is always better to sample more individuals than to score more loci.

Hypotheses testing

With unbiased dispersal, the type I error associated with Wilcoxon's signed-rank tests is less than or equal to 0.10 for the set of parameters considered in Table 1. This indicates that, with the sampling scheme of Table 1, the null hypothesis is rejected slightly more often than expected at the $\alpha = 0.05$ level. With sex-biased dispersal, the proportion of significant Wilcoxon's tests performed after dispersal (bold cells in Table 1) gives the power of the test. High power ($P \geq 0.87$) is obtained in situations where dispersal is strongly biased (at least 10:1) in the favour of a highly dispersing sex ($m_X \geq 0.10$). Low to medium power ($0.57 \leq P \leq 0.71$) is observed for intermediate situations where both the sex-bias for dispersal is less than or equal to a 10:1 ratio and the migration rate of the more dispersing sex is less than or equal to 0.10. Yet, very low power ($P \leq 0.42$) is observed in all other situations where sex-biased dispersal is unpronounced or the average migration rate is low. Therefore, the method of estimation performs well for marked sex-specific bias in dispersal and when the migration rate of the more dispersing sex is not too low.

Figure 6 shows the proportion of significant Wilcoxon's signed-rank tests, performed after dispersal for the same combinations of parameters as in Fig. 5. With very low migration rates ($m_X \leq 0.02$) and unpronounced bias, the power is very low even with large sampling effort ($P \leq 0.26$). Yet, provided that the mean migration rate is reasonably high, unpronounced bias (2:1) can be detected with 90 sampled individuals per subpopulation and at least 12 loci scored (Fig. 6). With strong dispersal biases (10:1, 20:1) and at least 60 individuals sampled per

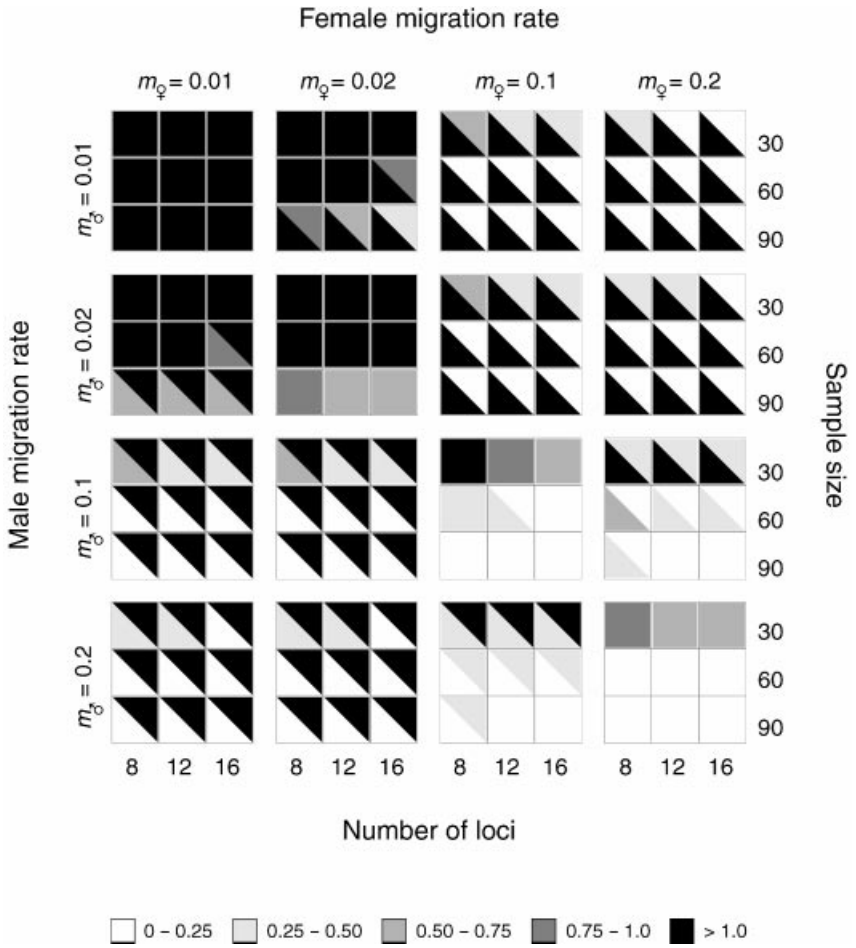


Fig. 5 Relative mean square error (MSE) of estimates of male and female migration rates for different combinations of sample sizes and numbers of loci and for different values of the estimated parameters. The relative MSE is the mean value of $[(\hat{m}_X - m_X)/m_X]^2$. Each square depicts a combination of the parameters. Within each square, the upper right triangle gives an interval for the MSE of female migration rate estimates and the bottom left triangle gives an interval for the MSE of male migration rate estimates. The sample size is given as the total number of individuals collected in each of the 10 subpopulations sampled, among 20 simulated subpopulations. One third of the total number of individuals is sampled before dispersal, two third of the total is sampled after dispersal. Thus, with unbiased sex ratio, the expected number of individuals of each sex sampled after dispersal equals the number of individuals sampled before dispersal. $N = 50$, $n = 20$ and $\mu = 10^{-3}$.

subpopulation typed at 8–16 loci, high power values are obtained ($0.77 \leq P \leq 1.00$). With $m_X \leq 0.1$ and lower dispersal bias (10:1, 5:1), the power of the tests attains 0.92 ($0.63 \leq P \leq 0.92$) with at least 60 individuals sampled per subpopulation and typed at 12 loci and more. Medium power values ($0.62 \leq P \leq 0.72$) are obtained when the actual bias was only 2:1 with 90 sampled individuals per subpopulation and at least 12 loci scored. However, the type I error is large ($0.05 \leq P \leq 0.18$) and, surprisingly, increases with the sample size and the number of scored loci.

Tests based on migration rate estimates

With only 30 individuals sampled per subpopulation and eight loci scored, the sex-specific migration rate estimates of the more philopatric sex is inferior to that of the more dispersing sex in more than 75.5% of cases. The only exception is for low migration rates of both sexes ($m_X \leq 0.02$), with only 68.9% reached. With unbiased dispersal, the proportion of cases where male migration rate is inferior to that of females is very close to 50% (between 46.3% and 52.9%).

There were but a few sets of parameter values for which strictly nonoverlapping confidence intervals were obtained in the great majority of cases (more than 78.1 cases out of 100). These corresponded to situations with highly biased dispersal (10:1 and 20:1) and a highly dispersing sex ($m_X \geq 0.2$), with at least 90 sampled individuals typed at 12 loci and more. Therefore, Wilcoxon's signed-rank tests seem to be more powerful to detect sex-biased dispersal than simple tests based on the crude comparison of migration rate estimates confidence intervals. Yet, with unbiased dispersal (H_0 true), the proportion of nonoverlapping confidence intervals (type I error) is comprised between 0.2 and 10.4. Thus, the type I error of Wilcoxon's test is higher than that of tests based on migration rate estimates.

Discussion

Direct studies of dispersal are often complicated by the difficulty to record migration events. In particular, long distance dispersal urges experimentalists to survey large study areas (Koenig *et al*. 1996). Moreover, direct methods cannot provide estimates of the rate at which individuals successfully reproduce outside their native group. Therefore, any attempt to estimate effective dispersal rates by indirect methods deserves careful examination.

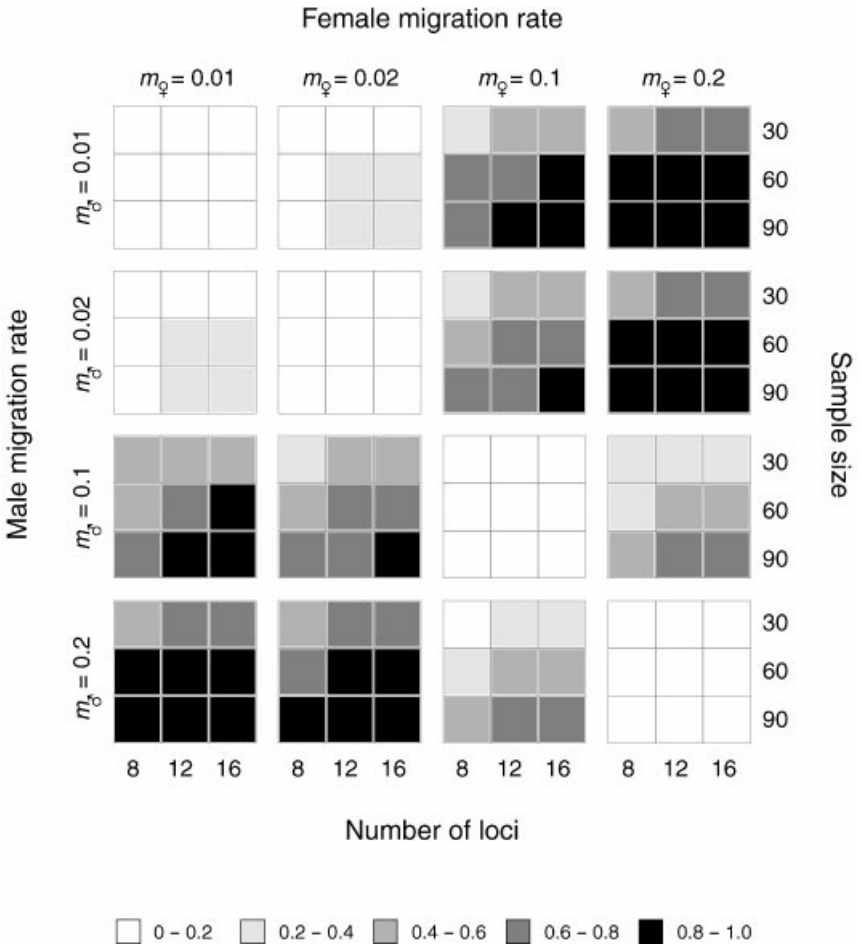


Fig. 6 Proportion of significant Wilcoxon's signed-rank tests among the tests performed after dispersal for a large set of parameter values. See Fig. 5 for additional legends.

In those species where a sex is thought to disperse more than the other, there would be a great advantage of obtaining indirect clues to support or weaken field data (see FitzSimmons *et al*. 1997). This challenge has prompted many authors to examine the potential of indirect methods to detect sexual differences in the ability to disperse, among many different species (Kawata 1985; Webb *et al*. 1995; Favre *et al*. 1997; Moritz *et al*. 1997; Balloux *et al*. 1998; Lyrholm *et al*. 1999; Mossman & Waser 1999; Surridge *et al*. 1999). From these studies it comes out that, at least in some cases, a 'genetic signature' of sex-biased dispersal could indeed be detected. Yet, only one study acknowledges the fact that it should not always be (Rassmann *et al*. 1997).

Discrepancies with previous results

In the recent years, Chesser (1991a,b) gave the recursion equations for the identity probabilities of pairs of genes sampled within and among social lineages, within populations. In its later version, this model assumed different migration rates of males and females, and various breeding strategies. Chesser *et al*. (1993) further extended the model to account for variance in progeny numbers, random sex of progeny, and Sugg & Chesser (1994) added multiple paternity. Recently, Wang (1997) corrected some incorrect probabilities used in Chesser and coworkers' expressions for *F*-statistics and effective population size. Wang (1997) predicted *F*-statistics equilibrium values and effective population size for dioecious subdivided populations with sex-dependent dispersal rates, arbitrary distribution of family size and biased sex ratio.

In both Wang's model and Chesser's model, genes are sampled immediately after reproduction and before dispersal. In such models, as Wang (1997) acknowledges, when the numbers of male and female progeny per parent are binomially distributed, F_{ST} takes the same value for males and females. Therefore, any difference in sex-specific F_{ST} values in those models is due to other factors than dispersal conditioned upon gender. In the present paper, however, I provided the evidence that sex-biased dispersal may induce differences in the gene frequency distributions between males and females, if and only if genes are sampled after dispersal and before reproduction (Figs 1 and 3). This is so because the pairs of genes sampled between subpopulations from the more philopatric sex have longer

coalescence times, on average, than such pairs of genes sampled from the more dispersing sex (Fig. 2).

Application to real data sets

The stochastic simulations I performed showed that the joint evaluation of F_{ST} before and after dispersal may be useful to infer sex-specific migration rates from eqn 17 (see Table 1 and Figs 4, 5). This method can apply to any species with separate sexes, provided that samples are taken before and after the individuals have migrated. This assumes that the timing of dispersal is known in the species. Although this may restrict the applicability of the method, I believe that for many species of birds and mammals, in which sex-biased dispersal has already been documented (Greenwood 1980; Greenwood & Harvey 1982), the biology is sufficiently known to allow the sampling of juveniles at the nest (before dispersal), and that of reproducing adults or young mature individuals (after dispersal).

Overall, the precision of estimates increases with the true migration rate (see, e.g. Table 1). For a given set of parameters, the migration rate of the more dispersing sex has always lower MSE than that of the more philopatric sex (Fig. 5). Yet, there is a close agreement between the estimations of sex-specific migration rates and their true values, for reasonable sample sizes and numbers of loci scored (Table 1, Fig. 5). Average bootstrap confidence intervals of sex-specific migration rate estimates do not overlap for biases larger than 2:1 (see Table 1). With strong dispersal biases (10:1), at least 90 sampled individuals per subpopulation scored at 12 loci and more, are necessary to have a good chance ($P \geq 0.69$) of obtaining nonoverlapping confidence intervals for male and female migration rates, i.e. to detect sex-biased dispersal. With less pronounced biases (5:1), as many as 90 sampled individuals per subpopulation scored at 16 loci are necessary to get nonoverlapping confidence intervals for male and female migration rates in 63.9 cases out of 100.

Significant differences between male and female one-locus $\hat{\theta}$ distributions may also be detected by Wilcoxon's signed-rank tests (see Table 1 and Fig. 6). When H_0 is true (i.e. when the test is performed before dispersal, or when dispersal is unbiased), the type I error of the test is superior to $\alpha = 0.05$. The power of the test may be at least equal to 0.78, provided that the sex-bias for dispersal is not too low (5:1 is a minimum), and that the average dispersal rate is not too small (i.e. the migration rate of the more dispersing sex is superior than or equal to 0.1). Yet, increasing the sample size (and, to a lesser extent, the number of loci) may increase the power.

Although the precision of the sex-specific migration rate estimates is satisfactory (Table 1, Fig. 5), nonparametric tests based on the comparison of sex-specific one-locus $\hat{\theta}$ distributions lack some power for unpronounced sex-biases for dispersal and low dispersal capabilities (Table 1, Fig. 6). Yet, previous attempts to detect sex-specific dispersal from genetic data aimed at providing qualitative information about relative movements of males and females (but see Petit *et al*. 2001) and the present approach is, to my knowledge, the first tentative to quantify sex-specific gene flow from polymorphic autosomal markers only. Indeed, Petit *et al*. (2001) computed the minimum sex-bias in dispersal from mitochondrial and nuclear microsatellite data.

Gene flow is generally estimated as the 'effective number of migrants per generation', namely the product of local effective population size and migration rate (Slatkin 1987; Beerli & Felsenstein 1999). Such estimates actually depend on the local effective population size. Here, I provided an estimator of the sex-specific immigration rate, i.e. an estimator of the proportion of individuals that immigrate and successfully reproduce in one subpopulation each year, on average. This parameter includes differential costs to dispersal among sexes.

Robustness to model assumptions

With biased sex-ratio, genetic drift is expected to be stronger. Therefore, all else being equal, F^{*}_{ST} value should be larger with biased than with unbiased sex-ratio, as should be the value of F^{XY}_{ST}. The ratio of these two parameters should not be different from that given in eqn 15. Yet, this method should also be tested against more complex life-cycles, such as overlapping generations, and more realistic models of population structure, such as isolation-by-distance models.

Acknowledgements

This work was initiated when I was a teaching assistant at the Institute of Ecology (University of Lausanne, Switzerland). J. Goudet and N. Perrin made helpful comments on a previous draft of this manuscript. I thank H. Fréville, B. Godelle, O. Ronce and F. Rousset for insightful comments on later versions of this manuscript. I am particularly indebted to F. Rousset for providing me with useful references on matrix analyses. I thank three anonymous reviewers, as well as L. Excoffier for suggestions and comments that helped in improving this article. I acknowledge financial support from the Fondation Sansouire and from the TMR Programme FRAGLAND. A user-friendly Microsoft Windows program (SEXDISPERSAL 1.0) that computes the estimates and confidence intervals discussed in this article is freely available by anonymous FTP login at isem.isem.univ-montp2.fr in the directory /pub/pc/estim/. This is contribution number 2001–068 of the Institut des Sciences de l'Évolution de Montpellier.

References

Balloux F, Goudet J, Perrin N (1998) Breeding system and genetic variance in the monogamous semi-social shrew, *Crocidura russula*. *Evolution*, **52**, 1230–1235.

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.
Chesser R (1991a) Gene diversity and female philopatry. *Genetics*, **127**, 437–447.
Chesser R (1991b) Influence of gene flow and breeding tactics on gene diversity within populations. *Genetics*, **129**, 573–583.
Chesser R, Rhodes O Jr, Sugg D, Schnabel A (1993) Effective sizes for subdivided populations. *Genetics*, **135**, 1221–1232.
Cockerham C, Weir B (1987) Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences of the USA*, **84**, 8512–8514.
DiCiccio T, Efron B (1996) Bootstrap confidence intervals. *Statistical Science*, **11**, 189–228.
Dobson F (1982) Competition for mates and predominant juvenile male dispersal in mammals. *Animal Behaviour*, **30**, 1183–1192.
Favre L, Balloux F, Goudet J, Perrin N (1997) Female-biased dispersal in the monogamous mammal *Crocidura russula*: evidence from field data and microsatellite patterns. *Proceedings of the Royal Society of London Series B*, **264**, 127–132.
FitzSimmons N, Limpus C, Normans J, Goldizen A, Miller J (1997) Philopatry of male marine turtles inferred from mitochondrial DNA markers. *Proceedings of the National Academy of Sciences of the USA*, **94**, 8912–8917.
Gandon S (1999) Kin competition, the cost of inbreeding and the evolution of dispersal. *Journal of Theoretical Biology*, **200**, 345–364.
Greenwood P (1980) Mating systems, philopatry and dispersal in birds and mammals. *Animal Behaviour*, **28**, 1140–1162.
Greenwood P, Harvey P (1982) The natal and breeding dispersal of birds. *Annual Review of Ecology and Systematics*, **13**, 1–21.
Horn R, Johnson C (1985) *Matrix Analysis*. Cambridge University Press, Cambridge, UK.
Hudson R (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.
Kawata M (1985) Sex differences in the spatial distribution of genotypes in the red-backed vole, *Clethrionomys rufocanus bedfordiae*. *Journal of Mammalogy*, **66**, 384–387.
Koenig W, Vuren DV, Hooge P (1996) Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution*, **11**, 514–517.
Lyrholm T, Gyllensten U (1998) Global matrilineal population structure in sperm whales as indicated by mitochondrial DNA sequences. *Proceedings of the Royal Society of London Series B*, **265**, 1679–1684.
Lyrholm T, Leimar O, Johanneson B, Gyllensten U (1999) Sex-biased dispersal in sperm whales: contrasting mitochondrial and nuclear genetic structure of global populations. *Proceedings of the Royal Society of London Series B*, **266**, 347–354.
McCracken G (1984) Social dispersion and genetic variation in two species of emballonurid bats. *Zeitschrift für Tierpsychologie*, **66**, 55–69.
Mendenhall W, Wackerly D, Scheaffer R (1990) *Mathematical Statistics with Applications*. PWS-KENT Publishing Co, Boston.
Moritz C, Heideman A, Geffen E, McRae P (1997) Genetic population structure of the Greater Bilby *Macrotis lagotis*, a marsupial in decline. *Molecular Ecology*, **6**, 925–936.
Mossman C, Waser P (1999) Genetic detection of sex-biased dispersal. *Molecular Ecology*, **8**, 1063–1067.
Motro U (1991) Avoiding inbreeding and sibling competition: the evolution of sexual dimorphism for dispersal. *American Naturalist*, **137**, 108–115.
Motro U (1994) Evolutionary and continuous stability in asymmetric games with continuous strategy sets: the parental investment conflict as an example. *American Naturalist*, **144**, 229–241.
Nagylaki T (1995) The inbreeding effective population number in dioecious populations. *Genetics*, **139**, 473–485.
Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.
Pardini A, Jones C, Noble L *et al.* (2001) Sex-biased dispersal of great white sharks. *Nature*, **412**, 140–141.
Perrin N, Mazalov V (1999) Dispersal and inbreeding avoidance. *American Naturalist*, **154**, 282–292.
Petit E, Balloux F, Goudet J (2001) Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. *Evolution*, **55**, 635–640.
Rassmann K, Tautz D, Trillmich F, Gliddon C (1997) The microevolution of the Galapagos marine iguana *Amblyrhynchus cristatus* assessed by nuclear and mitochondrial analyses. *Molecular Ecology*, **6**, 437–452.
Rice W (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.
Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, **142**, 1357–1362.
Rousset F (2001) Inferences from spatial population genetics. In: *Handbook of Statistical Genetics* (eds Balding DJ, Bishop M, Cannings C), pp. 179–212. John Wiley & Sons Ltd, Chichester.
Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.
Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.
Sokal RR, Rohlf FJ (1995) *Biometry: the Principles and Practice of Statistics in Biological Research*, 3rd edn. W.H. Freeman and Co., New York.
Sugg D, Chesser R (1994) Effective population size with multiple paternity. *Genetics*, **137**, 1147–1155.
Surridge AK, Ibrahim KM, Bell DJ, Rico C, Hewitt GM (1999) Fine-scale genetic structuring in a natural population of European wild rabbits (*Oryctolagus cuniculus*). *Molecular Ecology*, **8**, 299–307.
Wang J (1997) Effective size and *F*-statistics of subdivided populations. II. Dioecious species. *Genetics*, **146**, 1465–1474.
Webb N, Ibrahim K, Bell D, Hewitt G (1995) Natal dispersal and genetic structure in a population of the European wild rabbit (*Oryctolagus cuniculus*). *Molecular Ecology*, **4**, 239–247.
Weir B (1996) *Genetics Data Analysis II*. Sinauer Associates, Sunderland.
Weir B, Cockerham C (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Renaud Vitalis is a PhD student at the Laboratoire Génétique et Environnement (ISEM — UMR CNRS 5554), in Isabelle Olivieri's group, and at the Station Biologique de la Tour du Valat. He is interested in both applied and theoretical questions addressed by the conservation biology of rare and endangered species. He developed new methods for estimating population parameters such as effective population size and migration rate, from population genetics data sets.

