MINISTÈRE DE L'AGRICULTURE
ÉCOLE NATIONALE SUPÉRIEURE AGRONOMIQUE DE MONTPELLIER

THÈSE

présentée à l'École Nationale Supérieure Agronomique de Montpellier
pour obtenir le diplôme de Doctorat

Spécialité : Biologie de l'Évolution
Formation Doctorale : Biologie de l'Évolution et Écologie
École Doctorale : Biologie Intégrative

Estimation de paramètres de dispersion en populations structurées à partir de données génétiques

par

Raphaël LEBLOIS

Thèse co-dirigée par François Rousset et Arnaud Estoup

Soutenue le 3 mai 2004 devant le jury composé de

Laurent EXCOFFIER	Professeur, Université de Berne	Examinateur
Nicolas GALTIER	Chargé de Recherche, CNRS Montpellier	Examinateur
Marie-Laure NAVAS	Professeur, Agro-Montpellier	Président
François ROUSSET	Directeur de Recherche, CNRS Montpellier	Directeur de Thèse
Montgomery SLATKIN	Professeur, Université de Californie, Berkeley	Rapporteur
Xavier VEKEMANS	Professeur, Université de Lille I	Rapporteur

Membre invité au jury :

Arnaud ESTOUP	Chargé de Recherche, INRA Montpellier	Co-directeur de Thèse

Sommaire

Avant-propos **1**

1 Introduction **3**

1.1 Génétique des populations : applications en biologie de la conservation et des invasions 4

1.2 Estimation de paramètres démographiques : Quels outils pour quelles questions ? 6

1.3 Théorie de la coalescence : un apport majeur pour la génétique des populations 8

1.4 Ce que comporte ma thèse 10

2 Modèles démographiques et analyse statistique de la structuration génétique **13**

2.1 Notions d'identité génétique 13

2.1.1 Marqueurs génétiques : de l'identité par descendance à l'identité par états 13

2.1.2 F-statistiques 16

2.1.3 Relation entre identité génétique, F-statistiques et temps de coalescence 17

2.2 Vers une dispersion réaliste : du modèle en îles à l'isolement par la distance 19

2.2.1 Modèle en îles 20

2.2.2 Dispersion : caractéristiques et modélisation 22
2.2.3 Modèle en réseau : de la sous-population à la population continue 24
2.2.4 Structuration génétique en isolement par la distance . . 25
2.3 Coalescence en populations subdivisées 30
2.3.1 n-coalescent 30
2.3.2 Coalescent structuré 34
2.3.3 Simulation d'arbres de coalescence 37
2.4 Conclusion 40

3 Estimation par F-statistiques : précision et robustesse en populations continues sous isolement par la distance 41

3.1 Bases méthodologiques de l'étude 43
3.1.1 Algorithme de simulation 43
3.1.2 Analyse de la qualité des estimations 48
3.1.3 Intervalles de confiance 50
3.2 Influence de l'échelle d'échantillonnage et de la taille de l'habitat 52
3.3 Influence des processus de mutation 54
3.3.1 Modèles mutationnels 54
3.3.2 Taux de mutation (ou diversité génétique) 56
3.3.3 Effet d'une statistique prenant en compte les tailles des allèles 58
3.4 Influence d'hétérogénéités spatiales et temporelles 59
3.4.1 Variation temporelle de la dispersion 59
3.4.2 Variation de la densité dans le temps : goulet d'étranglement et explosion démographique 61
3.4.3 Expansion spatiale de la population à densité constante 63

3.4.4 Hétérogénéité spatiale de la densité : échantillonnage sur une zone de forte densité 65

3.5 Discussion 67

3.5.1 Processus mutationnels 67

3.5.2 Échelles d'échantillonnage et intervalles de confiance 68

3.5.3 Hétérogénéité spatiales et temporelles des paramètres démographiques 69

3.5.4 Interprétation de la robustesse générale de la méthode à l'aide des temps de coalescence 72

3.5.5 Implications quant aux études expérimentales 73

4 Estimation par maximum de vraisemblance : Où en est on ? 75

4.1 Principe 75

4.2 Méthodes permettant le calcul de la probabilité d'un échantillon de gènes 77

4.2.1 Approche de Felsenstein et collaborateurs 78

4.2.2 L'approche de Griffiths et collaborateurs 84

4.2.3 Vers des distributions d'échantillonnage pondéré plus efficaces pour les méthodes de Griffiths et collaborateurs 89

5 Précision et robustesse des estimations par maximum de vraisemblance 99

5.1 Approche de Felsenstein et collaborateurs : test du logiciel MIGRATE 99

5.1.1 Test sur jeu de données réel 100

5.1.2 Test sur jeux de données simulés 101

5.2 Précision des algorithmes de Griffiths et collaborateurs 105

5.2.1 Algorithme de Nath et Griffiths 1996 105

5.2.2 Algorithme de De Iorio *et al.* (2004) pour les modèles de mutation par pas 108

5.3 Conclusions 111

6 Implications en biologie de la conservation : contraction spatiale d'habitat en population continue sous isolement par la distance 117

6.1 Réduction d'habitat en isolement par la distance 119

6.1.1 Caractéristiques génétiques d'une population continue à l'équilibre 119

6.1.2 Influence d'une réduction de surface d'habitat 123

6.2 Application au criquet de la Crau 128

6.2.1 Présentation du modèle biologique et de son milieu . . 128

6.2.2 Caractéristiques de la population échantillonnée 129

6.3 Conclusions 132

7 Conclusions générales et perspectives 135

7.1 Validation du modèle d'isolement par la distance et de la méthode d'estimation de $D\sigma^2$ 136

7.2 Estimation par F-statistiques versus maximum de vraisemblance 137

7.3 Estimation séparée des tailles de population et des caractéristiques de dispersion 140

7.4 Vers des modèles plus réalistes et complexes 143

7.4.1 Limites de l'information génétique 143

7.4.2 Une alternative possible au maximum de vraisemblance 144

Bibliographie 147

Annexe A : annexes "mathématiques et algorithmiques" 161

A.1 Calcul des probabilités d'identité par état à partir de l'identité par descendance sous différents modèles mutationnels (F. Rousset) 163

A.2 Développements de l'algorithme d'échantillonnage pondéré de de Iorio *et al.* (2004) pour deux populations et une modèle mutationnel SMM 167

Annexe B : Publications **175**

B-1

Leblois R., Rousset F., Tikel D., Moritz C., Estoup A. 2000 **Absence of evidence for isolation by distance in an expanding cane toad (Bufo marinus) population : an individual-based analysis of microsatellite genotypes**. *Molecular Ecology.* 9 : 1905-1909. 177

B-2

Leblois R., Estoup A., Rousset F. 2003. **Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance.** *Molecular Biology and Evolution.* 20 :491-502. 185

B-3

Leblois R., Rousset F., Estoup A. 2004. **Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population from microsatellite data.** *Genetics* 166 : 1081-1092. 199

B-4

De Iorio M., Griffiths R., **Leblois R.**, Rousset F. 2004. **Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models** Soumis à *Theoretical Population biology* 213

Avant-propos

Cette thèse résume l'essentiel des travaux de recherche que j'ai réalisé depuis mon DEA (Diplôme d'études Approfondies en Écologie et Évolution) sous la direction d'Arnaud Estoup et François Rousset. Ces travaux ont été initialisés par des longues discussions avec Arnaud Estoup lors de mon stage de maîtrise réalisé à l'Université du Queensland en Australie dans le laboratoire de Zoologie dirigé par Craig Moritz.

Du point de vue de sa forme, cette thèse comprend trois parties. La première partie est un document de synthèse en Français, qui reprend les principaux arguments théoriques et expérimentaux qui ont servi de base à mes travaux. La seconde partie comprend des développement mathématiques que j'ai considérés trop complexes et non nécessaires à la première partie ; ceci dans un souci de concision et de clarté pour la première partie. La troisième partie est un recueil d'articles publiés, sous presse ou en préparation. Les parties 2 et 3 ne sont pas indispensables à la compréhension du document de synthèse qui peut être lu indépendamment des deux autres parties.

Précisons dès à présent quelques termes et notations utilisées dans ce documents :

On appellera *gène* la copie d'une information génétique. Cette définition a un sens immédiat lorsque l'on s'intéresse à la 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 (ou gènes) ne sont pas nécessairement identiques. Pour certains marqueurs génétiques dits évolutivement *neutres* (qui, par définition, 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 des 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*". Enfin, un *allèle* (ou *état allélique*) représentera une classe de gènes tous équivalents. Selon ces définitions, deux gènes sont donc dans le même état allélique, si l'information qu'ils portent est codée par la même séquence d'ADN, ou s'ils sont la copie exacte d'un ancêtre commun.

Dans ce qui suit, nous considérerons un *échantillon* de gènes constitué de plusieurs *sous-échantillons*. L'échantillon est pris dans ce qu'on appellera une *population*, potentiellement structurée en *sous-populations*, ou *dèmes*,

dans lesquelles sont pris les sous-échantillons. Cette structuration peut être la conséquence d'une dispersion localisée ou de barrières aux flux de gènes. On considérera que la reproduction est *panmictique* (i.e. union aléatoire des gamètes de la population) au sein de chaque sous-population et non dans la population entière.

Un modèle est défini par des *paramètres* tels que les tailles de populations, des taux de mutations et des taux des migrations. Nous considérerons que toutes quantités qui sont uniquement fonction des paramètres qui définissent le modèle (e.g. les probabilités d'identité ou les F-statistiques), sont aussi des paramètres. Les valeurs prises par ces paramètres dans une population "vraie" seront des *statistiques* ou des *estimations*, que l'on pourra mesurer dans cette population à l'aide d'*estimateurs*. L'estimateur et les statistiques sont alors traités comme des *variables aléatoires*. Les *espérances* de ces variables aléatoires sont fonctions des valeurs des paramètres du modèle. Si un estimateur est sans biais, son espérance correspond alors exactement aux valeurs des paramètres du modèle.

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^{ème} siècle, notamment Ronald A. Fisher, Sewall Wright et John B. S. Haldane, qui en sont les fondateurs. Ces derniers ont, chacun à leur manière, posé les bases conceptuelles et une formalisation mathématique de l'évolution de la variation génétique dans les populations. Bien que la description et la compréhension des mécanismes qui maintiennent la variabilité génétique au sein et entre les populations soient d'un grand intérêt en évolution, ce domaine est longtemps resté la préoccupation d'un petit nombre de biologistes. A l'heure actuelle, les outils et les concepts de cette discipline diffusent principalement dans la communauté scientifique et dans le grand public à travers trois champs distincts : la biologie de la conservation, la biologie des invasions et plus récemment, les phénomènes de résistance en agronomie et en médecine (aux insecticides et aux antibiotiques principalement). En effet, on observe une prise de conscience accrue de la nécessité de conserver les espèces en général et les ressources génétiques d'espèces d'intérêt agronomique en particulier. De ce fait, les notions de diversité génétique, de transfert de gènes, de barrières entre espèces, entre autres, 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 le brevetage du vivant ainsi que plus globalement sur la nécessité de préserver la faune et la flore et de minimiser les phénomènes de "bioinvasions". Ces exemples illustrent la place cruciale que tient aujourd'hui la génétique des populations dans l'évaluation, la compréhension et la résolution de ces questions. Dans ce contexte il est essentiel de développer des modèles de structuration des populations et de les tester avec des techniques appropriées, afin d'appliquer des outils

d'analyse performants permettant l'étude de modèles biologiques concrets.

1.1 Génétique des populations : applications en biologie de la conservation et des invasions

Nous assistons à une crise majeure dont les sciences de l'évolution sont les témoins depuis quelques dizaines d'années. Il s'agit de l'augmentation du nombre d'espèces qui disparaissent quotidiennement. En effet, même si les extinctions en masse ont ponctué l'histoire de la vie sur Terre, il n'en demeure pas moins que la phase actuelle d'extinction est caractérisée par une extrême rapidité, sans précèdent à l'échelle des temps géologiques (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 de nombreuses espèces et populations (Channell & Lomolino, 2000; Flannery, 1999). Dans une étude réalisée sur deux mille espèces en Amérique du nord, Wilcove *et al.* (1998) ont montré que les trois principaux facteurs exogènes menaçant les communautés naturelles sont la destruction et la dégradation de l'habitat, l'introduction d'espèces exotiques et la pollution. Bien sûr, il existe plusieurs niveaux de destruction d'habitat : (i) la réduction de la superficie totale occupée par les populations d'une espèces (qui suit la destruction de larges fragments d'habitat) ; (ii) la fragmentation de son aire de répartition et/ou de ses populations (qui accompagne, par exemple, la construction des routes) : ou encore (iii) la détérioration d'espaces situés entre les habitats favorables empêchant la dispersion des individus entre différents sites (par exemple la déforestation). L'ensemble des ces actions anthropiques entraîne une modification des connections entre les populations ainsi qu'une réduction de la taille des habitats favorables et donc une diminution des effectifs des populations naturelles.

Cette diminution des effectifs et l'isolement de petites populations constituent de réelles menaces pour la pérennité des espèces. En effet, les populations de petite taille subissent des effets d'échantillonnage à plusieurs niveaux. Trois au moins de ces niveaux sont facilement perceptibles. Il s'agit tout d'abord de l'évolution de la composition génétique des populations, caractérisée par le terme de *dérive*, qui décrit le phénomène de stochasticité génétique. La dérive définit le fait que les gènes d'une population ne se transmettent pas obligatoirement d'une génération à l'autre mais que cette

transmission se fait de façon aléatoire. Ainsi 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 probabilité d'espérance égale à un (i.e. de moyenne égale à un). La probabilité que certaines lignées s'éteignent étant non nulle, la dérive entraîne donc une diminution du niveau de variabilité génétique, simplement due au hasard. On peut aisément comprendre qu'elle est d'autant plus forte que la population est petite. Enfin, si ce phénomène touche des gènes dont les effets agissent sur la valeur sélective des individus (survie, fécondité), alors, des allèles délétères peuvent se fixer, par dérive, dans la population et diminuer ainsi la valeur sélective moyenne de la population. La dispersion des gènes entre populations peut limiter cet effet en apportant par migration de nouveaux allèles favorables. La forte dérive dans les petites populations, ainsi que l'isolement de ces mêmes petites populations, sont donc susceptibles d'augmenter le risque d'extinction. Le deuxième niveau concerne la démographie, c'est à dire la composition/structuration des populations en classes d'individus et leurs effectifs. Cette stochasticité démographique décrit un phénomène analogue au précédent mais à l'échelle des individus. A l'échelle de la population, il se peut qu'une année aucun individu ne se reproduise car si la probabilité que chaque individu ne donne aucun descendant est faible, elle n'est pas nulle. Cette stochasticité démographique est d'autant plus forte que les effectifs populationnels sont faibles. Un troisième niveau est lié à la stochasticité environnementale. Ce processus est souvent associé aux variations climatiques (sécheresses, intempéries) et aux catastrophes naturelles (incendies, cyclones, etc.). Les capacités de dispersion des individus leur permettent ou non de re-coloniser des sites vides. On comprend donc facilement que cette forme de stochasticité a d'autant plus d'impact sur les populations qu'elles sont de petite taille et isolées. Par conséquent, les populations de petite taille, et qui plus est isolées, sont plus à même de souffrir de problèmes démographiques et génétiques, et sont fortement sujettes aux changements plus ou moins brutaux de l'environnement. L'importance relative de ces trois niveaux de stochasticité dans les mécanismes d'extinction est mal connue et encore sujette à controverses (Lande, 1988).

Un autre phénomène majeur, auquel nous assistons depuis quelques décennies, est la forte augmentation de la fréquence des invasions biologiques. On considère aujourd'hui que la seconde cause d'extinction des espèces, après la destruction de l'habitat, est l'invasion de nouveaux habitats par des espèces exotiques. Une invasion biologique (ou *bioinvasion*) peut être définie comme l'apparition d'une espèce dans un habitat situé hors de sa zone naturelle de dispersion par une médiation d'origine humaine. L'invasion est distinguée de

la colonisation, souvent perçue comme une expansion naturelle de l'aire de répartition d'origine. Ces invasions biologiques posent des problèmes extrêmement variés comme l'extinction d'espèces endémiques, un impact négatif sur les agrosystèmes ou encore des problèmes de santé publique. On sait par exemple que les coûts associés à ces bioinvasions sont de plus de 100 milliard de dollars par an aux États Unis (Pimentel *et al.*, 2000). De plus, puisque le commerce international ne cesse de s'intensifier, il est prévisible que le nombre d'introductions, accidentelles ou intentionnelles, augmente encore dans les années à venir. Les bioinvasions ont donc reçu une attention toute particulière ces dernières années de la part de la communauté scientifique internationale et ont fait l'objet de nombreuses recherches en écologie (revue dans Keane & Crawley, 2002; Shea & Chesson, 2002). Mais, en dépit du fait que de nombreux paramètres démographiques et évolutifs paraissent cruciaux dans les dynamiques d'expansion de ces populations, ainsi que dans leur adaptation au nouvel environnement, les aspects évolutifs et de génétique des populations des bioinvasions ont été assez peu étudiés à ce jour (revue dans Lee, 2002).

On peut donc voir plusieurs intérêts à mesurer des tailles de population et des flux de gènes en populations naturelles : (i) évaluer les risques d'extinction associés à la petite taille des populations ; (ii) caractériser l'intensité de réductions de taille de population (*goulets d'étranglement*, de l'anglais *bottlenecks*) ou l'isolement de sous populations ; (iii) évaluer les aptitudes colonisatrices d'espèces invasives ; et plus généralement (iv) mieux appréhender les éléments favorisant ou non l'adaptation locale des populations à leur environnement (Gandon *et al.*, 1996; Ronce & Kirkpatrick, 2001; Lenormand, 2002).

1.2 Estimation de paramètres démographiques : Quels outils pour quelles questions ?

Comme nous venons de le voir, l'effectif des populations est un paramètre crucial en évolution puisqu'il détermine l'intensité des forces stochastiques en jeu (démographiques, génétiques et environnementales). De même, les paramètres de dispersion sont tout aussi importants, du point de vue de la conservation aussi bien que de la lutte contre les espèces invasives, puisqu'ils déterminent également la dynamique des individus et des gènes à l'échelle de la *métapopulation* (population de populations, Levins, 1968) et peuvent contrecarrer les effets des autres forces stochastiques. Ainsi, la connaissance

des effectifs des populations et des caractéristiques selon lesquelles les individus se déplacent ou dispersent leur progéniture sont autant de pré-requis nécessaires à une bonne compréhension des systèmes biologiques.

On distingue deux types d'approche pour aborder ces questions : les approches *directes* et *indirectes*. Les approches directes reposent sur le suivi démographique en populations naturelles, sur l'estimation de paramètres liés au cycle de vie et sur des méthodes matricielles de projection des effectifs sous différents modèles stochastiques. Lorsque les suivis individuels sont possibles (c'est à dire lorsque l'on peut marquer les individus ou bien les équiper de radio-émetteurs), les approches directes permettent d'estimer des paramètres d'échanges d'individus entre sites (voir, par exemple, Grosbois, 2001). 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 les chapitres suivants, il est possible de définir et d'estimer les paramètres qui décrivent la dynamique de l'évolution du polymorphisme génétique dans les populations. Ces approches indirectes permettent notamment, à travers l'étude de la distribution spatiale du polymorphisme, d'inférer (i.e. estimer) les mouvements des gènes (*flux de gènes*) (voir par exemple, Beerli & Felsenstein, 2001; Rousset, 2001a; Slatkin, 1987).

Toutefois, les paramètres estimés par des approches indirectes ne sont pas nécessairement identiques ou équivalents à ceux estimés par les approches directes. D'une part, les flux de gènes correspondent aux mouvements d'individus qui se sont reproduits avec succès hors de leur lieu de naissance, et non pas aux mouvements nets d'individus entre les différents sites d'études. D'autre part, l'échelle spatiale à laquelle l'estimation indirecte de la dispersion est pertinente dépend des facteurs démographiques (dispersion localisée ou non) des populations étudiées et des processus mutationnels des marqueurs génétiques utilisés (Rousset, 2001a,b, voir aussi annexes B-3). Enfin, les marqueurs moléculaires à partir desquels est mesurée la distribution du polymorphisme peuvent être soumis à des effets sélectifs. Ce dernier point ne sera pas abordé dans ma thèse et je considérerai tout au long de ce document que les marqueurs moléculaires utilisés sont évolutivement neutres. Notons à ce propos l'existence de méthodes pour détecter des locus sous sélection afin de les exclure des analyses fondées sur le polymorphisme neutre (voir, par exemple, Beaumont & Nichols, 1996; Galtier *et al.*, 2000 ; revue de Luikart *et al.*, 2003). Il est souvent considéré que les approches directes 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). Nous verrons dans le chapitre 3 qu'il n'est pas

certain, au moins sous certains modèles démographiques, que les deux types d'approche soient très différentes.

1.3 Théorie de la coalescence : un apport majeur pour la génétique des populations

Une caractéristique majeure de l'analyse du polymorphisme en populations naturelles, et plus généralement en génétique des populations et en évolution, est que l'on travaille sur des données "expérimentales" sans réplicats et pour lesquelles les conditions initiales de l'"expérience" ne sont pas connues. Ceci a des implications extrêmement importantes sur l'analyse des données. Ainsi, lorsqu'on étudie un échantillon d'individus d'une population génotypés à plusieurs locus, les états alléliques des différents locus peuvent être statistiquement dépendants si les locus sont proches sur le même chromosome (liaison génétique). De plus, pour chaque locus, les états alléliques des différents individus sont statistiquement dépendants du fait de l'histoire généalogique qu'ils partagent. Ces dépendances statistiques sont le résultat de l'histoire commune des événements de mutation, de recombinaison et de *coalescence* [1] des lignées. Ces facteurs doivent être intégrés dans l'analyse statistique des données. Une solution consiste à modéliser le passé à l'aide d'un modèle stochastique approprié, dont un exemple est *le coalescent*.

La théorie de la coalescence est une extension naturelle des modèles classiques de génétique des populations, découverte de façon indépendante, à la fin des années 80, par Kingman (1982a,b) et Tajima (1983). Cette théorie a reçu une attention accrue de la part des généticiens des populations à partir des années 90. Elle repose sur un principe simple, la simulation des généalogies (ou arbres de coalescences) possibles d'un échantillon de gènes d'une même espèce en remontant dans le passé jusqu'à l'ancêtre commun le plus récent des gènes de l'échantillon (MRCA, de l'anglais "most recent common ancestor"). Un intérêt majeur de la coalescence est qu'une telle simulation peut s'effectuer sans prendre en considération les autres gènes de la population (Fig.1.1). On applique ensuite sur ces arbres les effets des événements de mutation, ce qui détermine les états alléliques des gènes considérés. On peut ainsi dissocier, sous l'hypothèse de neutralité des marqueurs génétiques utilisés, le processus de mutation du processus généalogique. Les intérêts de

[1] la coalescence de deux ou plusieurs lignées (par extension on parle aussi de coalescence de gènes) correspond à la réunion de ces lignées en remontant dans le passé

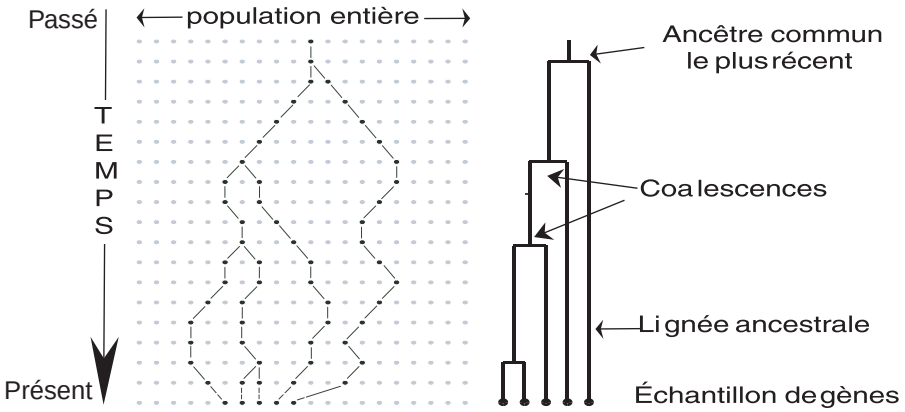

FIG. 1.1. Représentation du principe de la théorie de la coalescence. A gauche est représenté le "trajet" des lignée ancestrales de l'échantillon de gènes considéré au sein de la population totale. A droite est représenté l'arbre de coalescence de ce même échantillon.

ce modèle sont multiples : (i) la structure des données génétiques reflète, en grande partie, la généalogie sous-jacente aux données. De ce fait, l'étude de la généalogie permet une compréhension qualitative des patrons de variation des données génétiques (voir Nordborg & Tavare, 2002) ; (ii) les analyses quantitatives sont généralement plus faciles avec des méthodes généalogiques qu'avec les approches traditionnelles qui retracent la composition de la population entière en avançant dans le temps, au moins pour l'analyse du polymorphisme neutre ; (iii) l'utilisation de la théorie de la coalescence donne des méthodes de simulation extrêmement efficaces ; et (iv) la coalescence permet parfois l'utilisation complète de l'information des données génétiques (sous réserve d'utiliser des méthodes statistiques de maximum de vraisemblance et des algorithmes spécifiques).

L'arbre généalogique d'un échantillon de n gènes pris dans une population panmictique de taille constante N (N peut être exprimé en nombre de gènes ou en nombre d'individus) au cours du temps est modélisé par un processus stochastique connu sous le nom de *n-coalescent* (n pour souligner la dépendance vis à vis de la taille de l'échantillon). Ce processus a été introduit par Kingman (1982a) comme une approximation de la généalogie de gènes évoluant suivant le modèle neutre dit de "Wright-Fisher". Le modèle n-coalescent a une formalisation mathématique relativement simple que nous développerons dans la section 2.3.1. De nombreux modèles ont ensuite été développés pour s'ajuster à des situations démographiques plus complexes considérant de la recombinaison, de l'autofécondation ainsi que des variations de tailles

de populations au cours du temps. Certains auteurs ont aussi développés des modèles non neutres prenant en compte la sélection au locus considéré. Nous développerons quelques aspect du coalescent en population subdivisées dans la section 2.3.2. Pour d'autres développements, le lecteur pourra se référer à Hudson (1990), Neuhauser & Krone (1997) pour la sélection, et la revue de Nordborg (2001).

1.4 Ce que comporte ma thèse

Le propos principal de ma thèse est de montrer comment les modèles de populations subdivisées peuvent permettre non seulement l'estimation, à partir de données moléculaires, de paramètres d'intérêt pour la compréhension des systèmes biologiques, mais également que ces modèles sont indispensables à la compréhension et surtout à l'interprétation de la distribution du polymorphisme au sein des populations naturelles. Une des idées directrices de mon travail est que, bien souvent, lorsque l'on estime des paramètres démographiques en population naturelle, l'adéquation entre modèles et réalité est souvent incomplète et surtout incertaine. Ainsi, les multiples contradictions observées entre les estimations démographiques directes et les estimations indirectes à partir de données génétiques ont souvent été attribuées : (i) à une mauvaise description de la dispersion (par exemple, le modèle en îles) ; (ii) aux hypothèses de stabilité démographique dans le temps et dans l'espace ; (iii) aux hypothèses associées aux taux de mutation et aux processus mutationnels des marqueurs utilisés ; et (iv) à l'hypothèse de neutralité évolutive des marqueurs génétiques (voir Hastings & Harrison, 1994; Koenig *et al.*, 1996; Slatkin, 1994). Nous avons donc voulu évaluer dans quelle mesure ces facteurs, à l'exception du dernier, pouvaient influencer les résultats d'études du polymorphisme en populations naturelles. Pour cela, je me suis plus particulièrement intéressé à deux approches différentes : l'estimation par F-statistique et l'estimation par maximum de vraisemblance fondée sur la coalescence.

La suite de ce document de synthèse est structurée comme suit. Dans le second chapitre, je présenterai les différents modèles démographiques couramment utilisés en génétique des populations, et notamment ceux dits en populations structurées. Je discuterai de la pertinence de ces modèles et des différentes analyses statistiques qui peuvent en découler. Les différentes approches de l'analyse des modèles que j'ai choisi de présenter dans ce document proviennent en grande partie des travaux de Rousset (1997, 2000, 2004). Je

n'en présenterai que les grandes lignes et le lecteur pourra en trouver les détails dans Rousset (2004). Dans le troisième chapitre, je présenterai un ensemble de résultats portant sur des tests de précision et de robustesse d'une méthode d'estimation par F-statistiques de paramètres démographiques à partir de données génétiques sous isolement par la distance en population continue. Dans le quatrième chapitre, je discuterai de l'estimation de paramètres démographiques par des méthodes de maximum de vraisemblance fondées sur la théorie de la coalescence. J'en présenterai le principe ainsi que les principaux développements utilisés à ce jour en mettant l'accent sur les développements récents auxquels j'ai participé pendant ma thèse. Je présenterai ensuite, dans un cinquième chapitre, quelques tests préliminaires de précision et de robustesse de deux de ces méthodes d'estimation par maximum de vraisemblance. Dans le sixième chapitre, je présenterai les résultats d'une étude par simulation de l'implication, dans un contexte de biologie de la conservation, de la structuration en isolement par la distance sur l'analyse de la réduction d'habitat en population continue. Enfin, je présenterai les conclusions et perspectives générales de ce travail dans le septième chapitre.

Chapitre 2

Modèles démographiques et analyse statistique de la structuration génétique

2.1 Notions d'identité génétique

Dans cette partie, je m'attacherai à définir et discuter des concepts d'identité des gènes et des F-statistiques qui sont à la source de plusieurs modèles utilisés au cours de ma thèse. Dans ce document, nous n'envisagerons que l'identité génétique à un locus, pour une extension à plusieurs locus le lecteur peut lire par exemple Vitalis & Couvet (2001a). Décrire les différents états alléliques et déterminer l'identité exacte, au sens généalogique, des gènes d'un échantillon n'est pas tout à fait équivalent. En effet, seule la description des états alléliques est envisageable en pratique, tandis que la détermination de la vrai généalogie des gènes parait difficilement accessible. Il s'agit là de la différence entre l'identité par descendance (de l'anglais identity by descent (IBD)) et l'identité par état (de l'anglais identity in state (IIS)).

2.1.1 Marqueurs génétiques : de l'identité par descendance à l'identité par états

On dira que deux gènes sont *identiques par descendance* s'ils correspondent à deux copies exactes de leur ancêtre commun le plus récent (MRCA). 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. Considérons par exemple une population *panmictique* (i.e. dans laquelle tous les individus s'apparient au hasard) constituée de N individus diploïdes qui produisent chacun un nombre infini de gamètes subissant l'effet d'une mutation au locus considéré avec la probabilité μ. Les descendants de ces individus sont issus du croisement au hasard des gamètes dans la population. Chaque descendant a donc la même probabilité $\frac{1}{N}$ de descendre de n'importe lequel des N parents. La probabilité que deux individus partagent un même parent est donc $\frac{1}{N}$, et la probabilité qu'ils aient reçu une copie du même gène parental est alors $\frac{1}{2N}$ puisqu'ils sont diploïdes. Par conséquent, en considérant des générations non chevauchantes, on peut écrire la probabilité d'identité par descendance Q au temps $t+1$ (chez les descendants) de deux gènes pris au hasard dans la population, en fonction de cette même probabilité au temps t (chez les parents), comme

$$Q(t+1) = (1-\mu)^2\Big[\frac{1}{2N} + \Big(1 - \frac{1}{2N}\Big)\, Q(t)\Big]. \tag{2.1}$$

En notant $\gamma \equiv (1-\mu)^2$ la probabilité qu'aucun des deux gènes n'ait muté entre t et $t+1$, on obtient à l'équilibre

$$Q = \frac{\gamma}{2N(1-\gamma)+\gamma} \approx^{\mu\to 0} \frac{1}{1+4N\mu}. \tag{2.2}$$

On dira que deux gènes sont *identiques par état* s'ils appartiennent à la même classe allélique. Bien sûr, cette appartenance à une classe allélique donnée dépendra de notre perception de l'état des objets moléculaires. Prenons par exemple la cas d'un locus microsatellite, on dira que deux gènes sont identiques si les fragments d'ADN amplifiés qui les constituent migrent à la même distance sur un gel d'électrophorèse. Ceci traduit le fait que les fragments d'ADN amplifiés possèdent le même nombre de paires de bases. En revanche, les séquences de ces deux gènes peuvent être différentes. Cet exemple illustre le concept d'*homoplasie*, qui peut plus généralement se définir comme le fait que deux gènes sont identiques par état mais pas par descendance. Deux copies de gènes rigoureusement identiques sur le plan de leur séquences en nucléotides peuvent également ne pas être identiques par descendance si une mutation qui survient après une autre rétablit l'état initial (mutation reverse). L'homoplasie provient donc en partie de notre perception de l'état des objets que l'on manipule (isoformes enzymatiques, fragments amplifiés d'ADN, séquences d'ADN, etc.) mais aussi de la nature des mutations. Si l'on associe à un marqueur un mécanisme moléculaire tels que les mutations

ne créent pas nécessairement de nouveaux allèles dans la population mais peuvent générer des allèles déjà existants, alors seule l'identité par état est mesurable en pratique.

Pendant cette thèse, je me suis particulièrement intéressé aux marqueurs de type microsatellite. Ce sont des portions de chromosomes (que l'on peut définir comme locus) dont la séquence est constitué d'un certain nombre de répétitions à l'identique d'un motif nucléotidique (par exemple GTGTGTGT). Ce motif peut comporter deux, trois ou quatre nucléotides (deux, GT, dans notre exemple). Depuis leur découverte dans les années 80, ces marqueurs sont de plus en plus utilisés pour plusieurs raisons. D'une part, les progrès rapides des technologies de biologie moléculaire, dont la réaction de polymérisation en chaîne de l'ADN (PCR, de l'anglais Polymerase Chain Reaction), ont permis une utilisation facile et rapide de ces marqueurs. D'autre part, leur fort taux de mutation, probablement dû, au moins en partie, à des glissements de l'ADN polymérase sur ces motifs répétés, font que ces marqueurs ont progressivement remplacé, ou au moins complété, l'utilisation de marqueurs plus classiques tels que les allozymes dans de nombreuses applications en systématique moléculaire, génétique des populations et en écologie moléculaire (voir par exemple Estoup & Angers, 1998; Estoup *et al.*, 2002).

Plusieurs modèles ont été développés pour traiter du point du vue théorique les processus de mutation. Un premier modèle décrit l'identité entre gènes comme une identité par descendance : le modèle à nombre d'allèles infini (de l'anglais infinite allele model, IAM, Kimura & Crow, 1964). Selon ce modèle, chaque mutation crée un allèle différent de tous les allèles présents dans la population avant la mutation. Deux autres modèles, classiquement utilisés, ont été développés pour traiter le polymorphisme enzymatique puis utilisés pour d'autres marqueurs tels que les microsatellites. Ce sont le modèle à K allèles (K allele model en anglais, KAM, Crow & Kimura, 1970) et le modèle par pas (de l'anglais stepwise mutation model SMM, Ohta & Kimura, 1973). Sous le modèle KAM, la mutation engendre un allèle parmi K allèles possibles de façon équiprobable. Pour le modèle par pas, les possibilités de mutation sont beaucoup plus restreintes que pour les modèles précédents : une mutation diminue ou augmente, en proportions égales, le nombre de répétitions d'une unité. Un quatrième modèle mutationnel, spécifiquement adapté au cas des marqueurs microsatellites, est le GSM (de l'anglais generalised stepwise mutation model). Sous ce modèle, une mutation augmente ou diminue le nombre de répétitions d'un certain nombre d'unités, ce nombre d'unité est données par une loi géométrique. Ces quatre modèles présentent l'avantage de permettre le développement d'expressions mathématiques re-

lativement simples de paramètres populationnels. Sous ces modèles (KAM, SMM, GSM), seule l'identité par états sera mesurable en pratique. L'annexe B-1 de ce document présente le principe de base de l'intégration des modèles mutationnels dans les calculs d'identités par état à partir des identité par descendance. Par la suite, nous utiliserons le terme probabilité d'identité pour les probabilité d'identité par états.

Il est utile, pour décrire un modèle de population, subdivisé ou non, de définir des probabilités d'identité de paires des gènes à différents niveaux hiérarchiques. Il est 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 intra et inter-classes, ceci pour n'importe quel type de structure (voir par exemple Rousset, 1999a,b). Dans l'exemple précédent de la population panmictique isolée, on définira Q_0 la probabilité d'identité de paires de gènes pris au sein d'un individu diploïde et Q_1 la probabilité d'identité de paires de gènes pris dans deux individus différents de la population. Dans un modèle en population subdivisée, on ajoutera par exemple Q_2 la probabilité d'identité de paires de gènes pris dans deux sous-populations différentes. Ou encore, dans un modèle d'isolement par la distance, on considérera Q_r la probabilité d'identité de paires de gènes pris dans deux sous-populations séparées par une distance géographique r, et ce pour toutes les classes de distances. Nous reviendront plus en détails sur ces modèles démographiques dans la section 2.2.

2.1.2 F-statistiques

Pour quantifier la différenciation entre sous-populations, Wright (1943) a utilisé le rapport des variances des fréquences alléliques entre sous-populations sur les variances intra-populations, mieux connu sous le nom de F_{ST} (Wright, 1951). Depuis Wright, la définition et l'estimation des F-statistiques ont fait l'objet d'un vaste débat (Chakraborty & Danker-Hopfe, 1991; Excoffier, 2001; Rousset, 2001b; Weir & Cockerham, 1984). Nous retiendrons dans ce document uniquement l'approche développée par Cockerham (1969, 1973). Le point important de ces développements est qu'une décomposition de la variance totale du modèle considéré par Cockerham conduit naturellement à l'expression des F-statistiques en terme de rapport de probabilité d'identité par état (Rousset, 2001b). Ainsi, les paramètres F_{IS}, F_{ST} et F_{IT} sont défini par

$$F_{\mathrm{IS}} \equiv \frac{Q_0 - Q_1}{1 - Q_1};\ F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2};\ F_{\mathrm{IT}} \equiv \frac{Q_0 - Q_2}{1 - Q_2} \tag{2.3}$$

(voir Cockerham & Weir, 1987; Rousset, 1996). Ces expression mesurent la divergence entre gènes inter-classes relativement à la divergence intra-classe. Notons que les F-statistiques définies par les équations (2.3) sont des paramètres et non des statistiques. Cette écriture permet également de proposer des estimateurs de la forme

$$\hat{F}_{\mathrm{IS}} \equiv \frac{\hat{Q}_0 - \hat{Q}_1}{1 - \hat{Q}_1};\ \hat{F}_{\mathrm{ST}} \equiv \frac{\hat{Q}_1 - \hat{Q}_2}{1 - \hat{Q}_2};\ \hat{F}_{\mathrm{IT}} \equiv \frac{\hat{Q}_0 - \hat{Q}_2}{1 - \hat{Q}_2}. \tag{2.4}$$

Rousset (2001b) montre que ces estimateurs sont exactement identiques à ceux de Weir & Cockerham (1984).

Les F-statistiques ont fait et font encore l'objet d'une littérature très abondante. La première raison est que certains paramètres démographiques des populations peuvent être exprimés en fonction de F_{ST}. On connaît, par exemple, la formule de Wright de l'estimateur du *nombre efficace d'immigrants par génération*, dans le modèle à nombre d'îles infini, $Nm = (1/F_{\mathrm{ST}} - 1)/4$. Cet estimateur a été largement utilisé pour décrire la structuration 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 généralement pas remplies (Whitlock & McCauley, 1999). Nous étudierons plus en détails, ces modèles démographiques dans la section 2.2. La seconde raison, expliquant l'utilisation massive des F-statistiques, est qu'elles apparaissent également dans les modèles d'adaptation et de sélection de parentèle en populations subdivisées (Gandon & Rousset, 1999; Roze & Rousset, 2003; Whitlock, 2002, 2003).

2.1.3 Relation entre identité génétique, F-statistiques et temps de coalescence

Il existe une relation étroite entre les probabilités d'identité par état de paires de gènes, les F-statistiques et les temps de coalescence (Slatkin, 1991). La théorie de la coalescence s'intéresse à la généalogie des gènes. Comme on l'a vu en introduction, on dit que deux gènes coalescent à l'instant t dans le passé s'ils ont leur premier ancêtre commun (MRCA) à cet instant t (voir Fig.1.1). Un des intérêts majeurs de la théorie de la coalescence est que, sous l'hypothèse de neutralité des marqueurs, les processus de mutation peuvent être découplés des processus généalogiques (Hudson, 1990; Nordborg, 2001). 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. Par ailleurs, on peut facilement calculer les probabilités des temps de coalescence $C(t)$ de deux gènes pris dans différentes classes (Slatkin, 1991). $C(t)$ est plus précisément la probabilité que deux gènes coalescent à l'instant t. La probabilité que ces deux gènes n'aient pas muté pendant t générations étant γ^t, on a donc

$$Q_I = \sum_{t=1}^{\inf} \gamma^t C_I(t), \tag{2.5}$$

I prenant la valeur 0 si les gènes sont pris au sein d'un même individu, 1 s'ils sont pris dans deux individus distincts de la même population, 2 s'ils sont pris dans deux populations différentes et r s'ils sont pris dans deux populations séparées par une distance géographique r (Malécot, 1975; Slatkin, 1991). D'après cette définition des probabilités d'identité, et de la définition des F-statistiques (éq.2.3), on peut exprimer les F-statistiques en fonction des temps moyens de coalescence pour différentes paires de gènes. Ainsi, pour F_{ST}, on a

$$\lim_{\mu \to 0}(F_{\mathrm{ST}}) = \frac{T_2 - T_1}{T_2} \tag{2.6}$$

où $T_I = \sum_{t=1}^{\infty} tC_I(t)$ est le temps moyen de coalescence de paires de gènes dans la classe I. On peut alors mieux comprendre et interpréter les propriétés des F-statistiques à travers l'étude des probabilités de coalescence des paires de gènes pris dans différentes classes (Fig.2.1).

La figure 2.1 montre que, pour des temps anciens, la distribution des probabilités de coalescence dans une classe de gènes, par exemple $C_0(t)$, est proportionnelle à la distribution des probabilités de coalescence dans une classe de gènes moins apparentés, par exemple $C_1(t)$. Au contraire, dans une période de temps assez récente, les distributions diffèrent. Pour cette figure, nous avons volontairement considéré un taux d'autofécondation fort ($s = 0.5$) afin d'avoir des différences importantes entre les distributions de probabilités de coalescence de paires de gènes intra-individus et intra-dèmes. On peut donc décomposer la surface couverte par $C_0(t)$ en la somme de la surface couverte par $C_1(t)$ et une surface représentée par la région gris clair sur la figure 2.1. Cette surface gris claire représente une masse de probabilité équivalente à F_{IS} (Rousset, 2001b). De même, F_{ST} (région gris foncé sur la figure 2.1) est approximativement équivalent à la masse de probabilité correspondant à la différence des distributions $C_1(t)$ et $C_2(t)$. D'après cette figure, on comprend aisément que F_{IS} ne dépend que des événements récents

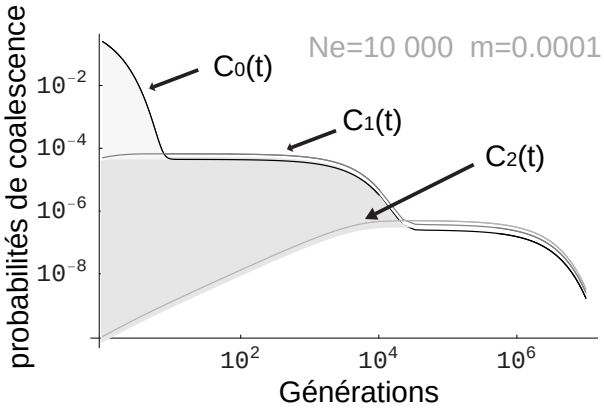

FIG. 2.1. Probabilité de coalescence en fonction du temps dans un modèle en îles. Les probabilités de $C_I(t)$ que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes : 0 pour des gènes intra-individu, 1 pour des gènes intra-dème mais inter-individu et 2 pour des gènes inter-dème. N=10000, m=0.0001 et n=100. Pour cette figure, nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t)$ et $C_1(t)$. L'aire gris clair correspond au F_{IS} et celle en gris foncée à F_{ST}. L'échelle des deux axes est logarithmique. D'après Rousset (2004).

de coalescence et qu'il est donc très peu influencé par la mutation (Rousset, 1996). Par contre, dans le modèle en îles considéré ici, F_{ST} dépend des événements de coalescence plus anciens. Il sera donc plus sensible à la mutation (i.e. taux de mutation et modèle mutationnel) que le F_{IS}. Nous verrons, dans la section suivante puis dans la section 3.5.4, comment différents facteurs démographiques (tailles de population et taux de migration) influencent ces courbes, et comment elles nous permettent de mieux appréhender les effets mutationnels et démographiques sur les F-statistiques.

2.2 Vers une dispersion réaliste : du modèle en îles à l'isolement par la distance

Comme nous l'avons vu en introduction, la distribution du polymorphisme peut nous renseigner sur les paramètres démographiques et génétiques influant sur la structuration des populations tels que les tailles (et/ou densité) des populations ainsi que sur les flux de gènes potentiels entre sous populations et/ou leurs patterns de dispersion. Ces patterns spatiaux de polymorphisme sont complexes et le développement de modèles en permet une

meilleure analyse. Le premier modèle, le plus simple, est le modèle en îles de Wright. Un des points faibles de ce modèle est la modélisation de la dispersion. En effet, l'hypothèse est faite que la dispersion se fait de façon équiprobable entre toutes les sous-populations, ce qui intuitivement semble en désaccord avec la réalité dans de nombreux cas. Malgré cette faiblesse, le modèle en îles a été, et est toujours, largement utilisé pour comprendre les conséquences évolutives de la dispersion.

Dans de nombreuses espèces, la dispersion est restreinte dans l'espace, et la différenciation génétique est plus faible à petite distance qu'à grande distance. Ceci est la pierre angulaire des modèles d'isolement par la distance introduits par Wright (1943, 1946). Dans cette section, je tacherai d'introduire ces différents modèles démographiques, en donnant les grandes lignes des analyses que l'on peut en faire et les principaux résultats qui en découlent, ceci sans entrer dans les détails mathématiques que le lecteur pourra trouver, par exemple, dans Rousset (2004).

2.2.1 Modèle en îles

On considère une population constituée de n_{d} sous-populations (ou dèmes), chacune de taille constante égale à $2N$ gènes (ou N individus adultes diploïdes). Comme pour le modèle panmictique de Wright-Fisher, chaque adulte produit une infinité de gamètes qui subissent l'effet de la mutation avec une probabilité μ. On considère ensuite qu'une proportion de ces gamètes migrent de leur dème d'origine vers un des $n_{\mathrm{d}} - 1$ autres dèmes avec la probabilité $m/(n_{\mathrm{d}} - 1)$. On peut noter ici que le nombre de migrants n'est pas strictement égal à Nm, mais est une variable aléatoire suivant une loi binomiale de moyenne Nm. Enfin, la compétition entre juvéniles ramène le nombre d'individus à N adultes.

Une analyse matricielle des probabilités d'identité dans ce modèle, analogue au raisonnement que l'on a eu en 2.1.1, permet d'exprimer, en fonction des paramètres du modèle, les ratios de probabilité d'identité par descendance suivant

$$\frac{Q_1}{1-Q_1} = \frac{1}{2Nn_d}\Big(\frac{\gamma}{1-\gamma} + (n_{\mathrm{d}}-1)\frac{\gamma(1-m\frac{n_{\mathrm{d}}}{n_{\mathrm{d}}-1})^2}{1-\gamma(1-m\frac{n_{\mathrm{d}}}{n_{\mathrm{d}}-1})^2}\Big), \qquad (2.7)$$

et

$$\frac{Q_2}{1-Q_1} = \frac{1}{2Nn_d}\Big(\frac{\gamma}{1-\gamma} - \frac{\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}{1-\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}\Big). \tag{2.8}$$

La différence entre ces deux expressions amène à l'expression suivante, d'une forme plus simple,

$$\frac{Q_1-Q_2}{1-Q_1} = \frac{F_\mathrm{ST}}{1-F_\mathrm{ST}} = \frac{1}{2N}\frac{\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}{1-\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}. \tag{2.9}$$

Contrairement aux expressions (2.7) et (2.8), cette expression a une limite finie quand $\mu \to 0$. On peut noter que toutes ces expressions sont fonctions du ratio $n_\mathrm{d}/(1-n_\mathrm{d})$ et ne dépendent donc que peu du nombre d'îles du modèle sauf si ce nombre est faible. Dans le cas limite où $n_\mathrm{d} \to \infty$, considéré par Wright dans son modèle à nombre d'îles infini, on a

$$\frac{F_\mathrm{ST}}{1-F_\mathrm{ST}} = \frac{1}{2N}\frac{\gamma(1-m)^2}{1-\gamma(1-m)^2} \approx \frac{1}{2N}\frac{\gamma(1-2m)}{1-\gamma(1-2m)}. \tag{2.10}$$

De cette expression, on obtient facilement le fameux résultat de Wright

$$F_\mathrm{ST} = \frac{1}{1+2N(2\mu+2m)} \approx_{\mu\to 0} \frac{1}{1+4Nm}. \tag{2.11}$$

On peut aussi retrouver l'expression correspondante pour un nombre d'îles fini, donnée par Li (1976),

$$F_\mathrm{ST} = \frac{1}{1+2N(2\mu+2\frac{n_\mathrm{d}}{n_\mathrm{d}-1}m)} \approx_{\mu\to 0} \frac{1}{1+4N\frac{n_\mathrm{d}}{n_\mathrm{d}-1}m}. \tag{2.12}$$

Les expressions (2.11) et (2.12) ont largement été utilisées pour caractériser des taux de migration entre sous-populations en calculant un $\hat{F}_\mathrm{ST}$ entre deux sous-populations et en exprimant le résultat en terme de nombre de migrants, $\hat{Nm} \approx (1/\hat{F}-1)/4$. Ce type de raisonnement n'est en aucun cas correct puisqu'un F_ST entre deux sous-populations n'est pas, dans le modèle en île, fonction du nombre de migrants entre ces deux sous-populations, mais du nombre de migrants moyen entre une sous-population et toutes les autres sous-populations du modèle. De plus, il est attendu que deux sous-populations, n'échangeant aucun migrant entre-elles mais échangeant des migrants avec d'autres sous-populations, aient entre-elles un F_ST non nul et donc en apparence un nombre de migrants, Nm, non nul.