Appendix I

Derivation of IBD probability equilibrium values. The notations are defined in the main text.

Post-dispersal sampling

Recursions from eqns 1–6 can be written in a matrix form. Let $\mathbf{Q} = (Q_0, Q_1^{\male\male}, Q_1^{\female\female}, Q_1^{\male\female}, Q_2^{\male\male}, Q_2^{\female\female}, Q_2^{\male\female})^T$ be the vector of IBD probabilities, for pairs of genes sampled after dispersal. The superscript T denotes the transpose of the vector. Then, the recurrence equations of sex-specific IBD probabilities for pairs of genes within and among subpopulations become

$$\mathbf{Q}(t+1) = \gamma\left[\mathbf{A}\mathbf{Q}(t) + \mathbf{D}\right] \tag{A1.1}$$

with $\mathbf{A}$ being a transition matrix and $\mathbf{D}$ a vector of constant terms in the recursions, i.e. the coalescent events occurring in one generation

$$\mathbf{A} = \begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\frac{a_{\male\male}}{2N} & \frac{a_{\male\male}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\male\male}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\male\male}}{2} & \frac{(1-a_{\male\male})}{4} & \frac{(1-a_{\male\male})}{4} & \frac{(1-a_{\male\male})}{2} \\
\frac{a_{\female\female}}{2N} & \frac{a_{\female\female}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\female\female}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\female\female}}{2} & \frac{(1-a_{\female\female})}{4} & \frac{(1-a_{\female\female})}{4} & \frac{(1-a_{\female\female})}{2} \\
\frac{a_{\male\female}}{2N} & \frac{a_{\male\female}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\male\female}}{4}\left(1-\frac{1}{N}\right) & \frac{a_{\male\female}}{2} & \frac{(1-a_{\male\female})}{4} & \frac{(1-a_{\male\female})}{4} & \frac{(1-a_{\male\female})}{2} \\
\frac{b_{\male\male}}{2N} & \frac{b_{\male\male}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\male\male}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\male\male}}{2} & \frac{(1-b_{\male\male})}{4} & \frac{(1-b_{\male\male})}{4} & \frac{(1-b_{\male\male})}{2} \\
\frac{b_{\female\female}}{2N} & \frac{b_{\female\female}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\female\female}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\female\female}}{2} & \frac{(1-b_{\female\female})}{4} & \frac{(1-b_{\female\female})}{4} & \frac{(1-b_{\female\female})}{2} \\
\frac{b_{\male\female}}{2N} & \frac{b_{\male\female}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\male\female}}{4}\left(1-\frac{1}{N}\right) & \frac{b_{\male\female}}{2} & \frac{(1-b_{\male\female})}{4} & \frac{(1-b_{\male\female})}{4} & \frac{(1-b_{\male\female})}{2}
\end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix}
0 \\
\frac{a_{\male\male}}{2N} \\
\frac{a_{\female\female}}{2N} \\
\frac{a_{\male\female}}{2N} \\
\frac{b_{\male\male}}{2N} \\
\frac{b_{\female\female}}{2N} \\
\frac{b_{\male\female}}{2N}
\end{pmatrix} \tag{A1.2}$$

At equilibrium, eqn A1.1 becomes

$$\mathbf{Q} = \gamma(\mathbf{I} - \gamma\mathbf{A})^{-1}\mathbf{D} \tag{A1.3}$$

with $\mathbf{I}$ being the identity matrix.

Pre-dispersal sampling

Let $\mathbf{Q}^* = (Q_0^*, Q_1^*, Q_2^*)^T$ be the vector of IBD probabilities of pairs of genes sampled before dispersal. Then, the recurrence

equations of IBD probabilities for pairs of genes sampled before dispersal are derived from eqns 1–6 as $\mathbf{Q}^*(t+1) = \gamma(\mathbf{A}^*\mathbf{Q}^*(t) + \mathbf{D}^*)$, where $\mathbf{A}^*$ is a transition matrix and $\mathbf{D}^*$ the vector of coalescent terms in the recursions

$$\mathbf{A}^* = \begin{pmatrix} 0 & a_{\sigma\varphi} & (1-a_{\sigma\varphi}) \\ \dfrac{1}{2N} & \left[\dfrac{a_{\sigma\varphi}}{N} + (1-\dfrac{2}{N})\bar{a}\right] & \left[\dfrac{(1-a_{\sigma\varphi})}{N} + (1-\dfrac{2}{N})(1-\bar{a})\right] \\ 0 & \bar{b} & (1-\bar{b}) \end{pmatrix} \text{ and}$$

$$\mathbf{D}^* = \begin{pmatrix} 0 \\ \dfrac{1}{2N} \\ 0 \end{pmatrix} \tag{A1.4}$$

At equilibrium, $\mathbf{Q}^* = \gamma(\mathbf{I} - \gamma\mathbf{A}^*)^{-1}\mathbf{D}^*$.

Appendix II

Relationship of IBD probabilities to coalescence times.

Since the probability that no mutation occurred in the past t generations is γ^t, the IBD probability for two genes taken h steps apart in the hierarchy ($h = 0$ for genes taken within individuals, $h = 1$ for genes taken among individuals in the same subpopulations and $h = 2$ for genes taken among subpopulations) is $Q_h = \Sigma_{t=1}^{\infty}\gamma^t C_h(t)$, with $C_h(t)$ being the probability that the two genes coalesce in generation t (Slatkin 1991; Rousset 1996). Since eqn A1.3 may be rewritten as

$$\mathbf{Q} = \sum_{t=1}^{\infty}\gamma^t\mathbf{A}^{t-1}\mathbf{D} \tag{A2.1}$$

then

$$\mathbf{C}(t) = \mathbf{A}^{t-1}\mathbf{D} \tag{A2.2}$$

gives the expected coalescence probabilities at t, with $\mathbf{C}(t)$ being the vector of probabilities of coalescence for pairs of genes drawn at different hierarchical levels. Using the same notations as for identity probabilities, $\mathbf{C} = (C_0, C_1^{\sigma\sigma}, C_1^{\varphi\varphi}, C_1^{\sigma\varphi}, C_2^{\sigma\sigma}, C_2^{\varphi\varphi}, C_2^{\sigma\varphi})^T$. Assuming $\mathbf{A}$ is diagonalizable, $\mathbf{A} = \mathbf{S}\Lambda\mathbf{S}^{-1}$ (see, e.g. Horn & Johnson 1985; page 46), where $\mathbf{S}$ is the matrix made of the right eigenvectors of $\mathbf{A}$, such that the right eigenvector associated with the kth eigenvalue is the kth column-vector of $\mathbf{S}$. $\mathbf{S}^{-1}$ is the matrix made of the left eigenvectors of $\mathbf{A}$, such that the left eigenvector associated with the kth eigenvalue is the kth row-vector of $\mathbf{S}$. Λ is the diagonal matrix with the eigenvalues $(\lambda_1, \ldots, \lambda_k)$, of $\mathbf{A}$ on the main diagonal. Therefore,

$$\mathbf{C}(t) = \mathbf{S}\Lambda^{t-1}\mathbf{S}^{-1}\mathbf{D} \tag{A2.3}$$

Since Λ^{t-1} is easily derived as diag $(\lambda_1^{t-1}, \ldots, \lambda_k^{t-1})$, the distribution of coalescence probabilities can readily be obtained from numerical computations of eqn (A2.3).

Annexe I

Détection de la sélection

Vitalis R., Dawson K. and Boursot P. (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics* **158** : 1811-1823.

Copyright © 2001 by the Genetics Society of America

Interpretation of Variation Across Marker Loci as Evidence of Selection

Renaud Vitalis,*,†,‡ Kevin Dawson*,§ and Pierre Boursot*

**Laboratoire Génome, Populations et Interactions, Université Montpellier II, 34095 Montpellier Cedex 05, France, †Laboratoire Génétique et Environment, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, 34095 Montpellier Cedex 05, France, ‡Station Biologique de la Tour du Valat, 13200 Arles, France and §I.A.C.R. Long Ashton Research Station, Department of Agricultural Science, University of Bristol, Bristol BS41 9AF, United Kingdom*

Manuscript received October 23, 2000
Accepted for publication May 14, 2001

ABSTRACT

Population structure and history have similar effects on the genetic diversity at all neutral loci. However, some marker loci may also have been strongly influenced by natural selection. Selection shapes genetic diversity in a locus-specific manner. If we could identify those loci that have responded to selection during the divergence of populations, then we may obtain better estimates of the parameters of population history by excluding these loci. Previous attempts were made to identify outlier loci from the distribution of sample statistics under neutral models of population structure and history. Unfortunately these methods depend on assumptions about population structure and history that usually cannot be verified. In this article, we define new population-specific parameters of population divergence and construct sample statistics that are estimators of these parameters. We then use the joint distribution of these estimators to identify outlier loci that may be subject to selection. We found that outlier loci are easier to recognize when this joint distribution is conditioned on the total number of allelic states represented in the pooled sample at each locus. This is so because the conditional distribution is less sensitive to the values of nuisance parameters.

PRESUMED neutral polymorphic loci are commonly used in making inferences about patterns of differentiation within or among populations of the same or closely related species. For this purpose, genetic distances (see, *e.g.*, Nei 1972) or Wright's (1951) *F*-statistics are estimated from allele-frequency data. Under particular models of population structure, these parameters are related to demographic or historical parameters, such as the effective population size, the rate of migration between populations, or the time since the populations diverged from their common ancestral population.

However, misinterpretations can occur if one is not able to clearly distinguish between the patterns generated by random genetic drift or by natural selection. The problem is that selective processes can also affect neutral loci. A locus that is neutral will respond to selection whenever it is in linkage disequilibrium (statistical association among allelic states at different loci) with other loci that are subject to selection. Such associations may arise by chance in small populations (Hill and Robertson 1966, 1968; Ohta and Kimura 1969). For example, stabilizing or balancing selection operating at a locus tends to maintain an elevated level of variation at closely linked neutral loci (Strobeck 1983; Hudson and Kaplan 1988). Selection acting on any locus has an effect on loosely linked loci, which resembles a reduction of effective population size (Robertson 1961; Barton 1995, 1998). Local adaptation tends to increase population differentiation at loci where selection acts, and very high F_{ST} values may be found at closely linked neutral loci (Charlesworth *et al.* 1997). The substitution of advantageous mutations at a locus may also reduce neutral variation at linked loci (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Barton 1995). Similarly, "background selection," caused by the selection against deleterious mutations (Charlesworth *et al.* 1993; Barton 1995) results in a reduced effective population size for neutral genes in the region of the chromosome where this selection is acting. Background selection may also increase the apparent population differentiation (Charlesworth *et al.* 1997).

Therefore, it is of prime interest to identify loci that are responding to selection to exclude them from the genetic analysis of population structure or history. It was recognized early on by Cavalli-Sforza (1966) that any form of selection will affect some regions of the genome more than others, whereas population history, demography, migration, and the mating system will affect the whole genome in the same way. Accordingly, Lewontin and Krakauer (1973) proposed two tests of selective neutrality. Both tests are based on the sampling distribution of a statistic $\hat{F}$, the standardized variance of gene frequency, which is an estimator of the parameter F_{ST}. Their first test is a goodness-of-fit test comparing

Corresponding author: Renaud Vitalis, Laboratoire Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France. E-mail: vitalis@isem.univ-montp2.fr

the observed distribution of $\hat{F}$ estimates (one estimate from each locus) to a χ^2 distribution with $(n - 1)$ d.f., where n is the number of populations sampled. The second test is based on the comparison of the observed variance of $\hat{F}$ (across loci) denoted $s^2_{\hat{F}}$, with the theoretical variance approximated as

$$\sigma^2 = \frac{k\bar{F}^2}{n - 1}, \quad (1)$$

where $\bar{F}$ is the mean value of $\hat{F}$ averaged across loci, and k is a constant that, according to Lewontin and Krakauer (1973), should not exceed 2 whatever the underlying distribution of allelic frequency. The ratio $s^2_{\hat{F}}/\sigma^2$ should be distributed approximately as a χ^2/d.f., the number of degrees of freedom being determined by the number of biallelic loci.

However, since populations of the same species share, to a certain extent, a common history and since populations are connected through the dispersal of individuals, $\hat{F}$ values will be correlated across loci. For example, the geographic and historical relationships between populations may have a hierarchical structure if populations have been derived from a common ancestral population by a sequence of successive splits. This is the pattern to be expected following the fragmentation of a species range. The effect of such a population history is always to increase the expected variance of $\hat{F}$ (Robertson 1975a,b). Moreover, even simple models of divergence by drift (Nei and Chakravarti 1977), island models (Nei *et al.* 1977), or stepping-stone models of dispersal (Nei and Maryuyama 1975) inflate the expected variance, making Lewontin and Krakauer's (1973) test unreliable in most cases (Lewontin and Krakauer 1975).

More recently, Bowcock *et al.* (1991) studied the worldwide human genetic differentiation based on DNA polymorphism. Simulating a reasonably well supported evolutionary scenario of divergence, they evaluated the theoretical distribution of F_{ST} conditional on initial gene frequencies. Among 100 nuclear RFLP markers a number of genes exhibited lower or, more often, higher variation than expected under neutrality. In an important article, Beaumont and Nichols (1996) proposed a method based on the analysis of the expected distribution of F_{ST} conditional on heterozygosity rather than allele frequency. The conditional distribution, constructed under an island model of population structure, is remarkably robust to a wide range of alternative models (colonization, stepping-stone). Interestingly, departures from equilibrium do not alter the expected distribution much whenever F_{ST} is <0.5. Yet, unequal numbers of immigrants per generation over the whole population generated some discrepancies with the symmetric island model for heterozygosities in the range [0.1, 0.5] (see Figure 3d in Beaumont and Nichols 1996).

Thus, their approach might be flawed whenever the true population history consists of repeated branching events or when the connectivity of populations is uneven. However, we cannot infer patterns of migration or historical branching and test for the homogeneity of the markers with the same data. This is what Felsenstein (1982) described as the "infinitely many parameters" problem. A solution to this problem is to restrict attention to simple but realistic scenarios that may apply to any *pair* of populations (Robertson 1975b; Tsakas and Krimbas 1976). This reduces the number of parameters in the model. Here, we develop a model of population divergence. We define population-specific parameters as functions of probabilities of identity for pairs of genes taken within or among populations. These parameters are simply related to the ratio of divergence time over effective population size. We construct simple estimators of these population-specific parameters. We then examine the expected joint distribution of these estimators under a wide range of neutral scenarios of divergence. This suggests a new method to assess the homogeneity of response of genetic markers to the historical processes, using empirical data. Finally, we apply our new method to a data set of allozyme loci from *Drosophila simulans* populations and compare our results to those obtained by using Beaumont and Nichols' (1996) method.

THE MODEL

We consider two haploid populations of constant sizes N_1 and N_2, which completely separated τ generations ago from a single population of stationary size N_0. By complete separation, we mean that the populations did not exchange any migrants between the time of the split and the present. We do not assume that the common ancestral population was at equilibrium when it split. Instead, we allow the ancestral population to have gone through a bottleneck τ_0 generations before present (with $\tau_0 > \tau$). Before this, the ancestral population was at mutation-drift equilibrium, with constant size N_e. Generations do not overlap. New mutations arise at a rate μ and follow the infinite allele model (IAM). This model of population divergence is illustrated in Figure 1.

Let $Q_{w,i}$ be the probability that two genes sampled at random within population i are identical by descent (IBD) and Q_a be the probability that a gene sampled at random from population 1 is IBD to a gene sampled at random from population 2. IBD probabilities are defined as the probabilities that two genes have not mutated since their most recent common ancestor (Malécot 1975). The probability that a pair of genes are IBD is equal to the probability that these genes are identical in state (IIS) whenever the mutation process follows the IAM.

More generally, let Q_h denote the IBD probability of any pair of genes: $h = (w, i)$ when two genes are sampled within population i, or $h = a$ when one gene is sampled

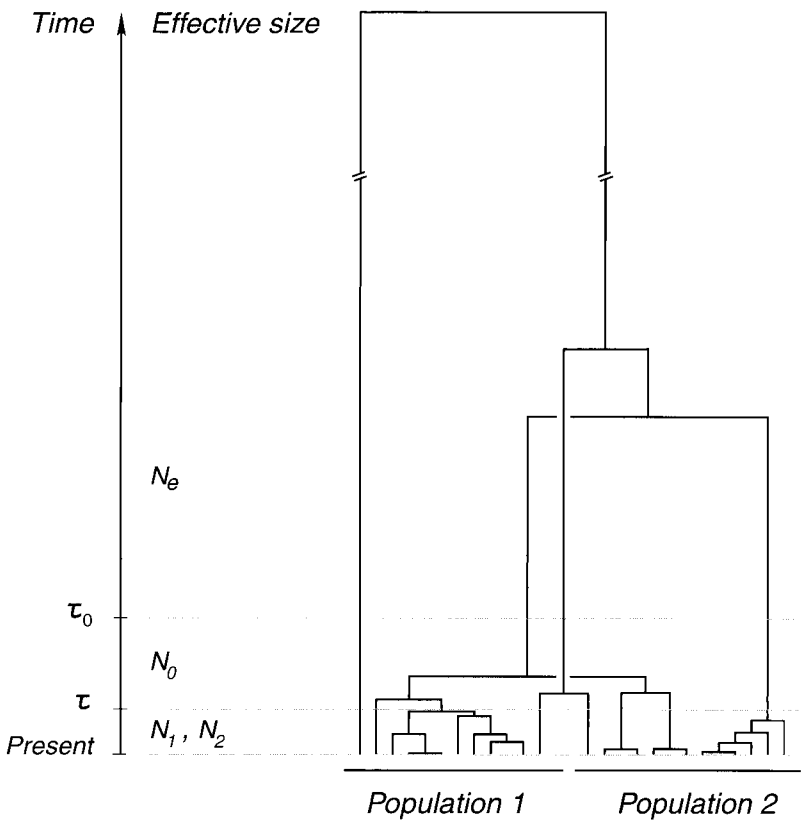


FIGURE 1.—A gene genealogy under our model for $n = 10$ genes sampled in each population. In this example, the parameter values are $N_1 = N_2 = 100$, $N_0 = 500$, $N_e = 1000$, $\tau = 50$, $\tau_0 = 150$, and $\mu = 10^{-3}$.

from each population. It is possible to give an expression for Q_h as a function of the coalescence time (SLATKIN 1991). Under a continuous time approximation

$$Q_h = \int_0^\infty \gamma^t c_h(t)\, dt \tag{2}$$

(HUDSON 1990), where $c_h(t)$ is the probability of coalescence at t for a pair of genes of type h, and $\gamma = (1 - \mu)^2$. The waiting time for a coalescent event in a population of size N_i has an exponential distribution with mean N_i. The IBD probability for a pair of genes in population i reduces to

$$Q_{w,i} = \int_0^\tau \frac{\gamma^t}{N_i} e^{-t/N_i} dt + (1 - C_i)\, Q_0, \tag{3}$$

where Q_0 is the IBD probability for two genes sampled at random from the common ancestral population at time τ (just before the split) and $(1 - C_i) = \gamma^\tau \cdot e^{-\tau/N_i}$ is the probability that the two genes neither coalesce nor mutate in the ith population in the time interval $0 < t \leq \tau$. The first term on the right-hand side of Equation 3 is the probability that the two genes coalesce in the time period $0 < t \leq \tau$ and are IBD. Following Equation 2, the IBD probability for a pair of genes sampled at random from the common ancestral population just before the split at time τ is given by

$$Q_0 = \int_r^{\tau_0} \frac{\gamma^{t-\tau}}{N_0} e^{-(t-\tau)N_0} dt + (1 - C_0) \int_{\tau_0}^{\infty} \frac{\gamma^{t-\tau 0}}{N_e} e^{-(t-\tau_0)/N_e} dt, \tag{4}$$

where $(1 - C_0) = \gamma^{\tau_0 - \tau} \cdot e^{-(\tau_0 - \tau)/N_0}$ is the probability that the two genes neither coalesce nor mutate in the time interval $\tau < t \leq \tau_0$. The first term on the right-hand side of Equation 4 averages over the coalescent events occurring during the population bottleneck. During this time interval ($\tau < t \leq \tau_0$) the waiting time for a coalescent event is exponentially distributed with mean N_0. The last term in Equation 4 averages over coalescent events occurring in the ancestral population at mutation-drift equilibrium. This last term represents the IBD probability for two randomly sampled genes in a stationary population of size N_e, which is $1/(1 + \theta)$, with $\theta = 2N_e\mu$. Solving the integrals in the low-mutation limit (where $\gamma^t \approx e^{-2\mu t}$), we find that the solution of Equation 3 is

$$Q_{w,i} \approx \frac{1}{\theta_i + 1}[1 - e^{-T_i(\theta_i + 1)}] + e^{-T_i(\theta_i + 1)} \cdot Q_0, \tag{5}$$

where $\theta_i = 2N_i\mu$ and $T_i = \tau/N_i$. The value of Q_0 is given by the solution of Equation 4,

$$Q_0 \approx \frac{1}{\theta_0 + 1}[1 - e^{-T_0(\theta_0 + 1)}] + e^{-T_0(\theta_0 + 1)} \left(\frac{1}{\theta + 1}\right), \tag{6}$$

where $\theta_0 = 2N_0\mu$ and $T_0 = (\tau_0 - \tau)/N_0$. The probability for a gene in population 1 to be IBD with a gene in population 2 is just given by

$$Q_a = \gamma^\tau Q_0. \quad (7)$$

Obviously, two such genes cannot coalesce during the τ generations between the moment of divergence and the present. They are IBD only if their respective ancestors are IBD when populations 1 and 2 diverge and, furthermore, if they do not undergo mutation during the divergence. Now, it is useful to consider the parameter

$$F_i = \frac{Q_{w,i} - Q_a}{1 - Q_a}. \quad (8)$$

It is worth noting that the weighted sum of F_i over the two populations gives the intraclass correlation for the probability of identity by descent for genes within populations relative to genes between populations. This is of particular interest, because the properties of the intraclass correlations for the probability of identity in state ("IIS correlations"; Cockerham and Weir 1987) can be deduced from the properties of the corresponding intraclass IBD correlations in the low-mutation limit (Rousset 1996). Indeed, such ratios of identity probabilities of the form of Equation 8 give the same low-mutation limit, whether one considers the infinite allele model or other mutation models (Rousset 1996, 1997).

If we neglect new mutations arising during the divergence process, Q_a reduces to Q_0 and $Q_{w,i} = C_i(1 - Q_0) + Q_0$. Thus

$$F_i \approx 1 - e^{-T_i}. \quad (9)$$

Note that Equation 9 gives a well-known result when both daughter populations are assumed to have the same size N, so that $F_1 = F_2 = F \approx 1 - e^{-\tau/N}$ (see, *e.g.*, Reynolds *et al.* 1983). Hereafter, the parameter T_i is referred to as the "branch length" of population i. An important result is that, in the low-mutation limit, the new parameters F_1 and F_2 do not depend on the "nuisance parameters" θ or T_0. This suggests that a simple moment-based estimator $\hat{T}_i$ of branch length can be derived as

$$\hat{T}_i = \ln(1 - \hat{F}_i), \quad (10)$$

where $\hat{F}_i$ is an estimator of F_i (see appendix for details).

PROPERTIES

Simulation procedure: For each set of parameter values, a sequence of artificial data sets was generated using standard coalescent simulations, as described by, *e.g.*, Hudson (1990). The simulations were performed as follows (see Figure 1 for an illustrated example of one simulated genealogy). For each population, the genealogy of a sample of n_i genes is generated for a period of time ranging from present to τ generations in the past. During this period, all the coalescent events are separated by exponentially distributed time intervals, with means $N_1/\binom{n_1}{2}$ in population 1 and $N_2/\binom{n_2}{2}$ in population 2 (see Equation 3). At time τ, the number n_0 of lineages that remain represents the ancestors of all the genes sampled in populations 1 and 2. The genealogy of these lineages is generated for the time period $[\tau, \tau_0]$, and all the coalescence events are separated by exponentially distributed time intervals, with mean $N_0/\binom{n_0}{2}$ (see the first term in the right-hand side of Equation 4). At time τ_0, the lineages that remain are the ancestors of all the genes sampled in populations 1 and 2. The genealogy of these n_e genes is generated for the period $[\tau_0, +\infty]$, with all coalescent events separated by exponentially distributed time intervals with mean $N_e/\binom{n_e}{2}$ (see the second term in the right-hand side of Equation 4). Once the complete genealogy is obtained, the mutation events are superimposed on the coalescent tree of lineages. In the results that follow, each artificial data set consisted of two (haploid) samples of size $n = 100$, one from population 1 and the other from population 2.

Simulation results: By calculating the estimators $\hat{F}_1$ and $\hat{F}_2$ for each of these artificial data sets, it was possible to obtain a close approximation to the expected distribution of these estimators (see appendix for details). Figure 2 shows this expected joint distribution of $\hat{F}_1$ and $\hat{F}_2$ for various combinations of the nuisance parameters θ and T_0. In this case, the "true" branch lengths were $T_1 = T_2 = 0.1$ (hence $F_1 = F_2 \approx 0.0953$). The expected value of the estimator $\hat{F}_1$ (respectively $\hat{F}_2$) was always close to the value of the parameter F_1 (respectively F_2). One can show that, by construction, the points $(\hat{F}_1, \hat{F}_2)$ lie within the upper-right triangle with vertices (1, 1), (−1, 1), and (1, −1). The joint distribution of these two statistics has a negative correlation. Most importantly, it is clear from this figure that the joint distribution of $\hat{F}_1$ and $\hat{F}_2$ depends strongly on the nuisance parameters, even though their expectations remain close to the true values of F_1 and F_2.

It can be seen that, for smaller values of T_0, the joint distribution becomes tighter as θ increases. On the other hand, for larger values of θ, the distribution is found to widen as T_0 increases. In both cases, it is the level of variation that remains before divergence that is crucial in shaping the joint distribution. With small θ and large T_0, the lineages coalesce rapidly before the divergence, and the number of distinct mutations (allelic states) that can be maintained is small. In this case, the variance of the estimates of populations branch lengths is large, as illustrated by the wide joint distribution of $\hat{F}_1$ and $\hat{F}_2$. Therefore, the joint distribution of $\hat{F}_1$ and $\hat{F}_2$ is not ideal for investigating the homogeneity of response of a set of molecular markers to the genealogical processes. Indeed, other factors such as heterogeneous mutation rates across loci may be invoked to explain disparities of branch length estimates among markers. Fortunately, this problem can be overcome by considering the joint distribution of $\hat{F}_1$ and $\hat{F}_2$, conditional upon the total

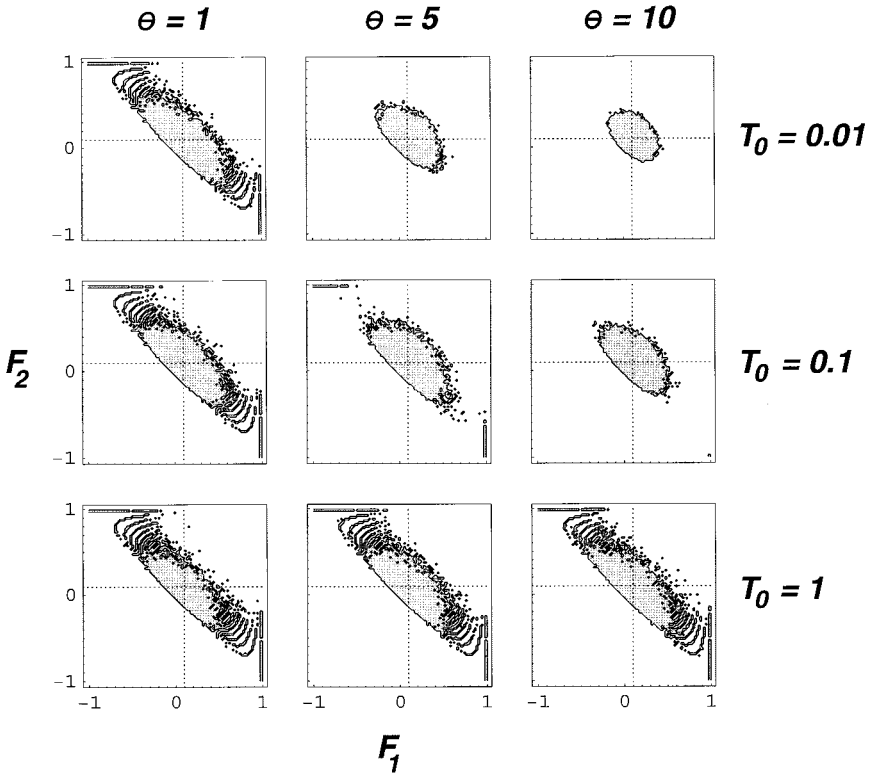


FIGURE 2.—Expected distribution of pairs of $\hat{F}_1$ and $\hat{F}_2$ estimates for wide ranges of values of the nuisance parameters $\theta = 2N_e\mu$ and T_0. $T_i = \tau/N_i$ is 0.10 for both daughter populations (with $\tau = 50$ and $N_1 = N_2 = 500$), giving an expected value $F_i \approx 0.0953$, as indicated by the dotted lines. For all parameter sets, $\mu = 10^{-4}$ and $N_0 = 1000$. One hundred individuals are sampled in each daughter population. The light gray area defines a region in which 95% of the simulated points are expected to lie (see APPENDIX for details).

number k of allelic states in the pooled sample at each locus. Figure 3 shows the estimated joint distribution for $T_1 = T_2 = 0.1$ (hence $F_1 = F_2 \approx 0.0953$), conditioned on $k = 4$. The combinations of nuisance parameter values are the same as in Figure 2.

The expected joint conditional distribution appears to be almost independent on the nuisance parameters. So, given the observed values for the parameters F_1 and F_2, and given the number of alleles in the sample, one can obtain the conditional joint distribution, and then a high probability region, that should contain 95% of the observed measures of pairwise $\hat{F}_i$'s values. This result provides the justification for using the conditional distributions to analyze the homogeneity in the patterns of genetic differentiation revealed by a (large) set of markers.

APPLICATIONS

In this section, we present a methodology for identifying outlier loci by a pairwise analysis of populations. For each pair of populations (i, j), we suggest the following protocol:

1. For all loci, the statistics $\hat{F}_i$ and $\hat{F}_j$ are computed (see APPENDIX).
2. The parameters F_i and F_j are estimated as the averages among loci weighted by the heterozygosities $(1 - \hat{Q}_i)$ and $(1 - \hat{Q}_j)$, respectively (see APPENDIX). This corresponds to the weighting of loci suggested by WEIR and COCKERHAM (1984) for the multilocus estimator of F_{ST}.
3. The expected joint distribution of $\hat{F}_i$ and $\hat{F}_j$ is generated by performing 10,000 coalescent simulations for a given set of nuisance parameter values. This is repeated using a wide range of values for the nuisance parameters. In the *D. simulans* data set discussed below, all the pairwise combinations for θ and T_0 were performed, with $\theta = 1$, 5, or 10 and $T_0 = 0.01$, 0.1, or 1. Thus, a total of 90,000 coalescent simulations were performed in this example. The simulated sample sizes are chosen to be representative of those actually realized in the real data set.
4. For each expected joint distribution of $\hat{F}_i$ and $\hat{F}_j$, we construct all the distributions, conditional on the number of allelic states k in the pooled sample, for $k = 2, 3, \ldots$ (the pooled sample is the sample obtained by pooling the samples from populations i and j). Remember, there is one expected distribution for each set of nuisance parameter values. For each conditional distribution, we identify the "high probability" or "high density" region, in the range of the points $\hat{F}_i$ and $\hat{F}_j$, where 95% of the data are expected to lie (see APPENDIX for the construction of this high probability region).
5. For each value of the number of allelic states in the pooled sample, we superimpose a scatter plot of the observed data points (pairs of $\hat{F}_1$ and $\hat{F}_2$ values) over an outline of the 95% high probability region to identify outlier loci.

***D. simulans* data set:** We applied this method to a *D. simulans* data set, described in SINGH *et al.* (1987) and CHOUDHARY *et al.* (1992). The raw data set was kindly

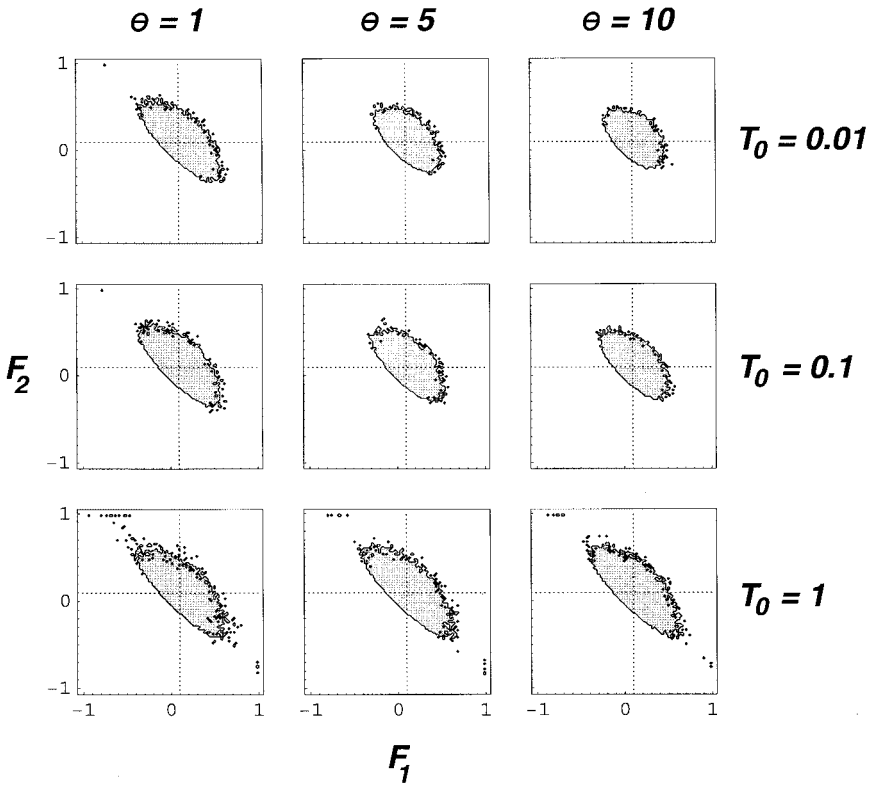


FIGURE 3.—Expected distribution of pairs of $\hat{F}_1$ and $\hat{F}_2$ estimates conditioned on a number of alleles in the sample equal to four. As in Figure 1, wide ranges of values were used for the nuisance parameters. The dotted lines indicate the expected values for F_1 and F_2.

provided by R. S. Singh and R. A. Morton. Among 111 allozyme loci, 43 were found to be polymorphic in the five populations studied in Europe and Africa. The samples consisted of isofemale lines maintained in the laboratory. The haploid sample sizes ranged from $n = 26$ to $n = 55$. Figure 4 shows the analysis performed on a particular pair of populations (France and Tunisia). The multilocus estimates of the parameters F_1 (French population) and F_2 (Tunisian population) were 0.0064 and 0.0617, respectively. The expected distributions with these averaged values, conditioned on the number of alleles in the pooled sample, are plotted with the actual monolocus pairwise $(\hat{F}_1, \hat{F}_2)$ estimates.

In the great majority of cases, the points fall within the 95% confidence region. With 43 loci we would expect two ($0.05 \times 43 \approx 2$) to lie outside the region by chance. But considering the joint distributions for loci with three or more alleles, we found 4 loci that clearly lie outside. Caution is required in the case of loci that lie on the borders of the possible range (Figure 4B). These correspond to loci that have an allele fixed in one population. Slight variations in the nuisance parameters can increase or decrease the relative proportion of loci that may fix one allele in a population. Indeed, we found some conditions under which the 95% envelope contained these 2 loci. This problem can remain even when we condition on the observed number of alleles. On the other hand, 2 other loci (coding for glutamate pyruvate transaminase and carbonic anhydrase-3) are clear outliers of the expected distributions (Figure 4, C and D). In all pairwise comparisons that included the French population, these two loci fell either outside, or on the edges of the 95% high probability region.

In all the pairs that included the population from Congo, two loci coding respectively for the larval protein-10 (Pt-10) and the phosphoglucomutase (PGM) were found to lie outside or on the limit of the 95% high probability region (Figure 5). The locus coding for the larval protein-10 systematically gives a longer estimated branch length for this African population than do all other loci, while it gives similar branch lengths to other loci for the other populations. This suggests that genetic variation was severely reduced by a factor other than genetic drift in this African population. The locus coding for phosphoglucomutase gives a longer branch length estimate than the other loci in three cases (Figure 5, A–C) and a shorter one in one case (Figure 5D). The locus coding for phosphoglucomutase was also found to lie outside the limit of the 95% high probability region in all the pairs that included the population from Seychelle Island (Figure 6). To strengthen our presumption that these loci were outside the limit allowed by a neutral model, we checked whether these loci also lie outside the limit of the 99% high probability region. The same results were obtained. For these loci, we did not find any plausible neutral scenario of divergence by drift that could provide such a scatter of points. We thus conclude that natural selection may have acted on these loci or on closely linked regions within the genome.

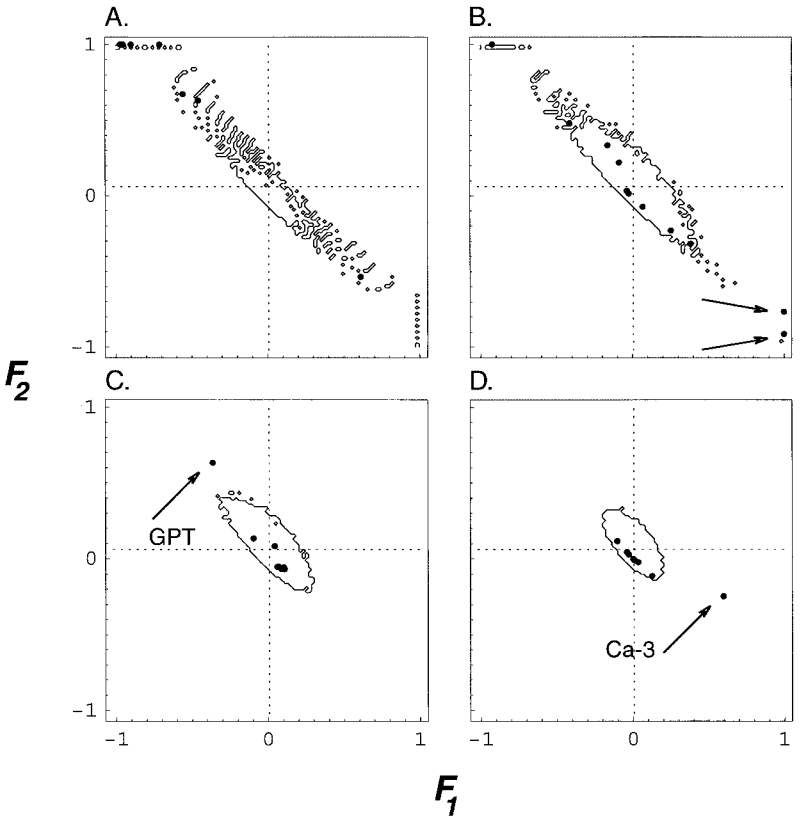


FIGURE 4.—$\hat{F}_1$ and $\hat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for the pairwise comparison of the populations from France ($n = 55$) and Tunisia ($n = 52$). n is the number of isofemale lines typed for each enzymatic system (haploid sample size). Each locus is represented with a solid dot. The averaged values are $\hat{F}_1 = 0.0064$ and $\hat{F}_2 = 0.0617$ as indicated by the dotted lines. Thin solid lines enclose a region in which 95% of the simulated data points are expected to lie. Four distributions are shown, conditioned on the number of allelic states k in the whole sample: (A) expected distribution of pairwise F_i estimates conditioned on $k = 2$; (B) with $k = 3$; (C) with $k = 4$; and (D) with $k = 5$. Solid arrows indicate outlier loci. The loci coding for glutamate pyruvate transaminase (GPT) and carbonic anhydrase-3 (Ca-3) are shown, respectively, in C and D.