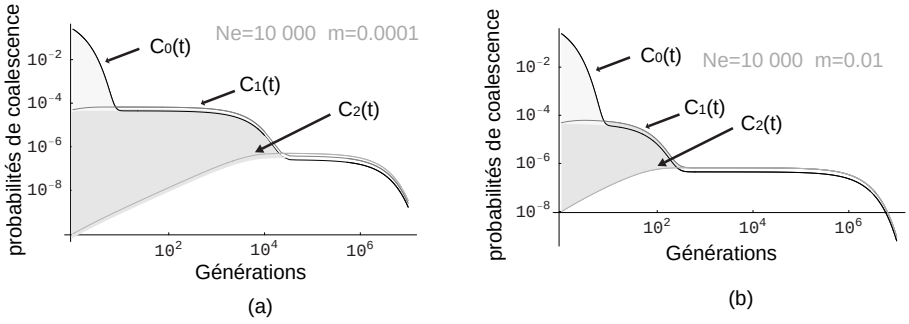

FIG. 2.2. Probabilités de coalescence dans un modèle en îles pour deux taux de migration : (a) migration faible $m = 0.0001$ (b) migration forte $m = 0.01$. Les probabilités, $C_I(t)$, que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes 0,1,2. Pour cette figure, $n_{\mathrm{d}} = 100$, $Ne = 10000$ et nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t)$ et $C_1(t)$. L'échelle des deux axes est logarithmique.

Le lien entre les F-statistiques, les probabilités d'identité et les temps de coalescence est utile pour une meilleure compréhension du modèle. Ainsi, la représentation des probabilités de coalescence en fonction du temps pour deux taux de migration différents (Fig.2.2a et 2.2b) nous renseigne sur quelques propriétés des F-statistiques dans le modèle en îles. On peut comprendre, par exemple, que le paramètre F_{ST}, correspondant à l'aire gris foncé, dépendra plus de la mutation que le F_{IS} (aire gris clair) et ce d'autant plus que la migration est faible. Si ce raisonnement est valide pour les événements de mutation agissant dans la zone de passée récente correspondant à la différence de probabilités de coalescence intra et inter-dèmes, il l'est aussi pour tout autre type d'événements affectant les probabilités de coalescence tels que des variations temporelles des paramètres démographiques (tailles de population et taux de migration). Ainsi, si les taux de migration sont forts et/ou, dans une moindre mesure, si les tailles de populations sont faibles, il est attendu que l'estimation de Nm par les F_{ST} corresponde plus à la valeur actuelle du paramètre qu'à une moyenne des valeurs passées et elle sera peu influencée par les processus de mutation des marqueurs génétiques (Rousset, 2004).

2.2.2 Dispersion : caractéristiques et modélisation

Une caractéristique majeure du modèle en îles est que les immigrants peuvent provenir, de façon équiprobable, de n'importe laquelle des sous-populations. Or, dans la réalité, la dispersion est le plus souvent localisée

dans l'espace. En d'autres mots, elle se fait préférentiellement entre deux sous-populations géographiquement proches. Puisque l'on considère ici la dispersion des gènes et non les mouvements nets d'individus, cela revient à dire qu'il y a une plus forte probabilité pour que des individus se reproduisent avec d'autres individus nés à proximité qu'avec des individus nés à plus grande distance.

Les jeux de données sur les distances de dispersion sont relativement rares, sans doute parce que la dispersion est un facteur difficile à estimer de manière rigoureuse. Endler (1977) a fait une revue extensive de la littérature de cette époque et a montré que généralement la dispersion est très limitée dans l'espace. Des analyses de clines d'allèles sélectionnés, par exemple, ont montré que la plupart des événements de dispersion était de l'ordre du kilomètre (Endler, 1977, pp.156-162). Quelques autres études ont aussi démontré des distances de dispersion restreintes chez les plantes (Fenster *et al.*, 2003; Crawford, 1984; Vekemans & Hardy, 2004) et chez les animaux (Rousset, 1997, 2000; Spong & Creel, 2001; Sumner *et al.*, 2001).

Les distributions de dispersion peuvent être caractérisées par leurs différents moments. Un moment M_k non centré d'ordre k est défini par $M_k \equiv E\left[X^k\right] = \sum_x x^k \Pr(X = x)$, un moment centré est défini par $M_k \equiv E\left[X - E(X)\right]^k$. Parmi les moments communément utilisés, la moyenne est le moment non centré d'ordre 1 et la variance est le moment centré d'ordre 2, $V(X) = E\left[X - E(X)\right]^2 = M_2 - M_1^2$. La kurtosis, définie par $M_4/M_3 - 3$, donne l'importance de la dispersion à courte et longue distance par rapport aux dispersions intermédiaires (le -3 est une convention pour que la kurtosis d'une loi normale soit nulle). Un autre moment qui apparaît souvent dans les modèles génétiques (par exemple lors de l'étude des clines, Nagylaki, 1975; Barton & Gale, 1993) est σ^2, le moment d'ordre 2 [1] de la distance axiale de dispersion, ou encore la moyenne des carrés des distances [2] de dispersion parents-descendants.

On a souvent voulu adapter des distributions théoriques pour modéliser la dispersion. Une des distributions les plus utilisées est la loi normale du

[1] Il s'agit du moment non centré d'ordre deux si la dispersion est mesurée en valeur absolue ; cependant puisque nous nous intéressons a la distance axiale (voir note suivante) la moyenne de la distribution de dispersion est nulle et les moments centrés sont alors identiques au moment non centrés.

[2] nous considérerons ici et dans le reste du document que les distances correspondent à de distances vectorielles (les coordonnées (x, y) d'une entité par rapport à une autre) et non les distances euclidiennes ($r \equiv \sqrt{x^2 + y^2}$), plus classiquement utilisées. x et y sont appelées distances axiales et peuvent être négatives, contrairement à la distance euclidienne.

fait de la convexité de la courbe à petites distances. Mais les distributions de dispersion sont souvent leptokurtiques, c'est à dire qu'elle ont un excès de dispersion à faible et longue distance par rapport à la dispersion à des distances intermédiaire (revue dans Endler, 1977; Kot *et al.*, 1996; Portnoy & Willson, 1993). On dit aussi qu'elles ont une longue queue, ou une forte kurtosis. Le problème de la loi normale est qu'elle ne prends pas en compte cette caractéristique majeure des distributions de dispersion. Une autre caractéristique des distributions de dispersion est qu'elles doivent avoir une forte kurtosis tout en ayant un taux de migration global assez fort (peu d'individus se reproduisent exactement où leurs parents se sont reproduits). Cette caractéristique fait que beaucoup de distributions communément utilisées, telles que les distributions exponentielles ou géométriques, sont inadaptées à la modélisation des processus de dispersion. Certaines familles de distribution permettent de combiner des taux de migration forts et une forte kurtosis, parmi celles-ci on peut citer les distributions discrètes de la forme $f_k = f_{-k} = M/k^n$, où f_k est la probabilité de migrer à une distance k. Pour ces distributions, M contrôle approximativement le taux de migration global et n la kurtosis. Elles correspondent à des distributions discrètes de Pareto (ou Zeta) tronquées (voir, par exemple, Patil & Joshi, 1968). Ce sont ces distributions que l'on utilisera dans le chapitre 3. En pratique, les distributions de dispersion sont extrêmement diverses et il parait donc préférable de ne pas se focaliser sur une famille de distributions, telle que la loi normale, mais de considérer des modèles généraux applicables à n'importe quels types de distributions.

2.2.3 Modèle en réseau : de la sous-population à la population continue

Les modèles considérés pour les analyses de l'isolement par la distance sont des modèles en réseaux. Par modèles en réseau, on entend un ensemble de sous-populations placées sur une grille régulière formant un cercle, en une

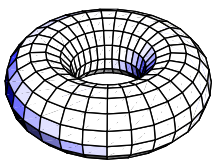

FIG. 2.3. Représentation graphique d'un tore

dimension, ou un tore, en deux dimensions. La modélisation en tore ou en cercle permet d'éviter les effets de bord et d'avoir ainsi une surface parfaitement homogène (un tore est représenté sur la Fig.2.3). N individus adultes composent chacune des sous-populations, dont la position sur le réseau est donnée par deux coordonnées, $r \equiv (x, y)$. L'unité de longueur est la distance inter-dèmes, c'est à dire la distance entre deux sous-populations adjacentes. Au cours du cycle de vie, chaque adulte produit une infinité de gamètes qui subissent les effets de la mutation avec une probabilité μ. Chaque gamète migre ou non de façon indépendante vers un autre dème et la compétition ramène le nombre d'adulte dans chaque dème à N.

Dans ce modèle, chaque dème se comporte comme une population panmictique dans lesquelles tous les individus sont équivalents. Dans un modèle plus réaliste, il n'y aurait pas forcément de structure en dèmes et les individus pourraient se placer n'importe où sur une surface continue définissant l'habitat. La position des individus de la population varierait, entraînant une variation des densités locales entre générations. De tels modèles de populations continues ont été développés (Wright, 1943, 1946; Malécot, 1967; Sawyer, 1977; Barton *et al.*, 2002) mais ne suivent pas un ensemble d'hypothèses biologiques (i.e. le fait qu'il n'y ait aucune régulation explicite des densités locales conduit une certaine forme d'agrégation des individus dans l'espace) et sont en ce sens non réalistes et difficilement utilisables (Maruyama, 1972; Felsenstein, 1975). Le meilleur modèle de population continue que nous avons à présent est donc le modèle en réseau avec un individu par nœud du réseau. Ce modèle peut être considéré comme une approximation de la population continue avec une régulation locale de la densité (par exemple lorsque la compétition locale est forte (Malécot, 1975; Rousset, 2000)).

2.2.4 Structuration génétique en isolement par la distance

Si les premières analyses des modèles d'isolement par la distance ont été faite par Wright (1943, 1946) et Malécot (1948), toutes les analyses rigoureuses découlent du modèle en réseau formulé par Malécot (1950). Dans ces modèles, toutes les distributions de dispersion peuvent être considérées, à condition d'avoir des moments finis jusqu'à l'ordre 3, impliquant une dispersion locale. On suppose également que la migration est homogène dans l'espace, c'est à dire que la distribution de dispersion est identique en tout point du réseau. L'outil principal dans l'analyse des modèles en réseau est l'analyse de Fourier. Un exemple d'une telle analyse est donnée dans les annexes A-1

et A-2. Pour l'application de l'analyse de Fourier aux modèles d'isolement par la distance dont nous ne donnerons ici que les principaux résultats, le lecteur pourra se référer à Rousset (1997, 2004) et Sawyer (1977).

Comme pour les précédents modèles, l'analyse mathématique du modèle d'isolement par la distance permet l'expression des probabilités d'identité par descendance en fonction des paramètres du modèle. Tous les résultats que l'on présentera dans cette partie sont valables pour des réseaux de taille infinie. On a alors, en une dimension,

$$\frac{Q_r}{1-Q_1} \approx \frac{e^{-\sqrt{2\mu}r/\sigma}}{4N\sigma\sqrt{2\mu}}, \tag{2.13}$$

où σ^2 est le moment d'ordre deux de la distribution de distance de dispersion parent-descendant. L'équation (2.13) est une approximation pour des grandes distances géographiques. Pour $r = 0$, on a alors

$$\frac{Q_1}{1-Q_1} \approx \frac{1}{4N\sigma\sqrt{2\mu}} + \frac{A_1}{4N\sigma}, \tag{2.14}$$

où A_1 est une constante déterminée uniquement par la distribution de dispersion. Sawyer (1977) en donne la définition suivante

$$A_1 \equiv 2\sigma\left(\frac{1}{\pi}\int_0^{\pi}\frac{\psi^2(x)}{1-\psi^2(x)} - \frac{1}{\sigma^2x^2}dx - \frac{1}{\pi^2\sigma^2}\right), \tag{2.15}$$

où ψ est la fonction caractéristique des probabilités de dispersion m_r, $\psi(z) \equiv \sum_r m_r e^{\imath rz}$. De telles fonctions caractéristiques sont couramment utilisées dans le cadre des analyses de Fourier (voir annexes A-1 et A-2).

En deux dimensions, pour deux gènes à la distance euclidienne $r \equiv \sqrt{x^2+y^2}$, on a

$$\frac{Q_r}{1-Q_1} \approx \frac{K_0(\sqrt{2\mu}r/\sigma)}{4N\pi\sigma^2}, \tag{2.16}$$

où K_0 est la fonction modifiée de Bessel de second type et d'ordre zéro (voir par exemple Abramovitz & Stegun, 1972). L'équation (2.16)est aussi une approximation pour des r grands et une autre expression doit être considérée pour $r = 0$:

$$\frac{Q_1}{1-Q_1} \approx \frac{-\ln(\sqrt{2\mu}+2\pi A_2)}{4N\pi\sigma^2}, \tag{2.17}$$

où A_2 est de la même nature que A_1. Le lecteur pourra en trouver une expression dans Sawyer (1977, éq.3.4). On définit le paramètre a_r, analogue à $F_{\mathrm{ST}}/(1-F_{\mathrm{ST}})$,

$$\begin{aligned} a_r \equiv \frac{Q_1 - Q_r}{1 - Q_1} &\approx \frac{1 - e^{-\sqrt{2\mu}r/\sigma}}{4N\sigma\sqrt{2\mu}} + \frac{A_1}{4N\sigma} \\ &\approx^{r \text{ et } \mu \text{ petit}} \frac{r}{4N\sigma^2} + \frac{A_1}{4N\sigma} \approx^{N \to D} \frac{r}{4D\sigma^2} + \frac{A_1'}{4D\sigma} \end{aligned} \tag{2.18}$$

en une dimension, et en deux dimensions on a

$$\begin{aligned} a_r &\approx \frac{-\ln(\sqrt{2\mu}) - K_0(\sqrt{2\mu}r\sigma) + 2\pi A_2}{4N\pi\sigma^2} \\ &\approx^{r \text{ et } \mu \text{ petit}} \frac{\ln(r\sigma) - 0.116 + 2\pi A_2}{4N\pi\sigma^2} \approx^{N \to D} \frac{\ln(r)}{4\pi D\sigma^2} + \frac{\ln(\sigma) - 0.116 + 2\pi A_2'}{4\pi D\sigma^2} \end{aligned} \tag{2.19}$$

Les dernières expressions correspondent au passage des tailles de populations N à la densité d'individus sur le réseau D. Dans ces expressions, peu importe l'unité de longueur utilisée, il suffit que la densité soit exprimée avec la même unité de longueur que σ. Une unité simple, que l'on utilisera ensuite, est la maille du réseau (la distance entre deux dèmes adjacents). C'est l'unité qui est utilisée pour σ lorsque l'on considère des tailles de populations (modèle avec structure démique). C'est à ce niveau que se fait le lien entre les modèles à structuration démique dans lesquelles les individus sont regroupés en dèmes et les modèles en populations continues dans lesquels les individus sont répartis de façon homogène sur toute la surface définissant l'habitat.

On a donc une relation linéaire entre a_r et la distance géographique en une dimension, et entre a_r et le logarithme de la distance géographique en deux dimensions (Fig.2.4). On remarquera que les approximations sont faites en considérant tout d'abord que les distances sont grandes (éq.2.13 et éq.2.16), et petites ensuite (éq.2.18 et éq.2.19). La relation linéaire sera donc valide pour des distance intermédiaires. Que ce soit en une ou deux dimensions, la pente de cette relation est fonction de $D\sigma^2$. Ces approximations permettent une description relativement simple de la différenciation attendue sous ce modèle et permettent d'envisager une estimation du produit $D\sigma^2$ par la pente de la relation entre la différenciation, observée sur des marqueurs génétiques et mesurée par une estimateur de a_r, et la distance géographique (ou le logarithme de la distance pour le modèle en deux dimensions). Nous verrons en détail les

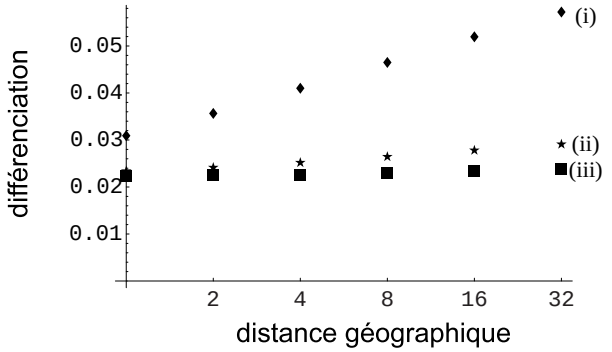

FIG. 2.4. Différenciation en fonction du logarithme de la distance en isolement par la distance en deux dimensions avec : (i) une structuration forte ; (ii) avec une structuration moins forte $\sigma_{ii}^2 > \sigma_i^2$; et (iii) dans un modèle en îles avec le même taux total de migration 4/9. Noter l'échelle logarithmique de la distance géographique. (d'après Rousset (2004))

caractéristiques d'une telle estimation dans le chapitre 3. On peut noter aussi que la différenciation sous ces modèles n'est pas uniquement fonction de σ^2 mais également d'autres caractéristiques de la dispersion "contenues" dans les constantes A. De plus, pour des taux de migration faibles, la différenciation entre deux sous-populations adjacentes est proche de la différenciation attendue sous un modèle en îles avec le même nombre d'émigrants (Fig.2.4). Ceci confirme que σ^2 n'est pas la seule caractéristique de dispersion jouant sur la différenciation génétique.

Sur la figure 2.5, l'aire gris foncé représente la masse de probabilité correspondant à a_r. Sous isolement par la distance, la relation entre les F-statistiques et les temps de coalescence nous indique les même tendances que pour le modèle en îles (Fig.2.5a et b). L'influence des mutations et des fluctuations démographiques passées sera d'autant plus faible que les taux de migration sont forts (Fig.2.5b) et, dans une moindre mesure que les tailles des dèmes (ou les densités) sont petites. On peut aussi souligner que cette influence sera d'autant plus faible que les sous-populations (ou les individus dans le modèle continu) comparées seront proches géographiquement (Fig.2.5c ; Slatkin, 1993).

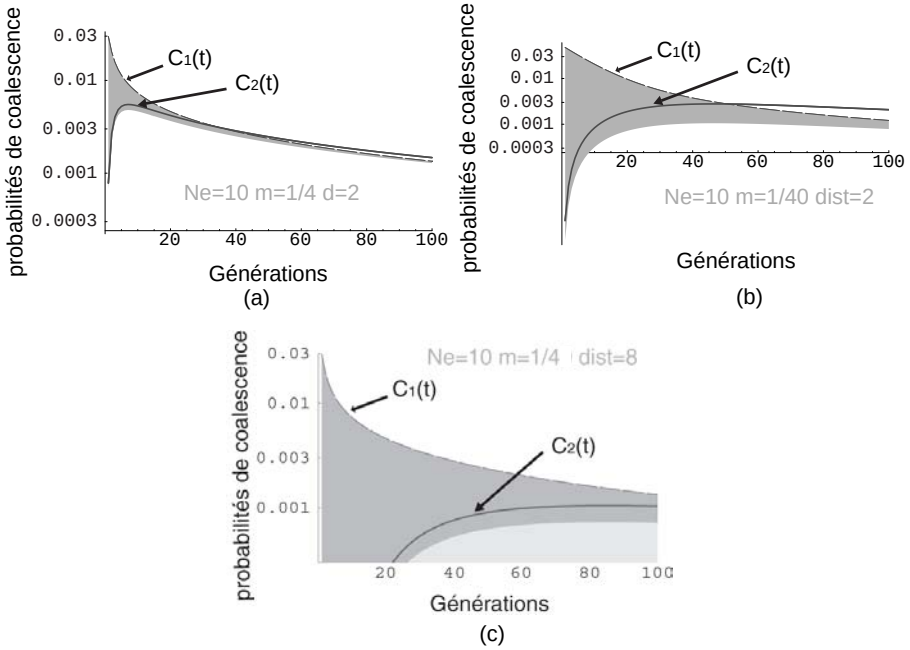

FIG. 2.5. Probabilité de coalescence sous isolement par la distance en fonction du taux de migration et de l'éloignement des gènes considérés. $C_1(t)$ correspond aux temps de coalescence de gènes intra-dèmes. $C_2(t)$ correspond aux temps de coalescence de gènes inter-dèmes situé à 2 unités du réseau pour les figures (a) et (b) et à 8 pas de distance pour (c). La migration se fait uniquement entre dèmes adjacents (stepping stone) avec un taux de migration de 1/4 pour (a) et (c); et 1/40 pour (b). Pour toutes ces figures, le réseau est en une dimension, le nombre de dèmes est 100 et chaque dème est constitué de 10 adultes haploïdes. L'aire gris foncée correspond à a_r. L'échelle de l'axe des ordonnées est logarithmique.

2.3 Coalescence en populations subdivisées

On peut distinguer deux approches principales pour l'analyse des modèles en génétique des populations : les méthodes des moments et les méthodes fondées sur la théorie du maximum de vraisemblance. Les méthodes des moments, dont nous avons donnés des exemples à travers les F-statistiques, sont fondées sur l'analyse des corrélations des fréquences alléliques entre différents niveau hiérarchiques définis au sein de la population étudiée. Les approches par maximum de vraisemblance nécessitent le calcul de la probabilité d'un échantillon en fonction des paramètres du modèle, appelée vraisemblance de l'échantillon. Le calcul de la vraisemblance d'un échantillon peut se faire par une approche analytique pour un petit nombre de modèles, ou par simulation. La plupart des développement récents de méthodes d'estimation de paramètres démographiques par maximum de vraisemblance utilisent une approche par coalescence pour estimer la vraisemblance d'un échantillon. Comme nous l'avons mentionné en introduction, un des intérêts majeurs de la coalescence est de prendre en compte l'histoire généalogique sous-jacente aux données.

Considérons un locus particulier dans le génome d'une espèce. Quel que soit l'échantillon que l'on considère, toutes les copies à ce locus sont reliées entre elles et à un ancêtre commun par leur histoire généalogique, que l'on peut représenter sous la forme d'un arbre généalogique (Fig.1.1 et 2.6).Le polymorphisme (les trois allèles a, b, c sur la figure 2.6) est dû aux mutations qui ont eu lieu le long des branches de cet arbre, et la fréquence de chaque allèle est déterminée par la fraction des branches qui portent ces allèles. Le pattern de polymorphisme à ce locus reflète donc l'histoire des coalescences de ces lignées, représentée dans l'arbre, et l'histoire des mutations. Tout comme les mutations sont distribuées de façon aléatoire dans les processus évolutifs, la généalogie de gènes l'est aussi et on doit donc considéré ces deux sources de variations dans les modèles. On a donc besoin de modèles qui nous permettent de décrire des généalogies aléatoires de gènes. La suite de cette section est une présentation succincte de quelques exemples classiques de tels modèles.

2.3.1 n-coalescent

Le n-coalescent (ou par extension le coalescent) est le modèle standard pour construire ces arbres de coalescence. Il a été développé pour décrire la généalogie d'un échantillon de n gènes dans une population de Wright-Fisher,

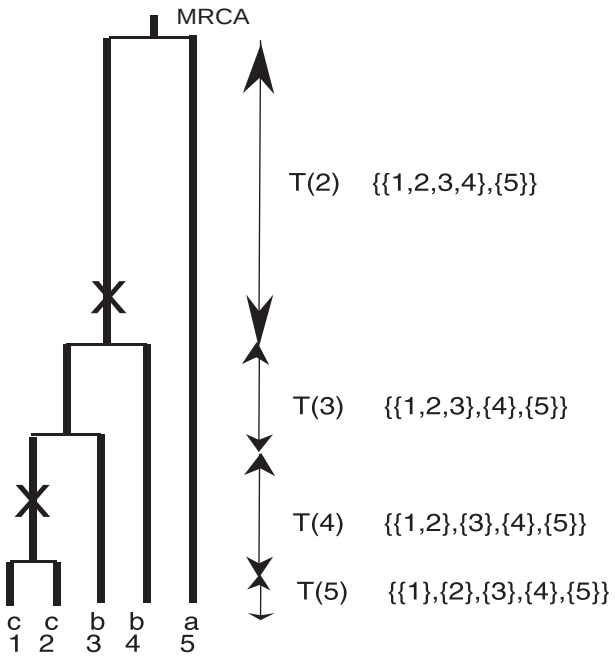

FIG. 2.6. La généalogie d'un échantillon de gènes peut être décrite en terme de topologie et de longueur de branche. La topologie peut être représentée comme des classes d'équivalence des lignées ancestrales. La longueur des branches correspond aux temps entre deux événements de coalescence.

c'est à dire dans une population panmictique, isolée, avec des générations non chevauchantes. Suivons deux lignées de gènes en remontant le temps. Dans une population de N gènes, la probabilité que ces deux lignées aient un ancêtre commun à la génération précédente et donc coalescent est $1/N$. La probabilité qu'elles restent distinctes est donc $1-1/N$. Puisque les générations sont indépendantes, la probabilité qu'elles restent distinctes plus de t générations dans le passé est donc $(1-1/N)^t$. De cette formule, on peut comprendre l'approximation standard en temps continue décrite ci-après. Cette approximation valable pour des grandes tailles de populations (N grand) est la base du n-coalescent et de nombreuses autres approximations en découlent. Considérons un changement d'échelle du temps tel que l'unité de temps correspond à N générations. La probabilité que deux lignées restent distinctes pendant plus de τ unité de temps est alors

$$(1-\frac{1}{N})^{\lceil N\tau\rceil} \approx^{N\to\infty} e^{-\tau}, \tag{2.20}$$

où $\lceil N\tau\rceil$ est le plus grand entier plus petit que $N\tau$. Le temps de coalescence (exprimé en N générations) d'une paire de gène suit donc une loi exponentielle, $e^{-\tau}$, de moyenne 1. Considérons maintenant k lignées, la probabilité qu'aucune de ces lignées ne coalesce à la génération précédente est

$$\prod_{i=1}^{k-1}(1-\frac{i}{N}) = 1-\frac{k(k-1)}{2N}+O(\frac{1}{N^2}). \tag{2.21}$$

Définissons $T(k)$, le temps de la première coalescence dans un échantillon de k lignées (voir Fig.2.6). On a alors, selon le même raisonnement que précédemment, $T(k) \approx^{N\to\infty} \frac{k(k-1)}{2}e^{-\tau\frac{2}{k(k-1)}}$. De plus, comme on le voit dans la formule (2.21), quand N est grand on peut négliger les coalescences multiples (c'est à dire la probabilité que plus de deux lignées coalescent simultanément à une génération donnée). Sous l'approximation en temps continu, le nombre de lignées d'un échantillon décroît donc pas à pas en fonction de $T(k)$, le temps nécessaire pour passer de k à $k-1$ lignées (Fig.2.6).

En résumé, le modèle du n-coalescent décrit la généalogie d'un échantillon de n gènes comme un arbre avec des bifurcations aléatoires, où les $n-1$ temps de coalescence, $T(n), T(n-1), .., T(3), T(2)$ sont des variables aléatoires mutuellement indépendantes suivant une loi exponentielle. Les mutations sont ensuite surimposée sur ces branches, en redescendant le temps (du MRCA jusqu'au présent), selon une loi géométrique avec comme paramètres le taux de mutation par unité de temps et la longueur de la branche.

On voit bien ici que la simulation d'un arbre de coalescence sous ce modèle est extrêmement efficace. Il suffit de simuler $n-1$ variables aléatoires selon une loi exponentielle, correspondant aux temps de coalescence, et de construire de façon indépendante une topologie aléatoire des bifurcations, les coalescences, en choisissant au hasard les paires de lignées qui coalescent. On ajoute ensuite sur l'arbre les mutations.

Quelques propriétés majeures en génétique des populations se comprennent facilement à l'aide de ce modèle. Ainsi, puisque les coalescences se font avec une probabilité proportionnelle à $k(k-1)/2$, elles sont d'autant plus rapides qu'il y a de lignées. Ainsi, le temps de coalescence attendu des k lignées (temps de l'ancêtre commun le plus récent, TMRCA), correspondant à la hauteur de l'arbre de coalescence, est donné par

$$TMRCA = E\left[\sum_{k=2}^{n} T(k)\right] = \sum_{k=2}^{n} E\left[T(k)\right] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2(1-\frac{1}{n}). \tag{2.22}$$

Or on a vu que l'espérance du temps de coalescence de deux lignées est $E\left[T(2)\right] = 1$. La portion de l'arbre pendant laquelle il n'a que deux lignées est donc plus grand que la moitié de sa longueur totale. De plus, Saunders *et al.* (1984) ont montré que la probabilité que le MRCA d'un échantillon de taille n soit le MRCA de la population entière est $(n-1)/(n+1)$. On voit donc bien qu'il n'est pas nécessaire de prendre un grand échantillon pour remonter le plus loin possible dans le passé. Ceci implique également que les inférences sur des processus démographiques anciens seront limitées du fait que le MRCA peut être récent notamment lorsque les tailles de populations sont faibles. Enfin, il est intéressant de noter que, dans la mesure où le nombre de mutations est proportionnel à la longueur des branches et que les copies d'un gène sont étroitement liées du fait de leur généalogie commune, une augmentation de la taille de l'échantillon (au dessus d'une valeur relativement faible d'environ 20 individus) n'accroîtra que peu la puissance des analyses de génétique des populations.

Le n-coalescent est donc une représentation naturelle des processus évolutifs dans une population panmictique. Il possède des propriétés mathématiques simples et permet de réaliser des simulations extrêmement efficaces. Enfin, la réelle importance du n-coalescent découle surtout du fait qu'il s'adapte facilement, grâce à des changements d'échelle judicieux, à de nombreux modèles neutres. Ainsi, il est possible de considérer que la variance du nombre de descendants n'est pas 1 comme dans le modèle de Wright-Fisher mais α. Dans ce cas, il suffit de considérer que l'échelle de

temps n'est plus N mais N/α et le modèle du n-coalescent peut s'appliquer. Ces changements d'échelle permettent de prendre en compte de nombreux phénomènes biologiques tels que des générations chevauchantes, des sexes séparés avec biais de sexe ratios, et d'autres systèmes de reproduction dans l'approche du n-coalescent. Ceci est dû au fait que ces phénomènes biologiques ne changent pas la topologie de l'arbre mais seulement la longueur des branches. La considération d'une échelle non linéaire avec le temps peut aussi permettre de prendre en compte certaines variations simples des tailles de population dans le temps (de nombreux exemples de ces changements d'échelle sont revus dans Nordborg, 2001; Rousset, 2004).

2.3.2 Coalescent structuré

Les populations naturelles étant le plus souvent structurées dans l'espace, il est nécessaire de considérer les applications de la théorie de la coalescence pour des modèles de populations structurées (on parle alors de coalescent structuré). Considérons une population d'individus haploïdes de taille N subdivisée en n_{d} sous-populations de tailles N_i telles que $\sum_i N_i = N$. A chaque génération, chaque adulte produit une infinité de gamètes qui subissent l'effet de la mutation avec une probabilité μ. On considère ensuite qu'une proportion de ces gamètes migre de leur dème d'origine i vers le dème j avec la probabilité m_{ij}. Enfin, la compétition entre juvéniles ramène le nombre d'individus à N_i adultes dans chaque sous-population. On définit les quantités $c_i \equiv \frac{N_i}{N}$ et $Bij \equiv Nb_{ij}$, ou les b_{ij} sont les probabilités de migration "arrière" (en remontant le temps). En d'autres termes b_{ij} est la probabilité qu'un gène du dème i ait son parents dans le dème j. Ces quantités sont défini en fonction des taux de migration "avant" m_{ji} par

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}. \tag{2.23}$$

On suppose que ces quantités sont constantes quand $N \to \infty$ et que le temps est mesuré en N générations. C'est à dire que l'on considère que les probabilités des événements de migration et de coalescence au sein d'un dème sont des $O(1/N)$. De plus, les probabilités que plusieurs événements de migration et/ou de coalescence se fassent dans une génération sont des $O(1/N^2)$. Dans la limite où $N \to \infty$, les seuls événements possibles sont donc une coalescence au sein d'un dème ou une migration entre deux dèmes. Le temps d'occurrence d'un de ces événements (i.e. le temps attendu avant qu'un événement se produisent) suit alors une loi exponentielle de moyenne, r, le taux d'événement

global, somme des taux de chaque événement possible,

$$r(\mathbf{n}) = \sum_i \Big(\frac{k_i(k_i-1)}{2c_i} + \sum_{j \neq i} k_i B_{ij} \Big), \tag{2.24}$$

où k_i est le nombre de lignées présentes dans le dème i à la génération considérée et $\mathbf{n} = \{k_i\}$, pour $i \in [1, .., n_{\mathrm{d}}]$, est la configuration globale des lignées dans les différents dèmes. Si un événement a lieu, c'est une coalescence dans le dème i avec la probabilité

$$\frac{k_i(k_i-1)/2c_i}{r(\mathbf{n})}, \tag{2.25}$$

ou c'est une migration ("arrière") du dème i vers le dème j avec la probabilité

$$\frac{k_i B_{ij}}{r(\mathbf{n})}. \tag{2.26}$$

Avec ces modèles de structuration, il n'y a pas de changement d'échelle possible pour se rapprocher du n-coalescent. Ceci traduit le fait que la structuration des populations ne change pas uniquement la longueur des branches mais aussi la topologie de l'arbre. Ainsi, si la migration est faible, les lignées échantillonnées dans un même dème vont coalescer rapidement entre elles et le temps nécessaire pour que deux lignées échantillonnées dans deux dèmes différents coalescent va être beaucoup plus long que dans le modèle panmictique. Intuitivement, on comprend que, si la migration est faible, ces lignées mettront un certain temps avant de se retrouver dans le même dème pour pouvoir coalescer. Cette caractéristique est illustrée par la figure 2.7. Une limite de ce modèle réside dans le fait que l'on peut considérer uniquement des grandes tailles de populations afin de ne pas avoir à considérer d'événements multiples pour avoir une expression simple de temps d'attente entre deux événements. Il est intuitif de dire qu'une faible migration aura un effet important sur la structure des généalogies. A l'inverse, une migration forte va, d'une certaine manière, nous rapprocher des généalogies obtenues sous un modèle panmictique. Une migration forte implique que les événements de migration sont beaucoup plus fréquents que les événements de coalescence et $\lim_{N\to\infty} Bij = \lim_{N\to\infty} Nb_{ij} = \infty$. Dans le cas limite où $N \to \infty$, il y aura donc une infinité d'événements de migration entre deux événements de coalescence. C'est ce que l'on appelle une séparation des échelles de temps : les événements de migration ayant lieu sur une échelle de temps beaucoup

plus rapide que les coalescences. Les événements de coalescence ayant lieu entre deux lignées présentes dans le même dème, ils sont alors fonction de la distribution stationnaire $\Pi = \{\pi_i, i \in [1, .., n_{\mathrm{d}}]\}$ des lignées entre les différents dèmes. Selon ces notations, la coalescence d'une paire de lignées a lieu dans le dème i avec la probabilité π_i^2/c_i, puisque la probabilité que deux lignées se retrouvent dans le dème i est π_i^2. En prenant comme échelle de temps N/α avec $\alpha \equiv \sum_i \pi_i^2/c_i$, le taux de coalescence total de l'échantillon, on se ramène alors au modèle du n-coalescent (Nagylaki, 1980; Notohara, 1993). Cette séparation des échelles de temps simplifie donc considérablement l'analyse mathématique du coalescent structuré. Elle permet entre autre de simplifier le calcul de la probabilité d'un échantillon sous ces modèles avec une forte migration. Toutefois cette séparation des processus de migration et de coalescence ignore certains effets spécifiques de la migration et, dans ce cas, l'utilisation du n-coalescent avec la mise à l'échelle N/α ne permet pas d'inférences sur les processus de migration.

Une notion similaire de séparation des échelles de temps peut être retrouvée dans le cadre des analyses des probabilités d'identité dans le modèle en îles. On s'intéresse ici à la séparation des événements de coalescence intra- et inter-dème. En effet, une augmentation du nombre de dèmes diminuera la probabilité de coalescence inter-dème tout en gardant une probabilité de coalescence intra-dème constante. Dans le cas limite où $n_{\mathrm{d}} \to \infty$, la probabilité de coalescence inter-dème ($C_2(t)$) va tendre vers zéro pour tout t. La figure 2.8 montre bien que lorsque l'on augmente le nombre de dèmes , $C_2(t)$ diminue pour tout t.

Une séparation similaire des échelles de temps est retrouvée pour une

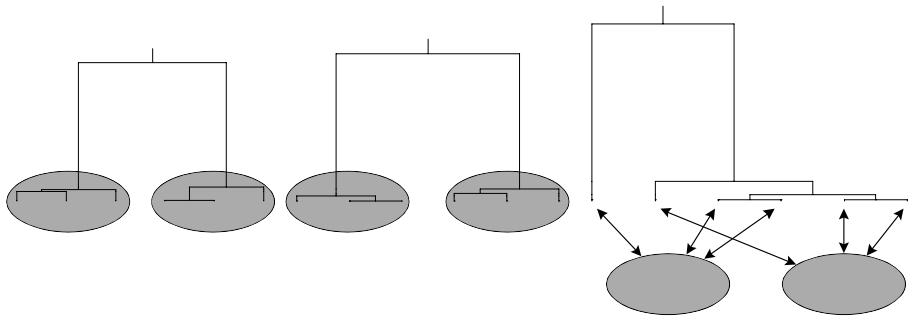

FIG. 2.7. Trois réalisations du coalescent structuré sous un modèle symétrique de migration à deux dèmes avec $N_i = 3$. Les lignées ont tendance à coalescer au sein d'un dème mais pas toujours comme le montre l'arbre de droite. (figure issue de Nordborg, 2001)

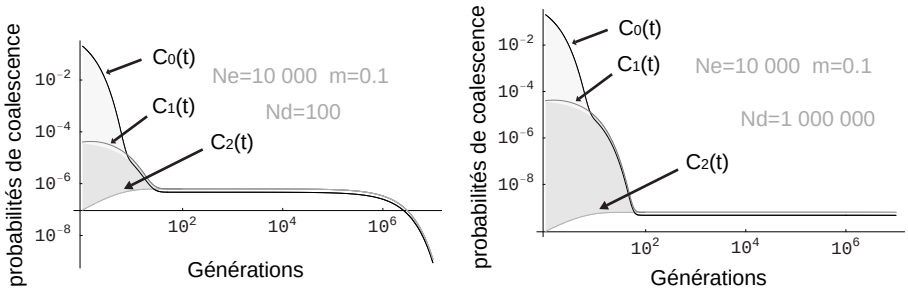

FIG. 2.8. Probabilité de coalescence dans un modèle en îles pour deux nombres de dèmes : (a) 100 dèmes (b) 1 000 000 dèmes. Les probabilités de $C_I(t)$ que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes $0, 1, 2$. Pour cette figure, nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t) et C_1(t)$. L'échelle est une double échelle logarithmique.

population panmictique. Lorsque la taille de la population tend vers l'infini, on a une séparation dans le temps des événements de coalescence intra- et inter-individus. Une telle séparation des événements de coalescence intra- et inter-classe permet une meilleur interprétation des propriétés de certains paramètres du modèle. Notons que ce raisonnement a déjà été appliqué précédemment de manière intuitive lorsque l'on a considéré que le F_{IS} était principalement déterminé par les probabilités de coalescence dans une zone du passé récent (section 2.1.3). On en avait déduit que le F_{IS} devrait être peu sujet à l'influence de la mutation et des processus démographiques anciens. Ainsi lorsqu'une séparation des échelles de temps est possible, certains paramètres s'inscrivent dans l'échelle de temps longs alors que d'autres sont dans l'échelle de temps courts et seront de ce fait moins influencés par des processus passés.

2.3.3 Simulation d'arbres de coalescence

Dans un nombre de situations très variées, la coalescence permet donc de développer un modèle simple pour l'analyse des propriétés d'un échantillon de gène. L'approche par coalescence permet aussi le développement d'un nouveau mode d'interprétation des propriétés d'un échantillon en génétique des populations par la prise en compte de la généalogie des gènes. On peut noter que ces propriétés peuvent le plus souvent être obtenues par d'autres approches plus classiques sans références explicites à la généalogie en utilisant par exemple des équations de diffusion ou des récurrences sur une génération (approche utilisée dans les sections 2.1 et 2.2). Toutefois, l'ap-

proche par coalescence fournit un cadre général relativement intuitif pour l'analyse des patterns de variation génétiques, sous de nombreux modèles démo-génétiques. Elle permet entre autre de décrire les propriété de certaines statistiques à l'aide de jeux de données génétiques obtenus par simulation. De plus, les méthodes d'estimation de paramètres démographiques par maximum de vraisemblance nécessitent le calcul de la vraisemblance d'un échantillon. Jusqu'à récemment, cela n'a été possible que pour un nombre limité de modèles correspondant globalement aux cas du n-coalescent ou du coalescent structuré. Dans le cas de modèles plus complexes, une approche alternative au maximum de vraisemblance est possible. Cette approche utilise des algorithmes de simulations de données génétiques. L'information présente dans ces données est ensuite résumée à l'aide d'un ensemble de statistiques, telles que le nombre d'allèles ou encore les F_{ST}, et comparée au même ensemble de statistiques estimées sur le jeux de données réel. Ces approches d'inférence sont dites approches par statistiques résumées (voir Beaumont *et al.*, 2002; Pritchard *et al.*, 1999; Estoup *et al.*, 2001; Marjoram *et al.*, 2003). Dans ce cadre, les simulations fondées sur le processus de coalescence jouent donc un rôle important dans la compréhension et l'interprétation des modèles évolutifs mais aussi, et surtout, dans le développement et le test de nouvelles méthodes d'inférence.

Trois différents principes de simulation permettent de générer des échantillons de gènes ayant évolué sous des modèles démo-génétiques variés, de la simple population de Wright-Fisher à des modèles complexes pouvant prendre en compte n'importe quels facteurs démographiques. La première méthode développée par Donnelly (1999) ne s'applique que dans le cadre du n-coalescent. C'est une méthode qui s'apparente à un tirage de boules de couleurs dans une urne avec les couleurs correspondant aux états alléliques des gènes de l'échantillon. L'avantage de cette méthode est son extrême rapidité, cet algorithme étant plus rapide que l'approche directe du n-coalescent mentionnée dans la section 2.3.1. L'inconvénient est qu'il ne s'applique que dans des cas simples.

La seconde méthode, développé par Hudson (1983, 1990), s'applique dans un cadre plus général pouvant comporter une structuration géographique et/ou des variations des paramètres démographiques (e.g. Cornuet & Luikart, 1996; Estoup *et al.*, 2001). Il est par exemple utilisable dans le cadre du coalescent structuré. Le principe est de remonter le temps événement par événement, sans considérer les mutations, selon les probabilités données par les équations (2.25) et (2.26). Pour chaque événement de coalescence ou de migration, on calcule le moment dans le passé auquel cet événement s'est pro-

duit. Ce temps entre deux événements suit une loi exponentielle de moyenne égale à l'inverse des probabilités mentionnées ci-dessus. Cela permet de décrire en même temps la topologie et la longueur des branches de l'arbre de coalescence. Lorsqu'on arrive au premier ancêtre commun de l'échantillon on a alors complètement décrit l'arbre de coalescence de notre échantillon. Il suffit alors d'ajouter les mutations sur les différentes branches en redescendant dans le temps (une description précise de l'ajout des mutation sur une généalogie est donnée en section 3.1.1). Le nombre de mutations par branche est donné par une loi Binomiale de paramètres (μ, t), où μ est le taux de mutation par génération et t est la longueur de la branche en nombre de génération. Dans la littérature, le loi Binomiale est souvent approximée par une loi de Poisson de paramètre (μt). Cette approximation est valide pour des grandes longueurs de branches, donc pour des taux de coalescence et de migration faibles et/ou des faibles taux de mutation.

La troisième méthode de simulation d'échantillons de gènes par coalescence correspond aux algorithmes dits génération par génération. Le principe est très simple et assez proche de la simulation en approximation continue de Hudson. La différence est que l'on ne remonte pas le temps événement par événement mais génération par génération. En d'autres termes, on envisage tous les événements possibles de coalescence ou de migration de tous les gènes de l'échantillon à chaque génération. Les mutations sont ensuite ajoutées sur l'arbre comme pour la méthode de Hudson. Un exemple très détaillé d'un tel algorithme sera développé dans le chapitre suivant.

Ces trois méthodes diffèrent essentiellement par leur rapidité et par la complexité des modèles démographiques qu'elles peuvent prendre en compte. La plus rapide, le modèle en Urne, ne s'applique que dans des situations simples s'apparentant au n-coalescent. La méthode la plus lente, l'algorithme génération par génération, peut s'appliquer sous n'importe quel type de modèle démographique. Enfin, l'algorithme en temps continu de Hudson peut s'appliquer à de nombreux modèles pour peu que l'on dispose d'une loi donnant l'expression des temps d'attente entre deux événements. Il correspond donc au cadre d'application du n-coalescent. Ceci n'est pas trivial dans de nombreux cas, par exemple lorsque l'on considère des petites populations, des taux de migration forts ou encore des démographies très variables dans le temps. Dans ces cas, seul l'algorithme génération par génération permettra de simuler des patrons génétiques de manière satisfaisante.

2.4 Conclusion

Ce chapitre a permis de poser les bases nécessaires au travail que j'ai effectué durant ma thèse. J'ai tout d'abord donné une description succincte des différents modèles classiquement utilisés en génétique des populations. J'ai ensuite montré comment la description des modèles évolutifs, que ce soit en terme de probabilités d'identité, de F-statistiques ou encore par des approches par coalescence, permettait une meilleur compréhension des systèmes biologiques et surtout de mieux appréhender les conséquences de différents processus évolutifs au niveau du polymorphisme des populations naturelles. Enfin, j'ai rapidement expliqué comment ces approches permettaient dans certains cas la mise en place de méthodes d'inférence, à partir de données moléculaires sur le polymorphisme des populations, de différents paramètres importants en évolution tels que des taux de migrations. Cependant, il est important de noter que les populations naturelles correspondent rarement aux descriptions que l'on peut en faire dans les modèles. Ainsi, si certains modèles, tels que les modèles d'isolement par la distance, semblent plus réalistes que d'autres, il parait extrêmement important de tester la précision mais surtout la robustesse des résultats par rapport à la violation des hypothèses que l'on fait généralement quand on applique ces modèles sur des données en populations naturelles. Le chapitre 3 de ma thèse développera un exemple de tests de la robustesse de l'estimation de paramètres démographiques sous un modèle d'isolement par la distance par rapport à divers facteurs mutationnels et démographiques.

Chapitre 3

Estimation par F-statistiques : précision et robustesse en populations continues sous isolement par la distance

Nous avons vu dans le précédent chapitre comment l'ajustement de modèles démo-génétiques pouvait permettre, au moins dans certains cas, l'estimation de paramètres démographiques à partir de données génétiques (cf. éq.2.11 et éq.2.19 par exemple). De nombreuses méthodes ont été développées dans ce but (revues dans Slatkin, 1994; Rousset, 2001b). Les multiples contradictions observées entre les estimations directes et indirectes à partir de données génétiques ont souvent été attribuées à l'inadéquation des hypothèses des modèles génétiques (Hastings & Harrison, 1994; Koenig *et al.*, 1996; Slatkin, 1994).

Dans ce chapitre, nous nous focaliserons sur une méthode d'analyse fondées sur l'augmentation de la différenciation génétique entre individus en fonction de leur distance (Rousset, 2000). Cette méthode, dite méthode des moments, repose sur l'analyse des modèles d'isolement par la distance introduite à la section 2.2.4. Une régression du paramètres a_r, analogue au paramètre $F_{\mathrm{ST}}/(1-F_{\mathrm{ST}})$, calculée entre paires d'individus et le logarithme de la distance géographique est utilisée pour estimer le paramètre $D\sigma^2$, ou D est la densité d'individus adultes et σ^2 le carré moyen de la distance de dispersion parent-descendant (voir la définition du paramètre a_r en section 2.2.4). Cette méthode est en principe plus performante que les méthodes pré-

cédentes traitant des modèles d'isolement par la distance pour au moins deux raisons : (i) le modèle démographique sous-jacent, qui est celui introduit en section 2.2.4, fait peu d'hypothèses quant aux distributions de dispersion et il est particulièrement robuste pour des distributions leptokurtiques, une caractéristique communément observée dans les populations naturelles (voir la section 2.2.2 relative aux distributions de dispersion) ; (ii) les variations des F-statistiques avec la distance géographique donnent des résultats plus facilement interprétables que les F-statistiques en elles-mêmes. Ce dernier point peut être illustré par les équations (2.18) et (2.19) de la section 2.2.4 dans lesquelles on voit bien que les constantes "A", entrant dans la définition du paramètres a_r, sont des fonctions complexes des distributions de dispersion alors que la variation de a_r avec la distance géographique n'est fonction que de $D\sigma^2$.

Un autre intérêt majeur de cette méthode est que l'analyse de la différenciation est faite à une petite échelle géographique (échelle géographique locale). Or les hypothèses de stabilité démographique sont moins critiques quand on considère des petites surfaces (Slatkin, 1993). Plus précisément, les études à des échelles géographiques locales donnent de meilleures estimations car l'hétérogénéité des paramètres démographiques (tels que la densité ou la dispersion) est réduite et leur influence sur l'hétérogénéité des processus génétiques tels que la dérive le sont aussi (Rousset, 2001a). Mais ceci ne fait qu'atténuer le problème de l'hétérogénéité potentielle des paramètres démographiques. Dans le cas de fortes variations dans le temps, il est légitime de se poser la question de la signification exacte des paramètres de dispersion et de densité estimés. Plus précisément, obtient-on un estimateur du paramètres actuel ou bien l'estimation subit-elle largement l'influence des variations antérieures ? D'autre part même si l'homogénéité spatiale est plus probable à des échelles démographiques locales, il est possible que la présence de zones présentant des densité plus fortes que d'autres ait une influence majeure sur les estimations du produit $D\sigma^2$, en particulier si l'échantillon a été récolté en partie ou totalement dans de telles zones. Enfin, comme nous l'avons vu en section 2.1.1, le processus mutationnel des marqueurs génétiques sont souvent complexes et mal connus, et leur influence sur l'estimation de paramètres démographiques dépend des méthodes utilisées. Aucune étude n'a pour l'instant évalué la robustesse de la méthode des moments vis à vis processus mutationnels des marqueurs.

Il nous a paru intéressant de quantifier l'effet de tels écarts par rapport aux hypothèses de base du modèle sur l'estimation de $D\sigma^2$ par la méthode des moments. Dans ce but, nous avons développé un algorithme de simula-

tion génération par génération fondé sur la théorie de la coalescence (voir section 2.3.3) pour tester l'influence sur l'estimation du produit $D\sigma^2$: (i) de l'échelle d'échantillonnage des individus, (ii) des processus mutationnels (taux et modalités de mutation) des marqueurs utilisés, avec une référence spéciale aux microsatellites, et (iii) d'hétérogénéités spatiales et temporelles des paramètres démographiques.