We are more cautious about claiming that the loci coding for glutamate pyruvate transaminase and carbonic anhydrase-3 were or are subject to selection. These loci are clear outliers in some pairwise comparisons involving the French population but fall just within the limits of the confidence region in other comparisons. Moreover, when considering 99% confidence regions instead of 95% confidence regions, some loci were no longer detected as outliers but rather as lying on the edges of the confidence limit. The locus coding for isocitrate dehydrogenase-1 was found to be an outlier in three (out of four) pairs that included the population from Seychelle Island. Overall, six more loci were detected as outliers in single pairwise comparisons only. Therefore, we should be very cautious about considering those latter loci as being under selection. Indeed, if a locus has responded to selection in one particular contemporary population since it became isolated, then we expect this locus to show up as an outlier in all (or most) comparisons involving this population. This pattern is exactly what we found for the two loci coding for larval protein-10 and phosphoglucomutase in the Congo and Seychelle Island populations.

Evaluating the robustness of this method to the assumptions of the model: In the data set discussed above, it is likely that the populations of *D. simulans* have exchanged migrants after divergence. More generally, one can wonder whether complete isolation and divergence by random drift accurately describes natural situations. An alternative approach would be to develop a new model of population divergence that allows subsequent migration after separation. But if we want to make inferences about a more realistic (and hence a more complex) model of divergence, then we need to distinguish between the pattern of genetic differentiation that results from (i) recent separation followed by very little migration or (ii) ancient separation followed by a moderate amount of migration. This is a difficult task, which would require more powerful methods for inferring parameter values (*e.g.*, maximum likelihood; see NIELSEN and SLATKIN 2000) that would be much more time consuming. Further, note that NIELSEN and SLATKIN (2000) assume that the mutation rate is zero.

So, we are interested in testing if our method (which assumes evolution in complete isolation after divergence) is undermined when applied to pairs of populations that still exchange genes after divergence. It should be borne in mind that gene flow, like genetic drift, affects the whole genome in the same way. We generated artificial data sets under neutral models of population divergence, including high mutation rates and moderate levels of migration between populations. We used a modified version of the algorithm described by HUDSON (1990), which accounts for symmetric migration between populations. For the period of time ranging from present to τ generations in the past, considering populations 1 and 2 altogether, the waiting time to the next event (coalescence or migration) is

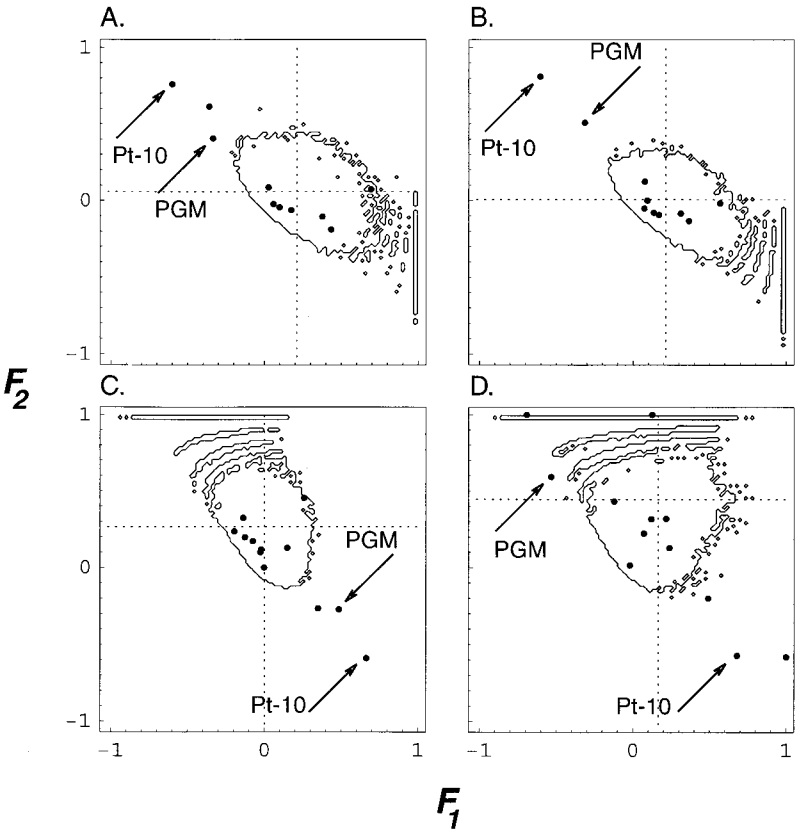


FIGURE 5.—$\hat{F}_1$ and $\hat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for all the pairwise comparisons involving the population from the Congo ($n = 45$). (A) Expected distribution for the populations from France ($n = 55$) and Congo. (B) Tunisia ($n = 52$) *vs.* Congo. (C) Congo *vs.* Cape Town, South Africa ($n = 32$). (D) Congo *vs.* Seychelle Island ($n = 26$). All distributions are conditioned on $k = 4$. Each locus is represented with a solid dot. Dotted lines give the expected values for $\hat{F}_1$ and $\hat{F}_2$. For each expected conditional distribution, solid arrows indicate the loci coding for the larval protein-10 (Pt-10) and phosphoglucomutase (PGM).

drawn from an exponential distribution with mean $N_1N_2/[N_2/\binom{n_1}{2} \cdot N_1/\binom{n_2}{2} + m(n_1 + n_2)N_1N_2]$, where m is the backward migration rate (NORDBORG 2001). Conditionally on the occurrence of one event, two genes coalesce in population 1 (respectively population 2) with probability $N_2/\binom{n_1}{2}/[N_2/\binom{n_1}{2} \cdot N_1/\binom{n_2}{2} + m(n_1 + n_2) N_1N_2]$ (respectively $N_1/\binom{n_2}{2}/[N_2/\binom{n_1}{2} \cdot N_1/\binom{n_2}{2} + m(n_1 + n_2)\ N_1N_2]$) or one gene migrates from population 2 to population 1 (respectively from population 1 to population 2) with probability $m \cdot n_1/[N_2/\binom{n_1}{2} \cdot N_1/\binom{n_2}{2} + m(n_1 + n_2)\ N_1N_2]$ (respectively $m \cdot n_2/[N_2/\binom{n_1}{2} \cdot N_1/\binom{n_2}{2} + m(n_1 + n_2)N_1N_2]$; see STROBECK 1987; TAKAHATA 1988; NORDBORG 2001). Then, for the period $[\tau, +\infty]$, the coalescent process was generated as previously described (see also Figure 1).

For each set of parameters, we generated 20 data sets composed of two samples ($n_1 = n_2 = 50$) of 50 loci each. The parameter values are given in Table 1. For each data set, we applied our method as described above. We generated joint distributions, conditional on the number of alleles, according to the actual numbers of alleles in each sample. For all sets of parameters, we grouped loci with eight alleles and more in a single class. The number of joint conditional distributions generated per artificial data set (*i.e.*, the number of classes for different numbers of alleles) ranged from three to seven. For each data set, over all the joint conditional distributions taken together, we expected to detect $0.05 \times 50 = 2.5$ outlier loci, just by chance. We performed Wilcoxon's signed-rank tests (see, *e.g.*, MENDENHALL *et al.* 1990) to determine if the distribution of the number of detected outlier loci was shifted to the right of 2.5 (one-tailed test).

Table 1 shows the total observed number of outlier loci (mean and median over 20 independent simulated data sets) detected for a range of nuisance parameter values (low and high mutation rates, short or long divergence by random drift, with or without migration). In no case could we reject the null hypothesis that the expected number of outlier loci detected by our method was equal to 2.5 (against the alternative hypothesis that the expected number of outliers was >2.5). Thus, our approach is conservative in the sense that the 95% confidence region contains at least 95% of the loci generated by a truly neutral model. At the level of 5% we do not (falsely) detect any more than 5% of outlier loci in a sample of neutral markers (type I error).

Comparison with BEAUMONT and NICHOLS' (1996) method: We also applied BEAUMONT and NICHOLS' (1996) procedure to the *D. simulans* data set. Based on a preliminary examination of the data, three loci (coding for α-fucosidase, dipeptidase-1, and mannose phosphatase isomerase) were found to lie outside the 95% confidence region of the conditional joint distribution of $\hat{F}_{ST}$ and mean heterozygosity. The percentiles were determined as described in BEAUMONT and NICHOLS (1996). Surprisingly, none of these three loci were detected as outliers using our method. There may be several reasons for this.

We suspect that, in the present case, the inclusion of

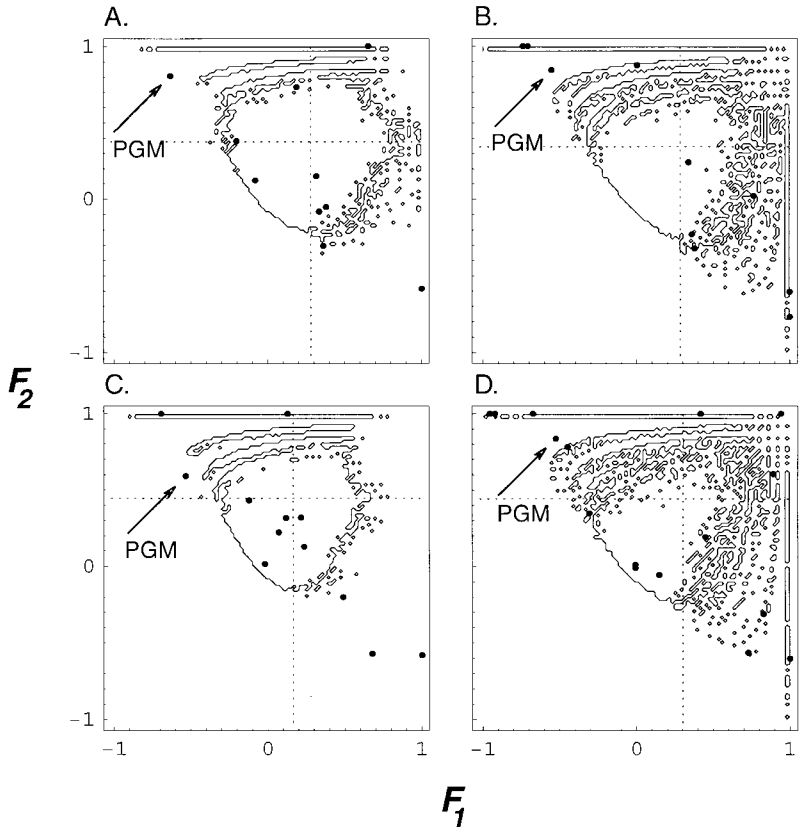


FIGURE 6.—$\hat{F}_1$ and $\hat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for all the pairwise comparisons involving the population from the Seychelle Island ($n = 26$). (A) Expected distribution for the populations from France ($n = 55$) and Seychelle Island. (B) Tunisia ($n = 52$) *vs.* Seychelle Island. (C) Congo ($n = 45$) *vs.* Seychelle Island. (D) Cape Town, South Africa ($n = 32$) *vs.* Seychelle Island. Distributions in A and C are conditioned on $k = 4$ and distributions in B and D are conditioned on $k = 3$. Each locus is represented with a solid dot. Dotted lines give the expected values for $\hat{F}_1$ and $\hat{F}_2$. For each expected conditional distribution, solid arrows indicate the locus coding for phosphoglucomutase (PGM).

a very distant insular population (Seychelle Island) may bias their analysis. Indeed, populations heterogeneous with respect to their demographic parameters (effective population sizes and migration rates) were shown to strongly affect their method (BEAUMONT and NICHOLS 1996). Isolation (low migration rates) together with population bottlenecks can introduce a further bias. Consider as an extreme case the fixation of a private allele at some locus in one population. This may be unexpected for a polymorphic locus in a mutation-migration-drift equilibrium model, unless there is a strong asymmetry, with some populations being smaller and receiving less immigrants than others. However, this is not unexpected for a model of separation and isolation, where there were population bottlenecks. This may boost the F_{ST} estimate at some locus and thus exclude it from the 95% high probability region. So, isolated populations should probably be excluded from BEAUMONT and NICHOLS' (1996) analysis.

Moreover, in general, the loci that were outliers in our analysis gave small values of (global) F_{ST}. But from the shape of the joint distribution of F_{ST} and heterozygosity, it seems that BEAUMONT and NICHOLS' (1996) analysis is likely to detect outlier loci that exhibit unusually large F_{ST} values. However, a process that would cause an apparent decrease of genetic variation at one locus in a single local population, without leading to a decrease of the variation over all populations, would not be detected in BEAUMONT and NICHOLS' (1996) procedure. In other words, if selection acts on one locus at a local scale, pairwise comparisons of populations are more likely to be efficient for detecting outlier loci.

DISCUSSION

Using population-specific estimators of branch lengths: Conventional pairwise genetic distances or pairwise measures of population differentiation are based on the assumption that the sizes of populations are equal and constant through time or that dispersal, if any, is symmetric. For example, the pairwise F_{ST} parameter is defined as a ratio of identity probabilities within and among populations. But the within-population term is taken as an average over the pair of populations. Thus, the definition of the parameter implicitly assumes that both populations share the same demographic parameters. WEIR and COCKERHAM's (1984) estimator θ of F_{ST} is constructed to have low bias and variance, assuming that the populations are independent replicates of the same stochastic process. This means that populations are supposed to have the same size and that they do not exchange migrants. Without these assumptions, θ would be a complex function of unequal (within-population) identity probabilities.

In contrast, the $\hat{F}_i$ parameters defined here make sense even when the populations are of unequal size. The only assumption we make is that when the two populations have separated, they remain completely isolated. From the estimation of F_i's for a pair of populations, we can infer the branch lengths. The ratio of

TABLE 1

Results from applications to various divergence scenarios

			Detected outliers		
μ	θ	T_0	Mean	Median	P value
No migration: $m = 0$					
10^{-5}	1	1	1.85	2.0	0.98
10^{-5}	10	10^{-2}	1.15	1.0	1.00
10^{-3}	1	1	2.60	3.0	0.28
10^{-3}	10	10^{-2}	1.75	2.0	0.76
Low migration: $m = 0.01$					
10^{-5}	1	1	2.30	2.5	0.79
10^{-5}	10	10^{-2}	2.25	2.0	0.77
10^{-3}	1	1	2.00	2.0	0.99
10^{-3}	10	10^{-2}	1.20	1.0	1.00
Moderate migration: $m = 0.1$					
10^{-5}	1	1	2.30	2.0	0.87
10^{-5}	10	10^{-2}	2.05	2.0	0.96
10^{-3}	1	1	2.25	2.0	0.89
10^{-3}	10	10^{-2}	1.85	2.0	0.98

For all sets of parameters, 50 loci were scored among 100 haploid sampled individuals (50 in each population). The mean (and median) number of outlier loci detected is tabulated. We provide the P values of Wilcoxon's signed-rank tests, performed on the distributions of detected outliers, to determine whether this distribution was shifted to the right of 2.5 (one-tailed test).

these branch length estimates is inversely proportional to the ratio of effective population sizes. Thus, these estimates may be seen as measures of the intensity of genetic drift that has occurred since population divergence. The main drawback to this approach is that when estimates of IIS probabilities are smaller within populations than among them (*i.e.*, $\hat{Q}_{w,i} < \hat{Q}_a$), $\hat{F}_i$ becomes negative, and the moment-based estimator of branch length fails. Although this can arise just by chance for some loci, averaging $\hat{Q}$ estimates over loci reduces the problem.

Provided that we obtain good estimates of branch lengths for a pair of populations (which requires the pooling of information from many independent loci), we may be able to evaluate the consistency of locus-specific estimates. Indeed, the joint distribution of branch length estimates, conditioned on the number of alleles in the pooled sample, depends only weakly on nuisance parameters of the simple model of divergence by drift. In particular, this conditional distribution is not sensitive to departures from mutation-drift equilibrium before isolation or to differences in mutation rates.

Detection of selection acting on genetic markers: We saw from the analysis of the *D. simulans* data set that the great majority of loci always fall in the confidence region of the conditional pairwise distributions of branch length estimates, while some loci do not. Overall, we identified two loci that were probably subject to selection in the population from Congo, one of which was also probably subject to selection in the population from Seychelle Island. We concluded that the distribution of variability at these loci may have been shaped by forces other than mutation and drift. Furthermore, we identified two other loci that either lie on the edges or fall just outside the high probability region of the expected conditional distribution in the French population, although we are more cautious about these latter loci. It is noteworthy that our estimation of the density of $\hat{F}_i$ parameters (see APPENDIX) is discontinuous, because of the discrete nature of the data (the allele counts). This is particularly true when the number of alleles on which the distribution is conditioned is small (for a given set of parameters, the lower the number of allelic states, the more discontinuous the null distribution; see Figure 4). Using discrete distributions is clearly preferable to using some (unnecessary) continuous approximations to it. Moreover, whenever the null distribution is based on the same number of allelic states and the same number of genes as in the sample, there is no tendency for loci to show up as outlier just because of the discrete nature of the distribution (*i.e.*, a locus cannot, by construction, show up *between* arc-shaped areas located at the edge of some distributions). Yet, when an apparent outlier lies very close to the 95% high probability region, it is highly advisable to check whether this locus also lies outside the 99% high probability region.

The main criticisms of LEWONTIN and KRAKAUER's (1973) attempts to interpret across-loci heterogeneity of F_{ST} values arose from their failure to consider allele frequencies as random variables, whose distribution depends on the underlying model of population structure and history. Indeed, uneven patterns of dispersal among populations (NEI and MARYUYAMA 1975) or sequences of population splits within the species (ROBERTSON 1975a,b) may strongly undermine the approach. LEWONTIN and KRAKAUER (1975) acknowledged that their tests might be limited to situations where the true population structure did not depart too much from the island model.

However, conditioning the distribution of F_{ST} on the heterozygosity (BEAUMONT and NICHOLS 1996) or on gene frequency for biallelic loci (BOWCOCK *et al.* 1991) was shown to give surprisingly robust results, in the sense that strong departures from the model assumptions do not alter the distribution very much. The strongest effect on the joint expected distribution of F_{ST} and heterozygosity occurs when populations are heterogeneous with respect to their demographic parameters (BEAUMONT and NICHOLS 1996), for example, when populations are founded by very different numbers of individuals or when populations are arranged in an irregular stepping-stone lattice. However, BEAUMONT and NICHOLS (1996) considered a large number d of subpopulations in the metapopulation ($d = 100$) and this parameter strongly

influences the expected heterozygosity [$H_e \approx 4Nd\mu/(1 + 4Nd\mu)$, for diploids]. In addition, at a local scale, F_{ST} is only weakly influenced by the total population size Nd (Rousset 2001). The number of populations has a stronger role than acknowledged by Beaumont and Nichols (1996) in determining whether mutation has an effect on F_{ST} or not. It was shown that, considering smaller numbers of populations, F_{ST} estimates may be reduced by mutation, especially with a stepwise mutation model (see Flint *et al.* 1999). With $d = 100$ islands, the sets of parameters used in Beaumont and Nichols (1996) did not account for any case where mutation may depress F_{ST}.

As already suggested by Tsakas and Krimbas (1976), restricting Lewontin and Krakauer's (1973) approach to pairs of populations removes all kinds of dependence on the unknown population structure. Indeed, whatever their history, two populations ultimately descend from a single ancestral one in the past. Still, nuisance parameters may broaden the joint distribution of pairwise F_i's (Figure 2). However, conditioning on the number of alleles (Figure 3) also gives distributions that are robust enough to variations in the values of nuisance parameters. It is obvious that, for each analysis of a pair of populations, we deliberately discard the information brought by other populations, which may decrease the power of the method (Tsakas and Krimbas 1976). But we believe that this enables us to explain a wider range of patterns than any symmetrical model, such as the island model. In this respect, our approach is conservative. Moreover, we found that low or moderate gene flow did not undermine our approach, in the sense that the probability of falsely detecting a neutral locus as an outlier (type I error) is no more than 5% (Table 1). We compared the performance of our method to that of Beaumont and Nichols (1996), using the empirical data from Singh *et al.* (1987) and Choudhary *et al.* (1992). We further tested whether our method would falsely reject neutral loci (type I error) any more than expected, under a wide range of nuisance parameter values (see Table 1). In particular, since the method assumes that the mutations arising after divergence can be neglected, we checked that high mutation rates do not weaken the approach.

We found that patterns such as those identified in, *e.g.*, the Tunisia *vs.* Congo data set as evidence of selection can be produced by "neutral models," where the coalescent process occurs independently at each locus. Indeed, similar scatters of points could be obtained whenever the parameters $\hat{F}_1$ and $\hat{F}_2$ vary across loci, having particularly high values at certain loci (results not shown). Models of this type provide a rough approximation to models of unlinked neutral loci, some of which were strongly influenced by selection (remembering that the effect of selection resembles a reduction in the effective population size experienced by these loci, as described by Robertson 1961; Barton 1995, 1998). So, it is certainly plausible that the patterns that we identified in, *e.g.*, the Tunisia *vs.* Congo data set were produced by selection. A thorough investigation of the conditions under which our method fails to identify selected loci (type II error) would be desirable. However, this is not feasible, as the range of models that incorporate selection is very large.

An important task for the future is to consider a more general neutral model of the divergence of two populations, where gene flow may continue after the moment of "separation." It is also desirable to extend this approach to more elaborate neutral models, incorporating recombination. More sophisticated estimators of the divergence parameters (branch lengths) would then be required. We assumed that the mutation process follows the IAM and we allowed a wide range of possible mutation rates. In the IAM, genes that are identical in state are also identical by descent. This may not be the case with other mutation models such as with the K allele or stepwise mutation processes, which can produce IIS genes that are not IBD (homoplasy). The IAM is probably an adequate model for allozyme data. It is certainly not so appropriate for potentially more variable markers, such as microsatellites. Recent studies revealed that the processes of mutation of microsatellite markers may be more complex than previously thought and may vary greatly among loci (Estoup and Angers 1998). Furthermore, the effect of homoplasy on measures of population subdivisions is not simple (Rousset 1996). Therefore, further studies should be conducted to test the application of our method across different classes of nuclear markers that differ in processes of mutation. Clearly, if a whole class of marker loci, which are known to have a very distinct mutation process, are identified as outliers by our analysis, then this class of markers should be interpreted with caution.

If we could identify those marker loci that responded to selection during the process of divergence, then we may be able to obtain improved estimates of the parameters of population structure and history by excluding these loci (Ross *et al.* 1999). Our method differs from previous ones in allowing selection to be detected in particular populations and in some pairwise comparisons but not others. This opens up the possibility that markers may be discarded only in the analysis of those populations where there is evidence that they have responded to selection. It is also of interest to use this approach to screen the genome for regions that have responded to strong selection in the recent past. If populations have diverged phenotypically and if this has been caused by selection, then it may even be possible to identify candidate regions for the quantitative trait loci underlying this adaptive divergence.

We are very grateful to R. S. Singh and R. A. Morton for providing the *Drosophila simulans* data set. We thank I. Olivieri for helpful comments on a previous draft of this manuscript and S. Billiard for valuable discussions about the structured coalescent. We are grateful to two

anonymous reviewers for their constructive comments. This work was funded by contract no. BIO4-CT96-1189 of the Commission of the European Communities (DG XII) to P.B., and R.V. was also partially funded by the Fondation Sansouire. This is publication no. 2001-045 of the Institut des Sciences de l'Évolution de Montpellier.

LITERATURE CITED

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics **140:** 821–841.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B **263:** 1619–1626.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto *et al.*, 1991 Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Genetics **88:** 839–843.

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B **164:** 362–379.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.

Choudhary, M., M. B. Coulthart and R. S. Singh, 1992 A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. VI. Patterns and processes of genic divergence between *D. melanogaster* and its sibling species, *D. simulans*. Genetics **130:** 843–853.

Cockerham, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 697–700.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology*, edited by G. R. Carvalho. IOS Press, Amsterdam.

Felsenstein, J., 1982 How can we infer geography and history from gene frequencies? J. Theor. Biol. **96:** 9–20.

Flint, J., J. Bond, D. C. Rees, A. J. Boyce, J. M. Roberts-Thomson *et al.*, 1999 Minisatellite mutational processes reduce F_{ST} estimates. Hum. Genet. **105:** 567–576.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269–294.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The hitchhiking effect revisited. Genetics **123:** 887–899.

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

Lewontin, R. C., and J. Krakauer, 1975 Testing the heterogeneity of *F* values. Genetics **80:** 397–398.

Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **8:** 212–241.

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

Mendenhall, W. M., D. D. Wackerly and R. L. Scheaffer, 1990 *Mathematical Statistics with Applications*. PWS-KENT Publishing Company, Boston.

Nei, M., 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

Nei, M., and A. Chakravarti, 1977 Drift variance of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. Theor. Popul. Biol. **11:** 307–325.

Nei, M., and T. Maryuyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

Nei, M., A. Chakravarti and Y. Tateno, 1977 Mean and variance of F_{ST} in a finite number of incompletely isolated populations. Theor. Popul. Biol. **11:** 291–306.

Nielsen, R., and M. Slatkin, 2000 Likelihood analysis of ongoing gene flow and historical association. Evolution **54:** 44–50.

Nordborg, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–55.

Reynolds, J., B. S. Weir and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics **105:** 767–779.

Robertson, A., 1961 Inbreeding in artificial selection programmes. Genet. Res. **2:** 189–194.

Robertson, A., 1975a Gene frequency distribution as a test of selective neutrality. Genetics **81:** 775–785.

Robertson, A., 1975b Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

Ross, K. G., D. D. Shoemaker, M. J. B. Krieger, J. DeHeer and L. Keller, 1999 Assessing genetic structure with multiple classes of molecular markers: a case study involving the introduced fire ant *Solenopsis invicta*. Mol. Biol. Evol. **16:** 525–543.

Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357–1362.

Rousset, F., 1997 Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics **145:** 1219–1228.

Rousset, F., 2001 Inferences from spatial population genetics, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Singh, R. S., M. Choudhary and J. R. David, 1987 Constrasting patterns of geographic variation in the cosmopolitan sibling species *Drosophila melanogaster* and *D. Simulans*. Biochem. Genet. **25:** 27–40.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

Takahata, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res. **52:** 213–222.

Tsakas, S., and C. B. Krimbas, 1976 Testing the heterogeneity of *F* values: a suggestion and a correction. Genetics **84:** 399–401.

Weir, B. S., and C. C. Cockerham, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: J. Hey

APPENDIX

Parameters estimation: For any given allele u, we use the indicator variable x_{iju} for describing the state of the jth gene in the ith population, with $i = (1, 2)$. $x_{iju} = 1$ if the allelic type is u, $x_{iju} = 0$ otherwise. Let p_{iu} be the frequency of allele u in the ith population. Then $p_{iu} = \varepsilon(x_{iju}|\mathbf{p})$, where $\varepsilon(\ |\mathbf{p})$ denotes the expectation, conditional on the array $\mathbf{p}$ of all the allele frequencies. Considering the second moments of the random variable x_{iju}, it follows that $\varepsilon(x_{iju}^2|\mathbf{p}) = p_{iu}$ and, since individuals are sampled independently from the ith population, $\varepsilon(x_{iju}\, x_{ij'u}|\mathbf{p}) = p_{iu}^2$ for $j' \neq j$. Then, summing over all alleles

gives the probability for two genes in population i to be identical in state (IIS),

$$Q_{w,i} = \varepsilon\left(\sum_{u=1}^{k} p_{iu}^2\right), \quad \text{(A1)}$$

where ε denotes now the expectation over the distribution of allele frequencies **p** and k is the number of alleles in the population. The IIS probability for two genes respectively taken in populations 1 and 2 is given by

$$Q_a = \varepsilon\left[\sum_{u=1}^{k} (p_{1u}p_{2u})\right]. \quad \text{(A2)}$$

An unbiased estimator of the frequency of allele u among n_i sampled individuals from the ith population is simply given by $\hat{p}_{iu} = \Sigma_{j=1}^{n} x_{iju}/n_i$. Expanding the square of this expression, and then taking expectation, gives $\varepsilon(\hat{p}_{iu}^2|\mathbf{p}) = [p_{iu} + n_i(n_i - 1)p_{iu}^2]/n_i$. Therefore,

$$\hat{Q}_{w,i} = \sum_{u=1}^{k} [\hat{p}_{iu}(n_i\hat{p}_{iu} - 1)]/(n_i - 1) \quad \text{(A3)}$$

is an unbiased estimator of the probability for two genes in population j to be identical in state, with k being the number of alleles in the sample. Similarly

$$\hat{Q}_a = \sum_{u=1}^{k} (\hat{p}_{1u}\hat{p}_{2u}) \quad \text{(A4)}$$

is an unbiased estimator of the IIS probability of two genes taken in the ancestral population, before divergence. Approximating the expectation of a ratio by the ratio of expectations, an estimator of F_i is given by

$$\hat{F}_i = \frac{\sum_{u=1}^{k}[\hat{p}_{iu}(n_i\hat{p}_{iu} - 1)/(n_i - 1) - \hat{p}_{1u}\hat{p}_{2u}]}{1 - \sum_{u=1}^{k}(\hat{p}_{1u}\hat{p}_{2u})}. \quad \text{(A5)}$$

When combining the information brought by all alleles at more than one locus, a multilocus estimator is defined as the ratio of the sum of locus-specific numerators over the sum of locus-specific denominators (see, *e.g.*, Weir and Cockerham 1984). It is worth noting that, when daughter population sizes are equal, this simple way to estimate parameters (*i.e.*, equating Q's to $\hat{Q}$'s in Equation 8 to get $\hat{F}$) directly yields Cockerham's estimators (Cockerham 1973; Weir and Cockerham 1984), developed with the methods of analysis of variance (see Rousset 2001 for a thorough demonstration of the equivalence between estimator formulas based on analyses of variance and expressions in terms of frequency of identical genes). Our estimator differs from previous ones (*e.g.*, Reynolds *et al.* 1983) in allowing separate parameters F_i's for each population.

Estimation of the density of F_i parameters: For each set of parameter values, coalescent simulations were performed, thus generating "artificial data sets." Each artificial data set yields a pair of estimates $\hat{F}_1$ and $\hat{F}_2$. An approximation to the expected joint distribution was obtained as follows. First, a two-dimensional histogram was constructed. Recall that the points $(\hat{F}_1, \hat{F}_2)$ are constrained to lie within the upper-right triangle of a square with vertices $(-1, -1)$, $(1, -1)$, $(-1, 1)$, and $(1, 1)$. The whole square region was covered by a two-dimensional array (or mesh) of 100×100 square cells. Each cell has thus sides of length 0.02. Each observation $(\hat{F}_1, \hat{F}_2)$ was binned in the appropriate cell. The cell counts were divided by the total number of observations to obtain a discrete probability distribution over the two-dimensional array. This discrete distribution is a close approximation to the expected joint distribution of the estimators $(\hat{F}_1, \hat{F}_2)$. The q-level "high probability region" ($q = 95\%$ or any other value) is constructed as follows. The cells are sorted in order of decreasing probability. Finally, starting from the cells with the highest associated probabilities, cells are sequentially added to the confidence region until the cumulative probability of the whole set of cells obtained is equal to (or just exceeds) the chosen q-value.

From this procedure, we obtain for each simulation a region within which a proportion q of the data lies. Note that this confidence region is not necessarily continuous. Constructing the high probability region using the discrete distribution is clearly preferable to using some (unnecessary) continuous approximation to it.

Annexe J

Taille efficace et banques de graines

Vitalis R., Glémin, S. and Olivieri I. (2002) F-statistics and effective population size with seed dormancy. En préparation.

F-statistics and effective population size with seed dormancy

Renaud Vitalis,[⋆†] Sylvain Glémin[⋆‡] and Isabelle Olivieri[⋆]

[⋆] *Laboratoire Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, 34095 Montpellier Cedex 05, France*

[†] *Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France*

[‡] *Station de Génétique et Amélioration des Plantes, Institut National de la Recherche Agronomique de Montpellier, Domaine de Melgueil, 34130 Mauguio, France*

Running head :
Seed bank population genetics
Keywords :
Seed dormancy, IBD probabilities, seed bank, F-statistics
Corresponding author :
Renaud VITALIS
Laboratoire Génétique et Environnement, C.C. 065
Institut des Sciences de l'Évolution de Montpellier
Université Montpellier II
Place Eugène Bataillon, 34095 Montpellier Cedex 05, France
Tel. : +33 4 67 14 32 50
Fax : +33 4 67 14 36 22
E-mail : `vitalis@isem.univ-montp2.fr`

INTRODUCTION

Seed dormancy defines the inability of mature seeds to germinate. Yet, this simple statement encompasses different processes (Leopold 1996) and this inactive phase may be the result of at least three phenomena : innate, secondary or enforced dormancy (Harper 1977). Innate dormancy prevents seeds from germinating onto the maternal plant ; secondary (induced) dormancy refers to the alteration of seed germination ability, induced by some environmental factor (*e.g.*, burial, chilling) experienced by seeds after maturation ; enforced dormancy characterizes the incapacity of seeds to germinate when one or several essential requirements for germination are lacking (e.g., light, water resource, oxygen) (Rees 1997). In plant populations, seed dormancy indubitably builds up the so-called « seed banks »(Harper 1977), i.e. pools of seeds that remain dormant in the soil. Commonly found among plant species, prolonged dormancy (beyond one breeding season) also exists in some insects, where diapausing eggs may play the same role as dormant seeds (De Stasio 1989; Weider *et al.* 1997).

Whatever the exact process involved, seed dormancy is an important life-history trait that profoundly influences plant population dynamics. What is the adaptive significance of delayed germination ? In plant communities living in habitats where many sites are available for colonization each year, germination rates have evolved to be higher and seeds rarely dorm beyond one breeding season (Rees 1996; Watkinson 1990). In those communities, species exist as seeds for only a short period of the year, and the resulting seed bank does not last more than a single year. But in other plant communities like those consisting of, *e.g.*, annuals in deserts (Venable *et al.* 1993) or aquatic plants in temporary marshes (Van der Valk and Davis 1978), prolonged seed dormancy has evolved so that seed banks may persist for more than one breeding season. Many models have been devoted to the evolution of seed dormancy (see Olivieri 2001, for a thorough review of the subject). Those models rely upon the rationale that dormancy, alike dispersal, prevents the risk of extinction, reduces crowding and enables seedlings to escape from sib-competition (Rees 1996; Venable and Brown 1988, 1993).

It is well understood that seed dormancy profoundly influences demography. Following a population crash, recruitment from the seed bank will prevent population extinction (Pacala 1986). Seed dormancy dampens oscillatory dynamics in variable environments (Crawley and Rees 1996; Thrall *et al.* 1989). Indeed, seed dormancy stabilizes population dynamics if the number of seed produced fluctuates from year to year. In years following low fecundity, recruitment from the seed bank increases population numbers. Furthermore, it has been shown that in grassland remnants in the

Swiss Jura mountains, species with longer-lived seeds in the soil had lower rates of local extinction than species with shorter-lived seeds, over a 35-years survey (Stöcklin and Fischer 1999).

However, seed banks may also destabilize population dynamics, and eventually lead to extinction, if more seeds enter the seed bank than germinate (McDonald and Watkinson 1981). Since seed dormancy delays the time before reproduction for at least one year, seed banks increase the generation time and generations may overlap. Also, longer generation times may slow the rate of population increase in growing populations, and may slow the rate of population decrease in declining populations (Caswell and Hastings 1980). This stabilizing effect of seed banks on population dynamics has been partially evidenced in the field (Jarry *et al.* 1995; Kalisz 1991; Kalisz and McPeek 1992).

What are the consequence of seed dormancy on population genetics ? The role of genetic differentiation over time was highlighted in a unique series of articles by McGraw *et al.* (1991) ,Vavrek *et al.* (1991) and Bennington *et al.* (1991). McGraw *et al.* (1991) sampled and dated seeds from an organic soil layer in Arctic toundra. They showed that viable buried seeds had persisted there for at least a century. A genetic change over time in *Luzula parviflora* « temporal »populations was found, the oldest population being estimated as old as 150 years (Bennington *et al.* 1991). Indeed, persistent phenotypic differences in common or multiple environments experiments showed genetic differences between different-aged seeds. Different responses to multi-environment conditions (different nutrient levels) were also shown among *Carex bigelowii* temporal populations (Vavrek *et al.* 1991). Adaptive genetic polymorphism is expected to be more easily maintained in temporally varying environments if the emergence of seedlings is delayed through time (Hedrick 1995). Yet, the presence of seed or egg banks generally slows down the rate of evolutionary change, thus acting as an « evolutionary load »(Brown and Venable 1986; Templeton and Levin 1979). This has been shown empirically by Hairston and De Statio, Jr. (1988) who observed a lag-time between the occurrence of intense predation events and the adaptive response (diapause) of the preys, the copepod *Diaptomus sanguineus*. Thus, the gain in genetic variance (Levin 1990) may increases the response to different selective regimes, but may also results in maladapted responses due to the emergence of genotypes issued from ancient environments.

Seed dormancy may also greatly influence the level and the distribution of neutral genetic variation in natural populations. This has been recognised recently through a series of articles that aim at comparing the levels of allozyme variation between seeds collected in the soil and flowering plants (Alvarez-

Buylla *et al.* 1996; Cabin 1996; Cabin *et al.* 1998; Del Castillo 1994; Mahy *et al.* 1999; McCue and Holtsford 1998; Tonsor *et al.* 1993). In this paper, we will develop a theoretical framework to examine the distribution of gene identity at different hierarchical levels. This will be used to derive the asymptotic effective population size, as well as the equilibrium values of F-statistics, within and among populations. There is a sound analogy between an annual life cycle with a seed bank and that of a perennial plant that lacks seed dormancy. Standing crops or rosettes replace the seed bank as the cause of the establishment of overlapping generations (Templeton and Levin 1979). We will therefore also examine the role of delayed maturation of perennial plants on the expected level and distribution of genetic variation.

THE MODEL

Demography : We consider a finite island model of population structure (WRIGHT 1951). Dispersal is achieved through the movements of pollen grains (pollen dispersal) and through the movements of seeds (seed dispersal). In any case, a dispersed pollen grain or a dispersed seed is equally likely to come from any of n populations.

[**Fig. 1** about here.]

Flowering plants produce f seeds *per capita* each generation. Seeds disperse with probability z and then enter a dormant stage with probability d. A fraction τ of the seeds die each generation in the seed bank, and a fraction c does not survive to dispersal. The maximum age a seed can attain is n_c, over which it dies. We assume that there is no local extinction, so that at any time the stable age-structure is the same in all populations. Following the life-cycle graph given in figure 1 we may obtain the transition matrix that gives the change of the number of seeds in each class, from one generation to the next. We note N_i the number of i years old seeds and N is the number of flowering plants that survive each generation (after regulation). Therefore

$$\begin{pmatrix} N \\ N_1 \\ N_2 \\ \vdots \\ N_{n_c} \end{pmatrix}^{t+1} = \begin{pmatrix} f[(1-d)(1-z)+z(1-c)] & (1-d) & (1-d) & \cdots & (1-d) \\ fd(1-\tau)(1-z) & 0 & 0 & \cdots & 0 \\ 0 & d(1-\tau) & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & d(1-\tau) & 0 \end{pmatrix} . \begin{pmatrix} N \\ N_1 \\ N_2 \\ \vdots \\ N_{n_c} \end{pmatrix}^{t} \tag{1}$$

We assume density dependent regulation in the surface population, so that the number of surviving seedlings is N. At equilibrium, the number of seeds in age-class i is given by

$$N_i = fN(1-z)\left[(1-\tau)d\right]^i \tag{2}$$

We assume that the seeds germinate from the seed bank in proportion to their abundance in each age-class. Thus, among the pool of seeds that germinate from the seed bank, the proportion of i-aged seeds is the ratio of the number of seeds in the ith age-class over the total number of seeds in the seed bank. The probability that a seed germinating from the seed bank is i years old is given by $p_i = N_i / \left(\sum_{j=1}^{n_c} N_j\right)$, i.e.

$$p_i = \frac{1-(1-\tau)d}{\left[1-(1-\tau)d\right]^n}\left[(1-\tau)d\right]^{i-1} \tag{3}$$

Let m_p be the probability that a pollen grain has immigrated and m_s the probability that a seed has immigrated. Assuming that dispersed seeds do not enter in a dormant state (plastic dormancy), then

$$m_s = \frac{(1-c)z}{(1-c\ z) - d(1-z)} \tag{4}$$

[**Table 1** about here.]

For the sake of clarity, all the notations used throughout the paper are given in Table 1.

After pollen dispersal, individuals reproduce. A fraction s of the offspring is produced by selfing. We define the selfing rate s as the conditional probability that two homologous gene copies of one individual come from the same parent, given that they both come from one population before pollen migration. The overall amount of selfing is thus $(1-m_p)^2 s$. Locally, in a completely random mating population of N hermaphroditic individuals, $s = 1/N$. After reproduction, individuals produce seeds that are dispersed according to an island model of dispersal. Genes are sampled among seedlings. Among those, a fraction $(1-k)$ comes from the seed bank, and a fraction k have been produced in the previous generation.

$$k = \frac{(1-d)(1-z)\left[1-(1-\tau)d\right]}{(1-d)(1-z)\left[1-(1-\tau)^{n+1}d^{n+1}\right] + z(1-c)\left[1-(1-\tau)d\right]} \tag{5}$$

Each generation, after pollen dispersal, the frequency of pairs of pollen grains taken at random in a population that come from a single population before dispersal is $\alpha = (1-m_p)^2 + m_p^2/(n-1)$ (see, e.g., Rousset 1996). Conversely, the frequency of pairs of seedlings taken at random from two populations that originate from the same population before dispersal is $\beta = (1-\alpha)/(n-1)$. In the following, we will note with a star (like in a^*) the backward probabilities of the origin of pairs of genes before pollen dispersal. No subscript will be used to not the backward probabilities of the origin of pairs of genes after seed dispersal. With probability 1/4, two sampled genes come from female gametes and are thus necessarily resident. With the same probability, two sampled genes come from male gametes and come from the same population with probability α. With probability 1/2, one gene comes from a male gamete and the other from a female gamete. Those come from the same population with probability $(1-m_p)$. Therefore, the frequency of pairs of genes sampled in one population that come from the same population before pollen dispersal is $a^* = (\alpha+1)/4 + (1-m_p)/2$. Conversely, the frequency of pairs of genes taken in distinct population that come from the same population before pollen dispersal is $b^* = \beta/4 + m_p/(2n)$.