3.1 Bases méthodologiques de l'étude

3.1.1 Algorithme de simulation

Modèle démographique et cycle de vie

Le modèle démographique considéré est un modèle d'isolement par la distance en population continue. Pour cela nous avons considéré un modèle en réseau avec un individu diploïde par nœud du réseau. Comme nous l'avons vu en section 2.2.3, ce modèle est la seule formalisation rigoureuse de population continue disponible à ce jour. Pour éviter les effets de bord, nous avons considéré un réseau en deux dimension formant un tore. Nous verrons dans ce chapitre et dans le chapitre 6 que la taille du réseau ainsi que les effets de bord n'ont que peu d'influence sur l'estimation du produit $D\sigma^2$ quand la taille de l'habitat est grande par rapport a la dispersion moyenne, ce qui est le cas ici. Enfin, nous avons considéré des individus diploïdes avec une phase de migration gamétique. Le cycle de vie, analogue aux cycles de vie introduits dans le chapitre 2, est composé de 5 étapes : (i) à chaque événement de reproduction, chaque adulte donne naissance à une infinité de gamètes et meurt ; (ii) les gamètes subissent l'effet de la mutation ; (iii) les gamètes dispersent ; (iv) dans chaque dème, des individus diploïdes sont formés à partir du pool de gamètes ; et (v) la compétition ramène le nombre d'adultes à N (le plus souvent $N = 1$, des précisions seront données lorsque $N > 1$).

Algorithme de coalescence génération par génération

Pour les modèles d'isolement par la distance en population continue, il n'existe pas de traitement analytique des probabilités et temps de coalescence pour plus de deux gènes. Les modèles classiques du n-coalescent et du coalescent structuré ne s'appliquent pas non plus puisque la migration est forte et les dèmes de très petite taille (un individu). Nous avons donc utilisé

un algorithme de simulation de coalescence génération par génération décrit ci-dessous :

Considérons, à un moment donné que nous appellerons présent, un échantillon de $n(0)$ gènes, numérotés de 1 à $n(0)$, répartis sur un réseau en deux dimensions. La position de chacun de ces gènes sur le réseau est décrite par une paire de coordonnées (x, y). L'ensemble des coordonnées des gènes échantillonnés est donné par les vecteurs $X(0) = (x_1(0), ., x_{n(0)}(0))$, $Y(0) = (y_1(0), ., y_{n(0)}(0))$, où $x_i(0)$ et $y_i(0)$ sont les coordonnées du gène i échantillonné à la génération $G = 0$. G corresponds au nombre de générations en remontant le temps depuis le moment de l'échantillonnage.

On se place sous l'hypothèse de reproduction à intervalles discrets ce qui permet de remonter le temps par pas d'une génération. Au cours du processus de coalescence, l'arbre de coalescence est constitué des parents des gènes considérés à $G = 0$ (gènes échantillonnés). A la génération $G = 1$, les parents de notre échantillon de $n(0)$ gènes ont pour coordonnées $x_i(1) = x_i(0) + dx$, $y_i(1) = y_i(0) + dy$, dx et dy étant des variables aléatoires représentants la distance de dispersion dans respectivement les dimensions x et y. La distance de dispersion est exprimée en nombre de pas du réseau. Sous un modèle à deux dimensions, la loi de la variable aléatoire (dx, dy) est une fonction $b_{dx,dy}$, que l'on appellera par la suite fonction de dispersion "arrière" (car on s'intéresse à la probabilité que le parent était à une distance (dx, dy), d'où le b pour "backward"). Cette fonction est calculée à partir de la fonction $f[orward]_{dx,dy}$, fonction de dispersion "avant" (dispersion parents-descendants) qui décrit à quelle distance un parent envoie ses juvéniles. Le passage des fonctions de dispersion "avant" aux fonctions de dispersion "arrière" ainsi que les distributions de dispersion en elles-mêmes sont détaillés ultérieurement.

Connaissant sur le réseau la position des parents, il s'agit à présent de trouver les événements de coalescence ayant eu lieu à $G = 1$. En d'autres termes, il s'agit de déterminer si certains gènes ont leur parent en commun à $G = 1$ (ceci correspond à leur premier ancêtre commun que l'on appellera par exemple $MRCA_{i,j,k}$ si i, j, k sont trois gènes coalesçant ensemble). Pour qu'une coalescence puisse se produire, il faut que les "nœuds" [1] parentaux soient au sein d'un même individu (à savoir dans le même dème dans le cas du modèle en réseau avec un individu par dème) et qu'ils soient issus du même gène de cet individu. Cette dernière condition a une expression simple à savoir une probabilité de 1/2 pour la coalescence d'une paire de gènes et

[1] un "nœud" de l'arbre correspond à un événement de coalescence ou à une extrémité de branche de l'arbre et ne doit pas être confondu avec les nœuds du réseau démographique que l'on appellera préférentiellement dèmes.

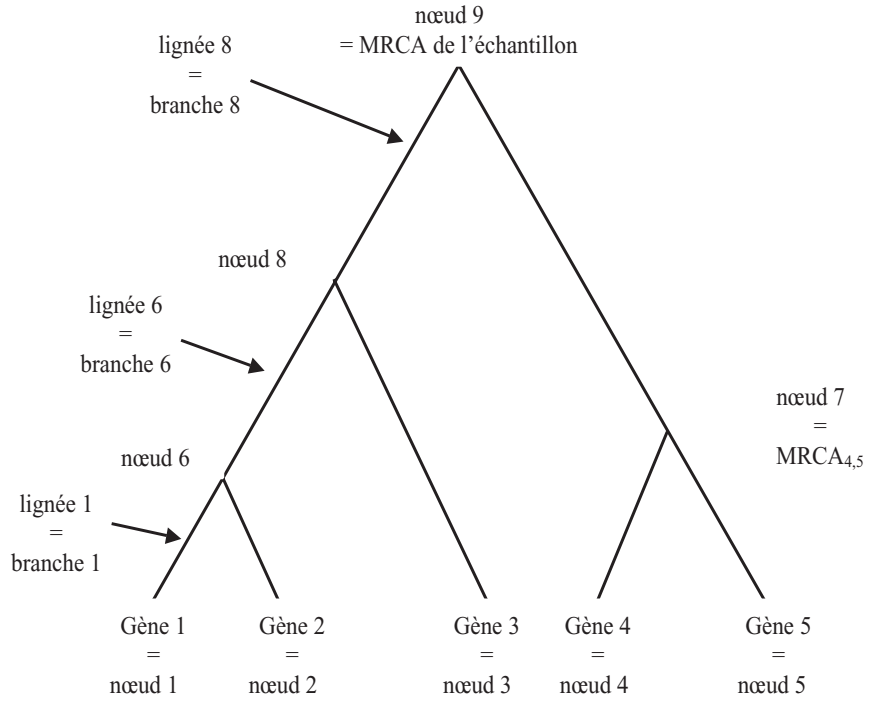

FIG. 3.1. Représentation d'un arbre de coalescence de cinq gènes. Les légendes correspondent aux notations utilisées pour la description de l'algorithme de coalescence génération par génération

$1/2^{n-1}$ pour une coalescence multiple de n gènes. Par commodité, on gardera les numéros ($i \in [1, .., n(0)]$) des gènes n'ayant pas coalescé pour leurs parents et on attribuera un nouveau numéro ($i \in [n(0)+1, .., n(1)]$) pour les gènes parents de gènes ayant coalescé (voir Fig.3.1). Ainsi la numérotation ci-dessus s'applique plus à la notion de branche de l'arbre de coalescence qu'à la notion de gène puisqu'un gène i à $G = 0$ et son parent à $G = 1$ peuvent avoir le même numéro i s'il n'y a pas eu d'événement de coalescence entre le gène i à $G = 0$ et un autre gène à $G = 0$. On a donc $X(1) = (x_1(1), ., x_{n(1)}(1))$, $Y(1) = (y_1(1), ., y_{n(1)}(1))$, l'ensemble des $n(1)$ coordonnées géographiques à $G = 1$ de chaque branche correspondant à une lignée de notre échantillon de départ. On garde en mémoire l'âge des nœuds de l'arbre ainsi que les numéros des branches issues de ce nœud, ces branches correspondent aux lignées descendant de ce nœud (voir Fig.3.1).

L'ensemble du processus est répété sur de nombreuses générations jusqu'à ce que l'on trouve le MRCA de l'ensemble des gènes de notre échantillon. On a ainsi l'ensemble des événements de coalescence présents dans la généalogie de

notre échantillon de gènes (ensemble des "nœuds" de l'arbre de coalescence), ce qui nous donne un arbre généalogique possible pour notre échantillon (voir Fig.3.1).

Ce modèle de base a un individu par dème. Certaines modifications sont nécessaires afin de modéliser la coalescence en population continue caractérisée par des déviations par rapport aux hypothèses de base du modèle en réseau. Plus spécifiquement : (i) le nombre de gènes par dème n'est plus constant ni dans le temps ni dans l'espace mais varie en chaque nœud du réseau selon une fonction de densité $N_{x,y,G}$; (ii) la fonction de dispersion avant peut aussi varier dans l'espace et dans le temps.

Notons que les changements de dispersion avant dans le temps ont une influence uniquement sur la détermination de la position sur le réseau du gène ancestral (la compétition ramène le nombre d'individu en chaque point du réseau à $N_{x,y,G}$). En revanche, les changements de densité ont une influence d'une part sur la détermination de la position sur le réseau du gène ancestral (migration arrière) mais aussi sur la probabilité de coalescence de deux gènes se retrouvant dans le même dème. Plus précisément la probabilité de coalescence de deux gènes d'un même dème (x, y) à la génération G n'est plus $1/2$ mais $1/(2N_{x,y,G})$.

Modélisation de la dispersion

Les distributions de dispersion que nous avons utilisées sont des distributions discrètes, de la forme

$$f_{dx} = M/dx^n \tag{3.1}$$

pour $0 < dx < K_{max}$ (voir section 2.2.2). Dans certains cas, afin de maintenir une dispersion totale forte, nous avons indépendamment fixé les p premiers termes de la distribution qui devient

$$\begin{aligned} f_1 = f_{-1} = M_1,\ f_2 = f_{-2} = M_2, .., \ f_p = f_{-p} = M_p \\ \text{et pour } p < dx < K_{max},\ f_{dx} = f_{-dx} = M/dx^n. \end{aligned} \tag{3.2}$$

On supposera pour la suite que la migration est indépendante dans chaque direction. On a alors $f_{dx,dy} = f_{dx} \cdot f_{dy}$. De plus lorsque l'on considère que la densité est homogène dans l'espace, la distribution de dispersion arrière est la même en tout point du réseau et est simplement égale à la distribution de dispersion avant, soit $b_{dx,dy} = f_{dx,dy} = f_{dx} \cdot f_{dy}$.

Plaçons-nous maintenant dans un cas plus complexe, à savoir un modèle avec variation de la densité dans le temps et dans l'espace et calculons les distributions de dispersion "arrière". Chaque point du réseau a alors une distribution de dispersion "arrière" propre, dépendant de la densité en chaque dème susceptible d'être le dème d'origine du gène considéré (à savoir l'ensemble des dèmes situés à une distance du dème d'origine inférieure ou égale à K_{max} pas). Soit $N_{x,y,G}$ le nombre d'individus au dème (x, y) à la génération parentale G. La probabilité de dispersion "arrière" de dx pas sur la première dimension et dy pas sur la deuxième dimension au dème (x_i, y_i) est alors de la forme

$$b_{dx,dy} = \frac{N_{x_i+dx,y_i+dy,G} \cdot f_{dx,dy}}{\sum_{dx,dy \leq K_{max}} N_{x_i+dx,y_i+dy,G} \cdot f_{dx,dy}}. \tag{3.3}$$

Ajout des mutations

Pour connaître l'état du gène du bas d'une branche de l'arbre généalogique en fonction de l'état allélique du gène du haut de cette branche, nous avons opté pour la procédure pas à pas suivante : prenons au hasard deux gènes i, j et leur premier ancêtre commun, le gène l ($MRCA_{i,j}$), et intéressons-nous à leurs états alléliques respectifs $état_i$, $état_j$, et $état_l$. Le nombre de mutations ayant eu lieu dans la lignée i est proportionnel à la longueur L_i de la branche i (qui va de i à l) et est donné par une loi Binomiale de paramètres (m, L_i). Soit m_i le nombre de mutations ayant eu lieu sur la branche, il suffit alors pour trouver $état_i$ de partir de $état_l$ en procédant par étapes, chaque étape correspondant à un événement de mutation sous le modèle mutationnel choisi. Ainsi en partant de $état_l$ on trouve l'état suivant correspondant à un événement de mutation sur $état_l$ sous le modèle mutationnel considéré. On part ensuite de cet état intermédiaire pour trouver le second état intermédiaire correspondant à un événement de mutation sur le premier état intermédiaire ou encore à deux événements de mutation sur $état_l$. On fait ceci m_i fois et on obtient ainsi $état_i$, l'état allélique de i descendant de l sur L_i générations. En partant d'un état donné pour l'ancêtre commun (tiré dans la distribution stationnaire des états alléliques sous le modèle mutationnel considéré) de l'ensemble des gènes de l'échantillon (racine de l'arbre) et en procédant comme ci-dessus en descendant le long de chacune des branches de l'arbre de coalescence jusqu'au "présent", on obtient alors les états alléliques de chaque gène de l'échantillon.

Validation des algorithmes et des programmes

L'algorithme ainsi que le programme de simulation utilisés pour cette étude ont été testés par comparaisons des probabilités d'identité de deux gènes obtenues par simulation avec celles obtenues analytiquement à l'aide des développements de Malécot (1975) et adaptées aux différents modèles mutationnels avec les méthodes générales de Rousset (1996)(voir annexe A-1). Ces comparaisons ont montré que les probabilité d'identité obtenues par simulation et les attendus analytiques diffèrent de moins de 1 pour mille pour des runs suffisamment longs (plus de 10^6 arbres échantillonnés).

3.1.2 Analyse de la qualité des estimations

Chaque simulation produit les génotypes à 7, 10, 13 ou 25 locus polymorphes, selon les situations, pour 100 (10×10) individus caractérisés par leurs coordonnées sur un réseau de taille (500×500) (sauf précision contraire). Sept à 25 locus et 100 individus représentent le nombre d'individus et de locus communément analysés dans des études expérimentales utilisant des marqueurs microsatellites. Enfin, des arbres de coalescence indépendants ont été utilisés pour simuler des données multilocus à des locus indépendants. En pratique, il est difficile d'échantillonner tous les individus sur une petite surface (échantillonnage quasi-exhaustif). Nous avons considéré un échantillon de (10×10) individus échantillonnés tous les deux dèmes, sur une aire de (20×20) dèmes du réseau. Nous nous rapprochons ainsi d'un échantillonnage expérimental habituel. Ce processus a été répété 1 000 fois pour chaque situation étudiée donnant 1 000 échantillons multilocus pour 100 individus partageant la même histoire démographique.

Pour chaque échantillon multilocus, un estimateur du paramètre $a_r \equiv \frac{Q_w - Q_r}{1 - Q_w}$ est calculé pour chaque paire d'individus, avec Q_w la probabilité d'identité par état de deux gènes pris dans le même individu, et Q_r la probabilité d'identité par état de deux gènes séparés par une distance géographique r (Rousset, 2000, voir éq.2.19). Un estimateur de a_r pour une paire ξ d'individus pris parmi les P différentes paires possibles est

$$\hat{a} \equiv \frac{\mathrm{SS}_{b(\xi)} P}{\sum_{k=1}^{P} \mathrm{SS}_{w(k)}} - \frac{1}{2}, \tag{3.4}$$

où $\mathrm{SS}_{b[etween](\xi)} \equiv \sum_{i,j,u} (X_{i:u} - X_{.:u})^2$ mesure la divergence entre deux gènes pris dans deux individus différents et $\mathrm{SS}_{w[ithin](\xi)} \equiv \sum_{i,j,u} (X_{ij:u} - X_{i:u})^2$

mesure la divergence entre deux gènes pris au sein d'un individu ($X_{ij:u}$ est une variable indicatrice prenant la valeur 1 si le gène i de l'individu j est de type allélique u ou 0 sinon)(Rousset, 2000). Pour chaque jeu de données simulé, on calcule la valeur de la pente de la droite de régression entre $\hat{a}$ et le logarithme de la distance géographique. Comme nous l'avons vu dans la section 2.2.4, sous certaines hypothèses, à savoir des distances géographiques intermédiaires et des taux de mutation faibles, l'inverse de cette pente est un estimateur du produit $D\sigma^2$ (éq.2.19).

La qualité d'un estimateur est généralement évaluée par le calcul de son biais et du carré moyen des erreurs (MSE de l'anglais mean square error). Ces mesures sont valides pour des estimateurs ayant une distribution approximativement normale mais pas pour des estimateurs prenant des valeurs infinies. Or, dans le cas présent, une pente négative doit être interprétées comme une estimation de $D\sigma^2$ infini. C'est pourquoi nous présenterons les résultats en terme de biais et de MSE de la pente de la droite de régression et non du produit $D\sigma^2$. Pour tout le reste de ce document nous appellerons valeurs attendues les valeurs des paramètres choisies pour les simulations. Nous avons donc calculé sur toutes les répétitions : (i) le biais moyen relatif de la pente par rapport à la valeur attendue de $1/(4\pi D\sigma^2)$ (i.e. moyenne de (pente observée-pente attendue)/pente attendue), (ii) l'erreur standard de ce biais, et (iii) le MSE relatif de la pente par rapport à la valeur attendue (i.e. moyenne de (pente observée-pente attendue)^2/(pente attendue)^2). On a aussi calculé la proportion d'estimations tombant dans un intervalle de facteur 2 par rapport à la valeur attendue (i.e. dans l'intervalle [pente attendue/2 ; 2$\times$ pente attendue]).

Il est important de noter que les biais observés dans nos simulations peuvent provenir : (i) d'un biais, inhérent à la méthode, dû à l'effet des forts taux de mutation sur la valeur du paramètre (que nous appellerons "biais paramétrique") ; (ii) d'un biais dû à la déviation des estimateurs par rapport à la valeur du paramètre lorsque l'on considère un échantillon fini d'individus et de locus ("biais d'échantillonnage") ; et (iii) d'un biais introduit par les écarts aux hypothèses du modèle démographique (variations spatialles et temporelles des paramètres démographiques). Pour plus de précision sur ces biais, le lecteur pourra se référer à l'annexe B-2 et B-3.

3.1.3 Intervalles de confiance

Afin d'avoir une estimation précise de l'incertitude associée à l'estimation du paramètre $D\sigma^2$, la procédure dite ABC bootstrap non-paramétrique de DiCiccio & Efron (1996) a été adaptée pour calculer des intervalles de confiance à 95% autour de la pente de la droite de régression. La méthode ABC (de l'anglais Approximate Bootstrap Confidence interval) est une approximation analytique d'un algorithme dit BCa (de l'anglais Bias Corrected and Accelerated bootstrap ; DiCiccio & Efron, 1996) qui permet de calculer un intervalle de confiance à partir d'une distribution artificielle du paramètre obtenue par bootstrap (i.e. par ré-échantillonnage à partir du jeu de données de départ, en réalisant pour chaque nouvel échantillon, un tirage avec remise dans le jeu de données initial). L'avantage de la technique ABC est que l'on calcule analytiquement, par 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, qui est ici limitant. Dans cette procédure, les données à chaque locus sont considérées comme des réplicats indépendants du processus généalogique. Des tests de cette procédure ont été réalisés en calculant la probabilité que la valeur $D\sigma^2$ attendue tombe dans l'intervalle de confiance à 95% (i.e. probabilité de recouvrement, coverage probability en anglais). Ceci a été fait sur 1000 jeux de données simulés à l'aide de l'algorithme décrit ci-dessus. On a choisi une distribution de dispersion arbitraire avec $\sigma^2 = 4$ (éq.3.2) dont les paramètres sont

$$M_1 = 0.06,\ M_2 = 0.03 \text{ et pour } 2 < dx < 49,\ M = 0.802,\ n = 2.518. \quad (3.5)$$

Les estimateurs de a_r et les intervalles de confiance ont été calculés pour 7, 13 et 25 locus sous un modèle mutationnel SMM avec un taux de mutation de $5 \cdot 10^{-4}$(voir section 2.1.1 pour les modèles mutationnels).

TAB. 3.1. Probabilité de recouvrement de l'intervalle à 95% autour de la pente de régression par ABC bootstrap.

	Taille de l'échantillon		
Proportion d'intervalles...	7 locus	13 locus	25 locus
...contenant la valeur attendue	0.842	0.885	0.90
...au dessous la valeur attendue	0.020	0.030	0.030
...au dessus de la valeur attendue	0.138	0.085	0.070

Les résultats présentés dans le Tableau 3.1 montrent que la procédure ABC bootstrap donnent des intervalles de confiance à 95% imparfaits en

TAB. 3.2. Probabilité de recouvrement de l'intervalle à 95% autour de la moyenne d'un échantillon tiré dans une loi très asymétrique (Fig.3.2) par ABC bootstrap.

	Taille de l'échantillon		
Proportion d'intervalles...	13	100	5000
...contenant la valeur attendue	0.80	0.891	0.917
...au dessous la valeur attendue	0.020	0.019	0.020
...au dessus de la valeur attendue	0.18	0.09	0.063

terme de probabilité de recouvrement même pour un grand nombre de locus (e.g. la probabilité de recouvrement est 0.90 au lieu de 0.95 pour 25 locus). La procédure ABC a une performance plus médiocre encore pour des petites tailles d'échantillon (e.g. 0.842 au lieu de 0.95 pour 7 locus, Tableau 3.1). La mauvaise performance de la méthode concerne surtout la borne inférieure de l'intervalle de confiance, dans la mesure où la proportion d'intervalles au dessus de la valeur attendue est 0.138 et 0.07 au lieu de 0.025 pour respectivement 7 et 25 locus. Quelle est l'origine de ce biais ? Il est intéressant de noter que la procédure ABC a été testée par ailleurs sur des données simulées correspondant à une tirage dans une loi normale et les probabilités de recouvrement correspondent exactement aux seuils choisis pour l'intervalle de confiance (résultats non montrés). Il est donc légitime de penser que l'asymétrie d'imprécision de l'intervalle de confiance reflète en fait l'asymétrie de la distribution de la pente de la droite de régression, cette distribution ayant une longue queue pour les faibles valeurs (i.e. grands $D\sigma^2$). Nous avons donc testé l'influence d'une asymétrie dans les distribution sur la procédure ABC en considérant un modèle statistique plus simple. Pour cela, nous avons calculé les intervalles de confiance sur la moyenne d'un échantillon aléatoire tiré dans une distribution de Student bivariée avec la distribution suivante

$$Pr(r) = 2\pi r \frac{\Gamma[1+p]}{\pi u \Gamma[p]} (1 + r^2/u)^{-1-p}, \tag{3.6}$$

avec $(p, u) = (1, 1)$ (Fig.3.2).

Cette distribution est asymétrique avec une kurtosis infinie. Même pour des tailles d'échantillon très grandes (e.g. 5000 points, Tableau 3.2), la procédure ABC donne une sous-estimation de la borne supérieure qui résulte en un intervalle de confiance trop étroit.

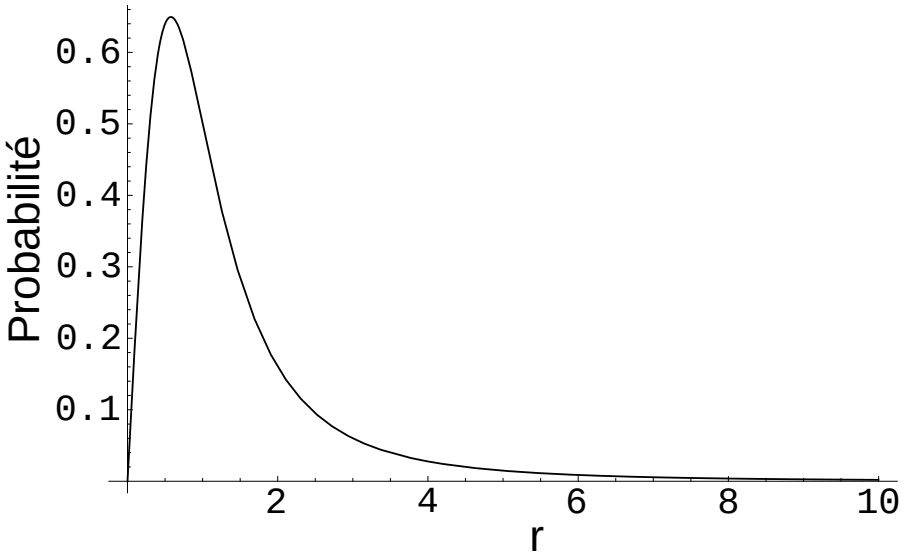

FIG. 3.2. Représentation de la distribution de Student bivariée (éq.3.6) utilisée pour tester la procédure ABC bootstrap

3.2 Influence de l'échelle d'échantillonnage et de la taille de l'habitat

Des simulations antérieures considérant des locus bi-alléliques (Rousset, 2000) ont suggéré que la méthode d'estimation de $D\sigma^2$ est optimale si l'on échantillonne $(100D\sigma^2)$ individus sur une surface d'environ $(10\sigma \times 10\sigma)$. On peut noter que si $D\sigma^2$ est supérieur à une certaine valeur, disons 5, il sera difficile en pratique d'échantillonner et de génotyper autant d'individus ($100D\sigma^2 = 500$ individus). Aussi, puisque le nombre d'individus à échantillonner est nécessairement limité, la méthode considérée ici sera moins efficace si $D\sigma^2$ est grand.

En pratique, les biologistes échantillonnent un relativement grand nombre d'individus (disons 100) sur une surface supérieure ou inférieure à la surface idéale de $(10\sigma \times 10\sigma)$. Afin de tester l'effet d'un échantillonnage à une échelle non idéale, nous avons considéré une distribution de dispersion avec $\sigma^2 = 4$ (éq.3.5) et 4 différents schéma d'échantillonnage. Cent individus sont pris : (i) tous les nœuds sur une aire de $(5\sigma \times 5\sigma)$, (ii) tous les deux nœuds sur une aire de $(10\sigma \times 10\sigma)$, (iii) tous les cinq nœuds sur une aire de $(25\sigma \times 25\sigma)$, et (iv) tous les dix nœuds sur une aire de $(50\sigma \times 50\sigma)$. Afin d'éviter de potentiels effets de bord, la taille du réseau est de (200×200) pour les

trois premiers échantillonnages et (500×500) pour les deux derniers. Ainsi l'échantillon ne représente jamais plus de la moitié de la taille du réseau. Il est important de souligner ici que la taille du réseau a peu d'effet sur l'estimation à condition que le réseau soit plus de 10 fois plus grand que la dispersion moyenne (Fig.3.3). En effet, à l'exception du cas où la taille du réseau est particulièrement petite (50×50), le biais et le MSE, représenté sur la figure 3.3 sont proches de ceux obtenus pour un très grand réseau (1000×1000).

L'échelle d'échantillonnage semble avoir un effet limité sur le MSE de l'es-

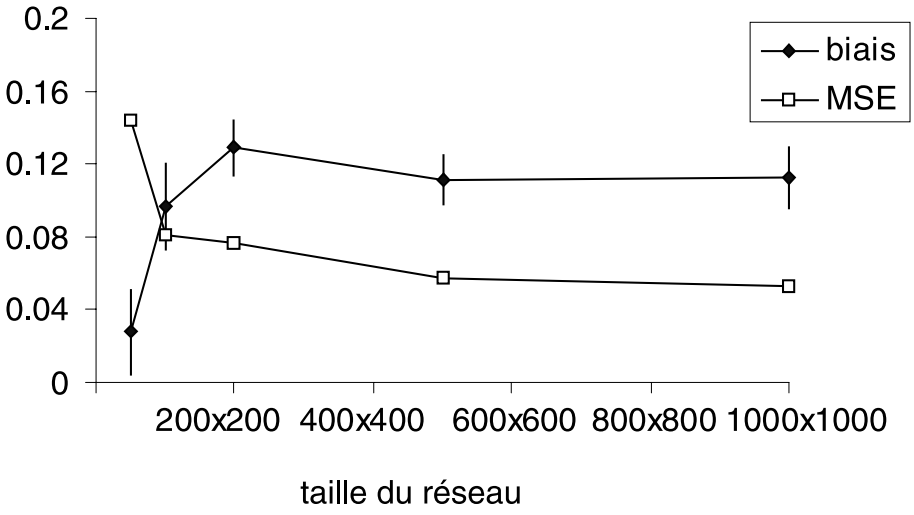

FIG. 3.3. Influence de la taille du réseau sur l'estimation du produit $D\sigma^2$. Les barres verticales représente l'erreur standard sur le biais. 500 répétitions par taille du réseau ont été utilisé pour faire cette figures. Les paramètres utilisés sont décrit dans la section 3.2 et l'échantillon est pris sur une surface de (20×20).

TAB. 3.3. Influence de l'échelle d'échantillonnage sur l'estimation de $1/(4\pi D\sigma^2)$. L'aire d'échantillonnage est exprimée en unité du réseau (voir le texte pour les détails).

	Échelle d'échantillonnage (surface)			
	1	2	5	10
	(10×10)	(20×20)	(25×25)	(50×50)
Biais	0.219	0.130	-0.056	-0.205
(erreur standard)	(0.0077)	(0.0077)	(0.0072)	(0.0064)
MSE	0.106	0.0763	0.0554	0.082
Facteur 2	0.999	0.996	0.967	0.93

timation de $D\sigma^2$ (Tableau 3.3). Quelle que soit l'échelle considérée, le MSE est petit (e.g. entre 5% et 12% dans les cas étudiés). Par contre, l'échelle d'échantillonnage a un effet important sur le biais. Un échantillon pris sur une aire deux fois plus petite que l'aire recommandée (première colonne du tableau 3) entraîne un fort biais positif (22%) et donc une sous-estimation de $D\sigma^2$. Le biais diminue ensuite lorsque la surface échantillonnée augmente et atteint de fortes valeurs négatives (i.e. sur-estimation de $D\sigma^2$) lorsque la surface recommandée est largement dépassée (-21% pour la troisième colonne du Tableau 3.3). L'estimation reste toutefois correcte même pour des échelles d'échantillonnage extrêmes puisque une large majorité des estimations tombent dans l'intervalle de facteur 2 (>93%).

3.3 Influence des processus de mutation

3.3.1 Modèles mutationnels

Pour étudier l'influence des processus de mutation des marqueurs utilisés, nous avons dans un premier temps considéré cinq modèles mutationnels différents (voir section 2.1.1 pour les définitions des différents modèles mutationnels) : (i) le modèle à nombre d'allèles infini (IAM), (ii) le modèle à K allèles (KAM), avec un choix arbitraire de $K = 10$, (iii) le modèle de mutation par pas (SMM), (iv) le modèle de mutation par pas généralisé (GSM), avec une variance de la loi géométrique de 0.36, correspondant à la valeur estimée sur le grand jeu de données de mutation microsatellitaires de Dib *et al.* (1996) chez l'homme, et (v) un GSM avec des contraintes sur les tailles des allèles (10 ou 20 allèles possibles). Pour plus de détails sur les modèles mutationnels et les caractéristiques mutationnelles des microsatellites, le lecteur pourra se référer à l'annexe B-2.

Les simulations ont été faites en considérant un échantillon de 100 individus pour 13 locus évoluant sur un tore de (100×100). La distribution de dispersion est celle considérées précédemment avec $\sigma^2 = 4$ (paramètres données dans l'éq.3.5). Les 100 individus sont pris sur une surface de $(10\sigma \times 10\sigma = 20 \times 20)$ afin de se rapprocher d'un échantillonnage expérimental classique.

Dans un premier temps, le taux de mutation des marqueurs a été fixé à $5 \cdot 10^{-4}$ pour tous les locus. Nos résultats montrent que la nature du modèle mutationnel a très peu d'influence sur l'estimation de $D\sigma^2$ (Tableau 3.4).

TAB. 3.4. Influence des modèles mutationnels sur l'estimation de $1/(4\pi D\sigma^2)$ à taux de mutation constant.

	Modèle mutationnel à taux de mutation constant $5 \cdot 10^{-4}$					
	IAM	KAM ($K = 10$)	SMM	GSM	GSM borné ($K = 10$)	GSM borné ($K = 20$)
Diversité génétique	0.787	0.711	0.703	0.772	0.679	0.720
Biais	0.109	0.0919	0.0917	0.104	0.0997	0.128
(erreur standard)	(0.0067)	(0.0088)	(0.0093)	(0.00863)	(0.0101)	(0.009)
MSE	0.057	0.0853	0.0953	0.0852	0.112	0.098
Facteur 2	0.998	0.982	0.975	0.987	0.976	0.984

Quel que soit le modèle mutationnel considéré, le biais est positif et d'environ 10%. Même si la précision de la méthode est maximale sous IAM (MSE de 6%) et minimale sous GSM avec des contraintes fortes sur les tailles alléliques ($K = 10$, MSE de 11%), ces différences sont mineures. Pour tous les modèles mutationnels considérés plus de 97% des estimations tombent dans l'intervalle de facteur 2.

TAB. 3.5. Influence des modèles mutationnels sur l'estimation de $1/(4\pi D\sigma^2)$ à diversité génétique constante.

	Modèle mutationnel à diversité génétique constante 0.68					
	IAM	KAM ($K = 10$)	SMM	GSM	GSM borné ($K = 10$)	GSM borné ($K = 20$)
Taux de mutation	0.0001	0.000218	0.000342	0.00012	0.0005	0.0002
Biais	0.111	0.104	0.118	0.121	0.0997	0.108
(erreur standard)	(0.01)	(0.01)	(0.015)	(0.012)	(0.0101)	(0.01)
MSE	0.119	0.109	0.119	0.159	0.112	0.121
Facteur 2	0.96	0.97	0.96	0.938	0.976	0.962

Pour un taux de mutation donné, la diversité génétique $(1 - Q_0)$, correspondant à la proportion d'individus hétérozygotes dans la population, varie en fonction du modèle mutationnel considéré. Puisque la diversité génétique peut avoir un effet important sur l'estimation, nous avons considéré tous les modèles mutationnels précédents à diversité génétique constante en ajustant

les taux de mutations selon les calculs de Rousset (1996) (Tableau 3.5). Les conclusions sont très similaires à celles pour un taux de mutation constant (Tableau 3.4). Pour une diversité génétique donnée, le biais et le MSE sur les estimations de $D\sigma^2$ montrent peu de variation en fonction du modèle mutationnel. On peut cependant remarquer qu'en considérant une diversité génétique constante, le MSE montre moins de variation entre différents modèles qu 'en considérant un taux de mutation constant.

3.3.2 Taux de mutation (ou diversité génétique)

L'influence du taux de mutation, ou de la diversité génétique, a été étudiée en considérant un des modèles mutationnels les plus réalistes par rapport aux processus mutationnels des microsatellites, le GSM (Estoup & Cornuet, 1999, voir section 2.1.1). L'interprétation des résultats est faite en terme de diversité génétique car c'est une mesure facile à obtenir en pratique alors que l'information sur le taux de mutation des marqueurs utilisés est rarement accessible. De plus, nos résultats précédents (Tableaux 3.4 et 3.5) ont montré que la diversité génétique a plus d'influence sur les performances de la méthode que le taux de mutation. Tout les paramètres de simulation sont ceux utilisés dans la section précédente traitant de l'influence des modèles mutationnels.

Nos résultats montrent que la diversité génétique a un effet important sur le biais et le MSE de l'estimation de $D\sigma^2$ (Tableau 3.6). Le MSE est plus fortement influencé par la diversité génétique que le biais. Pour des diversités génétiques de l'ordre de 0.5, le biais observé est positif et inférieur à 10% mais le MSE est fort (i.e. supérieur à 20%). Cependant, pour cette diversité génétique plus faible que la diversité génétique souvent observée aux locus microsatellites, 84% des estimations sont dans l'intervalle de facteur 2.

Pour des diversités génétiques fortes (i.e. 0.85), le biais devient fortement négatif et le MSE augmente rapidement avec la diversité génétique. Ce résultat traduit le fait que pour des diversité génétiques fortes, le "biais paramétrique", qui est négatif (Rousset, 1997), devient plus important que le "biais d'échantillonnage" et le biais global devient donc négatif.

Il est souvent considéré que les variations de taux de mutation entre locus peuvent influencer la précision des estimations en génétique des populations (Takezaki & Nei, 1996; Gonser *et al.*, 2000). Afin d'évaluer un tel phénomène, nous avons considéré 13 locus évoluant sous GSM avec un taux de mutation

TAB. 3.6. Influence du taux de mutation sur l'estimation de $1/(4\pi D\sigma^2)$. Le modèle mutationnel est le GSM.

	Taux de mutation				
	0.00005	0.00012	0.0005	0.005	0.05
Diversité génétique	0.56	0.68	0.77	0.82	0.85
Biais	0.0972	0.121	0.104	0.00946	-0.390
(erreur standard)	(0.016)	(0.012)	(0.0086)	(0.0062)	(0.0055)
MSE	0.268	0.159	0.0852	0.0380	0.182
Facteur 2	0.84	0.94	0.99	0.99	0.761

TAB. 3.7. Influence d'un taux de mutation variable sur l'estimation de $1/(4\pi D\sigma^2)$. Le modèle mutationnel est le GSM. * La variabilité intra-locus correspond à une augmentation du taux de mutation de 0.1% et de 1% par répétition, pour la variabilité intra-locus faible et forte respectivement.

		Variabilité		
	constant	inter-locus	intra-locus	
			faible*	forte*
Diversité génétique	0.77	0.77	0.77	0.77
Biais	0.104	0.114	0.0965	0.111
(erreur standard)	(0.0086)	(0.0096)	(0.00846)	(0.0081)
MSE	0.0852	0.105	0.0808	0.0778
Facteur 2	0.99	0.98	0.99	0.99

tiré pour chaque locus dans une distribution Gamma de paramètres égaux à $(2, 2.5 \cdot 10^{-4})$ dont la moyenne est $5 \cdot 10^{-4}$ et les quantiles à 2.5 et 97.5% sont respectivement $6 \cdot 10^{-5}$ et $1.4 \cdot 10^{-3}$ [la densité de la loi Gamma à deux paramètres (a,b) est donnée par $\Pr(x) = \frac{b^{-a}}{\Gamma(a)} \exp(-x/b) x^{a-1}$ pour $x \geq 0$ avec $\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x) dx$]. Tous les autres paramètres de simulation sont semblables à ceux utilisés précédemment. Nos simulations montrent qu'un taux de mutation variable a peu d'effet sur l'estimation du produit $D\sigma^2$ (Tableau 3.7 2ème colonne). Le biais et le MSE sont inférieurs à 12% ce qui diffère peu du cas avec un taux de mutation fixe ; d'autre part, plus de 98% des estimations sont dans l'intervalle de facteur 2. Enfin nos simulations montrent qu'un taux de mutation variable en fonction de la longueur de l'allèle a également peu d'influence sur l'estimation (Tableau 3.7). Pour cela, nous avons modélisé une augmentation linéaire du taux de mutation en fonction du nombre de répétitions de chaque allèle. Nous avons considéré une

augmentation faible et une forte, de 0.1% et 1% par répétition, proches des valeurs trouvées aux locus microsatellites étudiés par Brohede *et al.* (2002). Quelle que soit la variation, forte ou faible, le biais et le MSE sont proches de ceux obtenus avec un taux de mutation constant pour tous les locus et pour tous les allèles (environ 10%).

3.3.3 Effet d'une statistique prenant en compte les tailles des allèles

Comme des mutations par pas ont lieu aux locus microsatellites, il est tentant de prendre en compte ce processus mutationnel par le biais d'une statistique fondée sur les tailles alléliques. Afin de tester l'effet d'une telle statistique, nous avons considéré le paramètre b_r, analogue au paramètre a_r mais défini en terme de différence de tailles alléliques et non de probabilité d'identité. On a alors

$$b_r \equiv \frac{\mathrm{SD}_r - \mathrm{SD}_w}{\mathrm{SD}_w}, \tag{3.7}$$

où SD est l'espérance du carré de la différence des tailles alléliques de deux gènes pris dans un même individu (SD_w) ou séparés par une distance géographique r (SD_r). L'estimateur que nous avons considéré est de la forme

$$\hat{b} \equiv \frac{\mathrm{SSD}_{b(\xi)P}}{\sum_{k=1}^{P} \mathrm{SSD}_{w(k)}} - \frac{1}{2}, \tag{3.8}$$

où $\mathrm{SSD}_{b[etween](\xi)} \equiv \sum_{i,j,u}(S_{i.} - S_{..})^2$ mesure la divergence entre deux gènes pris dans deux individus différents et $\mathrm{SSD}_{w[ithin](\xi)} \equiv \sum_{i,j,u}(X_{ij} - X_{i.})^2$ mesure la divergence entre deux gènes pris au sein d'un individu ($X_{ij:u}$ est une variable représentant la taille de l'allèle du gène i de l'individu j, exprimée en fonction du nombre de répétition).

Le comportement d'une telle statistique a été étudié pour les modèles mutationnels SMM et GSM avec un taux de mutation constant de $5 \cdot 10^{-4}$. Tous les autres paramètres de simulation sont ceux utilisés précédemment. Les résultats de nos simulations, présentés dans le Tableau 3.8, montrent que la méthode d'estimation de $D\sigma^2$ est nettement moins performante lorsque l'on utilise b_r au lieu de a_r. Aussi bien sous SMM que sous GSM, l'augmentation du MSE est spectaculaire avec b_r. Le MSE est d'environ 10% quand on considère a_r et passe à plus de 100% avec la statistique b_r. On peut noter

que le biais est moins affecté par une telle statistique et reste inférieur à 20%.

TAB. 3.8. Estimation de $1/(4\pi D\sigma^2)$ utilisant le paramètre b_r prenant en compte les tailles alléliques.

	Modèle mutationnel			
	SMM	SMM	GSM	GSM
Paramètre estimé	a_r	b_r	a_r	b_r
Biais	0.0917	0.128	0.104	0.19
(erreur standard)	(0.0093)	(0.036)	(0.0086)	(0.034)
MSE	0.0953	1.13	0.0852	1.25
Facteur 2	0.98	0.518	0.99	0.497

3.4 Influence d'hétérogénéités spatiales et temporelles

On peut imaginer une infinité de scénarios intégrant des hétérogénéités spatiale et temporelles, nous avons choisi de focaliser notre étude sur des scénarios démographiques souvent rencontrés en biologie de la conservation ou lors de l'étude de bioinvasions. Dans ce contexte, nous avons évalué les effets sur l'estimation de $D\sigma^2$ (i) d'un changement au cours du temps de la dispersion, (ii) d'une réduction ou d'une augmentation de la densité au cours du temps, (iii) d'une expansion spatiale à densité constante, et (iv) d'un échantillonnage dans une zone de forte densité. Pour cette section, je ne présenterai pas les détails des simulations réalisées mais n'en présenterai que les principes et les résultats majeurs en découlant. Pour plus de détails, le lecteur pourra se référer à l'annexe B-3 de ce document.

3.4.1 Variation temporelle de la dispersion

Nous avons étudié l'effet d'une diminution des capacités de dispersion au cours du temps au travers d'une diminution du paramètre σ^2, le carré moyen de la distance de dispersion parent-descendant. Pour cela nous avons choisi deux distributions de dispersion avec des σ^2 très différents (e.g. 4 et 100), les autres paramètres des distributions étant constants. Nous avons

considéré que du présent jusqu'au moment du changement dans le passé (G_c), la distribution est celle avec $\sigma^2 = 4$. De G_c à TMRCA, la distribution de dispersion est celle avec $\sigma^2 = 100$. Le changement de dispersion intervient à trois moments différents dans le passé suivant les simulations ($G_c = 10$, 20, 100) et à un temps infini comme simulation témoin (aucun changement).

TAB. 3.9. Effet d'une diminution de la dispersion dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de dispersion.

	G_c			
	∞	100	20	10
Biais	0.444	0.0923	-0.0795	-0.234
(erreur standard)	(0.0062)	(0.0081)	(0.0076)	(0.0074)
MSE	0.228	0.0743	0.0642	0.109
Facteur 2	0.99	0.99	0.97	0.88

Nos simulations montrent que le biais dû à une réduction de la dispersion dans le temps est négatif (Tableau 3.9), et correspond donc à une surestimation du $D\sigma^2$ actuel. Ceci est en accord avec une transition d'une forte valeur de $D\sigma^2$ ($D\sigma^2 = 100$) pendant les génération passées (avant G_c) vers une valeur plus faible de $D\sigma^2$ ($D\sigma^2 = 4$) pour les générations récentes (i.e ; après G_c).Ainsi, la méthode d'estimation de $D\sigma^2$ a donc une certaine mémoire des caractéristiques de dispersion passées. Cependant, cette mémoire est faible puisqu'une réduction 100 générations dans le passé introduit seulement un biais très faible compensé dans les simulations par un "biais paramétrique et d'échantillonnage" (cf. première colonne du tableau 3.9). De plus, même pour une réduction récente ($G_c = 10$), le biais est inférieur à 25%, une valeur relativement faible par rapport à la forte amplitude de la variation de σ^2 modélisée (facteur 25). L'erreur standard sur l'estimation est faible quelque soit le moment dans le passé où a lieu la variation de dispersion et, pour des changement ayant lieu plus de 20 générations dans le passé, plus de 95% des estimations correspondent, à un facteur 2 près, à la valeur actuelle de $D\sigma^2$. Nos simulations montrent donc que globalement la précision de l'estimation du $D\sigma^2$ actuel est assez robuste aux changements temporels de dispersion.

3.4.2 Variation de la densité dans le temps : goulet d'étranglement et explosion démographique

Le deuxième type de fluctuations que nous avons étudié est une diminution et une augmentation de la densité dans le temps. Pour cela nous avons considéré quatre modèles de réseau avec chacun un nombre d'individu par dèmes différents (e.g. 1, 10, 1/9, 1/100). Les cas avec moins d'un individu par dème ont été modélisés en considérant qu'une certaine proportion des dèmes sont vides (e.g. une densité de 1/9 est obtenue avec 8/9 des dèmes vides). Les différentes densités utilisées sont résumées dans le tableau 3.10. Nous avons modélisé un goulet d'étranglement (i.e. diminution de la densité dans le temps) fort, d'un facteur 90, avec une densité passant de 10 à 1/9 d'individus par dème, et un faible, d'un facteur 10, avec une densité passant de 10 à 1 individu par dème. Nous avons aussi considéré une explosion démographique (i.e. augmentation de la densité dans le temps) forte, d'un facteur 100, avec une densité passant de 1/100 à 1 individu par dème, et un faible, d'un facteur 9, avec une densité passant de 1/9 à 1 individu par dème. Pour toutes ces situations, les changements se font à trois moments dans le passé, $G_c = 10$, 20, 100 et infini comme témoin.

Pour ce qui est du goulet d'étranglement, le biais négatif observé dans le tableau 3.11, correspondant à une sur-estimation de $D\sigma^2$, reflète les fortes densités passées. Pour une réduction d'un facteur 10, la méthode est assez robuste quand le changement a lieu à 20 ou plus de 20 générations dans le passé. Le biais et le MSE sont faibles (i.e. moins de 10%) et 99% des estimations correspondent bien à la valeur du paramètres $D\sigma^2$ actuel à un facteur

TAB. 3.10. modèles utilisés pour évaluer l'effet de changements de densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité. TMRCA correspond au temps de l'ancêtre commun le plus récent de l'échantillon.

Changement démographique		Densité (Nombre d'individus par nœud du réseau)		
		du présent à G_c	de G_c à TMRCA	Facteur
Goulet d'étranglement	faible	1	10	10
	fort	1/9	10	90
Explosion démographique	faible	1	1/9	9
	fort	1	1/100	100

2 près. Pour un changement plus récent (e.g. $G_c = 10$), le biais est nettement plus important. Toutefois le MSE reste petit et 92% des estimations tombent dans l'intervalle de facteur 2 par rapport à la valeur actuelle de $D\sigma^2$. L'effet

TAB. 3.11. Effet d'une diminution de la densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité.

		G_c			
Intensité		∞	100	20	10
	Biais	0.444	0.099	-0.063	-0.22
	(erreur standard)	(0.0062)	(0.0070)	(0.0064)	(0.0061)
Faible	MSE	0.228	0.0588	0.0449	0.0868
	Facteur 2	0.99	0.99	0.99	0.92
	Biais	-0.014	-0.074	-0.33	-0.53
	(erreur standard)	(0.0042)	(0.0027)	(0.0017)	(0.0012)
Forte	MSE	0.0175	0.0128	0.115	0.278
	Facteur 2	1	1	1	0.238

d'une réduction forte de densité (i.e. d'un facteur 90) est beaucoup plus marquée. Pour un changement récent (i.e. moins de 10 générations dans le passé), le biais atteint 50% et seulement 24% des estimations correspondent à la valeur actuelle de $D\sigma^2$ à un facteur 2 près. Pour $G_c = 100$, le biais et le MSE retrouvent un niveau proche de la simulation témoin. On peut noter que, même pour un fort goulet d'étranglement, si l'on considère un changement ayant eu lieu à 20 ou plus de 20 générations dans le passé, plus de 99% des estimations sont dans l'intervalle de facteur 2 par rapport à la valeur actuelle de $D\sigma^2$. Ainsi, même pour des goulets d'étranglement forts et relativement récents, la méthode des moments montre une bonne robustesse.

Dans le cas d'une explosion démographique, le biais positif observé dans le tableau 3.12, correspondant à une sous-estimation du $D\sigma^2$ actuel, reflète les faibles densités passées. Pour une faible augmentation de densité (facteur 9), le biais et le MSE sont forts, même pour une explosion démographique ancienne (e.g. 100 générations). La proportion d'estimations correspondant au $D\sigma^2$ actuel à un facteur 2 près reste faible (moins de 50%) même pour $G_c = 100$. Cet effet augmente considérablement avec l'intensité de la variation. Pour une explosion démographique d'un facteur 100 et pour $G_c = 10$, le biais atteint 390% et aucune des estimations ne tombent dans l'intervalle de facteur 2 (Tableau 3.12). Ainsi, en dépit du fait que le biais et le MSE diminuent lorsque G_c augmente, l'estimation reste incertaine dans les gammes

de temps étudiées que ce soit pour une variation forte ou faible. Ces résultats contrastent fortement avec les résultats obtenus pour les goulets d'étranglement et pour une diminution de la dispersion dans le temps.

TAB. 3.12. Effet d'une augmentation de la densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité.

		G_c			
Intensité		∞	100	20	10
	Biais	0.44	0.32	0.69	1.4
	(erreur standard)	(0.0062)	(0.040)	(0.043)	(0.046)
Faible	MSE	0.228	1.72	2.33	4.07
	Facteur 2	0.99	0.45	0.38	0.24
	Biais	0.43	0.65	2.2	3.9
	(erreur standard)	(0.0064)	(0.0094)	(0.015)	(0.019)
Forte	MSE	0.228	0.508	5.27	15.8
	Facteur 2	0.99	0.89	0.003	0

3.4.3 Expansion spatiale de la population à densité constante

Le quatrième type de situation étudié est une expansion spatiale de la population à densité constante (Fig.3.4). La population introduite dans le nouvel habitat vide est composé d'individus ayant évolué dans une population source à l'équilibre sous certaines caractéristiques démographiques (i.e. densité et distribution de dispersion). La population introduite s'étend en quelques générations (en 2 générations dans nos simulations) sur tout le nouveau territoire avec les même caractéristiques démographiques que celles de la population source. L'échantillon d'individus est pris sur le nouvel habitat à une distance de 50 dèmes de la zone où ils ont été introduits. Les simulations ont été faites pour une expansion à 10, 20 et 100 générations dans le passé, et à un temps infini pour le témoin.

Toutes les statistiques (Biais, MSE et proportion d'estimation dans l'intervalle de facteur 2) calculées sur ces simulations indiquent que l'estimation du $D\sigma^2$ actuel est bonne pour une expansion spatiale ayant eu lieu plus de 20 générations dans le passé (Tableau 3.13). Pour $G_c = 10$, le biais est négatif et de 8%, ce qui correspond à une faible sur-estimation du $D\sigma^2$ actuel. Le MSE

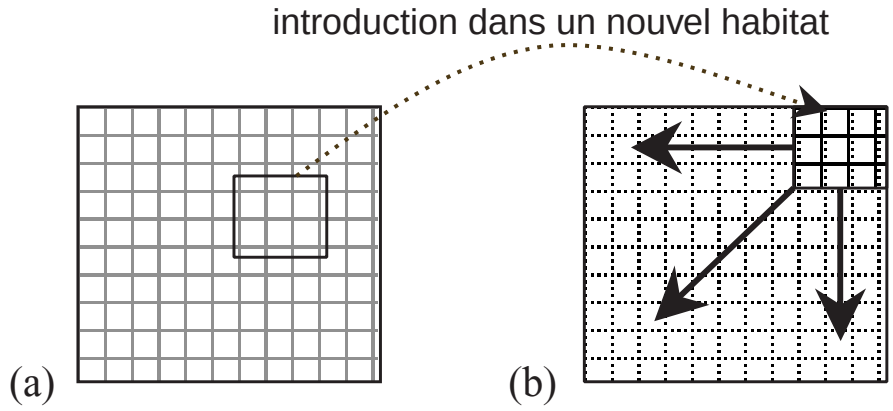

FIG. 3.4. Schéma d'une expansion démographique à densité constante telles que nous l'avons modélisé dans cette étude. La grille (a) correspond à une population source dont est issu un échantillon d'individus (petit cadre noir) introduits dans un nouvel habitat (flèche pointillée). La grille (b) correspond au nouvel habitat vide dans lequel l'échantillon introduit va s'étendre en quelques générations (flèches noires). Dans nos simulations, l'habitat est représenté sur un tore et non sur un habitat fini et plan comme sur cette figure.

TAB. 3.13. Effet d'une expansion spatiale à densité constante sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu l'expansion.

	G_c			
	∞	100	20	10
Biais	0.43	0.39	0.13	-0.0824
(erreur standard)	(0.0076)	(0.013)	(0.011)	(0.010)
MSE	0.243	0.23	0.08	0.0581
Facteur 2	0.99	0.98	0.99	0.97

est faible (10%) et 98% des estimations tombent dans l'intervalle de facteur 2. Ainsi, nos simulations montrent qu'une expansion spatiale, telle qu'elle est modélisée ici, a une influence limitée sur l'estimation de $D\sigma^2$. La méthode des moments est donc précise même pour des expansions récentes.

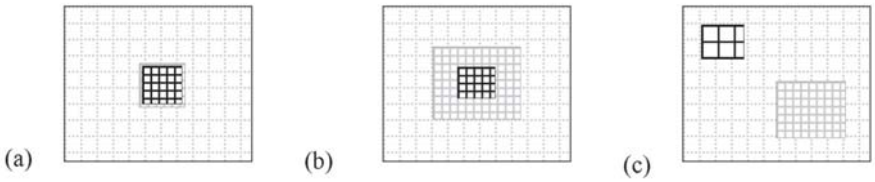

FIG. 3.5. Schéma d'hétérogénéités spatiales telles que nous les avons modélisées dans cette étude. (a) Petite zone de forte densité (grille gris foncé) correspondant strictement à la zone échantillonnée (grille noire) sur un réseau en deux dimensions de densité plus faible (grille gris claire). (b) Grande zone de forte densité (grille gris foncé) incluant la zone échantillonnée (grille noire) sur un réseau en deux dimensions de densité plus faible (grille gris claire). (c) Grande zone de forte densité (grille gris foncé) située hors de la zone d'échantillonnage (grille noire) sur un réseau en deux dimensions de densité plus faible (grille gris claire). Dans nos simulations, l'habitat est représenté sur un tore et non sur un habitat fini et plan comme sur cette figure.

3.4.4 Hétérogénéité spatiale de la densité : échantillonnage sur une zone de forte densité

La dernière situation que nous avons choisie d'étudier reflète le fait que les biologistes collectent généralement leurs échantillons sur des zones où l'espèce étudiée est facile à collecter, c'est à dire sur des zones de fortes densités. Pour cela nous avons considéré un modèle en réseau avec une densité homogène sauf sur une petite zone sur laquelle la densité est dix fois plus forte que sur le reste du réseau. Deux cas ont été considérés : (i) une petite zone de forte densité de (20×20) dèmes correspondant exactement à la surface sur laquelle on a échantillonné les individus, et (ii) une zone plus large de (40×40) dèmes qui inclut la zone de (20×20) dèmes de l'échantillon (Fig.3.5a et 3.5b). Nous avons alors évalué si l'estimation de la densité correspondait plutôt à la densité sur la zone échantillonnée ou si cette estimation était largement influencée par la densité autour de la zone échantillonnée. Dans le premier cas, la valeur attendue correspond à la valeur locale (i.e. sur la zone échantillonnée) et dans le deuxième cas à la valeur avoisinante (i.e. sur les zones autour de la zone échantillonnée). Enfin, une simulation complémentaire a été réalisée afin d'évaluer l'influence d'une zone de forte densité située hors de la zone échantillonnée (i.e. à une distance de 50 nœuds de l'échantillon) (Fig 3.5c).

Nos simulations montrent que l'estimation de $D\sigma^2$ n'est pas robuste

lorsque la zone de forte densité est petite et correspond strictement à la surface échantillonnée (Tableau 3.14). Les valeurs du biais et du MSE indiquent que, dans ce cas, la zone de faible densité autour de l'échantillon influence fortement l'estimation de $D\sigma^2$, qui devient alors une mauvaise mesure aussi bien de la densité locale (i.e. sur la surface échantillonnée) que de la densité avoisinante (i.e. autour de la zone d'échantillonnage). On peut toutefois remarquer que la proportion d'estimation tombant dans l'intervalle de facteur 2, même si elle est faible, est plus grande quand on se réfère à la densité locale plutôt qu'avoisinante comme valeur attendue pour $D\sigma^2$ (0.018 pour locale vs 0.001 pour avoisinante). Ceci suggère que la méthode a tendance à mesurer la densité locale plutôt qu'avoisinante. Cette tendance devient claire quand on regarde les résultats pour une zone de forte densité plus grande (Tableau 3.14). Dans ce cas, le biais et le MSE sont beaucoup plus faibles quand on se réfère à la densité locale plutôt qu'avoisinante pour la valeur attendue de $D\sigma^2$. Environ 90% des estimations correspondent, à un facteur 2 près, à la valeur locale de $D\sigma^2$, alors qu'aucune estimation ne tombe dans l'intervalle de facteur 2 par rapport à la valeur avoisinante de $D\sigma^2$. Nos simulations montrent donc que la méthode estime les paramètres démographiques

TAB. 3.14. Effet d'hétérogénéités spatiales de la densité sur l'estimation de $1/(4\pi D\sigma^2)$. Colonne 'Densité locale' : la densité attendue est la densité locale (i.e. sur la zone d'échantillonnage). Colonne 'Densité avoisinante' : la densité attendue est la densité avoisinante (i.e. autour de la zone échantillonnée). Les colonnes 'contrôle' correspondent à un réseau homogène avec une densité correspondant à la densité locale ou avoisinante pour les colonnes respectives de densité locale et avoisinante.

Hétérogénéité spatiale		Densité locale		Densité avoisinante	
		Estimation	Contrôle	Estimation	Contrôle
Petite zone	Biais	2.1	0.45	-0.69	0.43
de	(erreur standard)	(0.017)	(0.025)	(0.0017)	(0.0076)
forte densité	MSE	4.76	0.83	0.477	0.243
	Facteur 2	0.018	0.65	0.001	0.99
Grande zone	Biais	0.39	0.45	-0.86	0.43
de	(erreur standard)	(0.013)	(0.025)	(0.0013)	(0.0076)
forte densité	MSE	0.330	0.83	0.743	0.243
	Facteur 2	0.9	0.65	0	0.99
Grande zone de	Biais	0.45	0.43	13.5	0.45
forte densité	(erreur standard)	(0.0075)	(0.0076)	(0.075)	(0.025)
hors zone	MSE	0.256	0.243	187	0.83
d'échantillonnage	Facteur 2	0.99	0.99	0	0.65

locaux et qu'elle est robuste pour ces mesures quand la densité est relativement homogène autour de la zone échantillonnée (e.g. sur une surface environ égale à quatre fois la surface échantillonnée). Le troisième cas d'une zone de forte densité située hors de la zone d'échantillonnage confirme ce résultat puisqu'une zone de forte densité située à 50 nœuds de la zone échantillonnée n'a quasiment aucune influence sur l'estimation du $D\sigma^2$ local.

3.5 Discussion

3.5.1 Processus mutationnels

Une première conclusion générale de notre étude est que les modèles de mutation des marqueurs génétiques ont peu d'influence sur les performances de la méthode d'estimation du produit $D\sigma^2$ à partir de génotypes individuels. Ainsi, l'homoplasie de taille typiquement produite par des processus de mutation par pas (modèles SMM et GSM), caractéristique des marqueurs microsatellites, n'est pas un facteur préjudiciable pour la méthode étudiée. De plus, nos résultats sur les contraintes de tailles alléliques, aussi observées aux locus microsatellites et connues pour augmenter les phénomènes d'homoplasie, montrent que même des contraintes extrêmement fortes (e.g. $K = 10$) ont peu d'influence sur l'estimation de $D\sigma^2$.

Une seconde conclusion majeure de cette étude est que la diversité génétique (celle-ci étant largement influencée par le taux de mutation), a une influence importante sur l'estimation de $D\sigma^2$. Ceci est en accord avec des études antérieures démontrant que le taux de mutation est un facteur plus important que les processus mutationnels pour l'estimation de paramètres démographiques par F-statistiques (Rousset, 2001a; Estoup *et al.*, 2002). Il parait intéressant de noter que la diversité génétique typiquement observée aux locus microsatellites se situe généralement entre 0.5 et 0.8 (revue dans Estoup & Angers, 1998), ce qui correspond aux niveaux de diversité maximisant les performances de l'estimation de $D\sigma^2$. Enfin, les effets potentiels sur l'estimation de taux de mutation variables entre locus et entre allèles semblent faibles. Les marqueurs microsatellites sont donc plus appropriés pour l'estimation de $D\sigma^2$ que d'autres marqueurs moins polymorphes tels les allozymes.

L'importance du niveau de variabilité des locus utilisés pour l'estimation de paramètres démographiques a déjà été illustrée par plusieurs études expé-

rimentales et théoriques. Robertson & Hill (1984) ont, par exemple, montré que la précision de l'estimation du déficit en hétérozygotes (F_{IS}) augmente avec le niveau de variabilité des marqueurs. Goudet *et al.* (1996) ont aussi montré que la puissance des tests statistiques de différenciation augmente avec le nombre d'allèles. En pratique, même si une information précise sur les taux de mutation est difficile à obtenir, il est facile de calculer un indice de diversité génétique des marqueurs utilisés. Cette information pourra être utilisée pour en déduire un niveau de précision attendue de l'estimation de $D\sigma^2$. Nos simulations ont aussi montré qu'il est préférable d'éviter des locus avec une diversité génétique trop forte (i.e. >0.85) car ces locus biaiseront fortement l'estimation de $D\sigma^2$.

De nombreuses études ont mis en avant que les F_{ST} ne prennent pas en compte l'information présente dans les différences de taille des allèles microsatellites. Nos simulations montrent que l'utilisation d'une statistique prenant en compte les tailles alléliques (b_r) augmente au minimum d'un facteur 10 le MSE de l'estimation de $D\sigma^2$ par rapport à l'utilisation d'une statistique fondées sur les probabilités d'identité. Ces résultats confortent ceux de Gaggiotti *et al.* (1999) et Balloux & Goudet (2002) qui ont montré que, dans de nombreux cas, les F-statistiques fondées uniquement sur les fréquences alléliques sont plus fiables que des statistiques analogues spécifiquement adaptées aux mutations par pas des microsatellites. Takezaki & Nei (1996) ont également montré que, même dans le cas de locus évoluant sous un modèle de mutation par pas strict (SMM), les distances génétiques prenant en compte les tailles alléliques sont moins performantes pour l'inférence phylogénétique que celles fondées sur l'identité par état, et ce d'autant plus que la divergence entre populations est faible.

3.5.2 Échelles d'échantillonnage et intervalles de confiance

Il est attendu que les effets des processus mutationnels et des taux de mutation élevés sur l'estimation de $D\sigma^2$ soient plus importants à large échelle géographique (Rousset, 1997, voir aussi section 2.2 de ce document). En accord avec ces prédictions, nos résultats montrent qu'échantillonner à trop grande échelle entraîne une sur-estimation de $D\sigma^2$. Ainsi échantillonner sur des grandes distances diminue la probabilité de détecter un patron d'isolement par la distance. Au contraire, échantillonner sur une zone géographique trop petite entraîne une sous-estimation de $D\sigma^2$. Une explication possible est que la relation linéaire entre a_r et le logarithme de la distance est moins fiable

à petites distances (i.e. pour $r < \sigma$, Rousset, 1997). Cependant, l'utilisation d'échantillons non appropriés au cas biologique étudié (i.e. plus petit ou plus grand que la surface recommandée de $(10\sigma \times 10\sigma)$) donnent des estimations relativement robustes puisque dans la plupart des cas ces estimations correspondent, à un facteur 2 près, à la valeur attendue du paramètre $D\sigma^2$.

Nos résultats sur les intervalles de confiance par bootstrap montrent que, pour des échantillons de taille classique (i.e. 100 individus et 10 locus, e.g. Sumner *et al.*, 2001), la technique d'ABC bootstrap sous-estime la borne supérieure des intervalles de confiance sur $D\sigma^2$. La construction d'intervalles de confiance par des techniques de bootstrap est un problème récurrent, et ce spécialement dans des contextes comme celui-ci où le calcul d'un grand nombre de réplicats est très long et les distributions fortement asymétriques (DiCiccio & Efron, 1996). Quoiqu'il en soit la procédure d'ABC bootstrap testée demeure utile pour donner une idée de l'incertitude sur les estimations de $D\sigma^2$. Aussi, nous avons implémenté cette procédure dans la nouvelle version du logiciel GENEPOP (Raymond & Rousset, 1995).

3.5.3 Hétérogénéité spatiales et temporelles des paramètres démographiques

Dans la limite des situations étudiées ici, et à l'exception d'une explosion démographique, nos résultats montrent que des fluctuations temporelles et spatiales des paramètres démographiques, si elles ne sont pas trop importantes, ni trop récentes, ont une influence limitée sur l'estimation de $D\sigma^2$ actuel et local. Il est important de noter que nous parlons ici de changements sur une échelle de temps de l'ordre de quelques dizaines de générations. Si cette échelle peut paraître extrêmement récente en génétique des populations, elle reflète pour de nombreuses espèces (e.g. espèces menacées et bioinvasions) les changements démographiques liés à l'activité humaine. Soulignons également que les nombres de générations définissant les moments dans le passé auxquels ont eu lieu les changements doivent être considérés uniquement comme des indices approximatifs de la durée de l'effet des changements démographiques étudiés. En effet, la persistance dans le temps des effets des fluctuations démographiques dépend fortement de nombreuses caractéristiques des modèles démographiques (e.g. les valeurs de σ et de D) et des situations de déséquilibre. Il parait donc préférable de considérer des tendances globales que des nombres précis de générations pour chaque situation. Ces tendances sont résumées dans le tableau 3.15.

TAB. 3.15. Résumé qualitatif des effets d'hétérogénéités spatiales et temporelles des paramètres démographiques sur l'estimation de $1/(4\pi D\sigma^2)$. Intensité faible : biais relatif moyen inférieur à 50% ; Intensité forte : biais relatif moyen supérieur à 100%. Bon : proportion d'estimations tombant dans l'intervalle de facteur 2 supérieure à 80% ; Mauvais : proportion d'estimations tombant dans l'intervalle de facteur 2 inférieure à 80%. Durée courte : quelques générations (10-20) ; Durée moyenne : plus de 100 générations. N.A. : mesure non appropriée.

Hétérogénéité démographique		Effet sur l'estimation de $D\sigma^2$			
		Biais relatif		Facteur 2	Durée
		Signe	Intensité		
Temporelle	↗ Dispersion (facteur 25)	Positif	Moyen	Bon	Courte
	↘ Densité (facteur 10 à 90)	Positif	Faible à Moyenne	Bon à Mauvais	Courte
	↗ Densité (facteur 10 à 90)	Negatif	Forte	Mauvais	Moyenne
Spatiale	Zone de forte densité (facteur 10)	Negatif	Faible (locale) à forte (avoisinante)	Bon (locale) à Mauvais (avoisinante)	N.A.
Temporelle et Spatiale	Expansion spatiale	Negatif	Faible	Bon	Courte

La robustesse globale de la méthode des moments à diverses fluctuations des paramètres démographiques dans l'espace et dans le temps contredit de précédentes études sur les déséquilibres évolutifs. Dans leur commentaire, Koenig *et al.* (1996) ont conclu que l'estimation de paramètres de dispersion à partir de données génétiques donne des indications sur les flux de gènes passés plutôt qu'actuels, alors que les méthodes directes, telles que les techniques de capture-marquage-recapture, donne de meilleures estimations des paramètres de dispersion actuels. Boileau *et al.* (1992) ont montré de façon similaire que des centaines ou des milliers de générations étaient nécessaires pour effacer les effets de processus de colonisation sur des estimateurs de type F_{ST} à partir de données allozymiques dans de grandes populations. Ces auteurs concluaient que les estimations de flux de gènes à partir de données génétiques doivent être interprétées avec "précautions". Les fluctuations démographiques temporelles ont probablement un effet fort et persistant sur certains statistiques et méthodes. Néanmoins, notre étude montre que certaines méthodes indirectes, et certains marqueurs génétiques, donnent des estimations satisfaisantes de la densité et de la dispersion actuelles même si l'histoire démographique des populations étudiées inclut des fluctuations

démographiques relativement récentes.

Un examen plus précis de nos résultats montre que dans le cas d'études d'organismes envahissants, la méthode étudiée devrait donner des estimations précises de $D\sigma^2$ actuel, si il n'y a pas eu d'explosion démographique pendant l'expansion démographique. Cette caractéristique est intéressante car elle devrait permettre l'étude d'organismes envahissants pour lesquels les caractéristiques démographiques sur la zone envahie sont les mêmes que celles dans les populations sources. Nos simulations ont aussi montré que si un changement de dispersion a eu lieu pendant le processus de colonisation, cette méthode estimera bien les nouvelles caractéristiques de dispersion. D'un autre côté, les explosions démographiques (et dans une moindre mesure les goulets d'étranglements) peuvent influencer fortement l'estimation du $D\sigma^2$ actuel. Or de nombreuses populations envahissantes montrent pendant les processus de colonisation des fluctuations démographiques complexes incluant des goulets d'étranglement et/ou des explosions démographiques (Williamson, 1996; Estoup *et al.*, 2001). Il semble donc important de faire des simulations complémentaires afin d'évaluer correctement la robustesse de la méthode sur des scénarios complexes mais plus réalistes d'invasions.

Nos simulations montrent que les densités avoisinantes influencent fortement l'estimation du $D\sigma^2$ local quand l'échantillon est pris sur une petite zone de forte densité. Dans ce cas, les estimations ne correspondent ni à la densité locale ni à la densité avoisinante. Cependant, si l'échantillonnage est fait sur une zone de forte densité suffisamment large (e.g. sur une surface correspondant à quatre fois la surface échantillonnée), l'estimation correspond essentiellement à celle de la densité locale (i.e. la densité sur la zone échantillonnée). Nos simulations se sont focalisées sur l'étude de l'influence d'une zone de forte densité au milieu d'une large zone de densité homogène et plus faible. Cette situation correspond à un biais classique lors de l'échantillonnage sur le terrain (i.e. le fait de collecter des échantillons plutôt sur des zones où la densité en individus est forte). Cependant, de nombreuses situations biologiques d'hétérogénéités spatiales de la densité devraient plutôt correspondre à des fluctuations aléatoires de la densité en chaque dème du réseau. Il est alors attendu que, sous de tels scénarios, la différenciation soit fonction d'une "densité efficace" et du taux de dispersion. Le manque de formalisation analytique pour de telles quantités "efficaces" limite fortement l'interprétation d'étude par simulation de la performance des estimateurs. Il est toutefois peu probable que l'estimation du produit "$D\sigma^2$ efficace" soit plus affectée par des fluctuations aléatoires que par les hétérogénéités étudiées ici.

3.5.4 Interprétation de la robustesse générale de la méthode à l'aide des temps de coalescence

La robustesse générale de la méthode des moments aux facteurs mutationnels et aux hétérogénéités spatiales et temporelles des paramètres démographiques peut s'interpréter à l'aide des probabilités de coalescence, comme on l'a vu dans les sections 2.1.3, 2.2.1 et 2.2.4. On a vu en effet que les F-statistiques pouvaient être déduites des différences entre les distributions de probabilités de coalescence de différentes paires de gènes (Fig.2.1, 2.2 et 2.5). Dans le cas du F_{ST}, ou de a_r, ces distributions diffèrent essentiellement par un excès de probabilité de coalescence pour les gènes les plus apparentés ($C_1(t)$ sur la Fig.2.5). Pour les modèles d'isolement par la distance avec une forte migration, cet excès de probabilité est concentré sur une courte période τ dans le passé récent. Comme la sensibilité des F_{ST}, ou de a_r, aux processus mutationnels et démographiques dépend de la durée de cette période τ, ces processus devraient avoir d'autant moins d'influence que τ est petit. Or comme on l'a vu en section 2.2.4, cette période de passé récent est d'autant plus courte que les taux de migration sont forts, et dans une moindre mesure les tailles de dème petites. Il n'est donc pas étonnant que, sous les modèles d'isolement par la distance en population continue pour lesquels les taux de migrations sont forts (i.e. de l'ordre de 50%) et les tailles de dème faibles (i.e. un individu par nœud du réseau), l'influence des processus mutationnels et des fluctuations démographiques passées sur l'estimation de paramètres démographiques par F-statistique soit limitée. A l'inverse, sous le modèle classique en îles avec des grandes populations et des taux de migrations faibles, l'effet des mutations et des fluctuations démographiques doivent être plus problématiques, ce qui a été largement vérifié par ailleurs (Boileau *et al.*, 1992). De plus, puisque l'on s'intéresse à la différenciation à petite échelle géographique, ces effets seront d'autant plus faibles. En effet, comme on l'a vu sur la Fig.2.5, plus les gènes comparés sont distants géographiquement, plus la période τ s'étend dans le passé (Slatkin, 1994; Rousset, 2004).

Le même type de raisonnement peut être utilisé pour comprendre pourquoi la méthode donne des estimations de densité correspondant plutôt à la densité locale sur la zone échantillonnée qu'à la densité autour de l'échantillon. Puisque la période τ s'étend peu dans le passé, les F-statistiques, F_{ST} ou a_r, dépendent principalement des événements de coalescence, de migration et/ou de mutation ayant eu lieu dans le passé récent et à une échelle géographique locale, puisque la dispersion est localisée dans l'espace. Par conséquent, l'estimation de $D\sigma^2$ avec la méthode étudiée ici correspond à sa valeur locale sur la zone échantillonnée et devrait être peu influencée par des

caractéristiques démographiques de zones situées géographiquement éloignées de la zone échantillonnée.

3.5.5 Implications quant aux études expérimentales

Les principales implications de nos résultats quant à l'application de la méthode d'estimation de $D\sigma^2$ développée par Rousset (2000) sur des cas expérimentaux concrets sont les suivants : (i) il est conseillé d'utiliser des marqueurs fortement polymorphes. Dans la mesure où les processus mutationnels et notamment les phénomènes d'homoplasie ont très peu d'influence sur les estimations, l'utilisation de marqueurs microsatellites est un bon choix. Il est toutefois préférable d'éliminer les locus ayant des diversités génétiques trop élevées (i.e. supérieures à 0.85) ; (ii) l'utilisation d'une statistique prenant en compte les différences de tailles entre allèles microsatellites donne des estimations non fiables de $D\sigma^2$ à cause de la forte variance des estimateurs ; (iii) il est important de restreindre spatialement l'échantillonnage afin de rester à une échelle géographique très locale. Cependant une estimation précise demande aussi un échantillonnage plus large quand σ augmente, ce qui implique une connaissance à priori de la valeur de σ. Il paraît donc fortement conseillé de procéder en deux étapes en faisant une estimation préliminaire de σ permettant de définir une échelle d'échantillonnage adaptée qui permettra une estimation plus précise des paramètres de dispersion. En l'absence d'estimation préliminaire de σ, une estimation grossière déduite de la connaissance à priori de certaines caractéristiques de la dispersion semble utile pour définir une échelle minimale d'étude (e.g. la vitesse de colonisation du crapaud de la canne à sucre en Australie, Leblois *et al.*, 2000, voir annexe B-1) ; (iv) l'échantillonnage devra être fait préférentiellement au milieu d'une zone relativement large sur laquelle la densité est assez homogène ; Si les aspects (i) à (iv) sont globalement respectés, la méthode de la droite de régression devrait donner de bonnes estimations de $D\sigma^2$ avec un biais et un MSE faible ; enfin (v) dans les cas d'utilisation de la méthode sur des populations d'organismes envahissants montrant une forte augmentation de la densité par rapport à leur habitat d'origine, on pourra s'attendre à une sous-estimation de $D\sigma^2$ si les populations sont issues d'événements de colonisation récents. Au contraire, si la méthode est employée sur des espèces menacées ayant subi très récemment de fortes baisses de densité, il est prévisible que la méthode sur-estime le produit $D\sigma^2$, même si dans ce cas la mémoire de la densité passée est beaucoup plus courte que pour les explosions démographiques. Notons que dans les deux cas, les biais entraînent une "sous-estimation du

problème considéré" (e.g. la baisse ou l'augmentation des densités présentes par rapport aux densités passées). Enfin, en dépit d'une sur-estimation de la borne supérieure de l'intervalle, la procédure de construction d'intervalles de confiance par ABC bootstrap devrait être utile pour avoir une idée de l'incertitude sur l'estimation de $D\sigma^2$.

Chapitre 4

Estimation par maximum de vraisemblance : Où en est on ?

4.1 Principe

D'un point de vue purement statistique, un des aspects les plus intéressants de la théorie de la coalescence est de permettre une analyse par maximum de vraisemblance du polymorphisme de marqueurs neutres. Ce type d'analyse a l'avantage d'utiliser l'information complète des données génétiques. La démarche consiste simplement à considérer la vraisemblance $\mathcal{L}(\mathcal{P}; D) \equiv \Pr(D; \mathcal{P})$ d'un modèle défini par les valeurs prises par l'ensemble de ses paramètres $\mathcal{P}$, sachant les données $\mathcal{D}$.[1] Il faut ensuite trouver le jeu de paramètres $\mathcal{P}_{MLE}$ pour lequel la vraisemblance est maximum. $\mathcal{P}_{MLE}$ est alors l'estimateur par maximum de vraisemblance. Mis à part quelques cas spécifiques, il n'existe pas de formules explicites de la vraisemblance pour les modèles démo-génétiques évoqués dans le chapitre 2 de ce document.

De nombreuses méthodes ont été développées pour calculer la vraisemblance de différents types de données génétiques selon différents modèles de structuration de populations. Il est important de noter qu'il existe des méthodes d'estimation de paramètres démographiques par maximum de vraisemblance fondées sur les fréquences alléliques d'un échantillon n'utilisant

[1] Par extension (ou abus de langage), on parlera aussi bien de la vraisemblance d'un modèle sachant les données que de la vraisemblance des données sous un modèle, bien que la dernière formulation corresponde moins à l'esprit du maximum de vraisemblance. En effet, puisque l'on cherche les valeurs des paramètres qui maximisent la vraisemblance, ce sont les paramètres qui varient et non les données.

pas la coalescence (Rannala & Hartigan, 1996; Tufto *et al.*, 1996; Wang & Whitlock, 2003), cependant nous nous focaliserons dans ce document sur l'estimation par maximum de vraisemblance des paramètres d'un modèle démo-génétique à partir d'un échantillon de gènes par des méthodes fondées sur la théorie de la coalescence. Ces méthodes ne considèrent pas de formules explicites pour la vraisemblance mais utilisent la simulation d'arbres de coalescence pour obtenir une estimation de la vraisemblance d'un échantillon pour différentes valeurs des paramètres du modèle. Dans le contexte de la coalescence, on peut écrire la vraisemblance des paramètres d'un modèle en fonction des données comme

$$\mathcal{L}(\mathcal{P}; D) = \int_G \Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P}), \tag{4.1}$$

où l'intégrale représente une somme sur toutes les généalogies compatibles avec l'échantillon. Un estimateur non biaisé de la vraisemblance est alors

$$\mathcal{L}(\mathcal{P}; D) = E\left[\Pr(D|G; \mathcal{P})\right], \tag{4.2}$$

où E correspond à l'espérance d'une chaîne de Markov de distribution stationnaire $\Pr(G; \mathcal{P})$. En d'autre mots, l'estimation de $\mathcal{L}(\mathcal{P}; D)$ par l'équation (4.2) se fera en calculant la moyenne de $\Pr(D|G; \mathcal{P})$ sur un grand nombre de généalogie simulées selon la distribution stationnaire $\Pr(G; \mathcal{P})$. Cependant, simuler directement selon la distribution $\Pr(G; \mathcal{P})$ peut s'avérer difficile techniquement ou peu efficace (i.e. beaucoup de généalogies échantillonnées auront une probabilité $\Pr(D|G; \mathcal{P})$ très faible).

L'équation (4.1) peut être mise sous la forme

$$\mathcal{L}(\mathcal{P}; D) = \int_G \frac{\Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P})}{f(G)} f(G), \tag{4.3}$$

où $f(G)$ est ce que l'on appellera la *fonction d'échantillonnage pondéré* (de l'anglais Importance Sampling). Cette nouvelle distribution f impliquera que l'on simulera les généalogies selon une distribution stationnaire différente de $\Pr(G; \mathcal{P})$. Dans ce cas, un estimateur non biaisé de la vraisemblance est

$$\mathcal{L}(\mathcal{P}; D) = E\left[\frac{\Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P})}{f(G)}\right], \tag{4.4}$$

où E correspond à l'espérance d'une chaîne de Markov de distribution stationnaire $f(G)$. L'utilisation de fonctions d'échantillonnage pondéré permet

d'explorer l'espace des possibles (ici les différentes généalogies) par la simulation selon f de façon plus efficace que directement selon la distribution stationnaire $\Pr(G;\mathcal{P})$. C'est à dire que l'on choisira la fonction d'échantillonnage pondéré f telle que l'exploration de zones de forte probabilité sera favorisée au dépens des zones de faible probabilité. Dans le jargon des techniques d'échantillonnage pondéré, le rapport $\Pr(G;\mathcal{P})/f(G)$ est appelée *poids de l'échantillonnage pondéré* (de l'anglais, importance sampling weight) car il donne le poids statistique de chaque généalogie proposée par la fonction d'échantillonnage pondéré f. C'est en quelque sorte un facteur de correction dû à l'utilisation de la fonction d'échantillonnage pondéré f. Si l'on explore un grand nombre de généalogies $\{G_1, G_2, .., G_n\}$ selon la distribution f, on a alors

$$\hat{\mathcal{L}}(\mathcal{P};D) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(D|G_i;\mathcal{P})\Pr(G_i;\mathcal{P})}{f(G_i)}. \tag{4.5}$$

Il suffit alors pour estimer $\hat{\mathcal{L}}(\mathcal{P};D)$ de déterminer la probabilité $\Pr(D|G_i;\mathcal{P})\Pr(G_i;\mathcal{P})$ d'un échantillon de gène D pour chaque généalogie simulée G_i.

4.2 Méthodes permettant le calcul de la probabilité d'un échantillon de gènes

On peut distinguer deux approches pour l'estimation de paramètres démographiques par maximum de vraisemblance à l'aide de la coalescence. L'une, développé par Beerli, Felsenstein et collaborateurs (Kuhner *et al.*, 1995; Beerli & Felsenstein, 1999, 2001), utilise une fonction d'échantillonnage pondéré et l'équation (4.5) pour calculer la vraisemblance d'un échantillon sous différentes généalogies. Les généalogies sont explorées selon un algorithme de Metropolis-Hastings par chaînes de Markov. L'autre, initiée par Griffiths et collaborateurs (Griffiths & Tavaré, 1994; Nath & Griffiths, 1996; Bahlo & Griffiths, 2000; de Iorio & Griffiths, 2004a,b) utilise d'autres types d'algorithmes fondés aussi sur l'échantillonnage pondéré. Cette approche utilise des processus de simulation par chaînes de Markov absorbantes (i.e. avec un état final absorbant, ici les MRCA de l'arbre de coalescence) permettant d'approximer les solutions des équations de récurrence sur les généalogies ancestrales à l'échantillon considéré. La vraisemblance des paramètres $\mathcal{L}(\mathcal{P};D)$ est estimée en construisant un arbre de coalescence à partir de l'échantillon en remontant dans le temps événement par événement. Les principes de ces deux types d'algorithmes sont détaillés dans les sections suivantes.

4.2.1 Approche de Felsenstein et collaborateurs

Beerli & Felsenstein (1999, 2001) ont utilisé un algorithme de Monte Carlo par chaînes de Markov pour estimer $\mathcal{L}(\mathcal{P};\mathcal{D})$. Le terme Monte Carlo traduit le fait que l'on utilise un tirage aléatoire de généalogies selon une distribution d'échantillonnage f pour résoudre l'intégrale (4.1). Puisque, sous un modèle neutre, la généalogie ne dépend pas des paramètres mutationnels $\mathcal{M}$ mais uniquement des paramètres démographiques $\mathcal{D}$ (voir section 2.3), on a

$$\hat{\mathcal{L}}(\mathcal{P};D) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(D|G_i;\mathcal{M})\Pr(G_i;\mathcal{D})}{f(G_i)}, \tag{4.6}$$

avec $\mathcal{P}=(\mathcal{D},\mathcal{M})$. Beerli & Felsenstein (1999, 2001) ont utilisé la fonction d'échantillonnage pondéré suivante

$$f_{BF}(G) \equiv \frac{\Pr(G;\mathcal{D}_0)\Pr(D|G;\mathcal{M})}{\mathcal{L}(\mathcal{P}_0;D)} \tag{4.7}$$

pour un ensemble $\mathcal{D}_0$ de valeurs des paramètres démographiques, avec $\mathcal{P}_0=(\mathcal{D}_0,\mathcal{M})$. En remplaçant la fonction d'échantillonnage pondéré dans l'équation (4.6), on a alors

$$\frac{\mathcal{L}(\mathcal{P};D)}{\mathcal{L}(\mathcal{P}_0;D)} \simeq \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(D|G_i;\mathcal{M})\Pr(G_i;\mathcal{D})}{\Pr(G_i;\mathcal{D}_0)\Pr(D|G_i;\mathcal{M})} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(G_i;\mathcal{D})}{\Pr(G_i;\mathcal{D}_0)}, \tag{4.8}$$

où les généalogies G_i sont générées par une chaîne de Markov de distribution stationnaire f_{BF} avec les paramètres $\mathcal{P}_0$. L'équation (4.8) permet alors l'estimation du rapport des vraisemblances pour un ensemble de valeurs de $\mathcal{P}$ autour de $\mathcal{P}_0$ à partir de la réalisation d'une chaîne de Markov unique. Il suffit alors de trouver le jeu de paramètres $\mathcal{P}_{MLE}$ qui maximise ce ratio de vraisemblance. Nous allons maintenant voir rapidement comment on calcule la probabilité $\Pr(G_i;\mathcal{D})$ d'une généalogie connaissant les valeurs des paramètres démographiques puis comment la chaîne de Markov d'échantillonnage des généalogies est implémentée dans leurs algorithmes.

Densité de probabilité d'une généalogie connaissant les valeurs des paramètres démographiques

Considérons une population subdivisée en n_{d} sous-populations de taille $2N_i$ gènes échangeant des migrants avec un taux de migration "arrière" b_{ij} de

la sous-population i vers la sous-population j. Le cycle de vie est le même que celui considéré pour le coalescent structuré (voir section 2.3.2). La généalogie d'un échantillon de gènes est alors entièrement définie par (i) la séquence d'événements de coalescence ou de migration des lignées ancestrales, (ii) les intervalles de temps séparant ces différents événements et (iii) les lignées étant concernées par ces événements (Fig.2.6). On peut alors exprimer la probabilité d'une généalogie comme

$$\begin{aligned}\Pr(G;\mathcal{D}) = &\Pr(\text{intervalles de temps entre les événements constitutifs de G};\mathcal{D}) \\ &\times \Pr(\text{séquence d'événements de coalescence ou de migration};\mathcal{D}).\end{aligned} \tag{4.9}$$

Comme on l'a vu dans le cadre du coalescent structuré, si l'on considère que les tailles de populations sont grandes et les taux de migration petits, les temps d'attente u_τ entre deux événements en remontant dans le temps suivent une loi exponentielle de paramètres le taux d'événements global, somme des taux de chaque événement. On a donc

$$\begin{aligned}&\Pr(\{u_1,..,u_T\};\mathcal{D}) = \\ &\prod_{\tau=1}^{T}\left[\text{taux}_\tau \exp\left[\text{taux}_\tau u_\tau\right]\right]\end{aligned} \tag{4.10}$$

avec

$$\text{taux}_\tau = \sum_{i=1}^{n_\text{d}}\left[\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_\text{d}} n_{\tau i} b_{ki}\right]. \tag{4.11}$$

Le produit sur τ est le produit sur les T intervalles de temps composant la généalogie, u_τ est la durée de l'intervalle de temps τ et $n_{\tau i}$ les nombre de lignées présentes dans le dème i pendant l'intervalle de temps τ. Les autres notations correspondent aux notations utilisées dans tout ce document. Par ailleurs les probabilités des événements de coalescence et de migration sont identiques à ceux présentés dans le cadre du coalescent structuré, on a donc

$$\Pr(\text{coalescence}_\tau;\mathcal{D}) = \frac{\sum_{i=1}^{n_\text{d}} \frac{n_{\tau i}(n_{\tau i}-1)}{4N_i}}{\sum_{i=1}^{n_\text{d}}(\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_\text{d}} n_{\tau i} b_{ki})}, \tag{4.12}$$

et

$$\Pr(\text{migration}_\tau;\mathcal{D}) = \frac{\sum_{i=1}^{n_\text{d}}\sum_{k=1;k\neq i}^{n_\text{d}} n_{\tau i} b_{ki}}{\sum_{i=1}^{n_\text{d}}(\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_\text{d}} n_{\tau i} b_{ki})}. \tag{4.13}$$

Enfin, la probabilité qu'une lignée concernée par un événement de migration ou de coalescence appartienne à une sous-population donnée est proportionnelle au nombre de lignées présentes dans la sous-population dans l'intervalle de temps τ. La probabilité qu'une coalescence donnée se produise dans la sous-population v_τ à la fin de l'intervalle de temps τ est

$$\Pr(v_\tau;\mathcal{D}) = \frac{n_{\tau v_\tau}(n_{\tau v_\tau}-1)/4N_{v_\tau}}{\sum_{i=1}^{n_\mathrm{d}} n_{\tau i}(n_{\tau i}-1)/4N_i}\frac{2}{n_{\tau v_\tau}(n_{\tau v_\tau}-1)}. \tag{4.14}$$

Pour un événement de migration "arrière" de la sous-population w_τ vers la sous-population v_τ, la probabilité est

$$\Pr(v_\tau, w_\tau;\mathcal{D}) = \frac{n_{\tau v_\tau} b_{w_\tau v_\tau}}{\sum_{i=1}^{n_\mathrm{d}}\sum_{k=1;k\neq i}^{n_\mathrm{d}} n_{\tau i} b_{ki}}\frac{1}{n_{\tau v_\tau}}. \tag{4.15}$$

En posant $u'_\tau \equiv \mu u_\tau$, $\theta_{v_\tau} \equiv 4N_{v_\tau}\mu$, $b'_{ij} \equiv b_{ij}/\mu$ où μ est le taux de mutation des marqueurs, la probabilité d'une généalogie G sachant les paramètres démographiques $\mathcal{D}$ est alors

$$\Pr(G;\mathcal{D}) = \prod_{\tau=1}^{T}\left[\left(\delta_\tau b'_{w_\tau v_\tau} + (1-\delta_\tau)\frac{2}{\theta_{v_\tau}}\right)\exp\left(-u'_\tau\sum_{i=1}^{n_\mathrm{d}}\left(\frac{n_{\tau i}(n_{\tau i}-1)}{\theta_{v_i}} + n_{\tau i}\sum_{k=1;k\neq i}^{n_\mathrm{d}} b'_{ki}\right)\right)\right], \tag{4.16}$$

où δ_τ est une variable indicatrice prenant la valeur 1 si on a une migration ou 0 si on a une coalescence.

Échantillonnage des généalogies par chaînes de Markov

La topologie de la première généalogie est construite à partir de l'échantillon par la méthode UPGMA (de l'anglais Unweighted Pair Group Method with arithmetic averages, voir Swofford *et al.*, 1996) puis la méthode par parcimonie de Sankoff (1975) permet d'ajouter le nombre minimal d'événements de migration sur cette topologie. Nous considérerons pour la suite qu'une généalogie contient l'information sur les événements de migration ayant affecté le différentes lignées. Pour explorer les différentes généalogies compatibles avec l'échantillon, Felsenstein et collaborateurs ont utilisé des *chaînes de Markov*. Le terme chaîne de Markov traduit le fait que les généalogies vont être échantillonnées par transition d'une généalogie G_i à une autre G_{i+1} selon des probabilités de transition (de G_i à G_{i+1}) dépendant uniquement de ces

deux généalogies et non des généalogies échantillonnées précédemment (i.e. avant d'arriver à G_i). Dans l'approche de Felsenstein et collaborateurs , une nouvelle généalogie G_{i+1} est créée à partir de la généalogie G_i considérée par délétion et reconstruction d'une partie de cette généalogie en fonction des paramètres démographiques $\mathcal{D}$ du modèle. L'acceptation de la nouvelle généalogie est faite selon un algorithme de Metropolis-Hastings. En d'autres mots, l'algorithme de Metropolis-Hastings va donner les probabilités de transition entre les différentes généalogies explorées. C'est une méthode d'exploration d'un ensemble d'états pour laquelle les probabilités de transitions entre états sont choisies de telle manière à ce que l'espace soit exploré selon une fonction d'échantillonnage voulue, ici la fonction d'échantillonnage pondéré $f_{BF}(G)$. L'échantillonnage des généalogies par chaîne de Markov se fera donc en deux temps : (i) une nouvelle généalogie sera défini à partir de l'algorithme de délétion-reconstruction, et (ii) elle sera acceptée ou non selon la probabilité donnée par le critère de Metropolis-Hastings, $r = \min(1, r_m r_h)$, où le terme r_m correspond aux probabilités $\Pi(G_{i+1}|G_i)$ de définition de G_{i+1} à partir de G_i et r_m un facteur de correction pour que la chaîne de Markov ait comme distribution stationnaire la fonction d'échantillonnage pondéré $f_{BF}(G)$. Les termes r_h et r_m sont de la forme suivante

$$r_m = \frac{\Pi(G_i|G_{i+1})}{\Pi(G_{i+1}|G_i)} \text{ et } r_h = \frac{f_{BF}(G_{i+1})}{f_{BF}(G_i)} \tag{4.17}$$

(Hastings, 1970). Nous verrons que ce critère d'acceptation $r = \min(1, r_h r_m)$ de la nouvelle généalogie peut s'exprimer en fonction de probabilité facilement calculables dans notre modèle mais voyons tout d'abord comment se fait, techniquement, la transition vers une nouvelle généalogie G_{i+1} à partir de la généalogie courante G_i.

La construction de la nouvelle généalogie G_{i+1} à partir de G_i ce fait en quatre étapes : (i) un nœud de l'arbre (correspondant à une coalescence ou un gène de l'échantillon de départ) est choisi au hasard sur la généalogie (le nœud z sur la Fig.4.1) ; (ii) la lignée ancestrale de ce nœud est effacée pour obtenir une généalogie partielle G_p (étape B de la Fig.4.1) ; (iii) un nouvel intervalle de temps u est calculé, selon les probabilité des différents événements possibles considérée dans l'équation (4.11) mais conditionnellement au fait que la lignée ancestrale du nœud z est impliquée dans l'événement. Cet intervalle de temps u est donné par

$$u = -\frac{\ln(p_\tau)}{\sum_k b'_{ik} + \frac{2n'_{\tau i}}{\theta_i}}, \tag{4.18}$$

où $n'_{\tau i}$ est le nombre de lignées dans la sous-population i, qui contenait le nœud z (i.e. les lignées blanches sur la Fig.4.1), pendant l'intervalle de temps τ de la généalogie partielle G_p (qui ne contient plus le nœud z), et p_τ est une variable aléatoire tirée dans une loi uniforme [0,1] (étape C de la Fig.4.1, "1" correspond au nouvel intervalle de temps). On choisit ensuite un nouvel événement pour cet intervalle de temps selon la probabilité relative des événements. Par exemple pour une migration arrière de i vers l, la probabilité de cet événement est

$$\Pr(\text{migration de } i \text{ vers } l\ ) = \frac{b'_{il}}{\sum_k b'_{ik} + \frac{2n'_{\tau i}}{\theta_i}}. \tag{4.19}$$

Ces probabilités correspondent aussi aux probabilités des différents événements possibles considérés dans l'équation (4.11) mais conditionnellement au fait que la lignée ancestrale du nœud z est impliquée dans l'événement. Si c'est une migration, cet exemple est illustré sur la figure 4.1C, la lignée

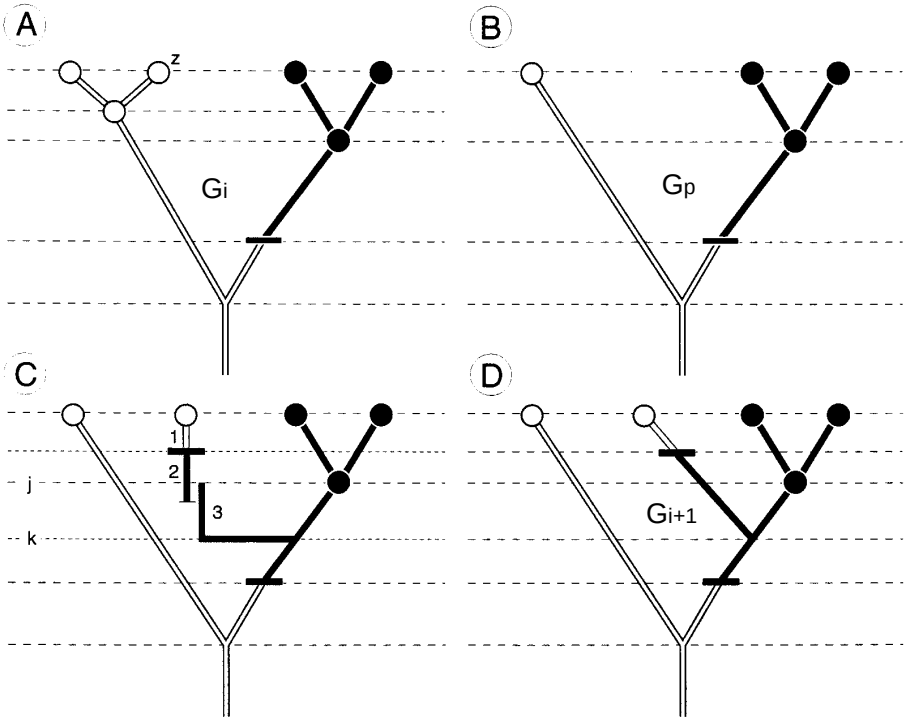

FIG. 4.1. Illustration du principe de délétion-construction d'une généalogie selon l'algorithme de Felsenstein et collaborateurs. Les lignées noires et blanches appartiennent à différentes sous-populations. Voir le texte pour les détails. D'après Beerli & Felsenstein, 1999.

change de population. Si c'est une nouvelle coalescence, la lignée coalesce alors avec une autre lignée de la même sous-population et une nouvelle généalogie G_{i+1} est ainsi formée. Dans le cas de la migration, des nouveaux intervalles de temps sont calculés (éq.4.18) pour finir avec une coalescence dans la bonne population (i.e. la population de la nouvelle lignée ancestrale du nœud z, Fig.4.1D).