Since genes are sampled among seedlings after seed dispersal, some other probabilities will be useful. Each generation, after seed dispersal, the frequency of pairs of seedlings in a population that come from one population before dispersal is $a = (1 - m_s)^2 + m_s^2/(n-1)$. Conversely, the frequency of pairs of seedlings sampled in distinct populations that originate from the same population before dispersal is $b = (1-a)/(n-1)$.

The number of offspring produced each generation is assumed to be very large (infinite) and the random number of offspring that descend from a single parent is binomially distributed as $\mathcal{B}\left(N, \frac{1}{N}\right)$. Therefore, an offspring is equally likely to descend from any of the N individuals in one population. Mutations arise during meiosis at rate μ and $\gamma = (1-\mu)^2$ is the probability that two genes have not undergone mutation in one generation. We assume an infinite model (IAM) of mutation.

Identity By Descent : Let Q_0^* be the IBD probability of two copies of a gene sampled after pollen dispersal (before seed dispersal) within an individual. Let Q_1^* be the corresponding probability for two gene copies sampled from two individuals in one population after pollen dispersal, and Q_2^* the IBD probability of two gene copies sampled among populations after pollen dispersal. Again, the superscript indicates that the profitabilities are evaluated after pollen dispersal. The corresponding probabilities for pairs of genes sampled after seed dispersal (Q_0, Q_1 and Q_2) have no superscript.

After pollen dispersal (before seed dispersal), the pairs of gene copies sampled within individuals may come from the same parent before pollen dispersal (selfing) with probability s or from different parents (outcrossing) with probability $(1-s)$. In this latter case, the gene inherited from the father may be resident with probability $(1-m_p)$ or migrant with probability m_p. Therefore, the IBD probability of a pair of genes taken within an individual after pollen dispersal is given by

$$Q_0^*(t) = \gamma \left[s\frac{Q_0(t)+1}{2} + (1-s)\left[(1-m_p)Q_1(t) + m_p Q_2(t)\right]\right] \tag{6}$$

(note that we consider that a generation ends up after seed dispersal, so that the following probabilities at generation t depend also on equivalent probabilities at generation t). With similar arguments,

$$Q_1^*(t) = \gamma \left[a^* \left[\frac{1}{N}\frac{Q_0(t)+1}{2} + \left(1-\frac{1}{N}\right) Q_1(t)\right] + (1-a^*)Q_2(t)\right] \tag{7}$$

and

$$Q_2^*(t+1) = \gamma \left[b^* \left[\frac{1}{N}\frac{Q_0(t)+1}{2} + \left(1-\frac{1}{N}\right) Q_1(t)\right] + (1-b^*)Q_2(t)\right] \tag{8}$$

Let $R_1^{(i)*}$ be the probability that a gene sampled among the seedlings after pollen dispersal is IBD to a gene sampled in an i-aged seed in the same population (i.e. a seed which is in the i^{th} age-class in the local seed bank). Let $R_2^{(i)*}$ be the corresponding probability for a gene sampled in a seedling of a given population after pollen dispersal and a gene sampled in an i-aged seed in a distant seed bank.

$$R_1^{(i)*}(t) = \gamma \left[\left(1 - \frac{m_p}{2}\right) R_1^{(i)}(t) + \frac{m_p}{2} R_2^{(i)}(t) \right] \tag{9}$$

and

$$R_2^{(i)*}(t) = \gamma \left[\frac{m_p}{2n} R_1^{(i)}(t) + \left(1 - \frac{m_p}{2n}\right) R_2^{(i)}(t) \right] \tag{10}$$

where R_1 is the probability that a gene sampled among the seedlings after seed dispersal is IBD to a gene sampled in an i-aged seed in the same population, and $R_2^{(i)}$ is the corresponding probability for a gene sampled in a seedling of a given population after seed dispersal and a gene sampled in an i-aged seed in a distant seed bank.

As noted above, Q_0 is the probability that the two copies of a gene sampled within an individual after seed dispersal are identical by descent. Q_1 is the IBD probability for two gene copies sampled from two individuals in one population after seed dispersal and Q_2 is the corresponding probability for two gene copies sampled among populations, after seed dispersal. Then

$$Q_0(t+1) = kQ_0^*(t) + (1-k) \sum_{i=1}^{n_c} p_i S_0^{(i)}(t) \tag{11}$$

and

$$\begin{aligned} Q_1(t+1) &= k^2 \left[aQ_1^*(t) + (1-a)Q_2^*(t) \right] \\ &+ 2k(1-k) \left[\sum_{i=1}^{n_c} p_i \left[(1-m_s) R_1^{(i)*}(t) + m_s R_2^{(i)*}(t) \right] \right] \\ &+ (1-k)^2 \sum_{(i,j)}^{n_c} p_i p_j S_1^{(i,j)}(t) \end{aligned} \tag{12}$$

and

$$\begin{aligned} Q_2(t+1) &= k^2 \left[bQ_1^*(t) + (1-b)Q_2^*(t) \right] \\ &+ 2k(1-gk) \left[\sum_{i=1}^{n_c} p_i \left[\frac{m_s}{n} R_1^{(i)*}(t) + \left(1 - \frac{m_s}{n}\right) R_2^{(i)*}(t) \right] \right] \\ &+ (1-k)^2 \sum_{(i,j)}^{n_c} p_i p_j S_2^{(i,j)}(t) \end{aligned} \tag{13}$$

where $S_0^{(i)}$ is the probability that two homologous gene copies within a i-aged seed are IBD, $S_1^{(i,j)}$ is the IBD probability of two gene copies sampled among one i-aged seed and one j-aged seed, in the same (local) seed bank, and $S_2^{(i,j)}$ is the corresponding probability for seeds sampled among distinct seed banks.

The newly produced seeds enter in the seed bank each generation. Therefore, $S_0^{(1)}(t+1) = Q_0^*(t)$, $S_1^{(1,1)}(t+1) = Q_1^*(t)$ and $S_2^{(1,1)}(t+1) = Q_2^*(t)$. As soon as they stand in the seed bank, pairs of genes can not coalesce. For seeds strictly older than one year, the IBD probability of a given pair of genes equals the same IBD probability for one-year-younger seeds, in the previous generation. This is true for homologous pairs of genes taken within seeds, *i.e.* $S_0^{(i)}(t+1) = \gamma' S_0^{(i-1)}(t)$, as well as for pairs of genes taken among local seeds, *i.e.* $S_1^{(i,j)}(t+1) = \gamma' S_1^{(i-1,j-1)}(t)$, or among distant seeds, *i.e.* $S_2^{(i,j)}(t+1) = \gamma' S_2^{(i-1,j-1)}(t)$. The IBD probability of a pair of genes taken among a one-year-old seed and a j-aged seed equals the probability that, in the previous generation, a gene taken in a $(j-1)$-aged seed is IBD with a gene sampled in a seedling. Thus, $S_1^{(1,j)}(t+1) = S_1^{(j,1)}(t+1) = R_1^{(j-1)*}(t)$ and $S_2^{(1,j)}(t+1) = S_2^{(j,1)}(t+1) = R_2^{(j-1)*}(t)$. Accordingly,

$$\begin{aligned} R_1^{(1)}(t+1) &= k\left[(1-m_s)Q_1^*(t) + m_s Q_2^*(t)\right] \\ &+ (1-k)\sum_{i=1}^{n_c} p_i R_1^{(i)*}(t) \end{aligned} \tag{14}$$

$$\begin{aligned} R_1^{(i)}(t+1) &= k\left[(1-m_s)R_1^{(i-1)*}(t) + m_s R_2^{(i-1)*}(t)\right] \\ &+ (1-k)\sum_{j=1}^{n_c} p_j S_1^{i-1,j}(t) \end{aligned} \tag{15}$$

$$\begin{aligned} R_2^{(1)}(t+1) &= k\left[\frac{m_s}{n} Q_1^*(t) + (1-\frac{m_s}{n})Q_2^*(t)\right] \\ &+ (1-k)\sum_{i=1}^{n_c} p_i R_2^{(i)*}(t) \end{aligned} \tag{16}$$

and

$$\begin{aligned} R_2^{(i)}(t+1) &= k\left[\frac{m_s}{n} R_1^{(i-1)*}(t) + (1-\frac{m_s}{n})R_2^{(i-1)*}(t)\right] \\ &+ (1-k)\sum_{j=1}^{n_c} p_j S_2^{(i-1,j)}(t) \end{aligned} \tag{17}$$

NUMERICAL ANALYSES : ANNUALS WITH A SEED BANK

Demography : We assume that the fecundity is large (infinite), so that the number of seeds in each age-class is potentially infinite. This amounts to assuming that there is no density dependence of any kind in the seed bank. The probability distribution of the age of germinating seeds is given in Figure 2 for various germination rates. The probability that a seed germinates exactly one year after it has been fecundated is k (see Eqn 5), while the other probabilities are given by $(1-k)p_i$'s, so that the bar lengths sum to one. For low germination rates, all age-classes contribute nearly equally to the recruitment of seeds. For high germination rates instead, the contribution of the seed pool (seeds older than a single year) is quasi absent.

[Fig. 2 about here.]

Effective population size : The effective population size was derived from the leading (real) eigenvalue of the transition matrix defined from the recursion equations of identity probabilities (see, e.g., Ewens 1982). The calculations were numerically performed using Mathematica 4.0 (Wolfram Research 1999). For a range of maximum seed longevity (i.e. the number of age-classes in the model), we plotted the effective size as a function of germination rate (Fig. 3). In a single population, effective population size is always greater than the census number of flowering plants in the population (Fig. 3A). The smaller the germination rate, the larger the effective population size. As can be seen from Figure 2, smaller germination rates are associated with recruitment from a greater number of age-classes. Thus, large effective sizes are the consequence of large (effective) contribution from the seed bank. For a given germination rate, the effective size increases with seed longevity.

[Fig. 3 about here.]

Temporal Wahlund's effect : We define F_{IS} as the intra-class correlation of IIS probabilities for pairs of genes sampled within individuals relative to genes sampled among individuals within the same population (Rousset 1996; Weir and Cockerham 1984)

$$F_{IS} = \frac{Q_0 - Q_1}{1 - Q_1} \tag{18}$$

F_{IS} measures the within-population differentiation among the flowering plants. Since we did not consider isolation by distance within populations, spatial Wahlund's effect (see, e.g., Hartl and Clark 1989, p. 282) cannot be invoked to explain any departure from Hardy-Weinberg expectation (HWE :

$F_{IS} = 0$). However, one might expect departure from HWE, the so-called « temporal Wahlund's effect »(Alvarez-Buylla *et al.* 1996).

Figure 4 shows the equilibrium value of F_{IS}, as a function of the number N of flowering plants and the germination rate $g = (1-d)$, for contrasted mating systems. Figure 4A shows the equilibrium value of F_{IS} of randomly mating flowering plants. With large number of flowering plants and germination rates, F_{IS} is close to zero. This indicates that the distribution of genotypes in the surface vegetation nearly follows that predicted by HW expectation. Note that with $g = 1$, the model reduces to a single random mating population with non overlapping generations, for which $F_{IS} = 0$ is a well known prediction. With smaller germination rates however, F_{IS} values depend on the number of flowering plants. As the number of flowering plants decreases, F_{IS} values get larger. This should be interpreted as follows. With fewer individuals in the surface population, genetic drift is stronger and gene frequencies will vary by chance from one generation to the next. Therefore, each new generation of plants, resulting from the recruitment of seeds from different ages, results in a population made of several « temporal »populations, which resembles a population made of several « spatial »populations. This effect has a greater amplitude with larger seed longevity and larger survival rate. Note however that this effect is negligible for reasonable surface population sizes.

Figure 4B shows the equilibrium value of F_{IS} of selfing flowering plants. With g tending towards one (remember that $g = 1$ corresponds to the completely annual situation), F_{IS} tends towards one for large (surface) population sizes, while it decreases as the surface population size decreases. This result, although derived already in some occasions (see, e.g., Rousset 1996), might sound a bit odd. Indeed it has not always been well recognized that small population sizes may indeed depreciate F_{IS} values in partially selfing populations. In a partially selfing population of finite size N with gametic migration $F_{IS} = \sigma/(2-\sigma)$ (see, e.g., Rousset 1996)), where $\sigma = s - 1/N$ represents the proportion of selfing that is not due to panmixia. For smaller germination rates however, this effect is counterbalanced by the « temporal Wahlund's effect »discussed above (see Figure 4A). This results in intermediate values of F_{IS}, that lack biological significance for the population biologist.

[Fig. 4 about here.]

Population differentiation : We define F_{ST} as the intra-class correlation of IIS probabilities for pairs of genes sampled within populations relative to genes sampled among populations (Rousset 1996; Weir and Cockerham

1984), i.e.

$$F_{ST} = \frac{Q_1 - Q_2}{1 - Q_2} \tag{19}$$

Figure 5A shows the equilibrium value of F_{ST} as a function of germination rate. We can see that the overall population differentiation increases with the germination rate. The effect is stronger as the maximum age a seed can reach is larger. This is to be understood as the consequence of the relationship between (local) effective population size and germination rate (Fig. 3). With lower germination rates, effective population size is *locally* larger. Thus, genetic drift is locally less intense and F_{ST}, which measures « the amount of differentiation among populations, relative to the limiting amount under complete fixation »(WRIGHT 1978). Therefore, the local increase of genetic diversity due to the seed bank decreases the equilibrium value of F_{ST} (see NAGYLAKI 1998). This suggests that F_{ST} cannot be useful to quantify gene flow in populations with active seed banks, unless the age-structure is known.

[FIG. 5 about here.]

Demography

Now, let us consider a slightly modified version of the seed bank model. So far we assumed that dispersed seeds could not become dormant in their target population. What happens if we relax this assumption ? This model could also be understood as fitting well with a monocarpic perennial life-cycle. Then, the seed stages could be interpreted as seedling stages of different ages. In this case, the proportion g reflects the yearly flowering fraction, N_i is the number of rosettes of age i and N is the number of flowering plants. The transitions between age-classes may be written as

$$\begin{pmatrix} N \\ N_1 \\ N_2 \\ \vdots \\ N_{nc} \end{pmatrix}^{t+1} = \begin{pmatrix} f(1-d)(1-cz) & (1-d) & (1-d) & \cdots & (1-d) \\ fd(1-\tau)(1-zc) & 0 & 0 & \cdots & 0 \\ 0 & d(1-\tau) & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & d(1-\tau) & 0 \end{pmatrix} . \begin{pmatrix} N_0 \\ N_1 \\ N_2 \\ \vdots \\ N_{nc} \end{pmatrix}^{t} \quad (20)$$

Yet, in monocarpic perennial plant populations, density dependent regulation must not be limited to juvenile stages (as in the seed bank model considered so far). Instead, density dependent regulation occurs at the scale of all the plants in the population, whatever their stage (including seedlings). It can be shown that the number of flowering plants is given by

$$N = \frac{(1-d)}{1+(1-d)(1-f(1-cz))} N_T \quad (21)$$

where N_T is the census number of plants (including seedlings) that can be maintained in the population at carrying capacity. The fecundity f shall not be lower than f^*, defined as

$$f^* = \frac{1-d(1-\tau)}{(1-d)(1-c\ z)\left[1-[d(1-\tau)]^{n+1}\right]} \quad (22)$$

Also, there is no reason to assume here that dispersed seeds must grow and flower in the next season (as it was previously assumed in the previous section, where the dormancy rate of dispersed seeds was set to zero). Therefore, the proportion k of individuals that have been produced in the previous generation is given by

$$k = \frac{1-(1-\tau)d}{1-[(1-\tau)d]^{n+1}} \quad (23)$$

and the backward seed dispersal rate becomes

$$m_s = \frac{(1-c)z}{(1-c\ z)} \tag{24}$$

Also, the following transition equations must replace the previous ones.

$$S_1^{1,1}(t+1) = (1-m_s)Q_1^*(t) + m_s Q_2^*(t) \tag{25}$$

$$S_2^{1,1}(t+1) = \frac{m_s}{n}Q_1^*(t) + \left(1 - \frac{m_s}{n}\right) Q_2^*(t) \tag{26}$$

$$S_1^{1,j}(t+1) = (1-m_s)R_1^{(j-1)*} + m_s R_2^{(j-1)*} \tag{27}$$

$$S_2^{1,j}(t+1) = \frac{m_s}{n}R_1^{(j-1)*} + (1 - \frac{m_s}{n})R_2^{(j-1)*} \tag{28}$$

$$\begin{aligned} R_1^{(1)}(t+1) &= k\left[(1-m_s)Q_1^*(t) + m_s Q_2^*(t)\right] \\ &+ (1-k)\sum_{i=1}^{n_c} p_i \left[(1-m_s)R_1^{(i)*}(t) + m_s R_2^{(i)*}(t)\right] \end{aligned} \tag{29}$$

$$\begin{aligned} R_2^{(1)}(t+1) &= k\left[\frac{m_s}{n}Q_1^*(t) + (1 - \frac{m_s}{n})Q_2^*(t)\right] \\ &+ (1-k)\sum_{i=1}^{n_c} p_i \left[\frac{m_s}{n}R_1^{(i)*}(t) + \left(1 - \frac{m_s}{n}\right) R_2^{(i)*}(t)\right] \end{aligned} \tag{30}$$

Effective size

Figure 6 shows the asymptotic effective population size of a monocarpic perennial plant population. With the flowering rate being below a certain threshold, effective population size is less than the census number of individuals (seedlings and flowering plants) in the population. This is because density dependent regulation at the whole population scale puts a constraint upon fecundity so that many individuals contribute to the total seed set. The effective number of individuals who reproduce is thus low. Beyond the threshold, effective population size gets larger than the census number of individuals in the population. Just as in the seed bank model, the absence of coalescent events in the seedling stages explain why effective size gets larger.

[Fig. 6 about here.]

Population differentiation

Figure 5B shows the equilibrium value of the parameter F_{ST} as a function of flowering rate. With low flowering rate, the local effective population size is smaller than the census number of flowering plants in the population (Fig. 6). Strong genetic drift occurring within each population increases the apparent population differentiation. With increasing flowering rates, the local effective population size gets larger and the apparent differentiation decreases.

DISCUSSION

Effective size in age-structured populations

Comparison with alternative age-structures : Effective size of age-structured populations has been investigated in a variety of ways and models (see Charlesworth 1994; Felsenstein 1971; Hill 1972, 1979; Johnson 1977; Orive 1993; Pollak 2000; Rousset 1999, among others). Felsenstein (1971) found that the effective population size of an age-structured population is equal to the number of newborns per generation times the generation time, and times the reciprocal of a term which can be strictly interpreted as the probability that an individual dies while it has still a positive reproductive value. The later terms have also been expressed by Hill (1972, 1979) as a function of the variances and covariances of life-time reproductive success (see also Pollak 2000, for more recent developments). Both formulas have been shown to be equivalent (Charlesworth 1994; Johnson 1977). In those models, effective size is expected to be less than the census number of individuals in the population, or equal to that number only if individuals all have the same reproductive value (Nagylaki 1980; Rousset 1999). Investigating the effects of various mating systems and age-structures, Nunney (1993) found that effective population size should rarely exceed one-half of the census number (but see Husband and Barret 1992, 1995; Nunney 1995, for a discussion about the model hypotheses).

Yet, with seed dormancy, genes cannot coalesce once they are engaged within the seed bank. Because such coalescent events cannot occur, the genes cannot become identical as a result of coalescence (Rousset 1999). Kaj *et al.* (2001) found that the rate of coalescence of sets of genes in a plant population with a seed bank was decreased by a factor equal to the squared inverse of the mean age T of seeds in the bank. Adding mutations to their model, they concluded that the presence of a seed bank, through the decrease of coalescence rates, eventually increases genetic variability. It is noteworthy that Kaj *et al.*'s (2001) results imply that the coalescent process they described reduces to the standard coalescent process with time measured in units of $N_e = N \cdot T^2$ generations (see, e.g., Norborg 2001).