Comme nous l'avons précisé en début de section, la nouvelle généalogie G_{i+1} est acceptée avec la probabilité de Metropolis-Hastings

$$r = \min(1, r_m r_h) = \frac{\Pi(G_i|G_{i+1})}{\Pi(G_{i+1}|G_i)} \frac{f_{BF}(G_{i+1})}{f_{BF}(G_i)}. \tag{4.20}$$

En remplaçant f_{BF} par son expression complète (éq.4.7), on obtient

$$r_h = \frac{f_{BF}(G_{i+1})}{f_{BF}(G_i)} = \frac{\Pr(G_{i+1};\mathcal{D}_0)\Pr(D|G_{i+1})}{\Pr(G_i;\mathcal{D}_0)\Pr(D|G_i)}, \tag{4.21}$$

et en considérant que la transition de G_{i+1} à G_i passe par l'étape intermédiaire de la généalogie partielle G_{p}, on a

$$r_m = \frac{\Pi(G_i|G_{i+1})}{\Pi(G_{i+1}|G_i)} = \frac{\Pi(G_i|G_{\mathrm{p}})\Pr(G_{\mathrm{p}}|G_{i+1})}{\Pi(G_{i+1}|G_{\mathrm{p}})\Pr(G_{\mathrm{p}}|G_i)}. \tag{4.22}$$

Par ailleurs, puisque la probabilité de transition entre la généalogie partielle G_{p} et les généalogies complètes G_i et G_{i+1} sont uniquement fonction du nombre de coalescences (i.e. les nœuds potentiellement choisis lors de l'étape de délétion d'une partie de la généalogie), on a donc $\Pi(G_{\mathrm{p}}|G_i) = \Pi(G_{\mathrm{p}}|G_{i+1})$. De plus, comme le processus de création de la nouvelle généalogie est uniquement dépendant des paramètres démographiques $\mathcal{D}$, $\Pi(G|G_{\mathrm{p}})$ est proportionnel à $\Pr(G|\mathcal{D}_0)$. Ceci simplifie largement l'expression du critère de Metropolis qui est alors

$$r_m = \frac{\Pr(G_i|\mathcal{D}_0)}{\Pr(G_{i+1}|\mathcal{D}_0)}, \tag{4.23}$$

et on a donc

$$r = \min\left(1, \frac{\Pr(D|G_{i+1})}{\Pr(D|G_i)}\right). \tag{4.24}$$

L'utilisation du critère de Metropolis-Hastings garanti le fait que la chaîne de Markov ait $f_{BF}(G)$ comme distribution stationnaire. On voit bien que,

jusque là, la transition de l'ancienne généalogie ne dépend pas des généalogies précédemment échantillonnées et correspond bien à une chaîne de Markov. L'ensemble de cette méthode a été implémentée dans le logiciel MIGRATE (Beerli & Felsenstein, 1999, 2001). Quelques tests de ce logiciel seront présentés dans la section 5.1.

4.2.2 L'approche de Griffiths et collaborateurs

La formulation de cette approche n'est pas triviale et j'essaierai ici d'en donner une explication la plus claire possible. Dans ce but, nous avons reformulé la manière de construire les récurrences à la base des algorithmes de Griffiths et collaborateurs. Le modèle démographique et le cycle de vie sont les mêmes que dans la section précédente et correspondent au modèle utilisé pour décrire le coalescent structuré (voir section 2.3.2).

Soit $\mathbf{n}(t)$ la configuration (i.e. les états alléliques et les sous-populations des gènes échantillonnés) d'un échantillon de $n(t)$ gènes pris au temps t. Griffiths et collaborateurs se sont intéressé à la probabilité de l'échantillon conditionnellement au nombre de gènes échantillonné $p(\mathbf{n}(t)) \equiv \Pr[\mathbf{n}(t)|n(t)]$. En considérant tous les états ancestraux de cet échantillon, nous avons la relations de récurrence suivante

$$p(\mathbf{n}(t)) = \sum_{\mathbf{n}(t-1)} \Pr[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)] \Pr[\mathbf{n}(t-1)|n(t)]\,. \tag{4.25}$$

En introduisant la probabilité de la taille de l'échantillon $n(t-1)$ à la génération précédente sachant la taille actuelle $n(t)$, ceci peut s'écrire

$$\begin{aligned} p(\mathbf{n}(t)) = \sum_{\mathbf{n}(t-1)} & \Pr[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)] \\ & \times \Pr[\mathbf{n}(t-1)|n(t), n(t-1)] \Pr[n(t-1)|n(t)]\,. \end{aligned} \tag{4.26}$$

On peut remarquer que le terme $\Pr[\mathbf{n}(t-1)|n(t), n(t-1)]$, ne dépend que de la taille de l'échantillon à $t-1$; il est donc égal à $\Pr[\mathbf{n}(t-1)|n(t-1)]$ qui correspond à $p(\mathbf{n}(t-1))$. On a donc

$$p(\mathbf{n}(t)) = \sum_{\mathbf{n}(t-1)} \Pr[n(t-1)|n(t)]\, p(\mathbf{n}(t-1)) \Pr[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)]\,. \tag{4.27}$$

Deux possibilités sont à envisager. La première est que $n(t-1) = n(t) - 1$, ce qui correspond à une événement de coalescence. La seconde est que $n(t-1) = n(t)$, c'est à dire qu'il n'y ait pas eu d'événement de coalescence mais uniquement de la mutation ou de la migration. Pour la suite, toutes les variables seront considérées au temps t , dans le cas contraire cela sera précisé. Ainsi pour un événement "avant" de mutation d'un gène de type allélique i vers un type allélique j, on a

$$\Pr\left[n(t-1)|n(t)\right] = \left(1 - \sum_a \frac{n_a(n_a-1)}{4N}\right) \tag{4.28}$$

et

$$\Pr\left[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)\right] = \left(1 - \sum_a \sum_{b,b\neq a} n_a m_{ab}\right) n_a \frac{n_{ai}+1}{n_a} \mu p_{ij}, \tag{4.29}$$

où n_a est le nombre de lignées dans la sous-population a, n_{ai} est le nombre de lignées de type allélique i dans la sous-population a, et p_{ij} la probabilité de muter d'un type allélique i vers un type allélique j sachant qu'il y a une mutation. De manière similaire, pour une migration "avant" de a vers b d'un gène de type allélique i, on a

$$\Pr\left[n(t-1)|n(t)\right] = \left(1 - \sum_a \frac{n_a(n_a-1)}{4N}\right) \tag{4.30}$$

et

$$\Pr\left[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)\right] = \left(1 - \sum_a n_a \mu\right) n_b m_{ab} \frac{n_{ai}+1}{n_a+1}. \tag{4.31}$$

Et enfin, pour un événement de coalescence dans le dème a, de deux lignées de types j, on a

$$\Pr\left[n(t-1)|n(t)\right] \Pr\left[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)\right] = \frac{n_a(n_a-1)}{4N} \frac{n_{aj}-1}{n_a-1}. \tag{4.32}$$

En négligeant les événements multiples, on a alors

$$
\begin{aligned}
& p(\mathbf{n}(t)=\eta) \\
&= \left(1-\sum_a \frac{n_a(n_a-1)}{4N}\right)\left(1-\sum_a n_a\mu\right)\left(1-\sum_a\sum_{b,b\neq a} n_a m_{ab}\right) p(\mathbf{n}(t-1)=\eta) \\
&+ \left(1-\sum_a \frac{n_a(n_a-1)}{4N}\right)\left(1-\sum_a\sum_{b,b\neq a} n_a m_{ab}\right) \\
&\quad \sum_a\left(n_a\mu\sum_i\sum_{j:n_{aj}>0,j\neq i}\frac{n_{ai}+1}{n_a}p_{ij}p(\mathbf{n}(t-1)=\eta-\mathbf{e}_{aj}+\mathbf{e}_{ai})\right) \\
&+ \left(1-\sum_a \frac{n_a(n_a-1)}{4N}\right)\left(1-\sum_a n_a\mu\right) \\
&\quad \sum_a\sum_{b,b\neq a}\left(n_b m_{ab}\sum_{i:n_{bi}>0}\frac{n_{ai}+1}{n_a+1}p(\mathbf{n}(t-1)=\eta-\mathbf{e}_{bi}+\mathbf{e}_{ai})\right) \\
&+ \sum_a\left(\frac{n_a(n_a-1)}{4N}\sum_{j:n_{aj}>1}\frac{n_{aj}-1}{n_a-1}p(\mathbf{n}(t-1)=\eta-\mathbf{e}_{aj})\right),
\end{aligned}
\tag{4.33}
$$

où $\mathbf{e}_{ai}$ est une matrice avec l'élément $(a,i)=1$, les autres éléments étant nuls. Plaçons nous maintenant à l'équilibre (i.e. $p\left[\mathbf{n}(t)=\eta\right]=p\left[\mathbf{n}(t-1)=\eta\right]$), en négligeant les termes d'ordre $o(\mu)$ et $o(\frac{1}{N})$, on a alors

$$
\begin{aligned}
p(\mathbf{n}(t)=\eta) =& \frac{1}{\left(\sum_a \frac{n_a(n_a-1)}{4N}+\sum_a n_a\mu+\sum_a\sum_{b,b\neq a} n_a m_{ab}\right)} \\
& \times\left[\sum_a\left(n_a\mu\sum_i\sum_{j:n_{aj}>0,j\neq i}\frac{n_{ai}+1}{n_a}p_{ij}p(\eta-\mathbf{e}_{aj}+\mathbf{e}_{ai})\right)\right. \\
& + \sum_a\sum_{b,b\neq a}\left(n_b m_{ab}\sum_{i:n_{bi}>0}\frac{n_{ai}+1}{n_a+1}p(\eta-\mathbf{e}_{bi}+\mathbf{e}_{ai})\right) \\
& \left.+\sum_a\left(\frac{n_a(n_a-1)}{4N}\sum_{j:n_{aj}>1}\frac{X_{aj}-1}{n_a-1}p(\eta-\mathbf{e}_{aj})\right)\right].
\end{aligned}
\tag{4.34}
$$

L'équation (4.34) peut être simplifiée en considérant $\theta = 4N\mu$, $\gamma_{ab} = 4Nm_{ab}$, $\gamma_a = \sum_{b,b\neq a} \gamma_{ab}$ et $\beta = \sum_a n_a(n_a - 1 + \gamma_a + \theta)$, on obtient alors

$$\begin{aligned} p(\mathbf{n} = \eta) =& \frac{1}{\beta} \sum_a \Bigg[\theta \sum_i \sum_{j:n_{aj}>0, j\neq i} (n_{ai}+1) p_{ij} p(\eta - \mathbf{e}_{aj} + \mathbf{e}_{ai}) \\ &+ \sum_{b,b\neq a} n_b \gamma_{ab} \sum_{i:n_{bi}>0} \frac{n_{ai}+1}{n_a+1} p(\eta - \mathbf{e}_{bi} + \mathbf{e}_{ai}) \\ &+ n_a \sum_{j:n_{aj}>1} (n_{aj}-1) p(\eta - \mathbf{e}_{aj}) \Bigg], \end{aligned} \tag{4.35}$$

ce qui correspond à l'équation (2.1) de Nath & Griffiths (1996) avec $p_{ii} = 0$ pour tout i. C'est la récurrence de base sur laquelle repose toute l'approche développée par Griffiths et collaborateurs.

Calcul de la probabilité d'un échantillon par simulation

La résolution de ces équations de récurrence n'est pas simple même pour des petites tailles d'échantillon. Griffiths & Tavaré (1994) ont proposé un algorithme fondé sur les techniques d'échantillonnage pondéré pour estimer les solutions de ces récurrences par simulation. On peut réécrire les récurrences sous la forme

$$\begin{aligned} p(\mathbf{n}) = w(\mathbf{n}) \Big(& \sum_{a,i,j:n_{aj}>0, j\neq i} \lambda_{aij}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{aj} + \mathbf{e}_{ai}) \\ &+ \sum_{a,b,b\neq a,i:n_{bi}>0} I_{abi}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{bi} + \mathbf{e}_{ai}) \\ &+ \sum_{a,j:n_{aj}>1} \mu_{aj}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{aj}) \Big) \\ &= \int_H W_{GT}(H) f_{GT}(H), \end{aligned} \tag{4.36}$$

où H correspond à l'histoire ancestrale d'un échantillon, c'est à dire à la généalogie de cet échantillon avec les événements de migration et de mutation. En partant de la configuration $\mathbf{n}$ de l'échantillon, on peut simu-

ler une histoire ancestrale complète (jusqu'au MRCA) en utilisant la fonction d'échantillonnage pondéré $f_{GT}(H)$ définie par λ_{aij}, I_{abi}, μ_{aj} correspondant respectivement aux probabilités de transition (i.e. probabilité de mutation, migration et coalescence) entre les différentes configurations $\mathbf{n}(\tau)$ le long de l'histoire ancestrale. On peut remarquer ici que l'histoire ancestrale de l'échantillon est construite en remontant le temps événement par événement, les transitions entre chaque configuration $\mathbf{n}(\tau)$ sont données par la fonction d'échantillonnage pondéré $f_{GT}(H)$. Avec ces notations, on a $H = \{\mathbf{n}(\tau)\},\ \tau \in [0, .., T]$, où T est le nombre d'événements dans l'arbre de coalescence simulé et $W_{GT}(H) = \prod_{\tau=0}^{T} w(\mathbf{n}(\tau))$. La fonction $W_{GT}(H)$ correspond aux poids de l'échantillonnage pondéré. On a alors

$$\begin{aligned} \mathcal{L}(\mathcal{P}; D) = p(\mathbf{n}) &= \int_H W_{GT}(H) f_{GT}(H) \approx \frac{1}{L} \sum_{h=1}^{L} W_{GT}(H_h) \\ &\approx \frac{1}{L} \sum_{h=1}^{L} \prod_{\tau=0}^{T} w_{GT}(\mathbf{n}_h(\tau)). \end{aligned} \tag{4.37}$$

$p(\mathbf{n})$ peut alors être estimé en calculant la moyenne de $\prod_{\tau=0}^{T} W_{GT}(\mathbf{n}(\tau))$ sur un grand nombre L de simulations d'arbres de coalescence indépendants partant de la configuration $\mathbf{n}$ et générés par la fonction d'échantillonnage pondéré f_{GT}. Par ailleurs, Griffiths & Tavaré (1994) ont montré comment une unique réalisation de la chaîne de Markov (création d'un seul arbre) peut être utilisée pour approcher la vraisemblance de l'échantillon pour différentes valeurs $\mathcal{P}_i$ des paramètres du modèle autour des valeurs $\mathcal{P}_0$ utilisées dans la chaîne de Markov (i.e. utilisées pour construire l'arbre). Cette approximation de la vraisemblance autour des valeurs utilisées pour simuler l'arbre de coalescence est similaire à la technique utilisée pour calculer le rapport de vraisemblance de l'équation (4.8). Le principe est de simuler les arbres de coalescence (avec la fonction d'échantillonnage pondéré) avec les paramètres centraux $\mathcal{P}_0$ et de calculer le poids $W_{GT,\mathcal{P}_i;\mathcal{P}_0}(G)$ pour différentes valeurs $\mathcal{P}_i$ des paramètres autour de $\mathcal{P}_0$. On a alors

$$\begin{aligned} \mathcal{L}(\mathcal{P}_i; D) &\approx \int_H W_{GT,\mathcal{P}_i;\mathcal{P}_0}(H) f_{GT,\mathcal{P}_0}(H) \\ &\approx \mathrm{E}\left[\frac{1}{L} \sum_{h=1}^{L} \prod_{\tau=0}^{T} w_{GT,\mathcal{P}_i;\mathcal{P}_0}(\mathbf{n}_h(\tau)) \right]. \end{aligned} \tag{4.38}$$

L'espace des histoires ancestrales H possibles est extrêmement large et de nombreuses histoires ancestrales ne contribueront que très peu dans le calcul

des équations (4.37) et (4.38). Le but des fonctions d'échantillonnage pondéré est de favoriser l'exploration de zones de forte probabilité au dépens des zones de faible probabilité. La fonction d'échantillonnage pondéré $f_{GT}(H)$ décrite ici (éq.4.36) n'est pas très efficace dans ce rôle car elle ne fait que choisir les événements de coalescence, de migration et de mutation en fonction de leurs probabilité respectives sans prendre en compte le fait que certains changements vont mieux correspondre à l'échantillon considéré. Autrement dit la fonction f_{GT} ne fait qu'explorer "uniformément" l'espace des histoires ancestrales possibles. Nous montrerons dans le chapitre suivant l'inefficacité relative de cet algorithme et les temps de calcul extrêmement longs qu'il engendre, ces deux caractéristiques rendant difficile l'utilisation en pratique de cet algorithme pour l'estimation des paramètres du modèle par maximum de vraisemblance.

4.2.3 Vers des distributions d'échantillonnage pondéré plus efficaces pour les méthodes de Griffiths et collaborateurs

Reprenons l'équation (4.36) sous la forme

$$\mathcal{L}(\mathcal{P};D) = \int_H \frac{\Pr(D|H;\mathcal{P})\Pr(H;\mathcal{P})}{f(H)} f(H), \tag{4.39}$$

dans laquelle les histoires ancestrales sont échantillonnées indépendamment par la fonction d'échantillonnage pondéré f. La fonction d'échantillonnage pondéré optimale f^* est la distribution conditionnelle des histoires ancestrales sachant les données $f^*(H) = \Pr(H|D;\mathcal{P})$. En effet, on peut alors vérifier que pour n'importe quelle histoire ancestrale générée par la fonction d'échantillonnage pondéré f^*, le produit des poids de l'échantillonnage pondéré est exactement la vraisemblance de l'échantillon : une seule histoire ancestrale est alors suffisante pour calculer la vraisemblance. Soit $\pi(\cdot|\mathbf{n})$ la distribution conditionnelle du type allélique d'un $n+1$ gène échantillonné sachant les états alléliques des n premiers gènes de l'échantillon avec $E = \{1,..,d\}$ l'ensemble des types alléliques possibles, on a alors

$$\pi(j|\mathbf{n}) = \frac{p(\mathbf{n},j)}{p(\mathbf{n})}. \tag{4.40}$$

La distribution optimale f^* pour le cas d'une population panmictique est alors générée par les transitions suivantes

$$\begin{aligned}
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{1}{\beta}\theta n_j \frac{\pi(i|\mathbf{n}(\tau)-\mathbf{e}_j)}{\pi(j|\mathbf{n}(\tau)-\mathbf{e}_j)} P_{ij} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_j + \mathbf{e}_i \text{ (mutation)}, \\
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{1}{\beta}\frac{n_j(n_j-1)}{\pi(j|\mathbf{n}(\tau)-\mathbf{e}_j)} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_j \text{ (coalescence)}, \\
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= 0 \qquad \text{dans les autres cas,}
\end{aligned} \tag{4.41}$$

où $\beta = (n(n-1+\theta))$, et n_j correspond, comme précédemment, au nombre de lignée de type allélique j dans $\mathbf{n}(\tau)$. Pour un modèle de population subdivisée tel qu'on l'a considéré dans les sections précédentes (section 2.3.2, 4.2.1 et 4.2.2), les probabilités de transitions définissant la fonction d'échantillonnage équivalente à l'équation (4.41) sont les suivantes

$$\begin{aligned}
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{\theta n_{aj}}{\beta}\frac{\pi(i|a,\mathbf{n}(\tau)-\mathbf{e}_{aj})}{\pi(j|a,\mathbf{n}(\tau)-\mathbf{e}_{aj})} P_{ij} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_{aj} + \mathbf{e}_{ai} \text{ (mutation)}, \\
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{n_{aj}}{\beta}\frac{\pi(j|b,\mathbf{n}(\tau)-\mathbf{e}_{aj})}{\pi(j|a,\mathbf{n}(\tau)-\mathbf{e}_{aj})}\gamma_{ab} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_{aj} + \mathbf{e}_{bj} \text{ (migration)}, \\
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{1}{\beta}\frac{n_{aj}(n_{aj}-1)}{\pi(j|a,\mathbf{n}(\tau)-\mathbf{e}_{aj})} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_{aj} \text{ (coalescence)}, \\
tr^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= 0 \qquad \text{dans les autres cas,}
\end{aligned} \tag{4.42}$$

où $\pi(\cdot|a,\mathbf{n})$ la distribution conditionnelle du type allélique d'un $(n+1)$ gène de la population a, sachant les types des n premiers gènes échantillonnés. Malheureusement, dans la majorité des cas, cette fonction f^* n'est pas connue car les probabilités conditionnelles π ne peuvent pas être calculées explicitement. Dans un premier temps, nous verrons une méthode générale proposée

par de Iorio & Griffiths (2004a) permettant une approximation des π. Nous verrons ensuite une application en population panmictique, que Stephens & Donnelly (2000) avaient proposé indépendamment de la méthode générale développée par de Iorio & Griffiths (2004a). Enfin, nous présenterons quelques applications en populations panmictiques.

Formalisation mathématique et développement du nouvel algorithme d'échantillonnage pondéré en population structurée

Il existe une connexion entre les équations de récurrence de Griffiths et collaborateurs (éq.4.35) et les équations de diffusion qui permet d'obtenir les équations de récurrence de la façon suivante. Pour un processus de diffusion, la densité de probabilité f des fréquences alléliques (ici dans différents dèmes) satisfait l'équation arrière de Kolmogorov, qui décrit les changements de f au cours du temps sous la forme

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \Phi(f), \tag{4.43}$$

où Φ est un opérateur différentiel qui prend ici la forme

$$\begin{aligned}\Phi =& \frac{1}{2}\sum_{i\in E}\sum_{j\in E} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j\in E}\Big(\sum_{i\in E} x_i r_{ij}\Big)\frac{\partial}{\partial x_j} \\ =& \sum_{j\in E}\Phi_j \frac{\partial}{\partial x_j}\end{aligned} \tag{4.44}$$

avec

$$R = \{r_{ij}\} \equiv \frac{\theta}{2}(P - I) \tag{4.45}$$

où $P = \{p_{ij}\}$ est la matrice de transition des états alléliques, communément appelée matrice de mutation, et I la matrice identité. Φ est aussi appelé "générateur" dans la littérature portant sur les processus de diffusion. L'équation arrière se généralise à toute fonction $u = E[g(X)]$ pour g bornée et continue (Karlin & Taylor, 1981, p.214), fonction des fréquences alléliques $\mathbf{x}$ dans la population

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \Phi(u). \tag{4.46}$$

Pour obtenir une récurrence sur les probabilité $p(\mathbf{n})$ d'échantillon, on écrit $p(\mathbf{n})$ sous la forme $E\left[g(\mathbf{x})\right]$

$$p(\mathbf{n}) = E\left[\binom{n}{\mathbf{n}} \prod_i x_i^{n_i}\right] \tag{4.47}$$

où

$$\binom{n}{\mathbf{n}} = \frac{n!}{\prod_i n_i!}. \tag{4.48}$$

On a donc

$$\frac{\mathrm{d}(p(\mathbf{n}))}{\mathrm{d}t} = \Phi\left[p(\mathbf{n})\right]. \tag{4.49}$$

A l'équilibre stationnaire, $\mathrm{d}(p(\mathbf{n}))/\mathrm{d}t$ est nulle. En développant l'expression pour $\Phi\left[p(\mathbf{n})\right]$, on retrouve alors la récurrence (4.35) entre les $p(\mathbf{n})$. On note que $\Phi\left[p(\mathbf{n})\right]$ peut s'écrire sous la forme

$$\sum_{j\in E} \Phi_j \frac{\partial}{\partial x_j}\left[p(\mathbf{n})\right], \tag{4.50}$$

pour un modèle à d états alléliques possibles, $E = [1, .., d]$, où $\mathbf{x} \in [0, .., 1]^d$ tel que $\sum_1^d x_i = 1$ est l'ensemble des fréquences alléliques de l'échantillon. La technique d'approximation développée par de Iorio & Griffiths (2004a) est d'approximer les $p(\mathbf{n})$, solutions de $\Phi\left[p(\mathbf{n})\right] = 0$, par les $\hat{p}(\mathbf{n})$ solutions de

$$E\left[\Phi_j \frac{\partial p(\mathbf{n})}{\partial x_j}\right] = 0, \text{ pour tout } j \in E, \tag{4.51}$$

i.e.

$$E\left[\Phi_j \frac{\partial}{\partial x_j}\binom{n}{\mathbf{n}} \prod_i x_i^{n_i}\right] = 0, \text{ pour tout } j \in E. \tag{4.52}$$

ce qui donne, pour une population panmictique, pour tout $j \in E$

$$n_j(n-1+\theta)\hat{p}(\mathbf{n}) = \\ n(n_j-1)\hat{p}(\mathbf{n}-\mathbf{e}_j) + \sum_{i\in E} \theta P_{ij}(n_i+1-\delta_{ij})\hat{p}(\mathbf{n}-\mathbf{e}_j+\mathbf{e}_i) \tag{4.53}$$

et pour un modèle de population subdivisée, un équivalent du système d'équation (4.53) est, pour tout $a \in [1, .., n_{\mathrm{d}}]$ et tout $j \in E$,

$$\begin{aligned} & n_{aj}\Big(n_a - 1 + \sum_{b\neq a}\gamma_a + \theta\Big)\hat{p}(\mathbf{n}) \\ & \quad = n_a(n_{aj}-1)\hat{p}(\mathbf{n}-\mathbf{e}_{aj}) + \sum_{i\in E}\theta P_{ij}(n_{ai}+1)\hat{p}(\mathbf{n}-\mathbf{e}_{aj}+\mathbf{e}_{ai}) \\ & \quad + \sum_{b\neq a}\gamma_{ab}\frac{n_a}{n_b+1}(n_{bj}+1)\hat{p}(\mathbf{n}-\mathbf{e}_{aj}+\mathbf{e}_{bj}). \end{aligned} \tag{4.54}$$

Les systèmes d'équations (4.53) et (4.54) peuvent être obtenus soit en considérant Φ comme ci-dessus, soit en considérant indépendamment chaque terme des sommes sur a (pour les population subdivisées uniquement) et sur j des récurrences (4.27) et (4.35). Cependant, on voit que les systèmes d'équations (4.53) et (4.54) ne contiennent pas les π que l'on cherche à déterminer. de Iorio & Griffiths (2004a) ont proposé la méthode suivante pour déterminer les $\hat{\pi}$, équivalent des π, mais calculés à partir de $\hat{p}(\mathbf{n})$ et non de $p(\mathbf{n})$.

Plaçons nous dans le cadre général d'une population subdivisée. Toutes les permutations de l'ordre de tirage des gènes de l'échantillon sont équiprobables, en effet l'ordre des gènes échantillonnés n'est pas pris en compte dans les calcul de $p(\mathbf{n})$. de Iorio & Griffiths (2004a) ont montré que cette notion d'équiprobabilité des permutations des gènes échantillonnés implique la relation, dite relation de symétrie, suivante

$$\pi(j|\alpha,\mathbf{n})p(\mathbf{n}) = \frac{n_{\alpha j}+1}{n_\alpha+1}p(\mathbf{n}+\mathbf{e}_{\alpha j}). \tag{4.55}$$

Si l'on considère que cette relation de symétrie est aussi valable pour les $\hat{\pi}$ et $\hat{p}$, ce qui d'ailleurs ne sera généralement pas le cas (de Iorio & Griffiths, 2004a), on a

$$\hat{\pi}(j|\alpha,\mathbf{n})\hat{p}(\mathbf{n}) = \frac{n_{\alpha j}+1}{n_\alpha+1}\hat{p}(\mathbf{n}+\mathbf{e}_{\alpha j}) \tag{4.56}$$

En intégrant la relation de symétrie (éq.4.56) pour les $\hat{p}$ et $\hat{\pi}$ dans le système d'équation (4.54), on a pour tout a et pour tout j

$$(n_a+\gamma_a+\theta)\,\hat{\pi}(j|a,\mathbf{n}) = n_{aj} + \sum_{i\in E}\theta P_{ij}\hat{\pi}(i|a,\mathbf{n}) + \sum_{b\neq a}\gamma_{ab}\hat{\pi}(j|b,\mathbf{n}). \tag{4.57}$$

Le système d'équations (4.57) donne théoriquement l'expression des $\hat{\pi}(\cdot|\cdot, \mathbf{n})$. Ce système d'équations est plus ou moins facile à résoudre selon les modèles considérés. de Iorio & Griffiths (2004b) donnent quelques exemples de résolution analytique de ce système d'équations dans des cas simples tels qu'un modèle de mutation indépendant du type allélique parentale (PIM, de l'anglais Parent Indépendant Mutation, le KAM et l'IAM décrit en section 2.1.1 en sont des exemples).

Approximation en population panmictique, lien avec l'algorithme de Stephens & Donnelly (2000)

Stephens & Donnelly (2000) ont proposé, dans le cas d'une population panmictique, une fonction d'échantillonnage pondéré s'approchant de $f^*(G)$, plus efficace que celle de de Griffiths & Tavaré (1994) correspondant à l'équation (4.36). Soit $\hat{\pi}(\cdot|\mathbf{n})$ la distribution définie par la probabilité de choisir aléatoirement un gène de $\mathbf{n}$ et de le faire muter selon la matrice de mutation P un certain nombre de fois, ce nombre étant donné par une loi géométrique de paramètre $\frac{\theta}{n+\theta}$, on a donc

$$\hat{\pi}(j|\mathbf{n}) = \sum_i \sum_{m=0}^{\infty} \frac{n_i}{n} \left(\frac{\theta}{n+\theta}\right)^m \frac{n}{n+\theta} (P^m)_{ij}. \tag{4.58}$$

En notant $\lambda_n = \frac{\theta}{n+\theta}$ et $M^{(n)} = (1-\lambda_n)(I - \lambda_n P)^{-1}$, où I est la matrice identité, on a

$$\hat{\pi}(j|\mathbf{n}) = \sum_i \frac{n_i}{n} M_{ij}^{(n)}. \tag{4.59}$$

Stephens & Donnelly (2000) définissent la fonction d'échantillonnage pondéré f_{SD} comme la distribution correspondant aux probabilités de transition de

l'équation (4.41) en substituant $\pi(\cdot|\cdot)$ par $\hat{\pi}(\cdot|\cdot)$. On obtient alors

$$\begin{aligned}
\hat{tr}^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{1}{\beta}\theta n_j \frac{\hat{\pi}(i|\mathbf{n}(\tau)-\mathbf{e}_j)}{\hat{\pi}(j|\mathbf{n}(\tau)-\mathbf{e}_j)} P_{ij} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_j + \mathbf{e}_i \text{ (mutation)}, \\
\hat{tr}^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= \frac{1}{\beta}\frac{n_j(n_j-1)}{\hat{\pi}(j|\mathbf{n}(\tau)-\mathbf{e}_j)} \\
&\qquad \text{pour } \mathbf{n}(\tau-1) = \mathbf{n}(\tau) - \mathbf{e}_j \text{ (coalescence)}, \\
\hat{tr}^*_{\mathcal{P}}(\mathbf{n}(\tau-1)|\mathbf{n}(\tau)) &= 0 \qquad\qquad \text{dans les autres cas.}
\end{aligned} \tag{4.60}$$

Bien que cette approximation ne corresponde à aucune approximation "analytique" du modèle mais plutôt à une "intuition", cet algorithme s'est révélé être extrêmement efficace dans de le cas d'une population panmictique (Stephens & Donnelly, 2000 ; Cornuet, communication personnelle).

Reprenons l'approche générale développé par de Iorio & Griffiths (2004a). Un équivalent du système d'équations (4.57) est, pour tout $j \in E$

$$\begin{aligned}
n(n-1+\theta)\hat{\pi}(j|\mathbf{n}-\mathbf{e}_j)\hat{p}(\mathbf{n}-\mathbf{e}_j) =& n(n_j-1)\hat{p}(\mathbf{n}-\mathbf{e}_j) \\
&+ n\sum_{i\in E}\theta P_{ij}\hat{\pi}(i|\mathbf{n}-\mathbf{e}_j)\hat{p}(\mathbf{n}-\mathbf{e}_j).
\end{aligned} \tag{4.61}$$

Les solutions du système d'équations (4.61) correspondent aux $\hat{\pi}(\cdot|\mathbf{n})$ définis par Stephens & Donnelly (2000). Stephens & Donnelly (2000) ont donc intuitivement trouvé l'approximation développée par de Iorio & Griffiths (2004a).

Application au modèle de mutation par pas en populations subdivisées

Pour des modèles de populations subdivisées et en considérant un modèle de mutation par pas non borné (SMM, voir section 2.1.1), le système d'équation (4.57) s'écrit alors

$$\left(\frac{n_{aj}}{q_a}+\gamma_a+\theta\right)\hat{\pi}(j|a,\mathbf{n}) = \frac{n_a}{q_a} + \frac{\theta}{2}\Big(\hat{\pi}(j-1|a,\mathbf{n})+\hat{\pi}(j+1|a,\mathbf{n})\Big)+\gamma_a\hat{\pi}(j|b,\mathbf{n}), \tag{4.62}$$

pour toute population a et pour tout état allélique j, où $q_a = N_a/N$ est la taille relative d'une sous-population par rapport à la taille totale N de la population.

Pour un modèle SMM non borné, il y a un une infinité d'équations de cette forme (i.e. une pour chaque allèle). On peut déterminer les $\hat{\pi}(\cdot|\cdot, \mathbf{n})$ en utilisant la technique de la transformée de Fourier. Comme le calcul ainsi que les solutions des $\hat{\pi}(\cdot|\cdot, \mathbf{n})$ sont assez lourds, je ne les présenterai pas ici (tous les détails de cet algorithme pour un modèle à 2 sous-populations sont présentés en annexe A-2). Le lecteur pourra aussi se référer à l'annexe B-4. Cet algorithme a été développé en collaboration avec M. de Iorio et R. Griffiths au cours de ma thèse. Quelques tests préliminaires de cet algorithme sont présentés dans la section 5.2.2.

Calcul de la probabilité d'un échantillon avec les nouvelles distributions d'échantillonnage pondéré

En partant de l'équations de récurrence (4.27), la probabilité d'un échantillon conditionnellement à sa taille est donnée par

$$p(\mathbf{n}(t)) \equiv \sum_{\mathbf{n}(t-1)} \Pr\left[n(t-1)|n(t)\right] \Pr\left[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)\right] p(\mathbf{n}(t-1)). \quad (4.63)$$

Si l'on note $tr_{GT}(\mathbf{n}(t)|\mathbf{n}(t-1)) = \Pr[n(t-1)|n(t)] \Pr[\mathbf{n}(t)|\mathbf{n}(t-1), n(t)]$ les probabilités de transition définies par l'équation (4.36) et $\hat{tr}^*(\mathbf{n}(t-1)|\mathbf{n}(t))$ les probabilités de transition définies par les équations (4.60), l'équation (4.63) peut être écrite sous la forme

$$p(\mathbf{n}(t)) = \sum_{\mathbf{n}(t-1)} \frac{tr_{GT}(\mathbf{n}(t)|\mathbf{n}(t-1))}{\hat{tr}^*(\mathbf{n}(t-1)|\mathbf{n}(t))} \hat{tr}^*(\mathbf{n}(t-1)|\mathbf{n}(t)) p(\mathbf{n}(t-1)). \quad (4.64)$$

On a alors un équivalent de l'équation (4.37)

$$\mathcal{L}(\mathcal{P}; D) \approx \frac{1}{L} \sum_{g=1}^{L} \prod_{\tau=T}^{1} \frac{tr_{GT}(\mathbf{n}_g(\tau)|\mathbf{n}_g(\tau-1))}{\hat{tr}^*(\mathbf{n}_g(\tau-1)|\mathbf{n}_g(\tau))} p(\mathbf{n}(0)), \quad (4.65)$$

où $\mathbf{n}(T)$ est la configuration de l'échantillon au moment de l'échantillonnage, et $\mathbf{n}(0)$ est la lignée ancestrale de l'échantillon, $p(\mathbf{n}(0))$ est donc la probabilité

du type allélique du MRCA de l'échantillon (tiré dans la distribution stationnaire des états alléliques) et les L généalogies sont générées indépendamment en partant de $\mathbf{n}(T)$ avec la distribution f_{SD} définie par les équations (4.60).

Chapitre 5

Précision et robustesse des estimations par maximum de vraisemblance

5.1 Approche de Felsenstein et collaborateurs : test du logiciel MIGRATE

La méthode développée par Felsenstein et collaborateurs est implémentée dans le logiciel MIGRATE (Beerli & Felsenstein, 1999, 2001). Ce logiciel est censé traiter des modèles très généraux de migration dans lesquels la migration est définie, comme pour le coalescent structuré, par une matrice de migration entre paires de sous-populations et chaque sous-population est de taille N_i (nombre d'individus diploïdes). Les résultats fournis par ce logiciel pour la traitement d'un échantillon de gènes pris dans différentes sous-populations sont les paramètres $\theta_a = 4N_a\mu$ pour chaque sous-population a échantillonnée et les taux de migration pour chaque paire de sous-population sous la forme $4N_a m_{ab}$ où m_{ab} est le taux de migration "avant" de la sous-population a vers b. Dans le chapitre 3 de ce document, nous avons vu qu'une méthode fondée sur les F-statistiques donne de bonnes estimations du produit $D\sigma^2$, où D est la densité d'individus et σ^2 le moment d'ordre deux de la distribution de dispersion parents-descendants, à partir de données génétiques sous isolement par la distance. Cependant, le fait que l'on estime uniquement σ^2 n'est pas tout à fait satisfaisant. En effet, il searit plus intéressant d'avoir une estimation de la distribution de dispersion en elle-même,

plutôt qu'une estimation de σ^2. Le logiciel MIGRATE aurait en théorie ce potentiel en considérant l'ensemble des estimateurs des taux de migration $4N_a m_{ab}$ répartis par classes de distances entre sous-populations. C'est pourquoi il nous est paru intéressant d'évoluer les potentialités de ce logiciel dans le cadre de modèles d'isolement par la distance sur un jeu de données réel ainsi que sur des jeux de données simulés.

5.1.1 Test sur jeu de données réel

Le jeu de données utilisé est constitué de données démographiques (Wood *et al.*, 1985) et des données génétiques de marqueurs allozymiques (Long *et al.*, 1986) obtenues sur des populations humaines de Papouasie-Nouvelle-Guinée. L'avantage de ce jeu de données est qu'il comporte à la fois des données démographiques et des données génétiques.

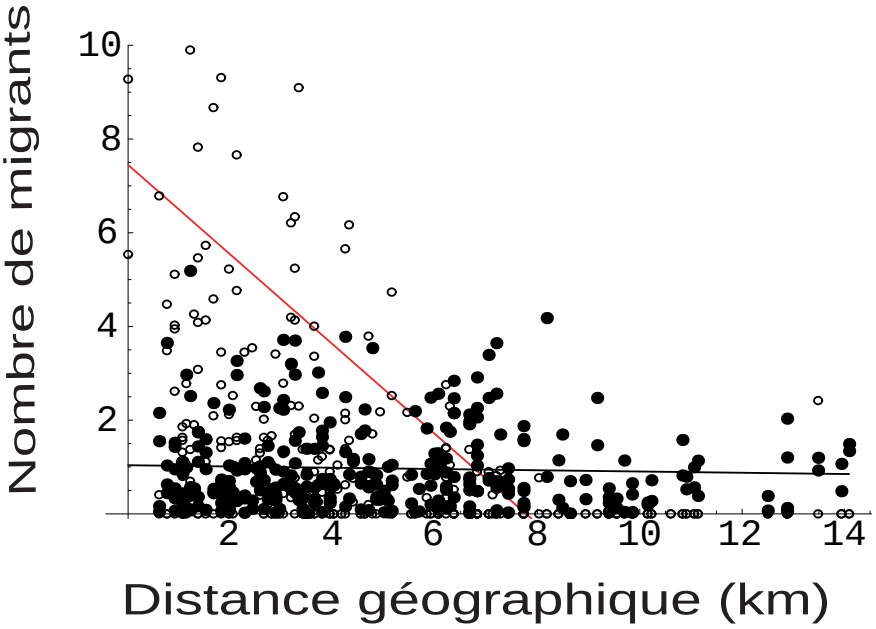

FIG. 5.1. Représentation du nombre de migrants en fonction de la distance géographique issu des données Humaine sur les villages de Papouasie Nouvelle-Guinée. Les cercles représentent les estimations à partir du jeu de données démographiques. Les points noirs sont les nombres de migrants estimés par MIGRATE à partir du jeu de données génétiques. Les droite correspondent aux droites de régression calculées sur chaque jeux de données.

Les données démographiques ont permis de calculer σ^2 à partir de la

distribution des distances des villages des parents par rapport aux villages des descendants. Cette estimation démographique donne une estimation de σ^2=1.93 km^2/génération (Rousset, 1997). Par ailleurs, les données génétiques ont été traitées avec la méthodes de Rousset (1997), analogue à la méthode de Rousset (2000) mais considérant un modèle d'isolement par la distance avec une structure en dèmes ; cette méthode donne une estimation indirecte de $\sigma^2 = 1.4$ km^2/génération (l'estimation de σ^2 est obtenue en considérant les tailles des populations issues de données démographiques), qui est proche de l'estimation démographique.

Ce même jeu de données génétique a été traité avec le logiciel MIGRATE et le calcul de σ^2 à partir de l'ensemble des taux de migration par paires de sous-populations donne une estimation indirecte de $\sigma^2 = 16.3$ km^2/génération. Ce résultat parait largement sur-estimer la dispersion réelle (i.e. d'un facteur 10). La figure 5.1 représente les estimations du nombre de migrants en fonction de la distance par la méthode démographique (points gris) et par MIGRATE (points noirs). On obtient une surestimation globale par MIGRATE de la migration pour chaque classe de distances, au point de ne plus observer de patron net d'isolement par la distance (Fig.5.1).

5.1.2 Test sur jeux de données simulés

Dans un deuxième temps nous avons testé l'estimation des taux de migration par MIGRATE sur des jeux de données simulés. Les simulations ont été faites en considérant un échantillon de 20 individus pris dans 11 sous-populations pour 5 locus. Les individus ont évolué sur un tore de (200×200) avec 20 individus par dème, le modèle mutationnel est le KAM à 10 allèles avec un taux de mutation de $5 \cdot 10^{-4}$. Enfin, la migration se fait uniquement entre dèmes adjacents (migration "stepping stone") avec un taux de migration total de 1/2. Trois jeux de données simulés ont été analysés avec MIGRATE. Nous n'avons analysé qu'un petit nombre de jeux de données pour seulement 5 locus car les temps de calcul demandés par MIGRATE sont longs et nous n'avions alors pas accès à des ordinateurs puissants. Les résultats de ces simulations sont présentés sur la figure 5.2.

On voit bien sur la figure 5.2 que MIGRATE sur-estime largement les nombres de migrants entre sous-populations. En effet, puisque la migration se fait, dans notre modèle de simulation, uniquement entre dèmes adjacents, on s'attend à avoir un nombre de migrants positif pour des distances de 1 pas sur le réseau et un nombre de migrants nul pour toutes les autres distances

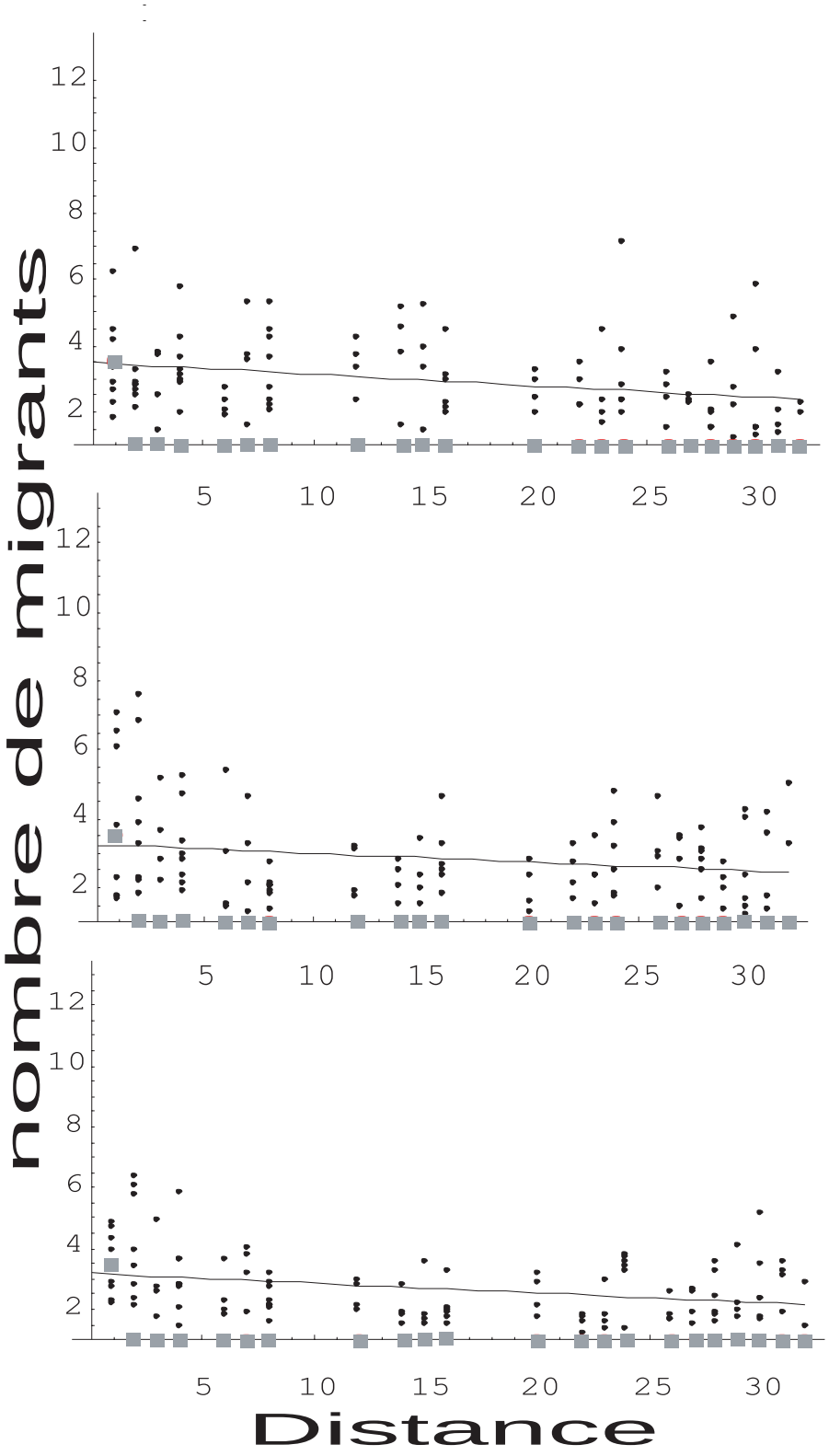

FIG. 5.2. Représentation du nombre de migrants en fonction de la distance géographique. Les carrés gris représentent les valeurs attendue (i.e. les valeurs avec lesquelles ont a simulé les données). Les points noirs sont les nombres de migrants estimés par MIGRATE sur ces jeux de données simulés.

(carrés gris Fig.5.2). On peut noter toutefois que l'estimation du nombre de migrants à une distance de 1 pas sur le réseau est bonne et correspond bien aux valeurs du modèle. Enfin, comme pour l'analyse du jeu de données réel, MIGRATE sous-estime largement l'isolement par la distance puisque le nombre de migrants estimé ne décroît que très peu avec la distance (droites de regression de la Fig.5.2), contrairement à ce qui est attendu.

Que ce soit pour le jeu de données réel ou pour les jeux de données simulés, les mauvais résultats obtenus avec le logiciel MIGRATE peuvent être dus à différents facteurs :

(i) le modèle mutationnel des marqueurs ne correspond pas exactement à celui assumé dans MIGRATE (IAM dans MIGRATE, KAM dans les simulations et inconnu pour le jeu de données réel). Une analyse de robustesse de la méthode aux processus mutationnels serait nécessaire. Si la méthode n'est pas robuste vis à vis des modèles mutationnels, cela risque de limiter considérablement son utilité en pratique du fait que l'on a rarement une idée précise des processus mutationnels des marqueurs utilisés lors d'études expérimentales.

(ii) le nombre de sous-populations échantillonnées (11 dans les simulations) ne correspond pas au nombre de sous-populations total du modèle démographique (40 000 dans les simulations). Or MIGRATE considère que le nombre de sous-populations de l'échantillon est égal au nombre total de sous-populations du système étudié. Une étude récente de Beerli (2004) a testé l'influence de sous-populations non échantillonnées sur l'estimation. Cette étude montre que lorsque la migration de la sous-population non échantillonnée vers les sous-populations échantillonnées est faible (i.e. du même ordre de grandeur que la migration entre les sous-populations échantillonnées), la présence d'une sous-population non échantillonnée a peu d'influence sur l'estimation des taux de migration et des tailles de populations. Par contre, si la migration entre la sous-population non échantillonnée et les sous-populations échantillonnées est forte, la présence d'une sous-population non échantillonnée a beaucoup d'influence sur l'estimation des taille de population mais peu sur l'estimation des taux de migration. Cette étude montre aussi que, dans le contexte d'un modèle en île, le nombre de sous-populations non échantillonnées a beaucoup d'influence sur l'estimation des tailles de populations mais peu sur l'estimation des taux de migration. La présence de sous-populations non-échantillonnées introduit une surestimation des tailles de populations, et ce d'autant plus que les taux d'émigration de ces sous-populations non-échantillonnées sont forts et que les sous-populations non-échantillonnées sont nombreuses. Le fait que l'estimation des taux de migration soit robuste à la

présence de sous-populations non échantillonnées est rassurant. Cependant, cette étude considère que les taux de migration entre sous-populations sont globalement faibles (de l'ordre de 0.0001 événements de migration par individus par génération), il est probable que la présence de sous-populations non échantillonnées ait plus d'influence quand les taux de migration sont forts, par exemple dans le cas des modèles d'isolement par la distance en populations continues. De plus cette étude considère des jeux de données sur 100 locus de 1000 paires de bases chacun, ce qui me parait extrêmement fort par rapport aux nombre de locus classiquement utilisés dans des études expérimentales (i.e. plutôt de l'ordre d'une vingtaine au maximum, et souvent moins pour des données de types séquences d'ADN). Des tests supplémentaires de l'influence de la présence de sous-populations non échantillonnées sur les estimations données par MIGRATE, avec moins de locus et des taux de migrations plus forts serait donc nécessaires pour conclure à la robustesse générale de cette méthode vis à vis de ce facteur.

(iii) la convergence est un problème récurrent dans le contexte des MCMC et la configuration par défaut de MIGRATE n'est peut être pas optimale de ce point de vue. Toutefois, une estimation avec des runs beaucoup plus longs (analyse de deux semaines par jeu de données) ont été faites sur les jeux de données simulés et aucune différence notable n'a été notée par rapport au runs "cours" (analyse de six jours par jeu de données) de la configuration par défaut.

Enfin (iv) il pourrait exister un biais inhérent à la méthode qui surestimerait les taux de migration. En effet, Beerli et Felsenstein ont observé sur des simulations un biais positif pour des paires de sous-populations n'échangeant aucun migrant (Beerli & Felsenstein, 2001). Ce biais pourrait être d'autant plus important que le nombre de paramètres estimés est élevé, réduisant ainsi la précision de l'estimation de chaque paramètre.

Tous ces facteurs nous poussent à conclure que les estimations de paramètres démographiques avec le logiciel MIGRATE doivent être interprétées avec beaucoup de précautions, notamment pour des populations en isolement par la distance. Des études approfondies de robustesse de ce logiciel vis à vis de différents facteurs mutationnels et démographiques seraient nécessaires.

5.2 Précision des algorithmes de Griffiths et collaborateurs

5.2.1 Algorithme de Nath et Griffiths 1996

Dans un premier temps, nous avons testé la précision de l'algorithme de Nath & Griffiths (1996) sur des jeux de données simulés. Cet algorithme correspond à la fonction d'échantillonnage pondéré définie par l'équation (4.36). Le but de cette étude est d'avoir quelques idées sur la précision de la méthode dans des cas simples ainsi que sur les temps de calcul nécessaires à une bonne estimation.

La méthode a été testée en considérant un modèle en îles à deux ou quatre dèmes suivant les situations. Les notations sont les notations utilisées tout le long de ce document. La migration est symétrique et se fait donc de façon équiprobable entre tous les dèmes du modèle avec le paramètre $\gamma = \gamma_{ab} = 4Nm/n_{\mathrm{d}} - 1$. Le modèle mutationnel est le KAM à 4 allèles (voir section 2.1.1) et le paramètre de mutation est $\theta = 4N\mu$ que l'on a fixé à 1.0 pour toutes les simulations. Les données ont été simulées en utilisant l'algorithme génération par génération décrit dans le chapitre 3 adapté aux modèles considérés pour cette étude. Les paramètres d'intérêt étant dans cette étude les paramètres de migration, nous considérons que le paramètre θ est connu et l'on estime donc uniquement le paramètre γ. Pour chaque simulation, 100 jeux de données avec 5 locus ont été analysés et le biais relatif et le MSE relatif de l'estimation du paramètre γ sont calculés à partir de l'estimateur de maximum de vraisemblance de chaque jeu de données. Dans notre étude, la vraisemblance est calculée pour 15 valeurs γ_i dans l'intervalle [0,..,1]. Un courbe de vraisemblance est ajustée sur ces 15 points à l'aide de la procédure de *Mathematica* d'ajustement de spline cubique. L'estimateur de maximum de vraisemblance est la valeur de γ de cette courbe pour laquelle la vraisemblance est maximale.

Enfin, comme nous l'avons vu lors de la description de l'algorithme dans la section 4.2.2, il est possible d'approximer la vraisemblance pour différentes valeurs de γ_i autour d'une valeur γ_0 que nous appellerons valeur centrale (en anglais driving value). Nous appellerons GPM (de l'anglais "Grid Point Method") l'estimation par le calcul de la vraisemblance non approchée pour chaque valeur γ_i, $i \in [1, .., P]$ et SSR (de l'anglais "Single Simulation Run") l'estimation par le calcul de la vraisemblance approchée pour les différentes valeurs γ_i autour de la valeur centrale γ_0. L'avantage de la méthode SSR

TAB. 5.1. Précision de l'estimation de $\gamma = 4N\frac{m}{n_d - 1}$ avec l'algorithme de Nath & Griffiths (1996) dans le cas d'un modèle à deux sous-populations. L est le nombre de généalogies échantillonnées pour estimer la vraisemblance (cf. éqs.4.37 et 4.38). e.s est l'erreur standard du biais relatif. GPM="Grid Point Method" et SSR= "Single Simulation Run" (voir texte pour les détails).

		GPM	SSR					
			$\gamma_0 = 0.2$			$\gamma_0 = 0.8$		
γ		L=50 000	20 000	50 000	100 000	20 000	50 000	100 000
	Biais	0.13	0.082	0.079	0.10	0.17	0.15	0.16
0.2	(e.s.)	(0.049)	(0.030)	(0.030)	(0.036)	(0.041)	(0.043)	(0.043)
	MSE	0.13	0.052	0.049	0.075	0.11	0.11	0.12
	Biais		-0.22	-0.19	-0.15	0.022	0.023	0.028
0.8	(e.s.)		(0.023)	(0.022)	(0.024)	(0.033)	(0.033)	(0.034)
	MSE		0.074	0.061	0.049	0.055	0.054	0.056

est qu'elle nécessite seulement des réalisations de l'histoire ancestrale H avec une valeur centrale γ_0 et non pas pour toutes les valeurs de γ. Le temps de calcul est donc diminué d'un facteur égal à P, le nombre de valeurs γ_i pour lesquelles on veut estimer la vraisemblance.

TAB. 5.2. Précision de l'estimation de $\gamma = 4N\frac{m}{n_d - 1}$ avec l'algorithme de Nath & Griffiths (1996) dans le cas d'un modèle à quatre sous-populations. L est le nombre de généalogies échantillonnées pour estimer la vraisemblance (cf. éq.4.37 et 4.38). e.s est l'erreur standard du biais relatif

		SSR					
		$\gamma_0 = 0.2$			$\gamma_0 = 0.8$		
		L=20 000	50 000	100 000	20 000	50 000	100 000
	Biais	0.031	0.011	0.026	0.27	0.20	0.23
$\gamma =$	(e.s.)	(0.015)	(0.014)	(0.016)	(0.023)	(0.022)	(0.022)
0.2	MSE	0.011	0.010	0.013	0.099	0.065	0.075
	Biais	-0.50	-0.45	-0.43	-0.047	-0.032	-0.018
$\gamma =$	(e.s.)	(0.012)	(0.013)	(0.013)	(0.031)	(0.031)	(0.032)
0.8	MSE	0.24	0.21	0.20	0.049	0.048	0.050

Les résultats pour les modèle à deux populations, présentés dans le tableau 5.1, montrent que l'estimation de γ est assez précise quand on utilise la méthode GPM ou la méthode SSR avec des valeurs centrales proches des valeurs attendues. Dans ces cas, le biais et le MSE sont inférieurs à 15% ce

qui montre bien que la valeur estimée est proche de la valeur attendue. Un résultat surprenant est que le nombre de généalogies échantillonnées pour estimer la vraisemblance (le paramètre L dans l'éq.4.38) a peu d'influence sur la précision de l'estimation avec la méthode SSR. Ce résultat est retrouvé avec la méthode GPM qui donne de bons résultats avec un nombre minimum de 50 000 généalogies échantillonnées ; au-dessus de ce nombre, l'échantillonnage de généalogies supplémentaires augmente peu la précision de la méthode (pour 200 000 généalogies échantillonnées, l'écart type du biais est de 0.036 (résultat non montré) ce qui n'est très différent de 0.049 (tableau 5.1). Les résultats pour un modèle à quatre populations montrent les même tendances (tableau 5.2). Un second résultat de cette étude est que la méthode SSR n'est précise que si l'on choisit des valeurs centrales proches des valeurs attendues ; dans le cas contraire le biais et le MSE sont plus forts, et ce d'autant plus que le nombre de généalogies échantillonnées est faible (tableaux 5.1 et 5.2).

Le résultat majeur de cette étude est que l'algorithme de Nath & Griffiths (1996) est extrêmement lourd à simuler et que les temps de calcul nécessaires à l'estimation des taux de migration, même pour les modèles simples tels que ceux considéré ici, sont très longs. La figure 5.3 illustre ce problème. Sur cette figure la complexité du modèle a été réduite au produit du nombre d'états alléliques possibles par le nombre de sous-populations. Nous avons considéré uniquement ces facteurs car, comme on le voit dans l'équation (4.36), ce sont ces paramètres qui déterminent le nombre de transitions possibles lors de la construction de l'arbre de coalescence (i.e. les paramètres qui déterminent l'espace des généalogies à explorer). L'estimation de paramètres démographiques avec cette méthode est donc possible uniquement pour des modèles simples avec moins de quatre populations et quatres allèles. Pour des modèles plus complexes, le temps de calcul de la vraisemblance d'un échantillon est si long que l'estimation des paramètres démographiques n'est pas envisageable avec l'algorithme de Nath & Griffiths (1996). La figure 5.3 illustre également le gain de temps dû aux nouvelles distributions d'échantillonnage pondéré de de Iorio & Griffiths (2004b). Ces nouvelles distributions d'échantillonnage pondéré, plus efficaces que les précédentes, permettent d'envisager l'estimation de paramètres démographiques pour des modèles plus complexes.

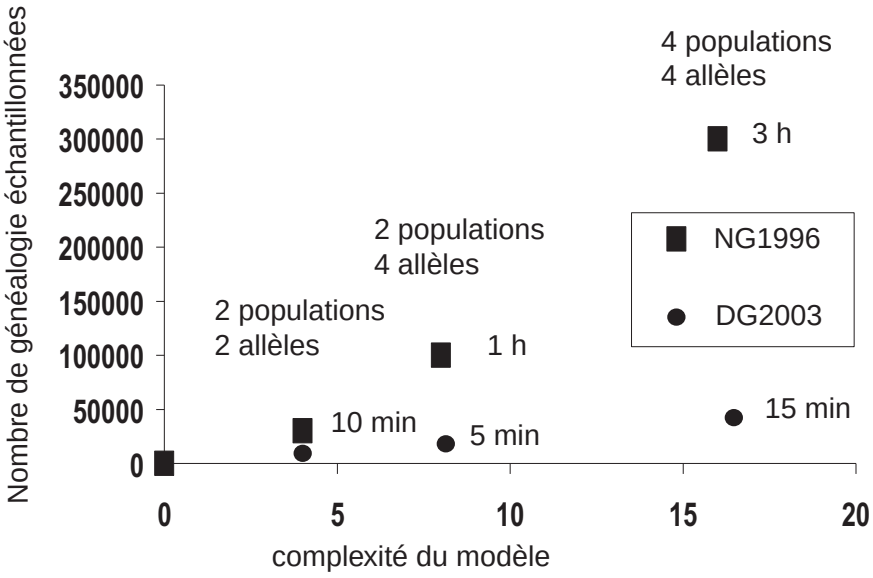

FIG. 5.3. Représentation du nombre de généalogies à échantillonner pour avoir une bonne estimation de la vraisemblance en fonction de la complexité du modèle. La complexité du modèle correspond sur ce graphique au produit du nombre de sous-populations du modèle démographique par le nombre d'allèles du modèle mutationnel. Les temps correspondent au temps nécessaire pour estimer la vraisemblance d'un échantillon pour un jeux de valeurs des paramètres avec un processeur d'1 Ghz. Les carré noirs (NG1996) correspondent à l'estimation avec l'algorithme de Nath et Griffiths (1996). Les points noirs (DG2003) à l'estimation avec l'algorithme de de Iorio & Griffiths (2004b) pour un modèle mutationnel où les mutation sont indépendantes du types allélique parental (PIM).

5.2.2 Algorithme de De Iorio *et al.* (2004) pour les modèles de mutation par pas

Pour des raisons de temps de calcul, nous avons testé uniquement les performances de l'algorithme de de Iorio *et al.* (2004) pour deux populations et un modèle de mutation par pas (SMM, voir section 2.1.1). Cet algorithme est décrit en détails dans les annexes A-2 et B-4.

Afin d'évaluer la précision des estimations et l'effet du nombre de locus sur l'estimation des paramètres (θ, γ), nous avons simulé 10 jeux de données pour cinq locus indépendants et 10 jeux de données pour 20 locus indépendants. Cette étude est très limitée du fait des contraintes de temps liées à la lourdeur des calculs (voir annexe A-2). En effet, pour le peu de données analysées, cette étude a nécessité l'utilisation de 50 processeurs pendant

un mois. Ces simulations ont été réalisées avec l'aide des super-ordinateurs du Centre Informatique National de l'Enseignement Supérieur (CINES). Les jeux de données ont été simulés avec l'algorithme génération par génération décrit dans le chapitre 3, adapté pour un modèle à deux sous-populations de tailles égales ($N_1 = N_2 = 1000$ gènes) et une migration symétrique avec le paramètre $\gamma = 4Nm_{12} = 4Nm_{21} = 2.0$. La vraisemblance est calculée pour chaque locus en échantillonnant 500 000 généalogies pour 50 valeurs (θ_i, γ_i) du vecteur de paramètres (θ, γ) dans l'intervalle $[0, .., 10]^2$. Un point important est que ces 50 valeurs du vecteur de paramètres (θ, γ) sont simulées de manière aléatoire dans l'intervalle considéré. Les vraisemblances à chaque locus sont ensuite multipliées entre elles pour avoir la vraisemblance multilocus à chaque point des valeurs de paramètres (i.e. aux 50 valeurs (θ_i, γ_i)). Enfin, le maximum de vraisemblance est calculé à partir de la surface de vraisemblance obtenue à partir des 50 points de vraisemblance. Je ne détaillerai pas ici l'algorithme permettant d'extrapoler une surface de vraisemblance à partir des quelques points de vraisemblance ; le lecteur pourra en trouver les détails dans l'annexe B-4.

TAB. 5.3. Précision de l'estimation de (θ, γ) avec l'algorithme de de Iorio *et al.* (2004) et avec le Logiciel MIGRATE dans le cas d'un modèle à 2 populations avec un modèle de mutation par pas. (e.s) est l'erreur standard de l'estimation.

		Algorithme			
		De Iorio *et al.* 2004		MIGRATE	
		5 locus	20 locus	5 locus	20 locus
θ	Biais	-0.036	0.039	0.23	-0.60
	(e.s.)	(0.32)	(0.20)	(0.49)	(0.18)
	MSE	0.10	0.041	0.30	0.40
γ	Biais	0.31	0.30	1.2	0.25
	(e.s.)	(0.67)	(0.32)	(1.2)	(0.62)
	MSE	0.54	0.21	3.1	0.46

Le biais relatif, l'écart type relatif et le MSE relatif des estimations avec l'algorithme de de Iorio *et al.* (2004) et avec le logiciel MIGRATE sont présentés dans le tableau 5.3. Nos résultats montrent de bonnes performances de l'algorithme de de Iorio *et al.* (2004) pour l'estimation de θ. Aussi bien pour 5 que pour 20 locus, le biais et le MSE sont inférieur à 10%. En revanche, nos résultats mettent en évidence une sur-estimation du paramètre de migration γ. Quoiqu'il en soit, le MSE n'est pas très fort ce qui montre que les estimations sont assez proches de valeurs des paramètres du modèle de simulation.

On peut aussi noter que la précision augmente (i.e. le MSE diminue) avec le nombre de locus pour le paramètre de mutation et le paramètre de migration mais le biais n'est pas très sensible au nombre de locus. Le petit nombre P de points (θ_i, γ_i) pour lesquels la vraisemblance est estimée peut expliquer, au moins en partie, ce biais des estimations. Il est important de noter ici qu'un problème de cette étude est le petit nombre de jeux de données analysés. Nos résultats donnent donc des tendances globales sur la précision des méthodes d'estimations mais il est difficile d'évaluer de manière plus fine l'influence de chaque facteur (e.g. nombre de locus et nombre de points échantillonnés). Ainsi, des simulations avec 50, 100 et 200 points montrent que le biais de l'estimation de γ semble diminuer lorsque l'on augmente le nombre de valeurs des paramètres (θ, γ) (tableau 5.4). Cependant l'imprécision des estimations (i.e. erreur standard forte) fait que le MSE de l'estimation de γ est toujours relativement fort. De même, l'influence du nombre de valeurs des paramètres (θ, γ) sur l'estimation de θ n'est pas claire.

L'analyse des jeux de données simulés prend deux fois moins de temps avec MIGRATE, en utilisant la configuration par défaut, qu'avec l'algorithme de de Iorio *et al.* (2004). Cependant, nos simulations montrent que pour les jeux de données considérés et pour la configuration par défaut de MIGRATE, l'algorithme de de Iorio *et al.* (2004) donne de meilleures estimations des paramètres démographiques que MIGRATE. En effet, la précision de l'estimation du paramètre de migration γ par MIGRATE est assez mauvaise pour un petit nombre de locus, mais la précision de l'estimation augmente avec le nombre de locus (tableau 5.3). Nos résultats sur l'estimation de θ par MIGRATE sont surprenants. En effet, pour un petit nombre de locus, MIGRATE sur-estime le paramètre θ et la précision n'est pas très bonne (MSE de 30%) ; pour un plus grand nombre de locus, l'erreur standard de l'estimation diminue mais θ est alors fortement sous-estimé (tableau 5.3) ce qui donne un MSE plus fort que pour un petit nombre de locus.

En contrepoids des bons résultats d'estimation obtenus par l'algorithme de de Iorio *et al.* (2004), notre étude par simulation a montré qu'un facteur fortement limitant de cet algorithme est le temps de calcul. En effet, le calcul de la vraisemblance une paire de paramètres (θ_i, γ_i) prend près d'une heure avec un processeur de 1 Ghz. C'est pour cette raison que cette étude par simulation a été limitée à un faible nombre de jeux de données dans des conditions correspondant strictement au hypothèses du modèle. C'est pour cette raison aussi que nous avons utilisé que 50 valeurs θ_i, γ_i pour estimer la surface de vraisemblance. Il aurait été plus judicieux de considérer plus de points vu la largeur de l'intervalle considéré (i.e. $(\theta_i, \gamma_i) \in [1, .., 10]^2$). Cet

TAB. 5.4. Influence du nombre de points P d'estimation de la vraisemblance sur la précision de l'estimation de (θ, γ) pour cinq locus avec l'algorithme de de Iorio *et al.* (2004) dans le cas d'un modèle à 2 populations avec un modèle de mutation par pas. (e.s) est l'erreur standard de l'estimation.

		Algorithme		
		De Iorio *et al.* 2004		
		50 points	100 points	200 points
θ	Biais	-0.036	0.13	0.14
	(e.s.)	(0.32)	(0.16)	(0.23)
	MSE	0.10	0.042	0.074
γ	Biais	0.31	0.30	0.14
	(e.s.)	(0.67)	(0.65)	(0.70)
	MSE	0.54	0.51	0.50

effet du petit nombre de points échantillonnés pour construire la surface de vraisemblance est peut être accentué par le fait que l'algorithme choisisse des points de façon aléatoire sur l'intervalle choisi. En effet, lorsque le nombre de points est petit, le hasard peut faire que certaines gammes de valeur des paramètres sont mal échantillonnées. Enfin, une étude de la robustesse de l'algorithme de de Iorio *et al.* (2004) à des écarts aux hypothèses du modèle, notamment les hypothèses sur le nombre de sous-populations du système et sur le modèle mutationnel, serait nécessaire pour avoir une meilleur idée des performances de l'algorithme. Cependant, une telle étude paraît difficilement réalisable avec les moyens informatiques disponibles aujourd'hui.

5.3 Conclusions

Nous avons vu dans ces deux derniers chapitres comment la théorie de la coalescence permet l'estimation par maximum de vraisemblance de paramètres démographiques à partir d'un échantillon génétique sous différents modèles démographiques. La conclusion majeure de ces études est que le temps de calcul est le principal facteur limitant pour ces approches. En effet, avec les moyens informatiques disponibles aujourd'hui, il est déjà long d'estimer des paramètres démographiques sur un seul jeu de données, ceci même pour des modèles populationnels et mutationnels relativement simples. Il semble donc à fortiori difficilement envisageable de mener des études de robustesse par rapport aux écarts aux hypothèses des modèles. Les popu-

lations naturelles correspondant rarement aux descriptions que l'on peut en faire dans les modèles, de telles études de robustesse sont pourtant nécessaire à l'évaluation des performances des méthodes d'estimation de paramètres démographiques à partir de données génétiques.

Des deux approches que nous avons considérées, l'approche de Felsenstein et collaborateurs et l'approche de Griffiths et collaborateurs, seule la première était censée pouvoir être applicable à un grand nombre de modèles démographiques (nous verrons plus loin pourquoi ce n'est plus vraiment le cas depuis très récemment). En effet, le logiciel MIGRATE peut potentiellement considérer n'importe quel modèle démographique de population structurée à condition que les tailles des sous-populations soient assez grandes et les taux de migration suffisamment petits pour que l'approximation des temps de coalescence et de migration par une loi exponentielle soit valide. Nos résultats ont toutefois montré que cette approche ne donne pas toujours des estimations satisfaisantes avec la configuration par défaut de MIGRATE, aussi bien pour des modèles d'isolement par la distance que pour le modèle très simple à deux populations considéré lors de l'étude de l'algorithme de de Iorio *et al.* (2004). Ces résultats montrent l'importance de tester la précision de toute méthode d'estimation sous différents modèles démographiques mais également de tester leur robustesse vis à vis de différents facteurs tels que les processus mutationnels des marqueurs et l'influence d'un échantillonnage non exhaustif des sous-populations du système étudié (i.e. nombre de sous-populations échantillonnées versus nombre de sous-populations totales du système). L'effet de ce derniers facteur sur les estimations a été partiellement fait récemment par Beerli (2004). Les mauvaises performances des algorithmes de Felsenstein et collaborateurs peuvent être dues au grand nombre de paramètres à estimer. En effet cette approche devrait être plus performante pour l'estimation d'un petit nombre de paramètres. La considération d'une migration homogène dans l'espace pourrait largement diminuer le nombre de paramètres et, de fait, pourrait peut-être améliorer les performances de ces algorithmes sous des modèles d'isolement par la distance. Enfin, si l'on s'intéresse à l'estimation des paramètres démographiques actuels et locaux, il semble également nécessaire de tester la robustesse des méthodes par rapport à des hétérogénéités spatiales et temporelles des paramètres démographiques.

Il est important de noter que l'approche de Griffiths et collaborateurs est actuellement en plein développement. Des nouvelles distributions d'échantillonnage pondéré ont été trouvées assez récemment (Stephens & Donnelly, 2000) et ont montré de bonnes performances en populations panmictiques. L'adaptation théorique de ces nouvelles distributions d'échantillonnage pon-

déré en populations subdivisées est encore plus récente (de Iorio & Griffiths, 2004a,b, voir section 4.2.3), et les systèmes d'équations définissant ces distributions d'échantillonnage pondéré (éq.4.54) semblent relativement difficiles à résoudre pour des modèles mutationnels et démographiques généraux. A ce jour, quelques algorithmes ont été développés pour des modèles mutationnels simples (i.e. uniquement pour le modèles de mutation PIM, pour l'instant) et un modèle démographique à deux populations ou des modèles démographiques pour lesquels $\gamma_{ab} = x_a y_b,\ a \neq b$ ou encore des modèles démographiques pour lesquels la matrice de migration est réversible (de Iorio & Griffiths, 2004b). Un algorithme a aussi été développé dans le cadre du modèle de mutation dit "modèle à nombre de site infini" pour lequel chaque mutation crée un nouvel allèle inexistant dans la population (de Iorio & Griffiths, 2004b). Ce modèle est un équivalent du modèle IAM pour des données de type séquences d'ADN. L'intérêt de ce modèle est que l'on peut retrouver directement et de manière presque certaine la topologie de l'arbre de coalescence à partir des séquences. Cela permet d'augmenter considérablement les performances des algorithmes puisque la topologie de l'arbre n'est plus un paramètre de nuisance du modèle (i.e. un paramètre que l'on doit considérer mais que l'on ne cherche pas à estimer). Un autre développement de l'approche de Griffiths et collaborateurs est l'algorithme pour le modèle de mutation par pas considéré dans la section précédente. Cet algorithme est caractérisé par des temps de calcul beaucoup plus long que les autres algorithmes précédemment cités, dus aux calculs complexes nécessaires à la détermination des $\hat{\pi}(\cdot|\cdot, \mathbf{n})$. Un point intéressant de ces développements récents est que les algorithmes adaptés au modèle de mutation PIM donnent de très bon résultats et sont relativement rapides. Dans le cas d'une population panmictique, la considération de ce modèle de mutation fait que les $\hat{\pi}(\cdot|\mathbf{n})$ correspondent exactement aux $\pi(\cdot|\mathbf{n})$ définissant la fonction d'échantillonnage pondéré optimale (éq.4.41). Le calcul de la vraisemblance d'un échantillon ne nécessite alors qu'une seule généalogie et le calcul est alors instantané. De façon similaire, le modèle de mutation PIM diminue considérablement le nombre de généalogies à échantillonner pour avoir une estimation précise de la vraisemblance dans un modèle démographique de population structurée relativement simple (cf ci-dessus). Il serait donc intéressant de tester la robustesse de l'estimation avec ces algorithmes fondés sur le modèle PIM par rapport aux processus de mutation. Si cette robustesse est acceptable, ces algorithmes sont très prometteurs puisque des temps de calcul très courts peuvent permettre de prendre en compte des modèles démographiques beaucoup plus réalistes. De plus, de nouveaux développements en cours devraient permettre de prendre en compte beaucoup plus facilement tous types de modèle démographique. Les systèmes d'équations définissant les distributions

d'échantillonnage pondéré (éq.4.54) sont des systèmes d'équations linéaires que l'on peut résoudre numériquement par des techniques classiques (e.g. méthode du pivot de Gauss). Pour des modèles de mutation PIM, la résolution numérique de systèmes d'équations linéaires donne de bons résultats en termes de temps de calcul et quelques simulations préliminaires, sous isolement par la distance (modèle de migration "stepping stone") montrent de bonnes performance pour l'estimation de la migration et des tailles de populations avec des temps de calcul relativement courts par rapports aux autres algorithmes développés jusqu'à présent (François Rousset, communication personnelle). Cette approche me parait être relativement intéressante puisque cela permettrait de se focaliser sur l'optimisation de la résolution des systèmes d'équations définissant ces distributions d'échantillonnage pondéré (éq.4.54), valables pour n'importe quel modèle démographique. Là encore, la robustesse des estimations par rapport aux modèles mutationnels est extrêmement importante mais n'est pas acquise.

Une caractéristique commune aux deux approches est que, du fait de temps de calcul très longs, on a cherché à estimer la vraisemblance pour un ensemble de valeurs des paramètres du modèle $\mathcal{P}_i$ autour de $\mathcal{P}_0$ à partir de généalogies construites avec les paramètres $\mathcal{P}_0$. Stephens (1999) met en avant le fait que bien que $f_{BF}(G) \equiv \Pr(G;\mathcal{D}_0)\Pr(D|G;\mathcal{M})/\mathcal{L}(\mathcal{P}_0;D)$ soit la distribution d'échantillonnage pondéré optimale pour estimer la vraisemblance à $\mathcal{P} = \mathcal{P}_0$, la variance de l'estimateur de $\mathcal{L}(\mathcal{P};D)/\mathcal{L}(\mathcal{P}_0;D)$ (éq.4.8) peut devenir très grande et même infinie pour des valeurs de $\mathcal{P}$ éloignées de $\mathcal{P}_0$. Par exemple, pour des écart d'un facteur supérieur à 2 par rapport à $\mathcal{P}_0$ la variance de l'estimateur $\mathcal{L}(\mathcal{P};D)/\mathcal{L}(\mathcal{P}_0;D)$ est infinie et les vraisemblances sont alors largement sous-estimées (Stephens, 1999). Trouver une fonction d'échantillonnage pondéré efficace pour un grand ensemble de valeur des paramètres autour de $\mathcal{P}_0$ peut poser les mêmes problèmes avec l'approche de Griffiths et collaborateurs.

En conclusion générale de ce chapitre on peut dire que l'estimation de paramètres démographiques par maximum de vraisemblance utilisant la coalescence est très prometteuse mais au stade de développement actuel les résultats donnés par les méthodes disponibles, par MIGRATE entre autre, qui est largement distribué et utilisé par des généticiens des populations de terrain connaissant peu les caractéristiques mathématiques des algorithmes sous-jacents, doivent être considérés avec précaution. On peut néanmoins espérer que l'augmentation des capacités informatiques, le développements de nouveaux algorithmes, et la publication de tests de robustesse des méthodes devraient permettre dans un futur proche de bien meilleures estimations que

celles présentées ici.

Chapitre 6

Implications en biologie de la conservation : contraction spatiale d'habitat en population continue sous isolement par la distance

L'évaluation des risques d'extinction des populations est un problème central en biologie de la conservation. Comme nous l'avons vu en introduction, les populations isolées de petites tailles sont fortement sujettes aux variations de l'environnement, et souffrent de problèmes démographiques et génétiques pouvant accroître leur risque d'extinction. Les problèmes génétiques sont liés au phénomène de dérive entraînant une perte de diversité génétique et une augmentation du fardeau de mutation. Le niveau de diversité génétique est souvent mesuré dans les populations naturelles par le niveau d'*hétérozygotie individuelle* et par le nombre d'allèles observés à des locus neutres (l'hétérozygotie individuelle correspond au paramètre de diversité génétique, $(1-Q_0)$, utilisé dans la section 3.3). De nombreux auteurs ont stipulé que des données sur la variabilité génétique à des marqueurs neutres peuvent fournir des indications sur les risques d'extinction (e.g. Dunham *et al.*, 1999). De plus, ces patrons de variabilité génétique, résultant des caractéristiques passées et présentes des événements démographiques, peuvent fournir des informations sur les risques d'extinction difficiles à obtenir par des analyses purement écologiques et/ou démographiques, par exemple en permettant la détection d'une diminution de taille de population dans le passé. Bien que ce lien entre niveau de variabilité et risque d'extinction ne soit pas encore clairement établi, quelques études récentes ont montré des corrélations négatives entre l'hé-

térozygotie et le risque d'extinction chez des papillons, des drosophiles et des souris (Frankham, 1995, 1996; Frankham & Ralls, 1998; Saccheri *et al.*, 1998). La relation entre le niveau d'hétérozygotie et les tailles de populations est bien connue sous le modèle panmictique de Wright-Fisher. Ainsi pour des marqueurs neutres suivant un modèle de mutation de type IAM (voir section 2.1.1), l'hétérozygotie dans une population de Wright-Fisher à l'équilibre mutation-dérive est donnée par la relation $H = \theta/(1+\theta)$, avec $\theta = 4N_e\mu$ où N_e est le nombre d'individus diploïdes de la population (Crow & Kimura, 1970). De même, la diminution d'hétérozygotie en fonction du temps dans une population panmictique de taille N_e, après réduction, est donnée par la relation $H_t = H_0(1 - 1/(2N_e))^t$, où H_0 est l'hétérozygotie initiale à $t = 0$ et H_t l'hétérozygotie t générations plus tard (Crow & Kimura, 1970). Ces relations entre diversité génétique et tailles de population, et par extension entre diversité génétique et risques d'extinction, peuvent être utilisées en pratique pour évaluer des risques d'extinction par des *études comparatives*. Par études comparatives on entend des études fondées sur la comparaison entre des statistiques calculées sur plusieurs populations, dont certaines seront considérées comme "témoins". Cela consistera par exemple à comparer la variabilité génétique de la population étudiée (i.e. pour laquelle on veut évaluer les risques d'extinction) avec la variabilité génétique de populations témoins pour lesquelles des informations sur leur démographie passée indiquent une stabilité démographique. Ce sont donc des approches inter-populationnelles. Le même type d'approche peut être utilisé sur une seule et même population pour laquelle on a des données à différents moments dans le temps, par exemple, avant et après une fluctuation démographique. Par ailleurs, un autre type d'approche, que l'on appellera approches intra-populationnelles, consiste à tenter de détecter, à l'aide de données génétiques intra-population, des événements démographiques ayant affecté une population donnée (e.g. un goulet d'étranglement) sans référence à des populations témoins. Plusieurs méthodes ont été développées pour détecter des goulets d'étranglement à partir de données génétiques sous un modèle de Wright-Fisher. Certaines d'entre elles utilisent la théorie de la coalescence et estiment l'intensité du goulet d'étranglement par maximum de vraisemblance (voir par exemple Beaumont, 1999; Kuhner *et al.*, 1998; Griffiths & Tavaré, 1994). D'autres sont basées sur les déviations par rapport aux attendus en population stable de diverses statistiques telles que le nombre d'allèles, l'hétérozygotie, ou la distribution des allèles d'un échantillon (e.g. Cornuet & Luikart, 1996; Luikart & Cornuet, 1998; Reich & Goldstein, 1998; Garza & Williamson, 2001). Cependant, comme on l'a vu dans la section 2.2.2, chez de nombreuses espèces la dispersion est limitée dans l'espace entraînant une structuration en isolement par la distance. Par ailleurs, nous avons vu dans

la section 2.2.4 puis dans le chapitre 3 que les processus démographiques et évolutifs affectant la diversité génétique des populations ont des conséquences relativement différentes selon les modèles de structuration considérés. Il est donc intéressant d'évaluer l'influence d'une structuration en isolement par la distance sur les conséquence génétiques d'une réduction d'effectif résultant d'une contraction de l'habitat. Une contraction d'habitat pour une population continue en isolement par la distance peut être assimilée d'un point de vue démographique à une baisse du nombre d'individus dans une population de Wright-Fisher (i.e. goulet d'étranglement). Cette étude a été réalisée en collaboration avec Réjane Streiff du Centre de Biologie et de Gestion des Populations (CBGP) dans le cadre d'une étude de biologie de la conservation portant sur une espèce de criquet à dispersion limitée et dont l'habitat a été fortement détruit et fragmenté au cours du $20^{\text{ème}}$ siècle.

6.1 Réduction d'habitat en isolement par la distance

6.1.1 Caractéristiques génétiques d'une population continue à l'équilibre

Considérons dans un premier temps une population continue sous isolement par la distance n'ayant subi aucune variation temporelle ni spatiale des paramètres démographiques. Plus spécifiquement, intéressons-nous aux valeurs dans de telles populations de différentes statistiques couramment utilisées en génétique des populations, avec comme point de comparaison clef une population panmictique de Wright-Fisher (i.e. sans isolement par la distance) composée d'un même nombre d'individus. Ces statistiques sont : l'hétérozygotie individuelle, le nombre d'allèles dans l'échantillon et le F_{IS}. Toutes ces statistiques sont calculées à partir d'un échantillon de 100 individus pour différentes valeurs de tailles de populations. L'échantillon est pris sur une petite surface de (10×10) au milieu de la population. Les simulations ont été réalisées avec l'algorithme génération par génération adapté aux modèles considérés. Pour toutes les simulations de ce chapitre, le modèle mutationnel est le GSM avec une variance de la loi géométrique de 0.36 (voir section 2.1.1) et un taux de mutation de $5 \cdot 10^{-4}$ pour tous les locus. Les modèles démographiques que nous avons considérés sont un modèle de population panmictique de Wright-Fisher et trois modèles d'isolement par la distance en population continue en deux dimensions avec différentes distributions de

dispersion. Les moments d'ordre 2 des distributions considérées sont $\sigma^2 = 1$ (éq.3.1 avec les paramètres $M = 0.55$ et $n = 3.8$), $\sigma^2 = 4$ (éq.3.5) et enfin $\sigma^2 = 20$ (éq.3.1 avec les paramètres $M = 0.72$ et $n = 2.03$). Ces différentes distributions de dispersion nous permettent d'explorer des modèles avec un structuration en isolement par la distance de très forte ($\sigma^2 = 1$) à très faible ($\sigma^2 = 20$), et un modèle sans structuration (population de Wright-Fisher). Nous avons considéré des tailles de populations variant de 400 (20×20) à 250 000 (500×500) individus. Notons dès à présent que pour des tailles de population réduites l'isolement par la distance en population continue ne sera pas détectable pour certaines gammes de valeurs de σ^2. En effet, pour qu'un isolement par la distance soit effectif, il faut que la taille de l'habitat soit largement plus grande que les distances moyennes de dispersion (i.e. plus grande que $10\sigma^2$). En d'autres mots, il est attendu que pour des surfaces d'habitat très petites, une dispersion limitée aura peu d'effet sur la structuration de la diversité génétique et l'on pourra considérer que ces populations fonctionnent comme des populations de Wright-Fisher. Pour cette raison, la distribution de dispersion avec $\sigma^2 = 20$ n'a pas été considérée pour les tailles de populations de (20×20) individus.

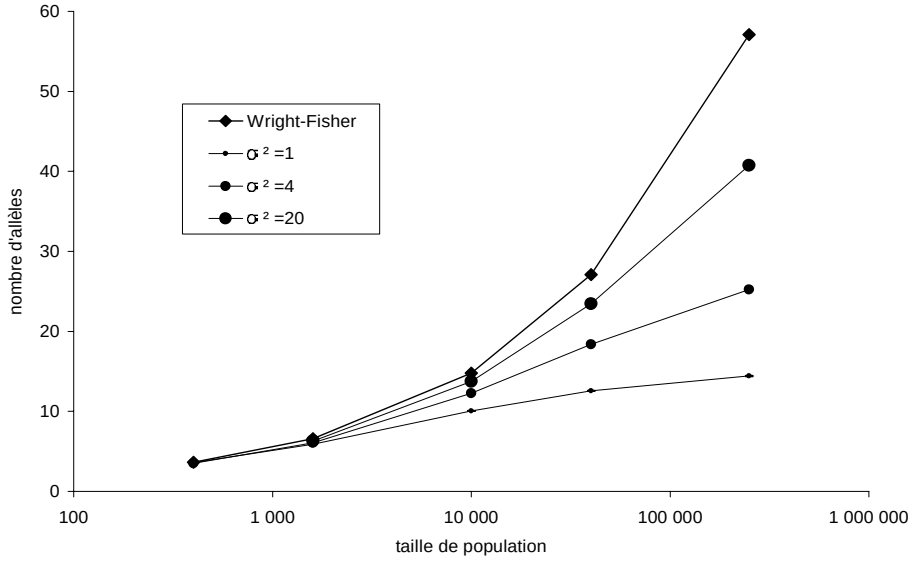

FIG. 6.1. Représentation du nombre d'allèles d'un échantillon local de 100 individus en fonction de la taille de la population et du modèle démographique considéré. Les écarts type des estimations sont représentés sur la figure mais ne se voient pas du fait de leurs faibles valeurs. L'échelle de l'axe des abscisses est logarithmique.

Comme le modèle en réseau sans effets de bord (i.e. sur un tore) n'est,

d'une manière générale, pas très réaliste, nous avons testé l'influence d'effets de bord de type réflectifs (i.e. effet miroir sur le bord de la population) et absorbant (i.e. la distribution de dispersion arrière est tronquée aux limites de la population). Nos simulations ont montré que les effets de bord ont peu d'effets sur les statistiques considérées, au moins pour les conditions étudiées ici, les courbes pour des bords réflectifs et absorbants étant surimposées (résultats non montrés). Nous présenterons donc dans la suite de cette étude uniquement les résultats pour des bords réflectifs. Les résultats de nos simula-

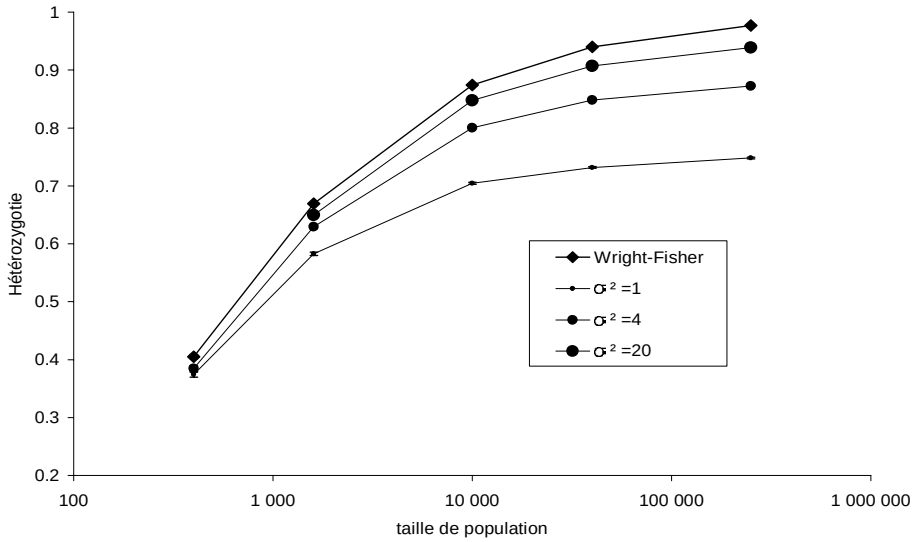

FIG. 6.2. Représentation de l'hétérozygotie d'un échantillon local de 100 individus en fonction de la taille de la population et du modèle démographique considéré. L'échelle de l'axe des abscisses est logarithmique.

tions sont présentés sur les figures 6.1, 6.2 et 6.3. Ces figures montrent qu'un isolement par la distance fort a un effet important sur les statistiques considérées. Ainsi, le nombre d'allèles dans un échantillon local sous isolement par la distance est, quel que soit σ^2, inférieur au nombre d'allèles d'un même échantillon pris dans une population de Wright-Fisher (Fig.6.1). Cependant, plus l'isolement par la distance est faible plus les valeurs des statistiques se rapprochent de celles obtenues sous un modèle de Wright-Fisher. Les même tendances sont observées pour l'hétérozygotie même si les résultats entre modèles sont moins contrastés (Fig.6.2). On notera également que les différences entre modèles se réduisent lorsque la taille totale de la population diminue. Ainsi, les valeurs du nombre d'allèles et de l'hétérozygotie sont très proches pour tous les modèles pour une taille de population de $(20 \times 20 = 400)$ individus.

L'ensemble de ces résultats est intuitivement attendu puisqu'un échantillon local sous isolement par la distance n'est pas représentatif de la population dans son ensemble mais uniquement de la zone échantillonnée, et ceci d'autant plus que la taille de la population est grande et σ^2 petit. A l'inverse, quel que soit l'échantillon pris dans une population de Wright-Fisher, cet échantillon, pour peu qu'il soit suffisamment grand (e.g. > 20 individus), sera représentatif de la population entière.

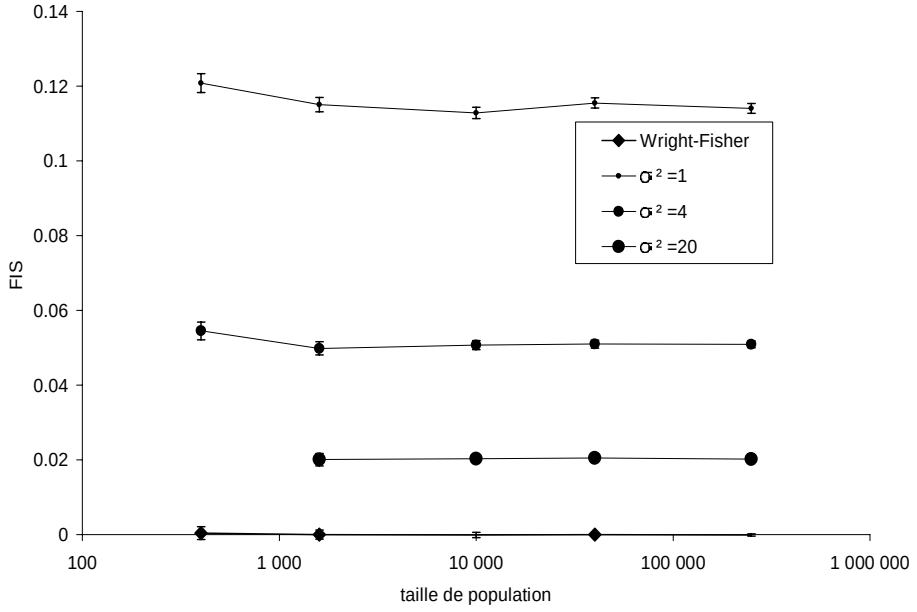

FIG. 6.3. Représentation du Fis d'un échantillon local de 100 individus en fonction de la taille de la population et du modèle démographique considéré. L'échelle de l'axe des abscisses est logarithmique.

Il est intéressant de noter que le nombre d'allèles et l'hétérozygotie à une échelle locale diminuent quand la structuration augmente (i.e. quand $D\sigma^2$ diminue). Ainsi, un petit nombre nombre d'allèles et une hétérozygotie faible peuvent être dus à une structuration forte couplée à un échantillonnage localisé et non pas nécessairement à des petites tailles de populations. Le fait de travailler sur des populations en isolement par la distance peut, par extension, avoir des conséquences en terme de stratégie d'échantillonnage. Ainsi, dans le cas d'isolement par la distance fort, si l'on veut avoir une idée de la diversité génétique d'une population, il est préférable d'échantillonner de façon homogène sur tout l'habitat occupé par la population, ou tout au moins sur une grande partie de l'habitat.

Nos résultats pour le F_{IS} montrent que plus l'isolement par la distance

est fort plus le F_{IS} est positif et élevé, indiquant une structuration au sein de l'échantillon d'autant plus forte que σ^2 est faible (Fig.6.3). Pour une population de Wright-Fisher, le F_{IS} est nul car on a aucune structuration au sein de l'échantillon. Il est intéressant de noter que la valeur du F_{IS} ne dépend pas de la taille de la population mais uniquement de $D\sigma^2$.

6.1.2 Influence d'une réduction de surface d'habitat

Considérons à présent une réduction dans le temps de la surface d'habitat. Cette réduction peut être considérée comme l'équivalent d'un goulet d'étranglement pour une population continue puisqu'elle se traduit par une diminution dans le temps de l'effectif total de la population. Les caractéristiques des modèles utilisés sont les mêmes que pour la section précédente. La réduction de la taille de la population se fait de manière instantanée en une génération à un moment G_c dans le passé (dans nos simulations G_c prend les valeurs 0, 10, 20, 100 et 200 générations). Nous avons considéré ici uniquement des réduction de surface récentes car nous nous sommes intéressés plus spécifiquement à des situations de réduction d'habitat correspondant aux périodes d'anthropisation majeures des milieux (i.e. deux derniers siècles). Nos simulations ont été restreintes à une réduction de surface d'un facteur 25 pour trois gammes de tailles de populations : (i) de très grandes populations (de 250 000 à 10 000 individus, respectivement avant et après réduction de la surface) ; (ii) des tailles de populations intermédiaires (de 40 000 à 1 600 individus) ; et (iii) des "petites" populations (de 10 000 à 400 individus).

Les statistiques calculées sont les mêmes que pour la section précédente. Cependant nous avons ajouté : (i) l'hétérozygotie attendue (i.e. la proportion attendue d'hétérozygotes en fonction des fréquences alléliques de l'échantillon, $\frac{n}{n-1}(1 - \sum_i p_i^2)$ où p_i est la fréquence de l'allèle i et n la taille de l'échantillon ; Nei, 1987) ; et (ii) la statistique ΔH qui calcul la différence entre l'hétérozygotie attendue calculée sur l'échantillon et l'hétérozygotie attendue pour un échantillon de la même taille avec le même nombre d'allèles dans une population de Wright-Fisher de taille quelconque (Cornuet et Luikart 1996). Nous avons considéré cette dernière statistique car elle est fréquemment utilisée pour détecter des goulets d'étranglement dans le cadre de populations panmictiques isolées (i.e. sans migration). En effet, Cornuet & Luikart (1996) ont montré qu'après un goulet d'étranglement récent la statistique ΔH prend des valeurs positives alors qu'elle a une valeur attendue nulle pour une population panmictique avec une démographie stable dans le temps.

TAB. 6.1. Effet d'une réduction de taille de population d'un facteur 25 sur le nombre d'allèles et l'hétérozygotie individuelle d'un échantillon en fonction du modèle démographique considéré et du moment G_c dans le passé auquel a eu lieu la réduction.

		G_c				
Structure		0	10	20	100	200
Réduction de 250 000 à 10 000 individus						
Wright-Fisher	Hétérozygotie	0.98	0.98	0.98	0.97	0.97
	(perte relative %)	0	0	0	1	1
	Nombre d'allèles	57	56	56	51	47
	(perte relative %)	0	1.8	1.8	11	18
$D\sigma^2 = 4$	Hétérozygotie	0.87	0.87	0.87	0.87	0.87
	(perte relative %)	0	0	0	0	0
	Nombre d'allèles	25	25	25	25	24
	(perte relative %)	0	0	0	0	4
$D\sigma^2 = 1$	Hétérozygotie	0.75	0.75	0.75	0.75	0.75
	(perte relative %)	0	0	0	0	0
	Nombre d'allèles	14	14	14	14	14
	(perte relative %)	0	0	0	0	0
Réduction de 40 000 à 1 600 individus						
Wright-Fisher	Hétérozygotie	0.94	0.94	0.93	0.92	0.89
	(perte relative %)	0	0	1.1	2.1	5.3
	Nombre d'allèles	27	26	25	20	17
	(perte relative %)	0	3.7	7.4	26	37
$D\sigma^2 = 4$	Hétérozygotie	0.85	0.84	0.84	0.83	0.80
	(perte relative %)	0	1.2	1.2	2.4	5.9
	Nombre d'allèles	18	17	17	15	13
	(perte relative %)	0	5.5	5.5	17	28
$D\sigma^2 = 1$	Hétérozygotie	0.73	0.73	0.73	0.72	0.71
	(perte relative %)	0	0	0	1.4	2.7
	Nombre d'allèles	13	12	12	12	11
	(perte relative %)	0	7.7	7.7	7.7	15.4
Réduction de 10 000 à 400 individus						
Wright-Fisher	Hétérozygotie	0.87	0.86	0.83	0.79	0.70
	(perte relative %)	0	1.2	4.6	9.2	20
	Nombre d'allèles	15	13	12	8.5	6.7
	(perte relative %)	0	13	20	43	55
$D\sigma^2 = 4$	Hétérozygotie	0.80	0.78	0.77	0.71	0.65
	(perte relative %)	0	2.5	4	11	19
	Nombre d'allèles	12	11	10	7.8	6.3
	(perte relative %)	0	8.3	17	35	48
$D\sigma^2 = 1$	Hétérozygotie	0.71	0.68	0.68	0.64	0.59
	(perte relative %)	0	4.2	4.2	10	17
	Nombre d'allèles	10	9.2	8.9	7.1	6.0
	(perte relative %)	0	8	11	29	40

Plaçons nous dans un premier temps dans le contexte d'étude comparative inter-population. Les résultats portant sur l'hétérozygotie attendue ne sont pas montrés car ils sont très proches de ceux obtenus pour l'hétérozygotie individuelle, $(1 - Q_0)$. De même, les résultats sur les F_{IS} ne sont pas présentés car, comme on l'a vu précédemment, cette statistique n'est pas influencée par des réductions de taille de population, que ce soit en population de Wright-Fisher ou en isolement par la distance (voir section 6.1.1). Les résultats portant sur les autres statistiques utilisée dans le contexte de comparaisons inter-populations sont présentés dans le tableau 6.1.

Nos résultats confirment qu'après une réduction d'effectif, l'hétérozygotie diminue plus lentement que le nombre d'allèles dans une population de Wright-Fisher (Cornuet & Luikart, 1996; Nei *et al.*, 1975). Cela est également le cas pour des populations en isolement par la distance, à condition que les populations ne soient pas trop grandes et que la structuration ne soit pas trop forte. La réduction du nombre d'allèles est globalement moins marquée en isolement par la distance qu'en population de Wright-Fisher. Ces résultats indiquent qu'il sera globalement plus difficile de mettre en évidence, lors d'études comparatives, des diminutions du nombre d'allèles en isolement par la distance avec un échantillon local qu'en population de Wright-Fisher. L'explication est à rechercher dans les conséquences d'un échantillonnage local en isolement par la distance. Un tel échantillonnage fait que le nombre d'allèles de l'échantillon sous-estime le nombre d'allèles total de la population en isolement par la distance. De plus, ce phénomène est d'autant plus fort que l'échantillon est petit et la population grande. Le nombre d'allèles de la population sera donc moins sous-estimé pour les populations après réduction que pour les populations stables, minimisant ainsi la diminution du nombre d'allèles après réduction par rapport aux populations stables sous isolement par la distance. Une autre façon de formaliser ce phénomène est de considérer qu'en isolement par la distance, la zone séparant la surface échantillonnée des limites de réduction spatiale d'habitat agit comme une "zone tampon". Cette notion de zone tampon traduit le fait qu'en isolement par la distance, le données génétiques d'un échantillon local n'ont de l'information uniquement sur ce qui se passe localement et pas sur des phénomènes évolutifs et démographiques ayant lieu à distance de la zone échantillonnée (voir également section 3.5.4).