Generation time : As highlighted by Rousset (1999), the effective size defined as the rate of coalescence of ancestral pairs of genes (Norborg 2001) is equal to the product of the inbreeding effective size of Felsenstein (1971) and the generation time. Note also that Felsenstein's (1971) inbreeding effective size is itself very close to Orive's (1993) coalescent-based definition of effective size. The generation time is usually defined as the mean age of

the parents of a newborn offspring (Charlesworth 1994). Yet, with more complex life cycles (e.g., stage-structured population models), such dependance poses the problem of how the generation time is defined (see Orive 1993, for the case of clonally reproducing organisms). Therefore, comparing effective population size from populations or species with different age or stage structures requires that effective size is measured in the same time-scale (Rousset 1999). There is a sound analogy with demography and matrix population models. Likewise, the asymptotic rate of increase (usually noted λ) of the population number is a parameter which expression is given per unit of time and not per generation (Caswell 1989, 2001).

Local *vs.* global effective size : We have seen that the presence of a seed bank increases the effective population size. In spatially structured populations, two effects should be accounted for. First, seed dormancy increases the local effective population size. Second, population subdivision increases effective size further (Wright 1939). Yet, the island model of population structure considered here may lack some generality. In the strong-mutation limit, where migration overcomes all the other evolutionary forces (Nagylaki 1980), the distribution of the place and time of coalescent events are independent of the demes sampled (Nagylaki 2000). This means that the distribution of gene lineages can be described as a coalescent process, in the appropriate time-scale (Charlesworth 2001). But in those situations where the deme size is generally not preserved from one generation to the next (Nagylaki 1980; Strobeck 1987), some demes contribute more to the future generations (i.e. they act as « source »populations), while others contribute less (« sink »populations). In this situation, gene lineages are not likely to be evenly distributed across the metapopulation as a whole. Rather, lineages are more likely to coalesce in source than in sink populations (Wakeley 2001). The variance among demes in their contribution to the total gene pool makes the global effective population size eventually smaller than expected in the classical (symmetric) island model (Whitlock and Barton 1997). Thus, further investigations are needed to evaluate whether the effect of variance of single demes contribution to the gene pool could counteract the local increase in effective populations size due to the presence of a seed bank.

Several authors have suggested that seed bank indeed increases global effective size (e.g., Tonsor *et al.* 1993), and have used this argument to explain why F_{ST} in species with a seed bank might be lower than in species without a seed bank. Our paper is the first to provide a quantitative description of the effect of seed bank on genetic differentiation among subpopulations.

Seed bank as a reservoir of allelic diversity

The idea that seed banks might have evolutionary consequences was formalized by Templeton and Levin (1979), and more recently by Hedrick (1995), with a special emphasis on loci at which fitness varies through time. Following this idea, several authors have tested the hypothesis that established plants represent a non-random sample of the seed bank. For instance, Tonsor *et al.* (1993) in *Plantago lanceolata* (Plantaginaceae), and Cabin *et al.* (1998) in *Lesquerella fendleri* (Brassicaceae) found overall differences of allele frequencies at allozyme loci between soil seeds and surface plants. However, the amount of genetic diversity among the surface plants was either similar (Tonsor *et al.* 1993) or larger (Cabin *et al.* 1998) than the genetic diversity among soil seeds. Other studies, like that of Mahy *et al.* (1999) in *Calluna vulgaris* (Ericaceae), found no difference in either allele frequencies or the amount of genetic diversity. Whenever allele frequencies differed between the surface plants and the soil seeds, the difference was attributed to natural selection, acting either directly (e.g., on PGI locus in *L. fendleri*) or through linkage with selected loci. This interpretation is further confirmed by the increase of observed heterozygosity (and a corresponding decrease in departure from Hardy-Weinberg expectation) when comparing soil seeds to seedlings (see last section).

More recently, empiricists started to address the hypothesis that seed banks also act as reservoirs for neutral genetic diversity. For instance McCue and Holtsford (1998) wrote that "if the seed bank has the potential to maintain genetic diversity we would expect to see as much or more diversity in the seed bank as in the standing crop of adults". They indeed found that in *Clarkia springvillensis* (Onagraceae), allozyme polymorphism was larger in the seed bank compared to the surface plants population. However, as the authors of this study suggested later in the paper, and ignoring selective effects following germination, the adults in fact represent a random sample of the seed bank, and so do the soil seeds that were collected and germinated from the seed bank for the allozyme analyses. There is thus no reason to expect that the genetic diversity of the seed bank as measured by germinating a random sample from soil seeds should be any larger than the genetic diversity measured from seedlings or established adults, provided sample sizes are similar. Similarly, there is no reason to expect a larger genetic differentiation among surface plants populations, than among seed bank from different populations. Whenever such differences have been found, they should be attributed to selective events following germination and not to the buffering effect of the seed bank.

In fact there are two ways one could test the hypothesis that seed banks

indeed represent reservoirs of neutral diversity as well. The first way would be to conduct a comparative analysis of the genetic diversity in species with a seed bank and species without a seed bank. Such comparative analysis unfortunately is likely to be difficult to conduct as there are several potentially confounding effects. An alternative procedure, concentrating on a species with a seed bank and experiencing local extinctions and recolonizations, would be to compare the genetic diversity among soil seeds between extinct (with respect to the surface population) and extant populations. Comparing the same life-history stage would allow to avoid the potentially confounding effects of selection following germination and establishment in the field. It is expected that, provided the seed bank of extinct populations has persisted for a long time, soil seeds from extinct populations should present rare alleles that might no longer occur within extant populations.

Temporal Whalund's effect will usually not explain positive F_{IS} in the seed bank

Three mechanisms have been offered by most authors to explain the often observed larger homozygosity in the seed bank, with significant departures from Hardy-Weinberg expectations, when compared to the larger observed heterozygotes frequencies among the surface plants (Alvarez-Buylla *et al.* 1996; McCue and Holtsford 1998; Tonsor *et al.* 1993).

The first mechanism is that the larger homozygosity in the seed bank compared with the adults could be caused by higher rates of inbreeding in previous years. But, as emphasized above, the adults are precisely coming from the seed bank. So there is absolutely no reason to expect that they were produced by a population experiencing a different mating system than the population which produced the seed bank.

The second mechanism that has been put forward is a temporal Wahlund effect. The classical (spatial) Wahlund's effect is the lower observed heterozygosity of a subdivided population with heterogeneous allele frequencies among subpopulations, compared to the heterozygosity in a single panmictic population with the same mean allele frequencies. This results in significant departure from Hardy-Weinberg genotypic genotypic proportions in a subdivided population compared to a single panmictic population. Such departure arises only when allelic frequencies vary among subpopulations (spatial Wahlund's effect, see Wahlund 1928). With overlapping generations, the deviation from Hardy-Weinberg genotypic proportions arises as the consequence of both the variance in gene frequency between resident and immigrant individuals and the variance between resident cohorts (Christiansen 1988).

As noticed by Christiansen (1988), in a population with overlapping generations, only each cohort in the population is initially in Hardy-Weinberg expected proportions.

In the particular case of seed bank induced age-structure, we have shown that a delayed germination does not in fact generate a temporal Wahlund's effect, unless the population size is dramatically small ($\leqslant 10$). Thus such potential Wahlund effect are unlikely to explain the significant homozygotes excess that were found by, e.g., Cabin *et al.* (1997) or McCue and Holtsford (1998). Moreover, for neutral loci, there is no theoretical reason to expect a larger departure from Hardy-Weinberg expectations in the seed bank compared to the surface plants. In only one study, was the F_{IS} of the seed bank compared with the F_{IS} among offspring of single trees (before dispersal). Alvarez-Buylla *et al.* (1996) studied allozyme polymorphism in *Cecropia obtusifiolia* (Moraceae). They sampled seeds from trees ("tree seeds"), seeds trapped from the overall seed rain ("seed rain", thus after dispersal), seeds from the soil (seed bank), seedlings, juveniles and adults. Interestingly, they found no departure from panmixia among tree seeds, a significant departure (positive F_{IS}) in seed rain and seed bank, and a trend towards increasing heterozygote frequencies from seedling to adults. A temporal Wahlund effect could explain the positive F_{IS} in the seed bank, especially given that the tree seed data suggested random mating. The most likely explanation for the seed rain positive F_{IS} is a spatial Wahlund effect, and certainly a temporal Wahlund effect could not possibly be involved. To explain the negative F_{IS} among adults, one must call for the third mechanism.

The third (and most likely) mechanism for decreasing F_{IS} from seeds to adults is natural selection. Provided homozygotes at allozyme loci are more likely to be inbred for most of their genome, compared to heterozygotes, and provided that there is some degree of inbreeding depression, it is expected that natural selection will progressively eliminate more inbred, and thus more homozygote, genotypes from seeds to adults. This is the most likely explanation for the results of McCue and Holtsford (1998) and Alvarez-Buylla *et al.* (1996).

A few studies have also compared the values of among-populations differentiation in the seed bank with that in the surface adults. For the same reason given above, that adults are issued from the seed bank, there is no reason to expect that F_{ST} will vary with life-stage, unless direct or indirect selection is involved.

To conclude, most data set suggesting that seed bank and surface adults differ with respect to the amount of polymorphism, allele frequencies, local departure from Hardy-Weinberg populations, or genetic differentiation among populations, cannot be explained by the storage effect of seed bank

or a temporal Wahlund effect. Natural selection acting upon the allozymes, or, rather on linked loci, is the most likely explanation for the observed patterns.

Références

Alvarez-Buylla, E. R., A. Chaos, D. Pinero and A. A. Garay, 1996 Demographic genetics of a pioneer tropical tree species : patch dynamics, seed dispersal, and seed banks. Evolution, **50** : 1155–1166.

Bennington, C. C., J. B. McGraw and M. C. Vavrek, 1991 Ecological genetic variation in seed banks. II. phenotypic and genetic differences between young and old subpopulations of *Luzula parviflora*. Journal of Ecology, **79** : 627–643.

Brown, J. S., and D. L. Venable, 1986 Evolutionary ecology of seed bank annuals in temporally varying environments. The American Naturalist, **127** : 31–47.

Cabin, R. J., 1996 Genetic comparisons of seed bank and seedling populations of a perennial desert mustard, *Lesquerella fendleri*. Evolution, **50** : 1830–1841.

Cabin, R. J., A. S. Evans and R. J. Mitchell, 1997 Genetic effects of germination timing and environment : an experimental study. Evolution, **51** : 1427–1434.

Cabin, R. J., R. J. Mitchell and D. L. Marshall, 1998 Do surface plant and soil seed bank populations differ genetically ? a multipopulation study of the desert mustard *Lesquerella fendleri* (brassicaceae). American Journal of Botany, **85** : 1098–1109.

Caswell, H., 1989 *Matrix population models*. Sinauer Associates Inc., Sunderland.

Caswell, H., 2001 *Matrix population models : Construction, Analysis, and Interpretation*. Sinauer Associates Inc., Sunderland, 2 edition.

Caswell, H., and A. Hastings, 1980 Fecundity, developmental time, and population growth rate : an analytical solution. Theoretical Population Biology, **17** : 71–79.

Charlesworth, B., 1994 *Evolution in age-structured populations*. Cambridge University Press, Cambridge, 2 edition.

Charlesworth, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genetical Research, **77** : 153–166.

Christiansen, F. B., 1988 The wahlund effect with overlapping generations. The American Naturalist, **131** : 149–156.

Crawley, M. J., and M. Rees, 1996 The balance of plant populations, pp. 191–212 in *Frontiers of Population Ecology*, edited by R. B. Floyd, A. W. Sheppard and P. J. De Barro. CSIRO Publishing, Melbourne.

De StasioB. T., Jr., 1989 The seed bank of a freshwater crustacean : copepodology for the plant ecologist. Ecology, **70** : 1377–1389.

Del Castillo, R. F., 1994 Factors influencing the genetic structure of *Phacelia dubia*, a species with a seed bank and large fluctuations in population size. Heredity, **72** : 446–458.

Ewens, W. J., 1982 On the concept of effective population size. Theoretical Population Biology, **21** : 373–378.

Felsenstein, J., 1971 Inbreeding and variance effective numbers in populations with overlapping generations. Genetics, **68** : 581–597.

Hairston, N. G., Jr., and B. T. De Statio, Jr., 1988 Rate of evolution slowed by a dormant propagule pool. Nature, **336** : 239–242.

Harper, J. L., 1977 *Population Biology of Plants*. Academic Press, London.

Hartl, D. L., and A. G. Clark, 1989 *Principles of population genetics*. Sinauer Associates, Inc., Sunderland, 2 edition.

Hedrick, P. W., 1995 Genetic polymorphism in a temporally varying environment : effects of delayed germination or diapause. Heredity, **75** : 164–170.

Hill, W. G., 1972 Effective size of populations with overlapping generations. Theoretical population biology, **3** : 278–289.

Hill, W. G., 1979 A note of ffective population size with overlapping generations. Genetics, **92** : 317–322.

Husband, B. C., and S. C. H. Barret, 1992 Effective population size and genetic drift in tristylous *Eichhornia paniculata* (pontederiaceae). Evolution, **46** : 1875–1890.

Husband, B. C., and S. C. H. Barret, 1995 Estimating effective population size : a reply to nunney. Evolution, **49** : 392–394.

Jarry, M., M. Khaladi, M. Hossaert-McKey and D. McKey, 1995 Modeling the population dynamics of annual plants with seed bank and density dependent effects. Acta Biotheoretica, **43** : 53–65.

Johnson, D. L., 1977 Inbreeding in populations with overlapping genrations. Genetics, **87** : 581–591.

Kaj, I., S. M. Krone and M. Lascoux, 2001 Coalescent theory for seed bank models. Journal of Applied Probability, **38** : 285–300.

Kalisz, S., 1991 Experimental determination of seed bank age structure in the winter annual *Collinsia verna*. Ecology, **72** : 575–585.

Kalisz, S., and M. A. McPeek, 1992 Demography of an age-structured annual : resampled projection matrices, elasticity analyses, and seed bank effects. Ecology, **73** : 1082–1093.

Leopold, A. C., 1996 Natural history of seed dormancy, pp. 3–16 in *Plant dormancy*, edited by G. A. Lang. CAB International, Oxon.

Levin, D. A., 1990 The seed bank as a source of genetic novelty in plants. The American Naturalist, **135** : 563–572.

Mahy, G., X. Vekemans and A.-L. Jacquemart, 1999 Patterns of allozymic variation within *Calluna vulgaris* populations at seed bank and adult stages. Heredity, **82** : 432–440.

McCue, K. A., and T. P. Holtsford, 1998 Seed bank influences on genetic diversity in the rare annual *Clarkia springvillensis* (onagraceae). American Journal of Botany, **85** : 30–36.

McDonald, N., and A. R. Watkinson, 1981 Models of an annual plant population with a seed bank. Journal of Theoretical Biology, **93** : 643–653.

McGraw, J. B., M. C. Vavrek and C. C. Bennington, 1991 Ecological genetic variation in seed banks. I. Establishment of a time transect. Journal of Ecology, **79** : 617–625.

Nagylaki, T., 1980 The strong-migration limit in geographically structured populations. Journal of Mathelatical Biology, **9** : 101–114.

Nagylaki, T., 1998 Fixation indices in subdivided populations. Genetics, **148** : 1325–1332.

NAGYLAKI, T., 2000 Geographic invariance and the strong-migration limit in subdivided populations. Journal of Mathelatical Biology, **41** : 123–142.

NORBORG, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of statistical genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. John Wiley & Sons, Ltd, Chichester.

NUNNEY, L., 1993 The influence of mating system and overlapping generations on effective population size. Evolution, **47** : 1329–1341.

NUNNEY, L., 1995 Measuring the ratio of effective population size to adult numbers using genetic and ecological data. Evolution, **47** : 389–392.

OLIVIERI, I., 2001 The evolution of seed heteromorphim in a metapopulation : interactions between dispersal and dormancy, pp. 245–268 in *Integrating Ecology and Evolution in a Spatial Context*, edited by J. SILVERTOWNand J. ANTONOVICS. Blackwell Science.

ORIVE, M. E., 1993 Effective population size in organisms with complex life-histories. Theoretical Population Biology, **44** : 316–40.

PACALA, S. W., 1986 Neighborhood models of plant population dynamics. 2. Multi-species models of annuals. Theoretical Population Biology, **29** : 262–292.

POLLAK, E., 2000 The effective size of some age-structured populations. Mathematical biosciences, **168** : 39–56.

REES, M, 1996 Evolutionary ecology of seed dormancy and seed size. Philosophical Transactions of the Royal Society of London Series B, **351** : 1299–1308.

REES, M., 1997 Seed dormancy, pp. 214–238 in *Plant Ecology*, edited by M. J. CRAWLEY. Blackwell Science, Oxford, 2nd edition.

ROUSSET, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics, **142** : 1357–1362.

ROUSSET, F., 1999 Genetic differentiation in populations with different class of individuals. Theoretical Population Biology, **55** : 297–308.

STÖCKLIN, J., and M. FISCHER, 1999 Plants with longer-lived seeds have lower local extinction rates in grasslands remnants 1950-1985. Oecologia, **120** : 539–543.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation : A test for population subdivision. Genetics, **117** : 149–153.

Templeton, A. R., and D. A. Levin, 1979 Evolutionary consequences of seed pools. The American Naturalist, **114** : 232–249.

Thrall, P. H., S. W. Pacala and Jr. Silander, J. A., 1989 Oscillatory dynamics in populations of an annual weed species *Abutilon theophrasti*. Journal of Ecology, **77** : 1135–1149.

Tonsor, S. J., S. Kalisz, J. Fisher and T. P. Holtsford, 1993 A life-history based study of population genetic structure - seed bank to adults in *Plantago lanceolata*. Evolution, **47** : 833–843.

Van der Valk, A. G., and C. B. Davis, 1978 The role of seed banks in the vegetation dynamics of prairie glacial marshes. Ecology, **59** : 322–335.

Vavrek, M. C., J. B. McGraw and C. C. Bennington, 1991 Ecological genetic variation in seed banks. III. Phenotypic and genetic differences between young and old seed populations of *Carex bigelowii*. Journal of Ecology, **79** : 645–662.

Venable, D. L., and J. S. Brown, 1988 The selective interactions of dispersal, dormancy, and seed size as adaptations for reducing risk in variable environments. The American Naturalist, **360-384** : 131.

Venable, D. L., and J. S. Brown, 1993 he population-dynamic functions of seed dispersal. Vegetatio, **108** : 31–55.

Venable, D. L., C. E. Pake and A. C. Caprio, 1993 Diversity and coexistence of sonoran desert winter annuals. Plant Species Biology, **8** : 207–216.

Wahlund, S., 1928 Zuzammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. Hereditas, **11** : 65–106.

Wakeley, J., 2001 The coalescent in an island model of population subdivision with variation among demes. Theoretical Population Biology, **59** : 133–144.

Watkinson, A. R., 1990 The population dynamics of *Vulpia fasciculata* : a nine-year study. Journal of Ecology, **78** : 196–209.

Weider, L., W. Lampert, M. Wessels, J. K. Colbourne and P. Limburg, 1997 Long-term genetic shifts in a microcrustacean egg bank associated with anthropogenic changes in the lake constance ecosystem. Proceedings of the Royal Society of London Series B, **264** : 1613–1618.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution, **38** : 1358–1370.

Whitlock, M. C., and N. H. Barton, 1997 The effective size of a subdivided population. Genetics, **146** : 427–441.

Wolfram Research, Inc., 1999 *Mathematica Edition : Version 4*. Wolfram Research, Inc., Champaign, Illinois.

Wright, S, 1939 Statistical genetics in relation to evolution, pp. 5–64In *Actualités scientifiques et industrielles*, Exposés de Biométrie et de la statistique biologique XIII. Hermann & Cie, Paris.

Wright, S., 1951 The genetic structure of populations. Annals of Eugenics, **15** : 323–354.

Wright, S., 1978 *Evolution and the Genetics of Populations,* Vol. IV. Variability *Within and Among Natural Populations*. University of Chicago Press, Chicago.

Tab. 1: Summary of main parameter notations.

N	Surface population size (carrying capacity)
f	Net fecundity
n	Number of populations
d	(Forward) dormancy rate
$g = (1 - d)$	(Forward) germination rate
z	(Forward) seed dispersal rate
k	(Backward) germination rate
s	Selfing rate
p_i	Probability that a seed germinating from the seed bank is i years old
m_s	Seed (backward) migration rate
m_p	Pollen (backward) migration rate
d	Proportion of seeds issued from the seed bank
τ	Mortality rate of seeds (or seedlings)
c	Cost to dispersal
Q_h^*	IBD probability of pairs of genes in adults after pollen dispersal
Q_h	IBD probability for pairs of genes in adults after seed dispersal
$R_h^{(i)*}$	IBD probability between a gene in an adult and a gene in a seed, after pollen dispersal
$R_h^{(i)}$	IBD probability between a gene in an adult and a gene in a seed, after seed dispersal
$S_h^{(i)}$	IBD probability for pairs of genes in seeds

Captions for figures

Fig. 1. Life cycle graph of the seed bank model. Each circle represents a given age-class, of size N_i. N is the maximum number of individuals among flowering plants (first class). The transitions from one class to another are also represented. See Table 1 for main parameter notations.