La diminution de l'hétérozygotie est moins affectée par la structuration en isolement par la distance que la diminution du nombre d'allèles. En effet, pour toutes les tailles de populations l'hétérozygotie diminue de manière similaire en isolement par la distance et en population de Wright-Fisher. On notera

cependant une légère exception à ce résultat, à savoir le cas avec $\sigma^2 = 1$ et 1600 individus pour lequel l'hétérozygotie diminue moins rapidement en isolement par la distance qu'en population de Wright-Fisher.

Plaçons nous maintenant dans le contexte d'études intra-populationnelles. Les valeurs de ΔH en dessous de 1% ayant peu de chance d'être détectées, nos résultats montrent que cette statistique détectera une réduction de taille de population uniquement pour les petites tailles de population (e.g. population passant de 10 000 à 400 individus) (tableau 6.2). Que ce soit en population de Wright-Fisher ou en isolement par la distance, la statistique ΔH prend, dans des petites populations ayant subi une réduction d'effectif, des valeurs positives, supérieures aux valeurs prises dans les populations stables. Ces valeurs augmentent avec l'ancienneté de la réduction jusqu'à 200 générations. Le signal maximum est donc atteint à un minimum de 200 générations ou plus tard dans le passé. Le signal maximum étant d'autant plus loin dans le passé que les tailles de populations sont élevées, ceci explique que pour des plus grandes tailles de population, aucun signal sur ΔH n'est observé pour les temps courts étudiés. Il est intéressant de noter que dans les cas de populations moyennes et grandes, la statistique ΔH prend des valeurs négatives pour des populations en isolement par la distance à l'équilibre et ce d'autant plus que l'isolement par la distance est fort. Dans la mesure où

TAB. 6.2. Effet d'une réduction de taille de population d'un facteur 25 sur la statistique ΔH en fonction du modèle démographique considéré et du moment G_c dans le passé auquel a eu lieu la réduction.

	G_c				
	0	10	20	100	200
Réduction de 250 000 à 10 000 individus					
Wright-Fisher	0.0002	0.0002	0.0002	0.0004	0.0012
$D\sigma^2 = 4$	-0.013	-0.013	-0.013	-0.012	-0.012
$D\sigma^2 = 1$	-0.020	-0.020	-0.019	-0.019	-0.0018
Réduction de 40 000 à 1 600 individus					
Wright-Fisher	0.0015	0.0017	0.0018	0.0013	0.0002
$D\sigma^2 = 4$	-0.0062	-0.0061	-0.0057	-0.0025	-0.0001
$D\sigma^2 = 1$	-0.012	-0.010	-0.010	-0.007	-0.0037
Réduction de 10 000 à 400 individus					
Wright-Fisher	0.0035	0.010	0.015	0.033	0.039
$D\sigma^2 = 4$	-0.0077	0.013	0.017	0.037	0.050
$D\sigma^2 = 1$	0.0039	0.0087	0.014	0.038	0.049

un signal de réduction d'effectif se traduit par des valeurs positives de ΔH, les tests fondés sur de telles statistiques auront plus de difficultés à détecter une réduction de surface en isolement par la distance qu'en population de Wright-Fisher.

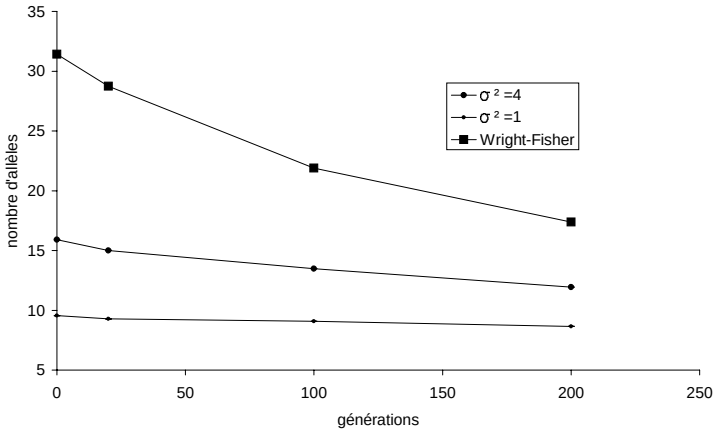

FIG. 6.4. Représentation du nombre d'allèles d'un échantillon local de 30 individus en fonction (i) du moment dans le passé auquel une réduction d'un facteur 100 a lieu et (ii) de la structuration de la population. La population passe d'une taille de 160 000 individus à 1 600 après réduction. Les écarts type des estimations sont représentés sur la figure mais ne se voient pas à cause de leur faibles valeurs.

D'une manière générale, nos simulations montrent qu'il sera globalement plus difficile de détecter, sur les statistiques considérées ici (i.e. hétérozygotie, nombre d'allèles et ΔH), un réduction d'effectif en isolement par la distance qu'en population de Wright-Fisher. Ce ne sont toutefois que des tendances et ces conclusions varient beaucoup en fonction des tailles de populations et de la force de la structuration en isolement par la distance. Ces résultats sont aussi fortement dépendants de la surface sur laquelle est pris l'échantillon en isolement par la distance et des statistiques utilisées. Ainsi, si l'on considère un échantillon encore plus petit, par exemple 30 individus sur une surface de (6×5) noeuds, le signal d'une réduction de surface sera encore moins marqué pour des populations en isolement par la distance, au moins pour les statistiques utilisées dans le cadre d'étude comparatives inter-populationnelles. Ceci est illustré par la figure 6.4 sur la quelle on voit que pour des populations de tailles intermédiaires, une forte réduction d'effectif (i.e. d'un facteur

100) n'a aucun effet sur le nombre d'allèles si la structuration en isolement par la distance est forte (cas du $\sigma^2 = 1$) alors que le nombre d'allèles décroît rapidement pour une population de Wright-Fisher.

Il est donc prévisible qu'un échantillonnage homogène sur la surface totale de la population, tout au moins sur une grande surface, devrait permettre de détecter plus facilement les effets d'une réduction d'effectif en isolement par la distance sur les statistiques considérées dans le contexte d'une étude comparative inter-populationnelle. En revanche, l'influence d'un échantillonnage à grande échelle est moins claire dans le contexte d'études intra-populationnelles, au moins pour celles fondées sur la statistique ΔH. Des simulations supplémentaires sont quoi qu'il en soit nécessaires pour évaluer clairement l'impact d'un échantillonnage à différentes échelles sur les deux types de statistiques.

6.2 Application au criquet de la Crau

Les données génétiques et démographiques présentées dans cette section ont été obtenue par Réjane Streiff et collaborateurs. Ma participation dans cette collaboration s'est faite uniquement en terme de simulations.

6.2.1 Présentation du modèle biologique et de son milieu

Le criquet de la Crau, *Prionotropis hystrix rhodanica* est un criquet endémique de la Crau, une plaine alluviale du sud de la France caractérisée par une faune et une flore originales (G., 1975; Foucart, 1997). *P.h. rhodanica* est une espèce univoltine (i.e. à une génération par an) dont les capacités de dispersion sont extrêmement limitées puisque que les deux sexes possèdent des ailes atrophiées et ne peuvent donc pas voler. Depuis sa première observation en 1919, *P.h. rhodanica* est considéré comme une espèce rare (R. & Rambier, 1950; Foucart & Lecoq, 1996; Vayssière, 1921). Son aire de distribution actuelle est restreinte à quelques kilomètres carrés dans la plaine de la Crau et il n'a jamais été observé ailleurs malgré de nombreuses prospections (Uvarov, 1923; Foucart & Lecoq, 1996). Son habitat, dénommé localement le "coussou", est une steppe semi-aride sur laquelle on a développé des élevages ovins extensifs. Cet habitat a été modifié depuis le 16$^{\text{ème}}$ siècle par la mise en place d'une agriculture irriguée. Du fait de ces modifications, le coussou

est passé d'une surface initiale estimée à 600 km^2 à une surface actuelle de moins de 100 km^2 (Wolff, 2001). De plus, cet habitat a été fortement fragmenté par la mise en place de parcelles cultivées, de canaux d'irrigation, de routes et par une urbanisation croissante (Wolff, 2002). L'effet de la fragmentation du coussou sur les espèces qu'il abrite ne dépend pas seulement des caractéristiques spatiales de cette fragmentation mais aussi des caractéristiques démographiques et évolutives de chaque espèce. Dans ce contexte, *P.h. rhodanica* parait spécialement concerné par cette évolution du coussou de par sa rareté, ses faibles capacités de dispersion et son étroite dépendance vis à vis de ce milieu.

6.2.2 Caractéristiques de la population échantillonnée

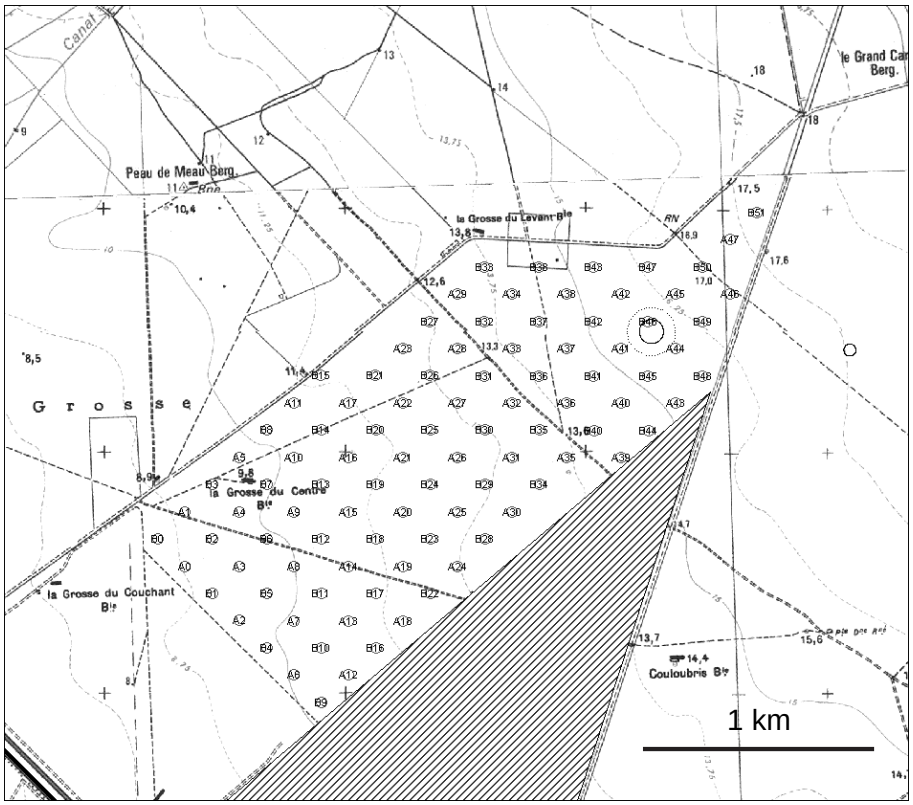

FIG. 6.5. Représentation de la parcelle échantillonnée. Les petits cercles correspondent aux deux grilles A et B échantillonnées.

Quatre vingt un disques de 25 m de rayon ont été échantillonnés de la façon la plus exhaustive possible. Ces 81 disques sont positionnés sur deux

grilles A et B légèrement décalées couvrant une surface totale d'environ 4 km^2 (Fig.6.5). Cette surface correspond approximativement à un des derniers fragments de coussou sur lequel des criquets ont été observés. Étant donné les capacités de dispersion très limitées de *P.h. rhodanica*, on peut faire l'hypothèse que cette surface correspond à une population isolée, car les habitats aux alentours ne sont pas propices au criquet. Lors de l'échantillonnage de la grille A, une zone "vide" (i.e. sans criquet) vers le sud-ouest a été repérée et la grille B a donc été échantillonnée seulement sur la zone nord-est où des criquets avait été trouvés. Le résultat final est illustré sur la figure 6.6 sur laquelle sont reportées les densités de criquets observées dans chaque cercle échantillonné. Cet échantillonnage quasi-exhaustif a permis de déterminer des densités de criquets observés par cercle. Ces densités ne correspondent pas exactement à la densité de criquet sur le coussou (i.e. probabilité de détection inférieure à 1) ni à des densités efficaces mais nous permettent d'avoir un ordre de grandeur des densités de la zone échantillonnée. La den-

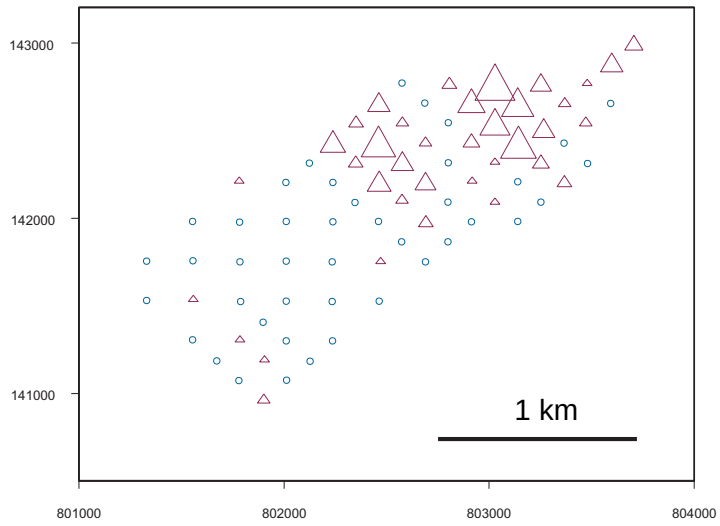

FIG. 6.6. Représentation des densités de criquets observés sur les cercles échantillonnés. Les triangles correspondent aux cercles sur lesquels au moins 1 criquet a été observé et leur taille est proportionnelle au nombre de criquets observés.

sité moyenne sur la zone d'échantillonnage est de 1400 individus/km^2. En considérant que la surface du fragment échantillonné est d'environ 4km^2, la taille de la population est alors d'environ 5600 individus. Cette taille de population se situe entre les tailles considérées pour des grandes et moyennes populations, respectivement 10 000 et 1 600 individus (tableau 6.1). On se rapprocherait toutefois plus des petites populations (1 600 individus) car il est probable que le nombre de criquets observés corresponde à une surestimation du nombre d'individus efficaces de la population. En considérant que l'habitat initial était moins fragmenté que maintenant, on peut considérer que la diminution de la surface d'habitat de cette population est d'un facteur proche de 25 ; le cas du criquet de la Crau correspond donc grossièrement aux cas des moyennes et grandes populations étudiés par simulation dans la section précédente.

Du point de vue génétique, 270 individus répartis sur toute la surface de la population ont été génotypés pour 10 locus microsatellites. Le nombre d'allèles varie de 4 à 44 selon les locus avec une moyenne de 19.8 allèles par locus et l'hétérozygotie moyenne est de 0.77. L'équilibre de Hardy Weinberg est rejeté au seuil de 5% et le F_{IS} est de 0.08. Enfin, un test de goulet d'étranglement avec le logiciel BOTTLENECK (Cornuet & Luikart, 1996) ne détecte aucune déviation significative de la statistique ΔH par rapport aux attendus (le ΔH moyen sur tous les locus est de -0.037 en faisant l'hypothèse d'un modèle GSM avec une variance de 0.36 pour tous les locus).

Pour ce qui est de l'isolement par la distance, un test d'isolement par la distance par test de Mantel sur les données de différenciation génétique (a_r, méthode de Rousset, 2000) et la distance géographique est hautement significatif ($p<10^{-5}$), montrant une forte structuration par la distance due à la dispersion limitée de cette espèce. Le traitement des données avec la méthode de Rousset (2000) donne une estimation ponctuelle de $D\sigma^2 = 3.58$ individus/m^2, ce qui donne, en prenant comme estimation de la densité 0.0014 criquets/m^2, un σ^2 de 2560 m^2 (ou encore $\sigma \approx 50$ m) . Ce résultat est en accord avec des données de marquage-recapture qui avait montré une dispersion moyenne de quelques dizaines de mètres par générations. On est donc dans un cas correspondant aux modèles simulés pour des populations moyennes et grandes avec un $D\sigma^2$ d'environ 4. Les valeurs du nombre d'allèles et de l'hétérozygotie correspondent en effet grossièrement à celles données dans le tableau 6.1 pour $D\sigma^2 = 4$ et des tailles de population après réduction de 1 600 ou 10 000. Pour ces caractéristiques démographiques, avec une réduction de surface ayant eu lieu entre quelques dizaines et quelques centaines de générations, il est attendu qu'aucun signal de goulet d'étranglement ne soit

détecté par la méthode de Cornuet & Luikart (1996). On est même dans le cas où, sous isolement par la distance, les valeurs de ΔH sont potentiellement négatives, réduisant d'autant plus la probabilité de détection du signal de réduction d'effectif (tableau 6.2). En accord avec nos simulations, le ΔH moyen calculé sur les dix locus génotypés chez *Prionotropis hystrix rhodanica* est négatif (i.e. -0.037).

En conclusion, nos simulations ont permis de montrer que, malgré des caractéristiques génétiques ne correspondant à priori pas aux attendus pour des espèces ayant subi de fortes réductions d'effectif (i.e. nombre d'allèles et hétérozygotie faible, détection de goulet d'étranglement par des méthodes génétiques), cette population de criquets de la Crau, par ses caractéristiques génétiques, peut parfaitement correspondre à une population structurée par un isolement par la distance fort et ayant subi une contraction d'habitat importante et relativement récente. Ces résultats illustrent le fait que l'idée, largement répandue, comme quoi des populations ayant subi des baisses d'effectifs importantes ont des diversités génétiques plus faibles que des populations ayant un effectif stable (Frankham, 1996), n'est pas toujours valable, surtout dans le cas d'espèces avec des grandes tailles de populations et un dispersion limitée entraînant une structuration en isolement par la distance. Ceci sera d'autant plus vrai que l'échantillon sera récolté localement sur une petite surface ne correspondant pas à la surface totale de la population étudiée. Pour ce qui est de la taille des populations, les insectes ont souvent des populations de très grandes tailles et certaines espèces peuvent être considérées comme menacées malgré des tailles de populations encore relativement grandes. C'est le cas du criquet de la Crau puisque la population échantillonnée contient encore quelques milliers d'individus maintenant ainsi une diversité génétique élevée malgré une très forte réduction de la surface de son habitat. Chez de telles espèces, les réductions de taille de population seront difficilement détectables avec les outils génétiques disponibles actuellement quelle que soit la structuration des ces populations, et ceci d'autant plus que les réductions de taille sont récentes (tableau 6.1 et 6.2).

6.3 Conclusions

Cette étude fondée sur la simulation nous a permis de mettre en évidence quelques conséquences de la structuration en isolement par la distance sur des statistiques communément utilisées en génétique des populations dans un contexte de biologie de la conservation. Un des résultats majeurs de notre

étude est que la structuration en isolement par la distance fait qu'un échantillon local n'est pas représentatif de la population entière et que les événements démographiques localisés loin de la zone échantillonnée auront peu de répercussions sur les caractéristiques génétiques de cette zone. Nos simulations ont montré que, dans certains cas au moins, cela pouvait avoir des conséquences relativement importantes sur la détection de réduction de tailles de populations via une réduction d'habitat. Ainsi si l'on s'intéresse au signal d'une réduction d'effectif de la population sur des statistiques telles que le nombre d'allèles ou l'hétérozygotie, un échantillonnage classique d'une trentaine d'individus sur une petite surface ne permettra pas de mettre en évidence cette baisse d'effectif si la population étudiée est relativement grande et fonctionne en isolement par la distance. Par conséquent, si l'on s'intéresse à la diversité génétique d'une espèce à dispersion limitée, il est fortement recommandé d'échantillonner sur une surface la plus large possible. Ces résultats ne sont valides que pour des grandes tailles de populations et une structuration forte. En effet, plus la population est petite, et/ou plus l'isolement par la distance est faible, plus l'échantillon est représentatif de cette population dans sa totalité. Dans le cas de petites populations, la prise en compte de la structuration en isolement par la distance n'est donc pas nécessaire.

Des outils pour détecter des réductions d'effectifs ont été développés dans le cadre de modèles de Wright-Fisher. Nos simulations montrent que leur utilisation sur des populations en isolement par la distance risquent de fortement limiter leur pouvoir de détection dans certaines circonstances (population moyenne ou grande et isolement par la distance fort). Pour ces situations, il parait nécessaire de développer des outils spécifiques aux populations structurées en isolement par la distance. Toutefois si l'on considère des petites populations et/ou un isolement par la distance faible, les différences avec une population de Wright-Fisher sont mineures et l'utilisation d'outils spécifiquement développés pour des populations de Wright-Fisher devrait donner des résultats corrects. Étant donné le nombre limité de situations et des statistiques considérées ici, il sera cependant nécessaire de tester plus précisément l'effet d'une structuration en isolement par la distance sur les différentes méthodes utilisées en prenant en compte les caractéristiques démographiques et génétiques des systèmes biologiques étudiés.

Chapitre 7

Conclusions générales et perspectives

Nous avons vu dans ce document comment la formalisation mathématique de modèles de génétiques des populations permet une meilleure compréhension des phénomènes évolutifs. Nous avons également vu comment, dans certains cas, cette formalisation mathématique permet l'estimation de paramètres démographiques à partir d'échantillons du polymorphisme neutre des populations naturelles. Tout au long de ce document je me suis intéressé uniquement à la variabilité génétique neutre, et non à celle sous sélection. D'un point de vue purement théorique, la notion de variabilité génétique neutre est utile pour comprendre l'effet des différentes forces évolutives que sont la mutation, la dérive, la migration, la recombinaison et les systèmes de reproduction sur la distribution du polymorphisme génétique. Je ne discuterai pas ici de l'utilisation de la variabilité génétique sélectionnée, même si cette dernière permet des inférences intéressantes, par exemple dans un contexte de biologie de la conservation, sur les niveaux de variabilité nuisible (i.e. associées aux effets délétères de certains allèles) ou adaptative (i.e. associé aux effets favorables de certains allèles) dans les populations naturelles. Le lecteur pourra trouver plus de détails sur l'utilisation de la diversité neutre versus sélectionnée dans Hedrick (2001) et Luikart *et al.* (2003).

Jusqu'à très récemment, la quasi-totalité des méthodes d'estimation de paramètres démographiques à partir de données génétiques se sont focalisées sur des modèles démographiques simples pour lesquelles les traitements mathématiques sont relativement faciles à développer. Ces modèles ont comme principal défaut de ne pas décrire de manière réaliste les caractéristiques

des populations naturelles, l'exemple le plus symptomatique étant le modèle en îles. Les modèles d'isolement par la distance sont des exceptions dans la mesure où ils permettent de prendre en compte deux caractéristiques fondamentales des populations naturelles observées chez de nombreuses espèces qui sont : (i) deux individus ont plus de chances de se reproduire entre-eux s'ils sont proches géographiquement et si la dispersion parents-descendants est limitée dans l'espace, et (ii) les individus ne forment pas forcément des groupes (que l'on assimile généralement à des populations) mais peuvent être répartis de façon plus ou moins homogènes sur l'habitat ; c'est la notion de population continue. Dans cette section je reviendrai, dans un premier temps, sur ces modèles d'isolement par la distance, puis je discuterai de l'intérêt des méthodes d'estimation par maximum de vraisemblance. J'envisagerai ensuite quelques alternatives méthodologiques possibles pour l'estimation jointe des paramètres de migration et des tailles de populations et je finirai par quelques perspectives pour l'estimation de paramètres démographiques dans des modèles plus réalistes et complexes.

7.1 Validation du modèle d'isolement par la distance et de la méthode d'estimation de $D\sigma^2$

Une formalisation mathématique rigoureuse des modèles d'isolement par la distance a été faites par Malécot (1950, 1967, 1975); Maruyama (1972); Nagylaki (1976, 1989) et Sawyer (1977). Celle ci a permis le développement de méthodes d'estimation de paramètres démographiques, tels que $D\sigma^2$, à partir de données génétiques selon une approche fondée sur les F-statistiques (cf. méthode des moments, Rousset, 1997, et voir Rousset, 2000 pour une adaptation aux modèles de populations continues). Les analyses par simulation présentées dans le chapitre 3 de ce document ont permis de montrer que la méthode des moments en population continue est relativement robuste à des écarts aux hypothèses du modèle sous-jacent en terme de processus mutationnels et d'hétérogénéités spatiales et temporelles des paramètres démographiques, tout au moins dans la limite des cas considérés. La robustesse globale de cette méthode peut être attribuée, entre autre, à certaines caractéristiques des modèles d'isolement par la distance et de la méthode elle même. Plus précisément, rappelons que la faible influence sur l'estimation de $D\sigma^2$ des processus mutationnels et des hétérogénéités spatiales et temporelles des paramètres démographiques peut être attribuée (i) aux taux de migration forts et aux petites tailles de dèmes des modèles d'isolement par la

distance en populations continues (i.e. généralement un individu par dème), et (ii) à l'échelle locale à laquelle est fait l'échantillonnage. De manière générale, les F-statistiques sont faiblement dépendantes des taux de mutation et des fluctuations démographiques passées à des échelles spatiales petites, ce qui avait déjà été noté par plusieurs auteurs dont Crow & Aoki (1984) et Slatkin (1993). Cette indépendance permet d'interpréter plus facilement les résultats d'une analyse par F-statistiques en s'affranchissant de l'histoire démographique passée et des processus mutationnels généralement mal connus.

Il est intéressant de noter que les bonnes performances de la méthode des moments sont confortées par plusieurs comparaisons entre des estimation indirectes (avec cette méthode) et directes (par des méthodes démographiques) (tableau 7.1). Ces comparaisons montrent que, sur ces jeux de données réels, les estimations directes et indirectes donnent des résultats similaires validant ainsi la méthode de Rousset (1997, 2000). Par ailleurs, ces comparaisons entre estimations directes et indirectes valident globalement les modèles d'isolement par la distance. En effet, pour la première fois, des estimations indépendantes par des méthodes génétiques et démographiques donnent des estimations similaires montrant, d'une certaine manière, que le modèle sous-jacent aux estimations génétiques est un modèle relativement réaliste.

TAB. 7.1. Comparaisons entre estimations directes et indirectes du produit $D\sigma^2$ sur différents jeux de données.

	estimation de $D\sigma^2$		référence
	directe	indirecte	
rats kangourou	1.43	2.58	Rousset (2000)
escargots	$6.0{\cdot}10^4$	$9.0{\cdot}10^4$	Rousset (1997)
humains	29.3	21.1	Rousset (1997)
lézards	11.5	5.5	Sumner *et al.* (2001)
légumineuse	9.6	13.9	Fenster *et al.* (2003)

7.2 Estimation par F-statistiques versus maximum de vraisemblance

En dépit des bonnes performances générales de la méthode des moments pour l'estimation de $D\sigma^2$, les approches par F-statistiques montrent certaines

limites, dont les deux principales sont :

(i) Lorsqu'on que l'on considère un ensemble de sous-populations de grandes tailles et/ou des taux de migration faibles, les F-statistiques ne sont plus indépendantes des processus mutationnels et des fluctuations démographiques démographiques passées. L'interprétation des paramètres démographiques estimés est alors problématique, notamment lorsque des variations temporelles des paramètres démographiques sont suspectées.

(ii) La formalisation mathématiques des modèles en termes de F-statistiques ne permet pas l'estimation de tous les paramètres du modèles. Par exemple, l'analyse des modèles d'isolement par la distance rend possible l'estimation du produit $D\sigma^2$ mais pas l'estimation de D et σ^2 indépendamment. De plus, même si une estimation indépendante de D est parfois possible, l'estimation des caractéristiques de la distribution de dispersion, autre que σ^2, semble difficile avec cette approche (voir éq.2.18 et 2.19).

L'estimation par maximum de vraisemblance fondées sur la coalescence ne devrait pas comporter, en principe, de telles limites. En effet, la prise en compte de fluctuations temporelles des paramètres démographiques est aussi théoriquement possible avec ces méthodes. De plus, comme on l'a vu dans le chapitre 4, l'estimation par maximum de vraisemblance permet théoriquement d'estimer tous les paramètres mutationnels et démographiques d'un modèle donné, pour peu que l'on sache estimer la vraisemblance des paramètres pour un jeux de données dans ces modèles et qu'il y ait l'information nécessaire dans les données génétiques. Un avantage important des méthodes par maximum de vraisemblance est qu'elles permettent une meilleure mesure de l'incertitude sur les paramètres estimés par des tests d'hypothèses (e.g. tests sur des valeurs de paramètres ou tests de modèles). En effet, sous l'hypothèse que la vraisemblance peut être exprimée comme un produit de variables aléatoires indépendants (i.e. les vraisemblances pour chaque locus), la distribution du logarithme de la vraisemblance peut être considérées comme une distribution normale dans la mesure où l'on a alors une somme de variables aléatoires indépendantes, et de fait les distributions du χ^2 peuvent être utilisées pour des tests de statistiques et des tests de modèles. Ces tests sont connus sous le nom de *tests de rapports de vraisemblances* (de l'anglais log-likelihood ratio tests, Cox & Hinkley, 1974).

Cependant, dans les modèles génétiques tels que ceux considérés dans ce document, la vraisemblance d'un échantillon est fonction des taux relatifs de coalescence, de migration et de mutation, fonction respectivement de la taille des populations N, des taux de migration m et des taux de mutation μ. De

fait, si un de ces trois paramètres n'est pas connu à priori, ces paramètres ne pourront pas être estimés indépendamment les uns des autres mais uniquement sous forme de deux produit (i.e. $(Nm, N\mu)$, $(N\mu, m\mu)$ ou encore $(Nm, m\mu)$). De plus, nos études par simulation ont montré que les performances des méthodes par maximum de vraisemblance sont encore loin des performances attendues. En effet, les méthodes de Felsenstein et collaborateurs fondées sur des algorithmes de Metropolis-Hastings (MH) implémentées dans le logiciel MIGRATE sont sensées pouvoir estimer des distributions de dispersion ; cependant ces méthodes montrent des performances médiocres sous des modèles d'isolement par la distance et nécessitent, d'autre part, des temps de calcul très longs. Il est possible que ces mauvaises performances soient liées au fait que l'on cherche à estimer un grand nombre de paramètres (François Rousset, communication personnelle). La considération de modèles d'isolement par la distance homogènes dans l'espace (i.e. dispersion et tailles de populations homogènes dans l'espace) pourrait réduire le nombre de paramètres à estimer et ainsi améliorer les performances de ces algorithmes par rapport aux méthodes de Rousset (1997, 2004).

Les algorithmes de Griffiths et collaborateurs fondés sur des techniques d'échantillonnage pondéré, constituant la seconde approche d'estimation par maximum de vraisemblance disponible actuellement, s'avèrent pour le moment limitées à des applications considérant des modèles mutationnels et démographiques relativement simples, du fait de temps de calcul extrêmement longs.

Les temps de calcul représentent donc un facteur limitant important pour ces méthodes par maximum de vraisemblance fondées sur la coalescence, même pour des modèles évolutifs relativement simples. L'estimation de paramètres démographiques sur de gros jeux de données demande des moyens informatiques importants (i.e. cluster d'ordinateurs et/ou stations multi-processeurs) qui ne sont généralement pas disponibles dans les laboratoires de biologie menant des études essentiellement empiriques. En particulier, la précision et la robustesse de ces nouvelles méthodes sont difficilement testables sur un grand nombre de jeux de données simulés. On notera que de plus en plus de méthodes, souvent complexes, d'estimation de paramètres démographiques sont publiées sans aucune étude de précision ni de robustesse ce qui nuit à l'interprétation des estimations obtenues avec ces méthodes (voir par exemple notre étude sur MIGRATE).

On peut concevoir plusieurs moyens de réduire les temps de calcul. Premièrement, les temps de calcul extrêmement longs des algorithmes de Griffiths et collaborateurs sont dus pour l'essentielle aux calculs lourds des fonc-

tions d'échantillonnage pondérés. Il faut donc favoriser des fonctions d'échantillonnage pondérés pour lesquelles les temps de calcul sont les plus courts possible. Une seconde manière de diminuer les temps de calcul des méthodes d'estimation par maximum de vraisemblance consiste à simplifier les modèles dans le passé. En effet, ces méthodes utilisent des arbres de coalescence dont la plupart remontent très loin dans le passé (e.g. au moins quelques dizaines de milliers de générations, ceci même pour des tailles de populations relativement faibles). Il semble intuitivement inutile de considérer un modèle démographique complexe sur de telles échelles de temps alors qu'il est probable que les caractéristiques démographiques dans le passé lointain aient peu d'influence sur les estimateurs des paramètres nous intéressant (taux de migration et tailles de populations actuels, i.e. au moment de l'échantillonnage), en particulier pour des modèles avec des petites tailles de populations et des forts taux de migration. Il semblerait donc intéressant de développer des méthodes d'estimation pour lesquelles la complexité du modèle démographique est réduite progressivement ou instantanément à partir d'un certain moment dans le passé. Cependant, cette simplification est susceptible d'entraîner une perte d'information et donc une diminution de la précision des estimations. C'est pourquoi la simplification du modèle dans le temps devra être couplée à une méthode d'évaluation de la précision d'estimation en fonction du niveau de simplification considéré. Seules les simplifications n'entraînant aucune perte d'information, ou une diminution faible de la précision des estimations, seront acceptables. Une telle approche de réduction de la complexité des modèles dans le passé est actuellement en cours de développement mais son efficacité n'a pas encore été testée (Stuart Baird, communication personnelle).

Bien que, dans une certaine mesure, les limitations dues aux temps de calcul semblent surmontables à court ou moyen terme grâce à l'utilisation de méthodes de plus en plus efficaces, à une simplification des modèles dans le temps et l'utilisation d'ordinateurs de plus en plus puissants, les difficultés liées à l'estimation d'un grand nombre de paramètres semblent plus difficiles à résoudre.

7.3 Estimation séparée des tailles de population et des caractéristiques de dispersion

Un limite générale des précédentes approches d'estimation de paramètres démographiques, que ce soient celles fondées sur les F-statistiques ou sur

des arbres de coalescence, est qu'elles permettent uniquement l'estimation du produit des taux de migration par les tailles de populations (i.e. Nm ou $D\sigma^2$). Or, dans de nombreux cas, il est intéressant de séparer l'information sur la dispersion de l'information sur les tailles de populations, permettant ainsi de distinguer les effets des flux de gènes des effets de la dérive sur les patrons de polymorphisme. Récemment, Wang & Whitlock (2003) ont développé une méthode par maximum de vraisemblance fondées sur les fréquences alléliques, utilisant sur un double échantillonnage dans le temps (i.e. sur différentes générations), pour estimer séparément des taux de migration et des tailles de populations. Quelques tests par simulation montrent de bonnes performances de cette méthodes aussi bien pour l'estimation des tailles de populations que pour les taux de migration. Cependant leur méthode n'est valide que pour un modèle en îles. De plus cette méthode ne semble valide que pour des échantillons n'ayant qu'un petit nombre d'allèles et ne permet donc pas de considérer des locus présentant un fort niveau de polymorphisme tels que les microsatellites. Des simulations complémentaires serait donc nécessaires pour mieux évaluer la précision et la robustesse de la méthode à différents facteurs tels que les processus de mutation et/ou des hétérogénéités spatiales et temporelles des paramètres démographiques.

Vitalis & Couvet (2001b) ont développé une méthode fondées sur des F-statstiques pour estimer séparément des taux de migration et des tailles de populations. Leur approche est fondée sur l'utilisation de l'information présente dans les déséquilibre de liaison en prenant en compte des probabilités d'identité à un et deux locus (Vitalis & Couvet, 2001a). Cette approche a été évaluée par simulation et semble donner de bons résultats dans un modèle en îles, au moins pour certaines valeurs des paramètres de migration correspondant à un niveau de migration moyen (Vitalis & Couvet, 2001b). Une conclusion majeure de leur étude par simulation est que les estimations des taux de migration m et des tailles de populations N restent fortement corrélées entre elles, montrant que l'information statistique permettant d'estimer N et m séparément est limitée, notamment dans le cas de locus indépendants. La raison majeure pour laquelle la majorité des méthodes développées jusqu'à présent ne permettent que l'estimation du produit de la migration par les tailles de populations est que la vraisemblance d'un échantillon dépend essentiellement du produit de la migration par les tailles de populations, et peu des valeurs des paramètres de migration et de tailles de populations considérés indépendamment. Toutefois, du fait des rapides progrès en génomique et du nombre grandissant de génomes séquencés, nous aurons de plus en plus accès à des données génétiques issues de marqueurs physiquement liés (i.e. proches sur le même chromosome) et pour lesquels les taux de recom-

binaison seront connus. Dans ce contexte, il paraît intéressant de développer des méthodes, telles que la méthode de Vitalis & Couvet (2001b), prenant en compte les déséquilibre de liaisons. En effet, l'information présente dans les déséquilibres de liaison permet, dans certains cas au moins, d'avoir plus de précision sur les estimations de tailles de populations (Vitalis & Couvet, 2001b; Waples, 1991) et sur les estimations de taux d'introgression lors de l'étude de populations hybrides (Jean-Marie Cornuet, communication personnelle).

Une approche alternative possible pour estimer uniquement les paramètres de dispersion (idéalement la distribution de dispersion complète) consisterait à utiliser une stratégie de double échantillonnage dans le temps, en échantillonnant la population avant et après les événements de dispersion. En effet, puisque la dispersion est l'unique force évolutive déterminant les changements de fréquences alléliques de la population pendant cet intervalle de temps, la signature de la dispersion dans les données génétiques devrait être plus forte sur de tels échantillons par rapport à un échantillonnage unique dans le temps. De telles stratégies de double échantillonnage dans le temps ont déjà été considérées dans de nombreuses études pour estimer des tailles de population dans le cadre de populations isolées et ont montré de bonnes performances par rapport à un échantillonnage classique (Beaumont, 2003; Wang, 2001; Waples, 1989). Dans ces approches, les échantillons sont récoltés à différentes générations et la migration n'est pas prise en compte. La seule approche considérant un double échantillonnage et de la migration est l'approche de Wang & Whitlock (2003) citée précédemment valide dans le contexte d'un modèle en îles. Cependant, le but de leur approche n'est pas d'estimer les caractéristiques de dispersion sur un ensemble de sous-populations, ni d'estimer des distributions de dispersion mais uniquement de déterminer le nombre d'individus arrivant par génération dans une sous-population focale parmi toutes les sous-populations du modèle (i.e. le nombre d'immigrants par génération d'une sous-population donnée).

Une approche originale, et potentiellement plus adaptée à l'estimation des distribution de dispersion sous isolement par la distance, consisterait à estimer, par une analyse de Fourier, les fonctions génératrices f des probabilités d'identité par état en fonction de la distance, avant et après dispersion, respectivement f_1 et f_2. Il devrait alors être possible d'estimer la distribution g permettant de passer de f_2 à f_1, c'est à dire g telle que $f_2 = g \circ f_1$. Cette distribution g correspondra alors à la distribution des distances de dispersion sur l'ensemble des sous-populations échantillonnées. Ce type d'approche, non explorée jusqu'à présent, semble séduisante dans la mesure où elle ne se foca-

lise pas sur une famille spécifique de distributions de dispersion, par exemple les distribution normales ou géométriques peu réalistes mais souvent utilisées pour modéliser la dispersion.

7.4 Vers des modèles plus réalistes et complexes

Le sujet initial de ma thèse était l'estimation de paramètres démographiques sous des modèles démographiques réalistes. Il s'agissait plus précisément de développer des méthodes d'estimation prenant en compte un certain nombre de variables écologiques, telles que la structure du paysage. Malheureusement, je n'ai pu réaliser cet objectif, le modèle le plus réaliste que j'ai pu considéré étant le modèle d'isolement par la distance en population continue. Le traitement de modèles plus complexes pour l'estimation de paramètres démographiques à partir de données génétiques pose, à ce jour, un certain nombre de problèmes dont nous allons voir quelques exemples.

7.4.1 Limites de l'information génétique

L'estimation indirecte de paramètres démographiques à partir de données génétiques est fondée sur le fait que des marqueurs génétiques neutres possèdent de l'information sur les caractéristiques démographiques des populations dans lesquelles ils ont évolué. Une limite de ces méthodes indirectes réside dans le fait que l'information présente dans ces marqueurs génétiques est limitée. Nous avons vu, par exemple, dans le chapitre 3 que l'utilisation de marqueurs montrant un niveau de polymorphisme faible entraîne une perte de précision lors de l'estimation de paramètres démographiques. D'autres études ont montré l'importance du niveau de variabilité des locus utilisés sur la précision et la puissance des analyses en génétiques de populations (Robertson & Hill, 1984; Goudet *et al.*, 1996). L'ensemble de ces résultats montrent que l'information que l'on peut extraire de données génétiques dépend fortement des marqueurs génétiques que l'on utilise. Dans certains cas, il ne sera pas possible d'extraire l'information voulue à partir des marqueurs disponibles. Cependant, il est important de noter que l'information présente dans les données génétiques ne dépend pas uniquement du niveau de variabilité et du nombre de marqueurs disponibles. En effet, bien que l'on puisse avoir accès à un nombre croissant de marqueurs génétiques présentant de forts niveaux de polymorphisme, l'information présente dans les données génétiques ne permettra l'estimation que d'un certain nombre

de paramètres évolutifs, et ceux de manière plus ou moins précise selon les paramètres considérés.

Une première illustration des limites de l'information génétique est, comme on l'a vu dans la section précédente, que lorsque l'on considère un échantillonnage classique unique dans dans le temps, les taux de mutation des marqueurs et les taux de migration entre sous-populations sont estimable uniquement sous forme de produits du taux de mutation par les tailles de populations et/ou des produits des taux de migration par les tailles de populations. Un deuxième exemple, mentionné dans le chapitre 2 lors de la présentation de la coalescence, est que même si l'on échantillonne tous les individus d'une population avec un grand nombre de marqueurs génétiques, il n'y aura jamais dans des données génétiques d'information sur des processus évolutifs ayant agi très loin dans le passé. En effet, l'inférence à partir de données génétiques pourra permettre de remonter dans le temps jusqu'au moment du premier ancêtre commun de la population actuelle mais pas plus loin. Un troisième exemple est que lorsque l'on cherche à estimer des temps de divergence entre sous populations, il est difficile à partir de données génétiques de différencier un scénarios dans lequel les sous-populations ont échangé de nombreux migrants dans le passé mais sont isolées depuis un temps t dans le passé, d'un scénario où les deux sous-populations échangent jusqu'au présent (i.e. au moment de l'échantillonnage) un nombre de migrants plus faible. Enfin, un dernier exemple est que lorsque l'on considère des fluctuations des tailles de populations, il est difficile d'estimer précisément les différentes tailles de la population et les moments des changements dans le passé car les données génétiques dépendent essentiellement de la moyenne harmonique des tailles de populations au cours du temps et peu des différentes fluctuations passées en elles-mêmes.

Lors de la considération de modèles de plus en plus réalistes, il semble donc nécessaire de s'assurer que les paramètres nous intéressant sont estimables en pratique. Il est fort probable qu'à partir d'un certain niveau de complexité l'information présente dans les données génétiques ne sera pas suffisante pour estimer précisément tous les paramètres d'un modèle.

7.4.2 Une alternative possible au maximum de vraisemblance

Dans le cas de modèles réaliste et donc complexes, l'estimation par maximum de vraisemblance n'est pas possible soit parce que la vraisemblance

sous ces modèles ne peut être estimée par les méthodes de simulation présentées dans le chapitre 4, soit parce que, pour ces modèles avec un très grand nombre de paramètres, l'estimation de la vraisemblance demande des temps de calcul trop longs. Une alternative possible aux méthodes d'estimation par maximum de vraisemblance est l'utilisation des méthodes d'estimation dites ABC (de l'anglais Approximate "Bayesian" Computation), utilisant des statistiques résumés et une approche par simulation fondée sur la coalescence (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999; Estoup *et al.*, 2001; Estoup & Clegg, 2003; Beaumont *et al.*, 2002; Marjoram *et al.*, 2003). A la place d'une évaluation numérique de la vraisemblance d'un échantillon, le jeu de données complet D est remplacé dans les méthodes ABC par un ensemble de statistiques résumées S, telles que le nombre d'allèles, l'hétérozygotie ou les F_{ST} entre sous-populations. Des jeux de données sont ensuite simulés selon un processus de coalescence, pour différentes valeurs des paramètres du modèle considéré. Pour chaque simulation, l'ensemble des statistiques résumées S sont calculées sur les données simulées. Chaque simulation (i.e. chaque jeu de valeurs des paramètres) est ensuite acceptée ou rejetée en fonction de l'écart entre les valeurs des statistiques résumées calculées sur les données simulées et leurs valeurs sur le jeux de données réels. Les distributions des valeurs de paramètres acceptées donnent la distribution à posteriori des paramètres sachant les données. Ces méthodes ont récemment été appliquées dans le contexte de scénarios évolutifs complexes (Estoup *et al.*, 2001; Estoup & Clegg, 2003). De plus, ces algorithmes ont étés récemment améliorés et des tests par simulation ont montré de bonnes performances de ces méthodes pour des modèles simples (Beaumont *et al.*, 2002).

La contrainte majeure de ces méthodes est le temps de calcul, où plutôt le temps nécessaire pour créer les jeux de données par simulation. En effet, ces méthodes nécessitent un grand nombre de simulations de jeux de données pour différentes valeurs des paramètres. Même si la théorie de la coalescence a permis le développement d'algorithmes de simulation très efficaces, la considération de modèles réalistes complexes implique des algorithmes de simulation relativement lents tels que des algorithmes génération par génération. Un point positif est que la parallélisation des méthodes ABC est facile et efficace puisque toutes les simulations de jeux de données sont indépendantes. De plus, des techniques de pondération des différentes statistiques résumées (Beaumont *et al.*, 2002), et une simplification des modèles dans le temps, telle qu'on l'a évoquée précédemment, peuvent réduire les temps de calculs. Un second facteur clé de la performance et de la rapidité des ces méthodes ABC est le choix des différentes statistiques résumées S. Une question essentielle ici est de savoir dans quelle mesure les valeurs des statistiques résumées choi-

sies saisissent l'information présente dans les données génétiques complètes, nécessaire à l'estimation des paramètres nous intéressant et dans quelle mesure ces statistiques jouent sur le taux d'acceptation des jeux de données simulés. Par exemple, dans le contexte de l'estimation de paramètres démographiques actuels en populations naturelles, il semble préférable d'utiliser des F-statistiques plutôt que des probabilités d'identité puisque ces dernières sont plus sensibles aux taux de mutation et aux fluctuations démographiques passées, au moins à une échelle géographique locale.

Ces méthodes n'ont jamais été appliquées pour l'estimation de paramètres de migration, ni pour l'estimation de distribution de dispersion. Il serait intéressant d'évaluer le potentiel des méthodes ABC pour l'estimation des distributions de dispersion ainsi que pour la détection de réduction de surface, et de manière plus générale pour la détection de changements temporels des caractéristiques démographiques, pour des populations sous isolement par la distance. Les méthodes ABC représentent, au moins provisoirement, une approche alternative possible pour le traitement de modèles complexes pour lesquels l'approche par maximum de vraisemblance est difficile. Il semble toutefois préférable de favoriser l'approche par maximum de vraisemblance sur les données génétiques complètes lorsque cela est possible (Beaumont *et al.*, 2002).

Il est prévisible que le développement de nouvelles méthodes d'estimation permettant la prise en compte de modèles démographiques plus réalistes (isolement par la distance, hétérogénéité spatiales et temporelles) ainsi que des systèmes biologiques plus complexes (données multi-locus avec de la recombinaison) soit, pour au moins quelques années encore, au centre des développements théoriques et expérimentaux dans le domaine de l'estimation de paramètres démographiques à partir de données génétiques.

Bibliographie

Abramovitz, M. & Stegun, I. A., éds. (1972) *Handbook of mathematical functions*. Dover, New York.

Bahlo, M. & Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology* **57** : 79–95.

Balloux, F. & Goudet, J. (2002) Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* **11** : 771–783.

Barton, N. H., Depaulis, F. & Etheridge, A. M. (2002) Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61** : 31–48.

Barton, N. H. & Gale, K. S. (1993) Genetic analysis of hybrid zones. In *Hybrid zones and the evolutionary process*, édité par Harrison, R. G., pp. 13–45. Oxford University Press, Oxford.

Beaumont, M. A. (1999) Detecting population expansion and decline using microsatellites. *Genetics* **153** : 2013–2029.

Beaumont, M. A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** : 1139–1160.

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

Beaumont, M. A., Zhang, W. Y. & Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162** : 2025–2035.

Beerli, P. (2004) Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology* **13** : 827–827.

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.

Beerli, P. & Felsenstein, J. (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 U.S.A.* **98** : 4563–4568.

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

Brohede, J., Primmer, C. R., Moller, A. & Ellegren, H. (2002) Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Research* **30** : 1997–2003.

Chakraborty, R. & Danker-Hopfe, H. (1991) Analysis of population structure : a comparative study of different estimators of Wright's fixation indices. In *Handbook of Statistics*, édité par Rao, C. R. & Chakraborty, R., vol. 8, pp. 203–254. Elsevier.

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

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

Cockerham, C. C. (1973) Analyses of gene frequencies. *Genetics* **74** : 679–700.

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

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

Cox, D. R. & Hinkley, D. V. (1974) *Theoretical statistics.* Chapman & Hall, London.

Crawford, T. (1984) The estimation of neighborhood parameters for plant populations. *Heredity* **52** : 273–283.

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 U.S.A.* **81** : 6073–6077.

Crow, J. F. & Kimura, M. (1970) *An introduction to population genetics theory.* Harper & Row, New York.

de Iorio, M. & Griffiths, R. C. (2004a) Importance sampling on coalescent histories. *Advances in applied Probability* **sous presse**

de Iorio, M. & Griffiths, R. C. (2004b) Importance sampling on coalescent histories in subdivided population models. *Advances in applied Probability* **sous presse**

de Iorio, M., Griffiths, R. C., Leblois, R. & Rousset, F. (2004) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *soumis*

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380** : 152–4.

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

Donnelly, P. (1999) The coalescent and microsatellite variability. In *Microsatellites : Evolution and Applications*, édité par Goldstein, D. & Schlotterer, C., pp. 116–128. Oxford University Press, Oxford.

Dunham, J., Peacock, M., Tracy, C. R., Nielsen, J. & Vinyard, G. (1999) Assessing extinction risk : integrating genetic information. *Conservation Ecology [online]* **3** : 2.

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

Endler, J. A. (1977) *Geographical variation, speciation, and clines.* Princeton University Press, Princeton.