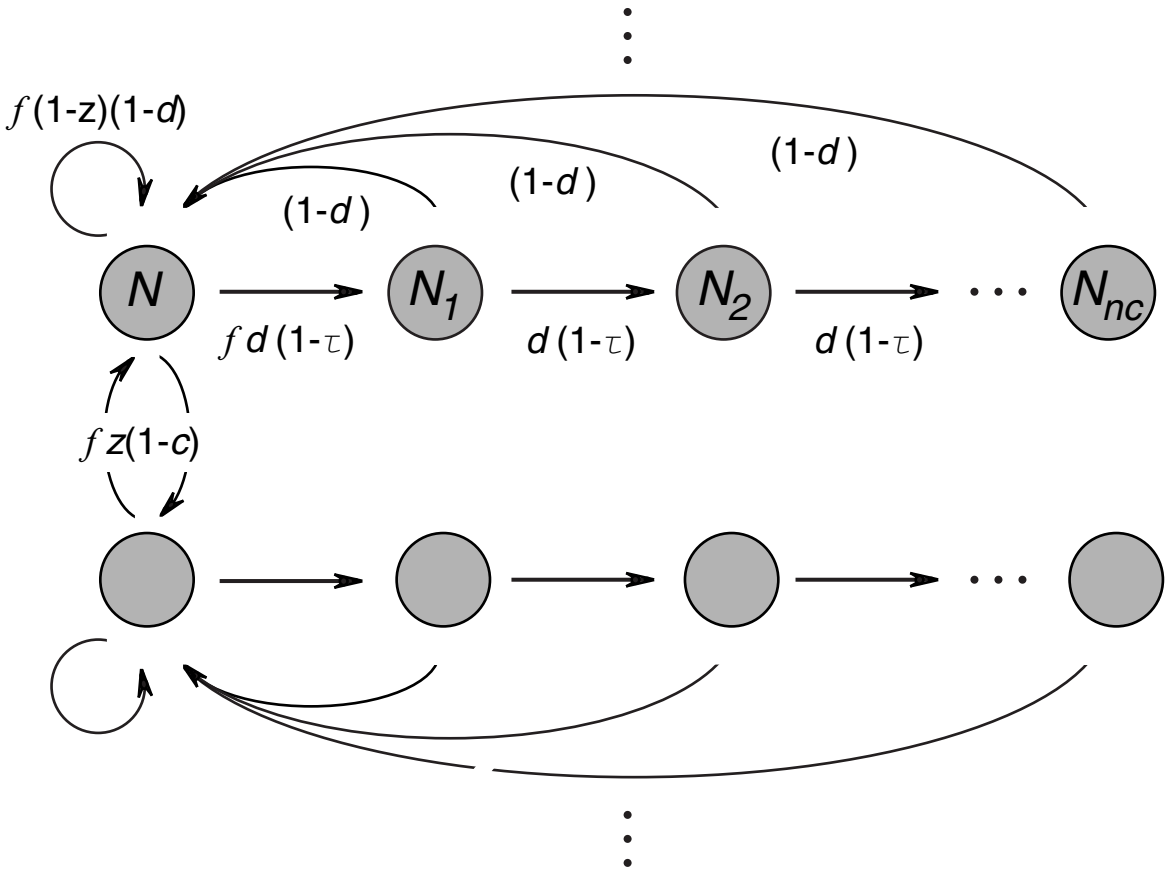
Fig. 2. Distribution of the probability that a germinating seed has been produced in the previous generation (top bar) or i generations before. The bar lengths sum to one. A. $d = 0.9$. B. $d = 0.7$. C. $d = 0.4$. D. $d = 0.1$.

Fig. 3. Effective population size of a population with a seed bank as a function of germination rate. A. A single population with $N = 50$, $\tau = 0.1$ and $N\ s = 1$ (random mating). B. A subdivided population, with $n = 20$, $N = 50$, $m_p = 0.02$, $z = 0.04$, $c = 0.5$, $\tau = 0.1$ and $N\ s = 1$.

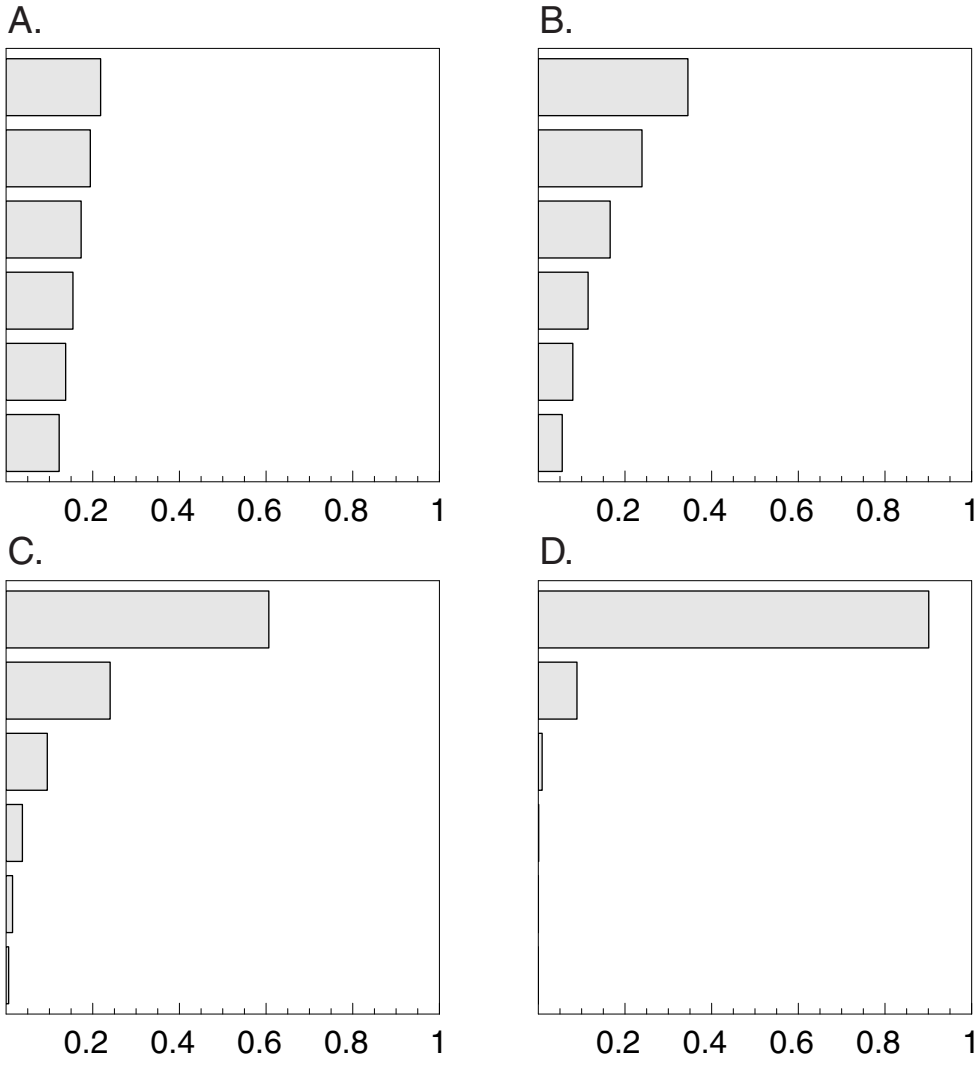
Fig. 4. Equilibrium value of F_{IS} among flowering plants, as a function of the surface population size N and the germination rate with $\tau = 0.1$ and $n_c = 10$ age classes in the seed bank. A. Random mating ($N\ s = 1$) B. Complete selfing ($s = 1$).

Fig. 5. Equilibrium value of F_{ST} among flowering plants, as a function of the germination (or flowering) rate with $\tau = 0.1$ and $N\ s = 1$ (random mating), $n = 20$, $m_p = 0.02$, $z = 0.04$, $c = 0.5$ and $\mu = 10^{-4}$. A. The case of annual plants with a seed bank ($N = 50$). B. The case of monocarpic perennials ($N_T = 50$).

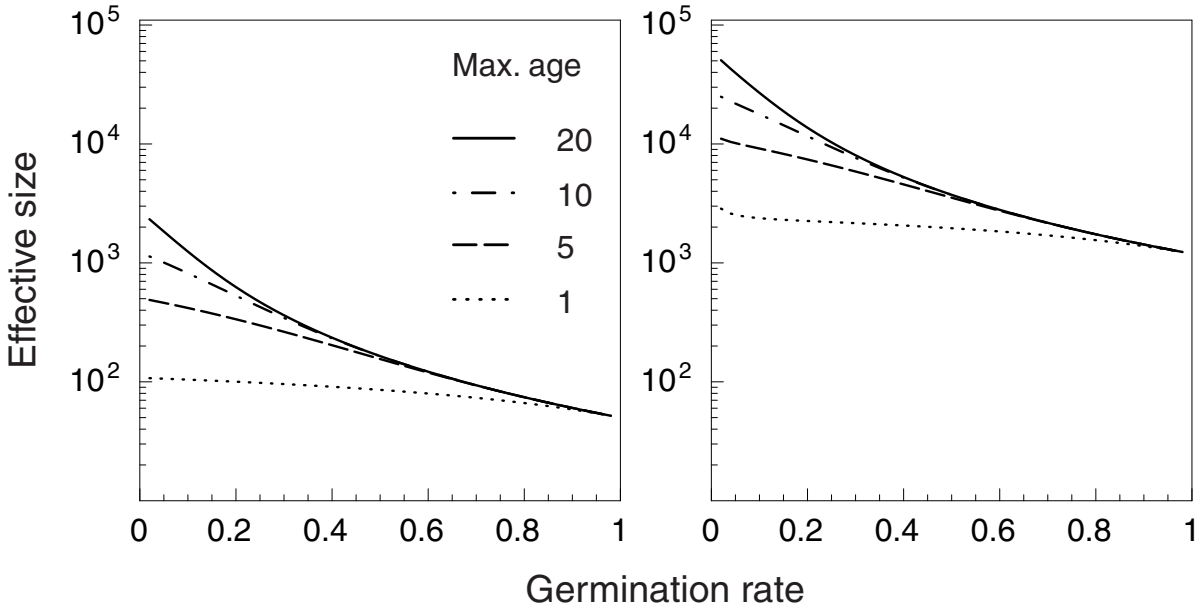
Fig. 6. Effective population size of a monocarpic perennial population as a function of flowering rate. A. A single population with $N = 50$, $\tau = 0.1$ and $N\ s = 1$ (random mating). B. A subdivided population, with $n = 20$, $N = 50$, $m_p = 0.02$, $z = 0.04$, $c = 0.5$, $\tau = 0.1$ and $N\ s = 1$.



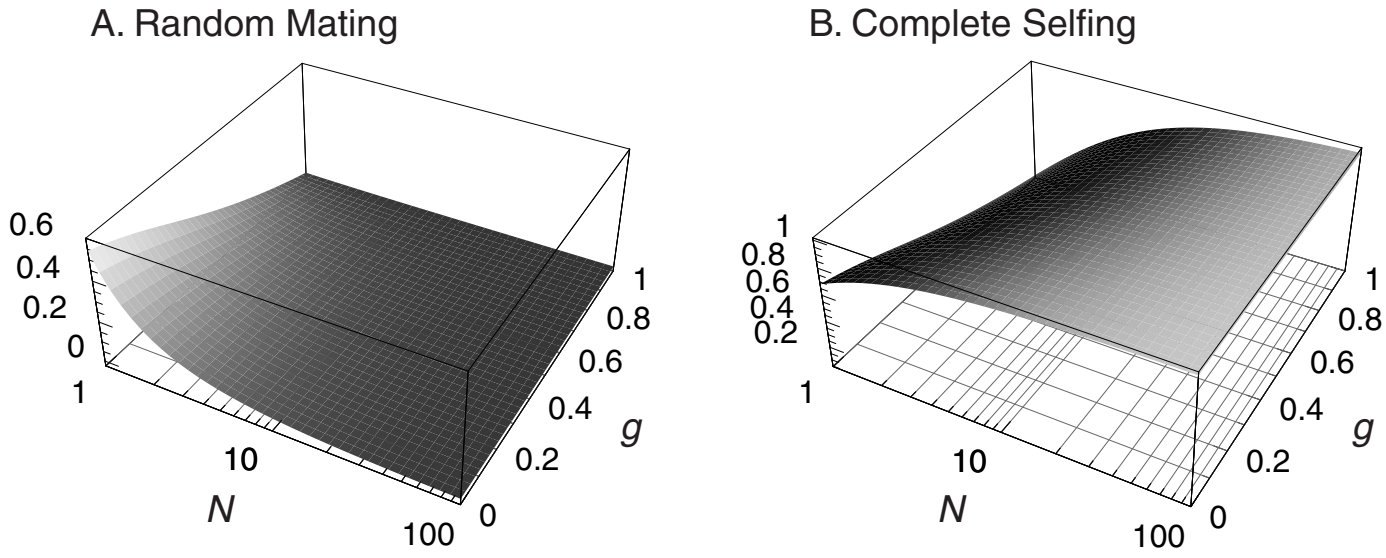
Vitalis *et al*. Figure 1



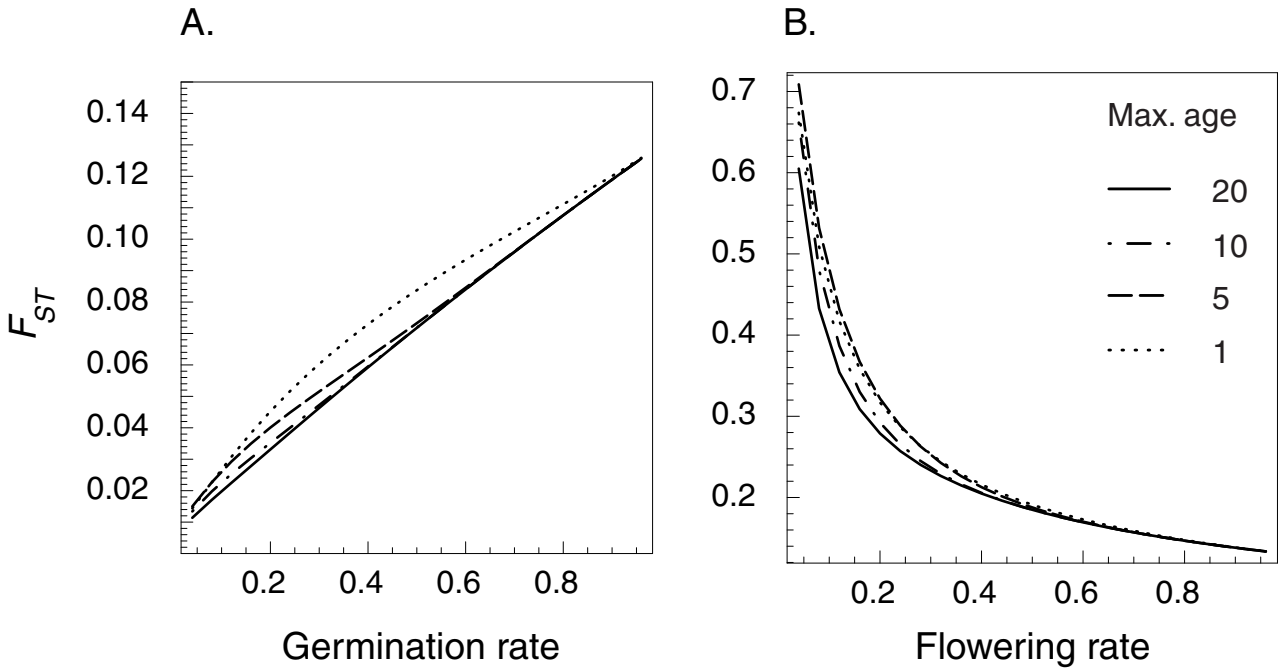
Vitalis *et al.* Figure 2



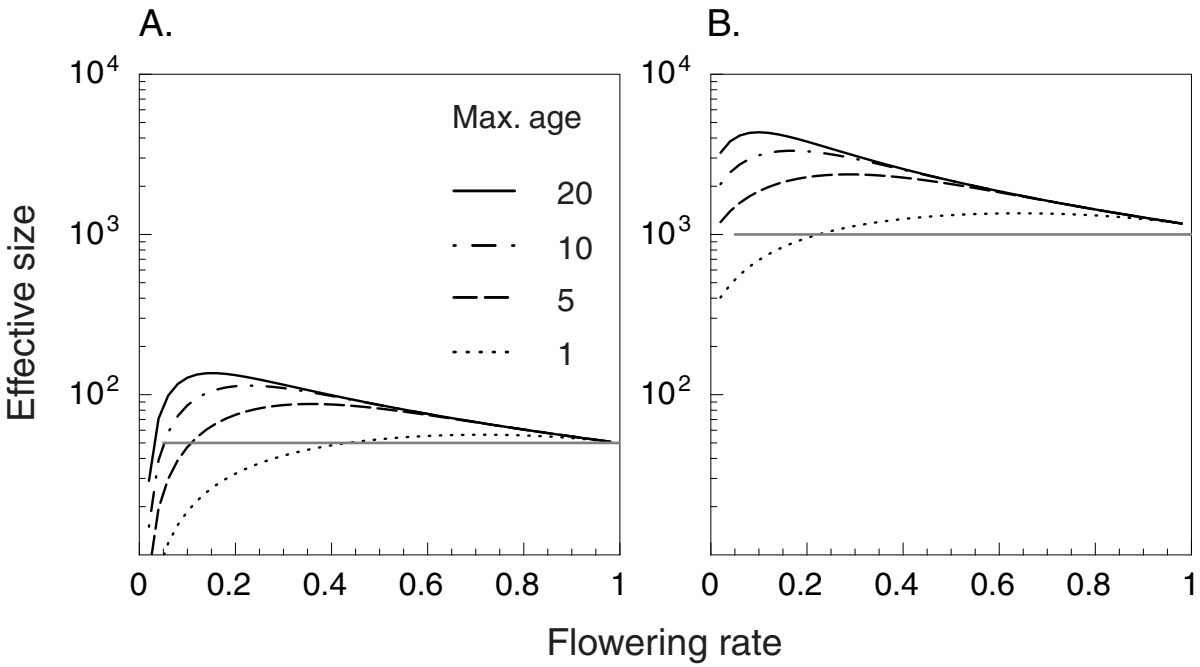
Vitalis *et al.* Figure 3



Vitalis *et al.* Figure 4



Vitalis *et al.* Figure 5



Vitalis *et al.* Figure 6

Summary

THIS thesis is concerned with the distribution of genetic variation within subdivided populations. Studying the rare aquatic fern *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta), I show how the concepts of *genetic identity* and *effective size* are crucial to understand which evolutionary forces shape the observed genetic variation. In particular, I developed theoretical models to show the influence of spatial and temporal structure on the distribution of genetic variation. From these models, three methods have been developed to analyse the genetic variation in natural populations. The first method enables to estimate the *effective size*, which determines the intensity of *genetic drift*. The second one concerns the estimation of sex-biased dispersal rates. Finally, a method has been developed to identify molecular markers that are potentially responding to selection.

Keywords : conservation biology, coalescence, population genetics, *Marsilea strigosa*, identity probabilities, effective size.
Research field : population biology and genetics.

Institut des Sciences de l'Évolution de Montpellier (UMR 5554)
Université Montpellier II F-34095 Montpellier Cédex 05, France.
and
Station Biologique de la Tour du Valat
Le Sambuc, F-13200 Arles, France.

Résumé

CETTE thèse a pour objet l'étude théorique de la distribution de la variation génétique dans les populations subdivisées. L'exemple de l'étude d'une fougère aquatique rare *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta) illustre l'importance des notions d'*identité génétique* et de *taille efficace* pour comprendre quelles sont les forces évolutives qui façonnent la variation génétique observée. Plus particulièrement, je me suis intéressé, au travers de modèles théoriques, à l'effet de la structure spatiale ou temporelle des populations sur la distribution de la variation génétique. A partir de ces modèles, trois méthodes d'analyse de la variation génétique dans les populations naturelles ont été développées. La première permet d'estimer la *taille efficace*, une quantité qui mesure l'intensité de la *dérive* génétique. La seconde méthode concerne l'estimation des taux de dispersion spécifiques aux individus de sexes différents. Enfin, une méthode de détection de marqueurs moléculaires potentiellement soumis à l'action de la sélection a été développée.

Mots-clés : biologie de la conservation, coalescence, génétique des populations, *Marsilea strigosa*, probabilités d'identité génétique, taille efficace.
Discipline : biologie et génétique des populations.

Institut des Sciences de l'Évolution de Montpellier (UMR 5554)
Université Montpellier II F-34095 Montpellier Cédex 05, France.
et
Station Biologique de la Tour du Valat
Le Sambuc, F-13200 Arles, France.