Estoup, A. & Angers, B. (1998) Microsatellites and minisatellites for molecular ecology : theoretical and empirical considerations. In *Advances in molecular ecology*, édité par Carvalho, G., pp. 55–86. IOS Press, Amsterdam.

ESTOUP, A. & CLEGG, S. M. (2003) Bayesian inferences on the recent island colonization history by the bird Zosterops lateralis lateralis. *Molecular Ecology* **12** : 657–674.

ESTOUP, A. & CORNUET, J.-M. (1999) Microsatellite evolution : inferences from population data. In *Microsatellites : evolution and applications*, édité par Goldstein, D. B. & Schlötterer, C., pp. 49–65. Oxford University Press, Oxford.

ESTOUP, A., JARNE, P. & CORNUET, J. M. (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11** : 1591–1604.

ESTOUP, A., WILSON, I. J., SULLIVAN, C., CORNUET, J. M. & MORITZ, C. (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. *Genetics* **159** : 1671–1687.

EXCOFFIER, L. (2001) Analysis of population subdivision. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 271–307. Wiley, Chichester, U.K.

FELSENSTEIN, J. (1975) A pain in the torus : some difficulties with models of isolation by distance. *American Naturalist* **109** : 359–368.

FENSTER, C. B., VEKEMANS, X. & HARDY, O. J. (2003) Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution* **57** : 995–1007.

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

FOUCART, A. (1997) Inventaire et dynamique annuelle du peuplement acridien de la plaine de la Crau sèche (Bouches-du-Rhône, France)(Orthoptera, Acridoidea). *Bulletin de la Société Entomologique de France* **102** : 77–87.

FOUCART, A. & LECOQ, M. (1996) Biologie et dynamique de Prionotropis hystrix rhodanica Uvarov, 1923, dans la plaine de la Crau (France) (Orthoptera, Pamphagidae). *Bulletin de la Société Entomologique de France* **101** : 75–87.

FRANKHAM, R. (1995) Effective population size/adult population size ratios in wildlife : a review. *Genetical Research (Cambridge)* **66** : 95–107.

FRANKHAM, R. (1996) Relationship of genetic variation to population size in wildlife. *Conservation Biology* **10** : 1500–1508.

Frankham, R. & Ralls, K. (1998) Conservation biology - Inbreeding leads to extinction. *Nature* **392** : 441–442. Material.

G., C. (1975) Esquisse écologique d'une zone semi aride : La Crau (Bouches du Rhône). *Alauda* **43** : 23–54.

Gaggiotti, O., Lange, O., Rassmann, K. & Gliddon, C. (1999) A comparison of two methods for estimating average levels of gene flow using microsatellites data. *Molecular Ecology* **8** : 1513–1520.

Galtier, N., Depaulis, F. & Barton, N. H. (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155** : 981–987.

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

Gandon, S. & Rousset, F. (1999) Evolution of stepping stone dispersal rates. *Proceedings of the Royal Society (London) B* **266** : 2507–2513.

Garza, J. C. & Williamson, E. G. (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10** : 305–318.

Gonser, R., Donnelly, P., Nicholson, G. & Di Rienzo, A. (2000) Microsatellite mutations and inferences about human demography. *Genetics* **154** : 1793–807.

Goudet, J., Raymond, M., de Meeüs, T. & Rousset, F. (1996) Testing differentiation in diploid populations. *Genetics* **144** : 1931–1938.

Griffiths, R. & Tavaré, S. (1994) Ancestral inference in population genetics. *Statistical Science* **9** : 307–319.

Griffiths, R. C. & Tavaré, S. (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society (London) B* **344** : 403–410.

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

Hastings, A. & Harrison, S. (1994) Metapopulation dynamics and genetics. *Annual Review of Ecology and Systematics* **25** : 167–188.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** : 97–109.

Hedrick, P. W. (2001) Invasion of transgenes from salmon or other genetically modified organisms into natural populations. *Canadian Journal of Fisheries and Science* **58** : 841–844.

Horn, R. A. & Johnson, C. R. (1991) *Topics in matrix analysis*. Cambridge University Press, Cambridge.

Hudson, R. R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23** : 183–201.

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

Karlin, S. & Taylor, H. M. (1981) *A second course in stochastic processes*. Academic Press, San Diego.

Keane, R. M. & Crawley, M. J. (2002) Exotic plant invasions and the enemy release hypothesis. *Trends in Ecology and Evolution* **17** : 164–170.

Kimura, M. & Crow, J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49** : 725–738.

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

Kingman, J. (1982b) On the genealogy of large populations. *Journal of Applied Probabilities* **19A** : 27–43.

Koenig, W. D., Vuren, D. V. & Hooge, P. N. (1996) Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution* **11** : 514–517.

Kot, M., Lewis, M. A. & Driessche, P. v. d. (1996) Dispersal data and the spread of invading organisms. *Ecology* **77** : 2027–2042.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140** : 1421–1430.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149** : 429–434.

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

Leblois, R., Rousset, F., Tikel, D., Moritz, C. & Estoup, A. (2000) absence of evidence for isolation by distance in an expanding cane toad (Bufo marinus) population : an indiviual-based analysis of microsatellite genotypes. *Molecular Ecology* **9** : 1905–1909.

Lee, C. E. (2002) Evolutionary genetics of invasive species. *Trends in Ecology and Evolution* **17** : 386–391.

Lenormand, T. (2002) Gene flow and the limits to natural selection. *Trends in Ecology and Evolution* **17** : 183–189.

Levins, R. (1968) *Evolution in changing environments. Some theoretical explorations.* Princeton University Press.

Li, W. H. (1976) Effect of Migration on Genetic Distance. *American Naturalist* **110** : 841–847.

Long, J. C., Naidu, J. M., Mohrenweiser, H. W., Gershowitz, H., Johnson, P. L. & Wood, J. W. (1986) Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. *American Journal of Physical Anthropology* **70** : 75–96.

Luikart, G. & Cornuet, J.-M. (1998) Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology* **12** : 228–237.

Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. (2003) The power and promise of population genomics : From genotyping to genome typing. *Nature Reviews Genetics* **4** : 981–994.

Malécot, G. (1948) *Les mathématiques de l'hérédité.* Masson, Paris.

Malécot, G. (1950) Quelques schémas probabilistes sur la variabilité des populations naturelles. *Annales de l'Université de Lyon A* **13** : 37–60.

Malécot, G. (1967) Identical loci and relationship. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, édité par Le Cam, L. M. & Neyman, J., vol. 4, pp. 317–332. University of California Press, Berkeley.

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

MARJORAM, P., MOLITOR, J., PLAGNOL, V. & TAVARE, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100** : 15324–15328.

MARUYAMA, T. (1972) Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics* **70** : 639–651.

MORAN, P. A. P. (1975) Wandering distributions and the electrophoretic profile. *Theoretical Population Biology* **8** : 318–330.

NAGYLAKI, T. (1975) Conditions for the existence of clines. *Genetics* **80** : 595–615.

NAGYLAKI, T. (1976) The decay of genetic variability in geographically structured populations. II. *Theoretical Population Biology* **10** : 70–82.

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

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

NATH, H. B. & GRIFFITHS, R. C. (1996) Estimation in an island model using simulation. *Theoretical Population Biology* **50** : 227–253.

NEI, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

NEI, M., MARUYAMA, T. & CHAKRABORTY, R. (1975) The bottleneck effect and genetic variability in populations. *Evolution* **29** : 1–10.

NEUHAUSER, C. & KRONE, S. M. (1997) The genealogy of samples in models with selection. *Genetics* **145** : 519–534.

NORDBORG, M. (2001) Coalescent theory. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 179–212. Wiley, Chichester, U.K.

NORDBORG, M. & TAVARE, S. (2002) Linkage disequilibrium : what history has to tell us. *Trends in Genetics* **18** : 83–90.

NOTOHARA, M. (1993) The strong-migration limit for the genealogical process in geographically structured populations. *Journal of mathematical Biology* **31** : 115–122.

Ohta, T. & Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research (Cambridge)* **22** : 201–204.

Patil, G. P. & Joshi, S. W. (1968) *A dictionary and bibliography of discrete distributions*. Oliver & Boyd, Edinburgh.

Pimentel, D., Lach, L., Zuniga, R. & Morrison, D. (2000) Environmental and economic costs of nonindigenous species in the United States. *Bioscience* **50** : 53–65.

Portnoy, S. & Willson, M. F. (1993) Seed dispersal curves : behavior of the tails of the distribution. *Evolutionary Ecology* **7** : 25–44.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. (1999) Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16** : 1791–8.

R., D. & Rambier, A. (1950) Notes orthoptérologiques. *Bulletin de la Société Entomologique de France* **29** : 35–40.

Rannala, B. & Hartigan, J. A. (1996) Estimating gene flow in island populations. *Genetical Research (Cambridge)* **67** : 147–158.

Raymond, M. & Rousset, F. (1995) An exact test for population differentiation. *Evolution* **49** : 1283–1286.

Reich, D. E. & Goldstein, D. B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **95** : 8119–8123.

Robertson, A. & Hill, W. G. (1984) Deviations from Hardy-Weinberg proportions : sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107** : 703–718.

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

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 classes of individuals. *Theoretical Population Biology* **55** : 297–308.

ROUSSET, F. (1999b) Genetic differentiation within and between two habitats. *Genetics* **151** : 397–407.

ROUSSET, F. (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13** : 58–62.

ROUSSET, F. (2001a) Genetic approaches to the estimation of dispersal rates. In *Dispersal : individual, population and community*, édité par Clobert, J., Danchin, É., Dhondt, A. A. & Nichols, J. D., pp. 18–28. Oxford University Press, Oxford.

ROUSSET, F. (2001b) Inferences from spatial population genetics. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 239–269. Wiley, Chichester, U.K.

ROUSSET, F. (2004) *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton, New Jersey.

ROZE, D. & ROUSSET, F. (2003) Diffusion approximations for selection and drift in subdivided populations : a straightforward method and examples involving dominance, selfing and local extinctions. *Genetics* **165** : 2153–2166.

SACCHERI, I., KUUSSAARI, M., KANKARE, M., VIKMAN, P., FORTELIUS, W. & HANSKI, I. (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature* **392** : 491–494.

SANKOFF, D. (1975) Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* **28** : 35–42.

SAUNDERS, I. W., TAVARE, S. & WATTERSON, G. A. (1984) On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16** : 471–491.

SAWYER, S. (1977) Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probabilities* **9** : 268–282.

SHEA, K. & CHESSON, P. (2002) Community ecology theory as a framework for biological invasions. *Trends in Ecology and Evolution* **17** : 170–176.

Sih, A., Jonsson, B. G. & Luikart, G. (2000) Habitat loss : ecological, evolutionary and genetic consequences. *Trends in Ecology and Evolution* **15** : 132–133.

Slatkin, M. (1987) The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32** : 42–49.

Slatkin, M. (1991) Inbreeding coefficients and coalescence times. *Genetical Research (Cambridge)* **58** : 167–175.

Slatkin, M. (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47** : 264–279.

Slatkin, M. (1994) Gene flow and population structure. In *Ecological Genetics*, édité par Real, L. A., pp. 3–17. Princeton University Press, Princeton.

Smith, F. D. M., May, R. M., Pellew, R., Johnson, T. H. & Walker, K. S. (1993) How much do we know about the current extinction rate ? *Trends in Ecology and Evolution* **8** : 375–378.

Spong, G. & Creel, S. (2001) Deriving dispersal distance from genetic data. *Proceedings of the Royal Society (London) B* **268** : 2571–2574.

Stephens, M. (1999) Problems with computational methods in population genetics. *Bulletin of the 52nd Session of the International Statistical Institute (invited paper)*

Stephens, M. & Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *Journal of the Royal Statistical Society* **62** : 605–655.

Sumner, J., Estoup, A., Rousset, F. & Moritz, C. (2001) 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology* **10** : 1917–1927.

Swofford, D., Olsen, G., Waddelland, P. & Hillis, D. (1996) Phylogentic inference. In *Molecular Systematics*, édité par Hillis, D., Moritz, C. & Mable, B., pp. 407 – 514. Sinauer Associates, Sunderland, MA.

Tachida, H. (1985) Joint frequencies of alleles determined by separate formulations for the mating and mutation systems. *Genetics* **111** : 963–974.

TAJIMA, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** : 437–460.

TAKEZAKI, N. & NEI, M. (1996) Genetic distances and reconstruction of phylogenetic tree from microsatellites DNA. *Genetics* **144** : 389–399.

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. & DONNELLY, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145** : 505–518.

TUFTO, J., ENGEN, S. & HINDAR, K. (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* **144** : 1911–1921.

UVAROV, B. P. (1923) Sur les races géographiques du Prionotropis hystrix Germ. *Annales de la société entomologique de France* **XCI** : 245–248.

VAYSSIÈRE, P. (1921) La lutte contre le criquet marocain (Dociostaurus maroccanus Thunb.) en Crau en 1920. *Annls. epiphyties* pp. 117–167.

VEKEMANS, X. & HARDY, O. J. (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13** : 921–921.

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 (Cambridge)* **77** : 67–81.

WANG, J. (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research (Cambridge)* **78** : 243–257.

WANG, J. & WHITLOCK, M. C. (2003) Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163** : 429–446.

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 population size of cetacean populations. In *Genetic ecology of whales and dolphins, vol. Special Issue 13*, édité par Hoelzel, A. R., pp. 279–300. International Whale Commission, London.

WEIR, B. S. & COCKERHAM, C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38** : 1358–1370.

WHITLOCK, M. C. (2002) Selection, load and inbreeding depression in a large metapopulation. *Genetics* **160** : 1191–1202.

WHITLOCK, M. C. (2003) Fixation probability and time in subdivided populations. *Genetics* **164** : 767–779.

WHITLOCK, M. C. & MCCAULEY, D. E. (1999) Indirect measures of gene flow and migration : $F_{\mathrm{ST}} \neq 1/(4Nm+1)$. *Heredity* **82** : 117–125.

WILCOVE, D. S., ROTHSTEIN, D., DUBOW, J., PHILLIPS, A. & LOSOS, E. (1998) Quantifying threats to imperiled species in the United States. *Bioscience* **48** : 607–615.

WILLIAMSON, M. (1996) *Biological invasions*. Chapman and Hall, London.

WOLFF, A. (2001) The benefits of extensive agriculture to birds : the case of the little bustard. *Journal of Applied Ecology* **38** : 963–975.

WOLFF, A. (2002) Landscape context and little bustard abundance in a fragmented steppe : implications for reserve management in mosaic landscapes. *Biological Conservation* **107** : 211–220.

WOOD, J. W., SMOUSE, P. E. & LONG, J. C. (1985) Sex-specific dispersal patterns in two human populations of highland New Guinea. *American Naturalist* **125** : 747–768.

WRIGHT, S. (1943) Isolation by distance. *Genetics* **28** : 114–138. Republié dans Wright (1986), pp. 401–425.

WRIGHT, S. (1946) Isolation by distance under diverse systems of mating. *Genetics* **31** : 39–59. Republié dans Wright (1986), pp. 444–464.

WRIGHT, S. (1951) The genetical structure of populations. *Annals of Eugenics* **15** : 323–354. Republié dans Wright (1986), pp. 580–611.

WRIGHT, S. (1986) *Evolution : selected papers*. University of Chicago Press, Chicago.

Annexe A : annexes “mathématiques et algorithmiques”

A.1 Calcul des probabilités d'identité par état à partir de l'identité par descendance sous différents modèles mutationnels (F. Rousset)

The probability of identity between two genes may be written

$$Q = \sum_t c_t Q_{|t} \tag{A.1}$$

in terms of the probability of coalescence c_t at time t and the probability $Q_{|t}$ of identity given coalescence occurs at t. For identity by descent, the conditional value is $Q_{|t} = (1-u)^{2t}$ in terms of the mutation rate u, and the unconditional one is $\sum_t c_t(1-u)^{2t} \equiv \dot{Q}((1-u)^2)$. More generally, the conditional probability is given by a Markov chain model. For example for a two allele model with mutation rate u, the transition probabilities between the two allelic states $1, 2$ along a gene lineage are described by the matrix

$$\mathbf{U} \equiv (u_{ij}) = \begin{pmatrix} 1-u & u \\ u & 1-u \end{pmatrix}. \tag{A.2}$$

Thus, if the gene was initially of type 1, the probabilities that its descendant t generations later is of type 1 or 2 are the elements of the vector

$$\mathbf{U}^t \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{A.3}$$

If we follow two genes lineages we have four possible states $11, 12, 21, 22$. The transition probabilities between these four states are given by the 4×4 matrix

$$\mathbf{U} \otimes \mathbf{U} = \begin{pmatrix} u_{11}\mathbf{U} & u_{12}\mathbf{U} \\ u_{21}\mathbf{U} & u_{22}\mathbf{U} \end{pmatrix}. \tag{A.4}$$

$\mathbf{U} \otimes \mathbf{U}$ is the direct, or Kronecker, product of $\mathbf{U}$ with itself.

Thus, if the common ancestor was initially of type 2, the probabilities that the two descendants t generations later are both of type 1 is the element

11 of the vector

$$(\mathbf{U} \otimes \mathbf{U})^t \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{A.5}$$

Here the value of the vector $\mathbf{v}$ is a vector having all elements being 0 except element ij being 1, where ij is the allelic state of the ancestral lineage(s) of the pair of genes. Write $\boldsymbol{\delta}_{ij}$ such a vector; e.g. the above expression is $(\mathbf{U} \otimes \mathbf{U})^t \boldsymbol{\delta}_{22}$.

These expressions are conveniently expressed in terms of the eigenvalues λ_{ij} and eigenvectors $\mathbf{e}_{ij}$ of the matrix $\mathbf{U} \otimes \mathbf{U}$: for any vector $\mathbf{v}$,

$$(\mathbf{U} \otimes \mathbf{U})^t \mathbf{v} = \sum_{ij} \lambda_{ij}^t \frac{\mathbf{e}_{ij} \cdot \mathbf{v}}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij} \tag{A.6}$$

where $\mathbf{e}_{ij} \cdot \mathbf{v}$ is the inner product $\sum_{kl} (\mathbf{e}_{ij})_{kl} \mathbf{v}_{kl}$. Hence the vector of probabilities of allelic states given coalescence occurred t generations ago and given that the common ancestor was of state 2 is

$$(\mathbf{U} \otimes \mathbf{U})^t \boldsymbol{\delta}_{22} = \sum_{ij} \lambda_{ij}^t \frac{\mathbf{e}_{ij} \cdot \boldsymbol{\delta}_{22}}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij}. \tag{A.7}$$

The probabilities of identity not conditional on the coalescence time t may be computed without any explicit knowledge of the conditional probabilities given t. For example, the probability that two genes are both 1 given their common ancestor was 2 is obtained by summing the above expression over the distribution of coalescence times, so it is element 11 of

$$\sum_t c_t \sum_{ij} \lambda_{ij}^t \frac{\mathbf{e}_{ij} \cdot \boldsymbol{\delta}_{22}}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij} \tag{A.8}$$

where $\boldsymbol{\delta}_{ij}$ is the vector with ijth element 1 and other elements 0. This can be written

$$\sum_{ij} (\sum_t c_t \lambda_{ij}^t) \frac{\mathbf{e}_{ij} \cdot \boldsymbol{\delta}_{22}}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij} = \sum_{ij} \dot{Q}(\lambda_{ij}) \frac{\mathbf{e}_{ij} \cdot \boldsymbol{\delta}_{22}}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij}. \tag{A.9}$$

Thus such probabilities can be expressed as function of probabilities of identity by descent for various mutation rates. In practice one only needs to find $\dot{Q}$ as a function of $(1-u)^2$ for the given demographic model, and to find the eigenvectors and eigenvalues of the mutation matrix $\mathbf{U}$. This is the separation of the mating and mutation systems noted by Tachida (1985).

Summing over possible ancestral states l each with probability $\pi(l)$, the probability that two genes are both of type 1 is element 11 of

$$\sum_{ij} \dot{Q}(\lambda_{ij}) \frac{\mathbf{e}_{ij} \cdot (\sum_l \pi(l)\boldsymbol{\delta}_{ll})}{\mathbf{e}_{ij} \cdot \mathbf{e}_{ij}} \mathbf{e}_{ij} \tag{A.10}$$

and the probability that two genes are identical is the sum of elements kk of this vector.

Computations are made easier by the fact that the eigenvalues λ_{ij} and eigenvectors $\mathbf{e}_{ij}$ are the products $\lambda_i\lambda_j$ and direct products $\mathbf{e}_i \otimes \mathbf{e}_j$ of the eigenvalues and eigenvectors of $\mathbf{U}$ (Horn & Johnson, 1991, Theorem 4.2.12).

For the two-allele model, $\lambda_1 = 1$, $\lambda_2 = 1-2u$; $\mathbf{e}_1 = {}^T(1/2, 1/2)$ and $\mathbf{e}_2 = {}^T(1/2, -1/2)$. Thus $\lambda_{12} = 1-2u$, $\mathbf{e}_{12} = {}^T(1/2(1/2,-1/2), 1/2(1/2,-1/2))$, and so on. By applying the above formulas one can check that

$$Q = \frac{1}{2}\dot{Q}(1) + \frac{1}{2}\dot{Q}((1-2u)^2) = \frac{1}{2}(1 + \dot{Q}((1-2u)^2)) \tag{A.11}$$

the more general formula for a symmetric KAM being

$$Q = \frac{1}{K}\left(1 + (K-1)\dot{Q}((1 - \frac{Ku}{K-1})^2)\right). \tag{A.12}$$

Now consider a model with say four alleles with transition matrix

$$\mathbf{U} = \begin{pmatrix} 1-u & u/2 & 0 & u/2 \\ u/2 & 1-u & u/2 & 0 \\ 0 & u/2 & 1-u & u/2 \\ u/2 & 0 & u/2 & 1-u \end{pmatrix} \tag{A.13}$$

i.e. mutation is possible towards each of the two "neighbouring" states 4 and 2 for allele 1, 1 and 3 for allele 2, ... 1 and 3 for allele 4. This has

the same structure as the migration matrix for a stepping stone model of 4 demes on a circle. Since this matrix is a circulant matrix (i.e. where each row is the previous one cycled forward one step), the general formulas above reduce to the formulas of discrete Fourier analysis, which can be independently derived by consideration of characteristic functions (Moran, 1975). We directly generalize to K alleles. For $\imath \equiv \sqrt{-1}$ the eigenvectors are $\mathbf{e}_j = {}^T(1, \ldots, e^{\imath 2\pi jk/K}, \ldots, e^{\imath 2\pi j(K-1)/K})$ and the eigenvalues are $l_j = \psi(e^{\imath 2\pi j/K})$ where $\psi(e^{ix}) = 1-u(1-\cos(x))$ (e.g. 1, $1-u$, $1-2u$ and $1-u$ in the above example). Thus the probability that two genes are identical is

$$\frac{1}{K}\sum_{k=0}^{K-1} \dot{Q}(\psi^2(e^{\imath 2\pi k/K})). \tag{A.14}$$

Note that eq. (A.12) for the KAM in a specific case of this one. Likewise the probability that two genes are of states "j steps apart" is

$$\frac{1}{K}\sum_{k=0}^{K-1} \dot{Q}(\psi^2(e^{\imath 2\pi k/K}))e^{\imath 2\pi kj/K} \tag{A.15}$$

and the result for the unbounded stepping stone model is the limit when $K \to \infty$ of the above expression, which is

$$\frac{1}{\pi}\int_0^{\pi} \dot{Q}(\psi^2(e^{ix}))\cos(jx)\,dx. \tag{A.16}$$

These formulas hold for more general unbounded stepwise mutation models, e.g. the TPM has

$$\psi(e^{ix}) = 1 - u\left(1 - p\cos(x) - (1-p)\frac{(1-q)(\cos(x)-q)}{1-2q\cos(x)+q^2}\right) \tag{A.17}$$

where p and q are as in Rousset (1996, p. 1359).

A.2 Développements de l'algorithme d'échantillonnage pondéré de de Iorio *et al.* (2004) pour deux populations et une modèle mutationnel SMM

The system of equations (4.57) defining the $\hat{\pi}$'s considering two populations and a strict stepwise mutation model without bounds is : for pop α and for each type j

$$\Big(\frac{n_{\alpha j}}{q_\alpha}+m_\alpha+\theta\Big)\hat{\pi}(j|\alpha,\mathbf{n}) = \frac{n_\alpha}{q_\alpha}+\frac{\theta}{2}\Big(\hat{\pi}(j-1|\alpha,\mathbf{n})+\hat{\pi}(j+1|\alpha,\mathbf{n})\Big)+m_\alpha\hat{\pi}(j|\beta,\mathbf{n}) \tag{A.18}$$

similarly we have the same set of equations for pop β and for each type j replacing α by β in the equation above. The aim is to determine the $\hat{\pi}(\cdot|\cdot,\mathbf{n})$'s. Denote the Fourier transform of $\hat{\pi}(j|\alpha,\mathbf{n})$

$$\tilde{\pi}(\xi|\alpha,\mathbf{n}) \equiv \sum_{j=-\infty}^{\infty} e^{\imath\xi j}\hat{\pi}(j|\alpha,\mathbf{n}) \tag{A.19}$$

and the Fourier transform of n_α

$$\tilde{n}_\alpha(\xi) \equiv \sum_{j=-\infty}^{\infty} e^{\imath\xi j} n_{\alpha j}. \tag{A.20}$$

From eq.A.18 we have for pop α

$$\Big(\frac{n_\alpha}{q_\alpha}+m_\alpha+\theta\Big)\tilde{\pi}(\xi|\alpha,\mathbf{n}) = \frac{\tilde{n}_\alpha(\xi)}{q_\alpha}+\theta\cos(\xi)\tilde{\pi}(\xi|\alpha,\mathbf{n})+m_\alpha\tilde{\pi}(\xi|\beta,\mathbf{n}) \tag{A.21}$$

and the equivalent for pop β

$$\Big(\frac{n_\beta}{q_\beta}+m_\beta+\theta\Big)\tilde{\pi}(\xi|\beta,\mathbf{n}) = \frac{\tilde{n}_\beta(\xi)}{q_\beta}+\theta\cos(\xi)\tilde{\pi}(\xi|\beta,\mathbf{n})+m_\beta\tilde{\pi}(\xi|\alpha,\mathbf{n}). \tag{A.22}$$

This can be written in the following matrix form

$$\begin{pmatrix} \frac{\tilde{n}_\alpha(\xi)}{q_\alpha} \\ \frac{\tilde{n}_\beta(\xi)}{q_\beta} \end{pmatrix} = \begin{pmatrix} \frac{n_\alpha}{q_\alpha}+m_\alpha+\theta(1-\cos(\xi)) & -m_\alpha \\ -m_\beta & \frac{n_\beta}{q_\beta}+m_\beta+\theta(1-\cos(\xi)) \end{pmatrix} \begin{pmatrix} \tilde{\pi}(\xi|\alpha,\mathbf{n}) \\ \tilde{\pi}(\xi|\beta,\mathbf{n}) \end{pmatrix} \tag{A.23}$$

or

$$\begin{pmatrix}\tilde{\pi}(\xi|\alpha,\mathbf{n})\\ \tilde{\pi}(\xi|\beta,\mathbf{n})\end{pmatrix} = (\mathcal{A}(\xi))^{-1}\begin{pmatrix}\frac{\tilde{n}_\alpha(\xi)}{q_\alpha}\\ \frac{\tilde{n}_\beta(\xi)}{q_\beta}\end{pmatrix}. \tag{A.24}$$

Thus the solution to eq.A.18 is

$$\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-\imath\xi j}\begin{pmatrix}\tilde{\pi}(\xi|\alpha,\mathbf{n})\\ \tilde{\pi}(\xi|\beta,\mathbf{n})\end{pmatrix} d\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-\imath\xi j}(\mathcal{A}(\xi))^{-1}\begin{pmatrix}\frac{\tilde{n}_\alpha(\xi)}{q_\alpha}\\ \frac{\tilde{n}_\beta(\xi)}{q_\beta}\end{pmatrix} d\xi. \tag{A.25}$$

Note that

$$e^{-\imath\xi j}\begin{pmatrix}\frac{\tilde{n}_\alpha(\xi)}{q_\alpha}\\ \frac{\tilde{n}_\beta(\xi)}{q_\beta}\end{pmatrix} = \begin{pmatrix}\frac{\sum_{k=-\infty}^{\infty} e^{\imath\xi(k-j)} n_{\alpha k}}{q_\alpha}\\ \frac{\sum_{k=-\infty}^{\infty} e^{\imath\xi(k-j)} n_{\beta k}.}{q_\beta}\end{pmatrix} = \sum_{k=-\infty}^{\infty} e^{\imath\xi(k-j)}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix} \tag{A.26}$$

Then

$$\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{A}^{-1}(\xi)\sum_{k=-\infty}^{\infty} e^{\imath\xi(k-j)}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix} d\xi. \tag{A.27}$$

Griffiths proposed the following computation to analytically solve eq.A.27. The determinant of $\mathcal{A}$ is

$$|\mathcal{A}| = (\frac{n_\alpha}{q_\alpha} + m_\alpha + \phi)(\frac{n_\beta}{q_\beta} + m_\beta + \phi) - m_\alpha m_\beta \tag{A.28}$$

where $\phi \equiv \theta(1-\cos(\xi))$. Thus $|\mathcal{A}| = (\phi-\lambda_1)(\phi-\lambda_2)$, where λ_1, λ_2 are the roots of

$$\phi^2 + \phi(x_\alpha + x_\beta) + x_\alpha x_\beta - m_\alpha m_\beta, \tag{A.29}$$

with $x_\alpha \equiv \frac{n_\alpha}{q_\alpha} + m_\alpha$, and similarly for x_β. Result (proof later) : $\lambda_1, \lambda_2 < 0$, $\lambda_1 \neq \lambda_2$ and both are real. Then

$$|\mathcal{A}|^{-1} = \frac{C}{\phi-\lambda_1} + \frac{D}{\phi-\lambda_2} \tag{A.30}$$

where $C = \frac{1}{\lambda_1 - \lambda_2}$ and $D = \frac{1}{\lambda_2 - \lambda_1}$. We can write

$$\begin{aligned} 2(\phi - \lambda) &= \theta(2 - e^{\imath\xi} - e^{-\imath\xi}) - 2\lambda \\ &= e^{-\imath\xi}(2(\theta - \lambda)e^{\imath\xi} - \theta e^{2\imath\xi} - \theta) \\ &= -\theta e^{-\imath\xi}(e^{2\imath\xi}\frac{2(\theta - \lambda)}{\theta} + 1) \\ &= -\theta e^{-\imath\xi}(e^{\imath\xi} - u(\lambda))(e^{\imath\xi} - v(\lambda)) \end{aligned} \tag{A.31}$$

where $u(\lambda), v(\lambda)$ are the roots of $\Psi^2 + 2(\frac{\lambda}{\theta} - 1)\Psi + 1 = 0$, namely

$$-(\frac{\lambda}{\theta} - 1) \pm \sqrt{\Big((\frac{\lambda}{\theta} - 1)^2 - 1\Big)} = -(\frac{\lambda}{\theta} - 1) \pm \sqrt{\Big(-\frac{\lambda}{\theta}(2 - \frac{\lambda}{\theta})\Big)}. \tag{A.32}$$

Since $\lambda_1, \lambda_2 < 0$, $u(\lambda)$ and $v(\lambda)$ are real and not identical. Moreover, the limit of $\Psi^2 + 2(\frac{\lambda}{\theta} - 1)\Psi + 1$ at infinity is infinite and positive,$1^2 + 2(\frac{\lambda}{\theta} - 1)1 + 1 = 2\frac{\lambda}{\theta}$ is negative and $0^2 + 2(\frac{\lambda}{\theta} - 1)0 + 1 = 1$. Thus we have one root, say $u(\lambda)$, between 0 and 1 and the other, $v(\lambda)$, greater than 1. Writing eq.A.31in the form

$$\frac{1}{(\phi - \lambda)} = -\frac{2}{\theta}e^{\imath\xi}\frac{1}{u(\lambda) - v(\lambda)}\Big(\frac{1}{e^{\imath\xi} - u(\lambda)} - \frac{1}{e^{\imath\xi} - v(\lambda)}\Big), \tag{A.33}$$

eq.A.30 takes the form

$$\begin{aligned} |\mathcal{A}|^{-1} = &- \frac{2}{\theta(\lambda_1 - \lambda_2)}e^{\imath\xi}\frac{1}{u(\lambda_1) - v(\lambda_1)}\Big(\frac{1}{e^{\imath\xi} - u(\lambda_1)} - \frac{1}{e^{\imath\xi} - v(\lambda_1)}\Big) \\ &- \frac{2}{\theta(\lambda_2 - \lambda_1)}e^{\imath\xi}\frac{1}{u(\lambda_2) - v(\lambda_2)}\Big(\frac{1}{e^{\imath\xi} - u(\lambda_2)} - \frac{1}{e^{\imath\xi} - v(\lambda_2)}\Big) \quad . \\ = &- \frac{2}{\theta}e^{\imath\xi}\sum_{i=1}^{2}\frac{1}{\lambda_i - \lambda_{j\neq i}}\frac{1}{u(\lambda_i) - v(\lambda_i)}\Big(\frac{1}{e^{\imath\xi} - u(\lambda_i)} - \frac{1}{e^{\imath\xi} - v(\lambda_i)}\Big) \end{aligned} \tag{A.34}$$

Then eq.A.27 can be evaluated in a straightforward way using the above developments. Let I_2 be the 2 dimensional identity matrix and

$$\frac{1}{D_i} = \frac{2}{\theta}\frac{1}{\lambda_{j\neq i} - \lambda_i}\frac{1}{u(\lambda_i) - v(\lambda_i)}. \tag{A.35}$$

we can rewrite equation (A.27) as

$$
\begin{aligned}
\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix} &= \sum_{k=-\infty}^{\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{A}^{-1}(\xi)e^{\imath\xi(k-j)}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&= \sum_{k=-\infty}^{\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{\imath\xi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{1}{e^{\imath\xi}-u(\lambda_i)}-\frac{1}{e^{\imath\xi}-v(\lambda_i)}\Big)\\
&\quad\begin{pmatrix}\frac{n_\beta}{q_\beta}+m_\beta+\theta(1-\cos(\xi)) & m_\alpha\\ m_\beta & \frac{n_\alpha}{q_\alpha}+m_\alpha+\theta(1-\cos(\xi))\end{pmatrix}e^{\imath\xi(k-j)}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&= \sum_{k=-\infty}^{\infty}\Bigg[\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-v(\lambda_i)}\Big)\\
&\quad\begin{pmatrix}\frac{n_\beta}{q_\beta}+m_\beta+\theta & m_\alpha\\ m_\beta & \frac{n_\alpha}{q_\alpha}+m_\alpha+\theta\end{pmatrix}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&\quad-\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-v(\lambda_i)}\Big)\begin{pmatrix}\theta e^{\imath\xi} & 0\\ 0 & \theta e^{\imath\xi}\end{pmatrix}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&\quad-\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-v(\lambda_i)}\Big)\begin{pmatrix}\theta e^{-\imath\xi} & 0\\ 0 & \theta e^{-\imath\xi}\end{pmatrix}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\Bigg]\\
&= \sum_{k=-\infty}^{\infty}\Bigg[\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-v(\lambda_i)}\Big)\\
&\quad\begin{pmatrix}\frac{n_\beta}{q_\beta}+m_\beta+\theta & m_\alpha\\ m_\beta & \frac{n_\alpha}{q_\alpha}+m_\alpha+\theta\end{pmatrix}\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&\quad-\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{\theta e^{\imath\xi(k-j+2)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{\theta e^{\imath\xi(k-j+2)}}{e^{\imath\xi}-v(\lambda_i)}\Big)I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\\
&\quad-\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{2}\frac{1}{D_i}\Big(\frac{\theta e^{\imath\xi(k-j)}}{e^{\imath\xi}-u(\lambda_i)}-\frac{\theta e^{\imath\xi(k-j)}}{e^{\imath\xi}-v(\lambda_i)}\Big)I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}d\xi\Bigg]
\end{aligned}
\tag{A.36}
$$

Considering that for γ in the trigonometric circle

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{\imath\xi k}}{e^{\imath\xi}+\gamma}d\xi=\begin{cases}0, & k<1\\ \gamma^{k-1}, & k\geq 1.\end{cases}, \tag{A.37}$$

and for γ not in the trigonometric circle

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{\imath\xi k}}{e^{\imath\xi}+\gamma}d\xi=\begin{cases}-\gamma^{k-1}, & k<1\\ 0, & k\geq 1.\end{cases}, \tag{A.38}$$

we have to consider equation (A.36) as a sum of two terms, one with the $u(\lambda_i)$ terms and the second with $v(\lambda_i)$ terms. Then

$$\begin{aligned}\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_u =&\sum_{i=1}^{2}\frac{1}{D_i}\Bigg[\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-u(\lambda_i)}\Big)d\xi\Big]\\ &\times\begin{pmatrix}(\frac{n_\beta}{q_\beta}+m_\beta+\theta)\frac{n_{\alpha k}}{q_\alpha}+m_\alpha\frac{n_{\beta k}}{q_\beta}\\ m_\beta\frac{n_{\alpha k}}{q_\alpha}+(\frac{n_\alpha}{q_\alpha}+m_\alpha+\theta)\frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j+2)}}{e^{\imath\xi}-u(\lambda_i)}\Big)d\xi\Big]\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j)}}{e^{\imath\xi}-u(\lambda_i)}\Big)d\xi\Big]\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\Bigg],\end{aligned} \tag{A.39}$$

$$\begin{aligned}\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_v =&\sum_{i=1}^{2}\frac{1}{D_i}\Bigg[\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j+1)}}{e^{\imath\xi}-v(\lambda_i)}\Big)d\xi\Big]\\ &\times\begin{pmatrix}(\frac{n_\beta}{q_\beta}+m_\beta+\theta)\frac{n_{\alpha k}}{q_\alpha}+m_\alpha\frac{n_{\beta k}}{q_\beta}\\ m_\beta\frac{n_{\alpha k}}{q_\alpha}+(\frac{n_\alpha}{q_\alpha}+m_\alpha+\theta)\frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j+2)}}{e^{\imath\xi}-v(\lambda_i)}\Big)d\xi\Big]\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{\infty}\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{e^{\imath\xi(k-j)}}{e^{\imath\xi}-v(\lambda_i)}\Big)d\xi\Big]\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\Bigg],\end{aligned} \tag{A.40}$$

and

$$\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix} = \begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_u - \begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_v \tag{A.41}$$

Using equations (A.37) and (A.38), we have

$$\begin{aligned}\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_u =& \sum_{i=1}^{2}\frac{1}{D_i}\Bigg[\sum_{k=j}^{\infty}u(\lambda_i)^{k-j+1}\begin{pmatrix}(\frac{n_\beta}{q_\beta}+m_\beta+\theta)\frac{n_{\alpha k}}{q_\alpha}+m_\alpha\frac{n_{\beta k}}{q_\beta}\\ m_\beta\frac{n_{\alpha k}}{q_\alpha}+(\frac{n_\alpha}{q_\alpha}+m_\alpha+\theta)\frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=j-1}^{\infty}u(\lambda_i)^{k-j+2}\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}-\sum_{k=j+1}^{\infty}u(\lambda_i)^{k-j}\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\Bigg],\end{aligned} \tag{A.42}$$

and

$$\begin{aligned}\begin{pmatrix}\hat{\pi}(j|\alpha,\mathbf{n})\\ \hat{\pi}(j|\beta,\mathbf{n})\end{pmatrix}_v =& \sum_{i=1}^{2}\frac{1}{D_i}\Bigg[\sum_{k=-\infty}^{j}v(\lambda_i)^{k-j+1}\begin{pmatrix}(\frac{n_\beta}{q_\beta}+m_\beta+\theta)\frac{n_{\alpha k}}{q_\alpha}+m_\alpha\frac{n_{\beta k}}{q_\beta}\\ m_\beta\frac{n_{\alpha k}}{q_\alpha}+(\frac{n_\alpha}{q_\alpha}+m_\alpha+\theta)\frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{j-1}v(\lambda_i)^{k-j+2}\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\\ &-\sum_{k=-\infty}^{j+1}v(\lambda_i)^{k-j}\theta I_2\begin{pmatrix}\frac{n_{\alpha k}}{q_\alpha}\\ \frac{n_{\beta k}}{q_\beta}\end{pmatrix}\Bigg]\end{aligned} \tag{A.43}$$

✓**Proof** that the roots λ_1, λ_2 are < 0 and not identical.

The roots of $\phi^2 + \phi(x_\alpha + x_\beta) + x_\alpha x_\beta - m_\alpha m_\beta = 0$ are

$$\begin{aligned}&\Big(-(x_\alpha+x_\beta)\pm\sqrt{(x_\alpha+x_\beta)^2-4(x_\alpha x_\beta-m_\alpha m_\beta)}\Big)/2\\ =&\Big(-(x_\alpha+x_\beta)\pm\sqrt{(x_\alpha-x_\beta)^2+4m_\alpha m_\beta}\Big)/2.\end{aligned} \tag{A.44}$$

The roots are thus different since $m_\alpha m_\beta > 0$ and

$$\begin{aligned}
&\Big(-(x_\alpha + x_\beta) + \sqrt{(x_\alpha - x_\beta)^2 + 4m_\alpha m_\beta}\Big)/2 \\
&\quad < \Big(-(x_\alpha + x_\beta) + \sqrt{(x_\alpha - x_\beta)^2 + 4x_\alpha x_\beta}\Big)/2 \\
&\quad < \Big(-(x_\alpha + x_\beta) + \sqrt{(x_\alpha + x_\beta)^2}\Big)/2 \\
&\quad < 0.
\end{aligned} \tag{A.45}$$

Both roots are therefore real, negative and not equal.

Annexe B : Publications

B-1

Leblois, R., Rousset F., Tikel D., Moritz C., Estoup A. 2000.
Absence of evidence for isolation by distance in an expanding cane toad (Bufo marinus) population : an individual-based analysis of microsatellite genotypes.
***Molecular Ecology*. 9 : 1905-1909.**

Molecular Ecology (2000) **9**, 1905–1909

SHORT COMMUNICATION

Absence of evidence for isolation by distance in an expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes

RAPHAEL LEBLOIS,*† FRANÇOIS ROUSSET,‡ DANI TIKEL,* CRAIG MORITZ* and ARNAUD ESTOUP*†

*Department of Zoology and Entomology, University of Queensland, Queensland 4072, Australia, †Laboratoire Modélisation et Biologie Evolutive, CBGP-INRA, 34090 Montpellier, France, ‡Laboratoire Génétique et Environnement, CNRS-UMR 5554, 34095 Montpellier, France

Abstract

The cane toad (*Bufo marinus*) was introduced in 1935 in Australia, where it spread rapidly. We have tested for isolation by distance by analysing at a local geographical scale a continuous population using seven microsatellite markers and an individual-based method. The matrix of pairwise individual differentiation was not significantly correlated with that of geographical distance. Regression analyses gave a low positive slope of 0.00072 (all individuals) or a negative slope of 0.0017 (individuals with a distance higher than the previously estimated mean dispersal distance). The absence of evidence for isolation by distance favours the hypothesis that the substantial differentiation and autocorrelation previously observed at enzyme loci, mainly results from discontinuities in the colonization process with founder effects occurring at the time of the establishment of new populations.

Keywords: individual based analysis, invading species, isolation by distance, local scale differentiation, microsatellite DNA, neighbourhood size

Received 31 March 2000; revision received 29 June 2000; accepted 29 June 2000

Introduction

Understanding the evolutionary dynamic of invasive species may help to construct predictive models for future spread and design measures of biological control. The cane toad (or giant toad *Bufo marinus*) is by far the most widely and successfully introduced amphibian species (Sabath *et al*. 1981). This species is native to the American tropics and was deliberately introduced in 1935 as a biocontrol agent in Australia, where it spread across more than one million km^2 and continues to colonize new areas. This strong invading potential translates into high colonization rates ranging from one to 30 km per year (Van Beurden & Grigg 1980; Sabath *et al*. 1981; Easteal & Floyd 1986). The rapid expansion of *B. marinus* suggests that the species is very mobile, and as a result, there could be a large amount of gene flow between its populations, reducing the rate at which genetic differentiation can occur between them. Paradoxically, substantial genetic differentiation was found among Australian cane toad populations sampled sometimes over relatively short distances (e.g. $\approx 50 \times 80$ km for the Moreton bay region, Australia) (Easteal 1985). Moreover, spatial analysis of populations differentiation revealed significant autocorrelation over various distance classes at most enzyme loci (Easteal *et al*. 1985).

It remains unclear whether the geographical pattern of variation observed at enzyme loci has been shaped predominantly by an isolation by distance process due to limited dispersal, or by complex demographic events which occurred during the recent range expansion of the species in Australia (e.g. discontinuities in the colonization dynamics with founder events occurring at the time of the establishment of new populations), and to what degree this pattern has been influenced by natural selection at some enzyme loci (Easteal 1985, 1988; Guinand & Easteal 1996). Beside the possible problem of selection, the main difficulties in interpreting the previous enzyme data sets are the mixture of geographical scales and the

Correspondence: Arnaud Estoup. Department of Zoology and Entomology, University of Queensland, Queensland 4072, Australia. Fax: 61 7-33 65 16 55; E-mail: aestoup@zoology.uq.edu.au

different establishment times for populations. This confuses the effect of recent demographic history and isolation by distance. A clear reference to a specific model of population structure (e.g. island or isolation by distance models) is also often missing.

In this paper, we have specifically tested the occurrence of isolation by distance by analysing at a local geographical scale a continuous population of *B. marinus* established for a relatively long time. Such analysis was achieved using microsatellite markers and an individual-based method which formally refers to isolation by distance models (Rousset 1997, 2000a).

Materials and methods

Statistical treatments

In a continuous population, genetic differentiation among individuals is expected to increase with their geographical distance measured at a local scale when isolation by distance occurs (Wright 1943). This can be quantified by a generalization of the theory of *F*-statistics. Here we consider the statistic a_r, a generalization of $F_{ST}/(1 - F_{ST})$ between pairs of individuals (Rousset 2000a). When the continuous population is represented by a two dimensional lattice (i.e. fixed individual positions and no spacial density heterogeneity) and when applied on a small geographical scale, a_r is approximately linearly related to the logarithm of distance, $a_r \approx (\ln(d)/4\pi D\sigma^2) + \text{constant}$, where d is the geographical distance between two individuals, D is the density of effective individuals, σ^2 is the second moment of the dispersal distance (i.e. the mean squared parent–offspring distance), and the constant is the value of the linear approximation at d = 1 length unit. Thus, the inverse of the regression slope provides an estimate of the 'neighbourhood size' $S = 4\pi D\sigma^2$. The statistic $\hat{a}$, a multilocus estimate of a_r computed for each pair of individuals, was regressed against the logarithm of the geographical distance between these individuals, as described in Rousset (2000a) for a two dimensional model. Note that the individual-based method used here is conceptually similar to the 'subpopulation'-based isolation by distance method described in Rousset (1997). Ninety-five per cent confident intervals around the regression slope value were computed using the ABC bootstrap procedure described in DiCiccio & Efron (1996), using code written in *Mathematica* (Wolfram 1999) after the S procedures available at http://www.stat.cmu.edu. All other analyses were performed using the version 3.2 of the package GENEPOP (Raymond & Rousset 1995).

The above method presents four interesting features: (i) It avoids the arbitrary setting of geographical limits for the sampling of subpopulations, a feature particularly useful when populations are mostly continuous as is the case for the cane toad; (ii) the variation of *F*-statistics with geographical distance gives more easily interpretable information than a *F*-statistic value computed over all units; (iii) the demographic model on which this method is based makes only weak assumptions on the distribution of dispersal distances and is robust for distribution of dispersal more leptokurtic than normal, a feature commonly observed in natural populations (Rousset 1997; 2000a); and (iv) studies at a local geographical scale are more likely to yield valuable estimates because heterogeneity of demographic parameters (e.g. spatial and historical variation in the dispersal or density of individuals) are reduced, as are their influences on heterogeneity of genetic parameters such as the ones considered here (Slatkin 1993; Rousset 2000b).

Sampling and microsatellite analysis

The *Bufo marinus* continuous population studied here is located in the region of Byron Bay (28°39′00″ S 153°37′00″ E, Australia), an area located 150 km south from the populations of the Moreton Bay region studied by Easteal (1985). Cane toads were introduced near Byron Bay in 1964. The generation time being approximately equal to one year (Easteal & Floyd 1986), this corresponds to a relatively long period of establishment (≈25–35 generations). The colonization front is currently (February 1999) at Woodburn, 90 km south from Byron Bay (A. Estoup, personal observation).

For application of the above individual based regression method, it is preferable to select individuals separated by distances shorter than $\approx 20\,\sigma$, with σ^2 the average squared axial parent–offspring distance (Rousset 1997; 2000a). Assuming that the rate of growth of newly formed populations (α) was large enough to consider that the rate of colonization ρ is equal to 2σ (Easteal & Floyd 1986), σ-values can be estimated from ρ values by using the relationship $\rho = \sqrt{2\alpha}$ (Skellam 1951). Mean rate of colonization in the Byron Bay region was estimated to be 1.07 km/year (Van Beurden & Grigg 1980) and 2.5 km/year (Easteal & Floyd 1986), which translate into parental dispersal rates of 0.535 and 1.25 km/generation, respectively. A total of 120 toads (90 matures and 30 immatures with snout-urostyle length > and <90 mm, respectively) were randomly collected in February 1999 along a 20-km transect. For each capture geographical coordinates were recorded and a toe was clipped and stored in 95% ethanol. DNA extractions were performed using a Chelex-based protocol (Estoup *et al*. 1996). Seven microsatellite loci (BM101, BM121, BM229, BM235, BM239, BM279 and BM322) were genotyped using fluorescent polymerase chain reaction (PCR) and an ABI sequencing machine (Applied Biosystem, Perkin Elmer) as described in Tikel *et al*. (2000). BM101 is an unpublished locus with 5′–3′ primer sequences GTTTCAGTAGGCAGGTGAAGA and ACCCATCCTCACAAAGGTC, allelic size range between

170 and 180 bp and PCR conditions similar to those of the locus BM128 (see Tikel *et al*. 2000).

Results and Discussion

Low level of isolation by distance

The matrix of pairwise multilocus a_r values estimated for all pairs among the 120 individuals was not significantly correlated with that of geographical distance (Mantel's test, $P = 0.82$). The regression method gave a low positive slope of 0.00072, which translates into an estimate of 'neighbourhood size' S of 1389 individuals. The ABC bootstrapping procedure gave a large 95% confidence interval with 0.025 and 0.975 threshold values being S = 90 and infinity. As the linear relationship is expected to be poor for individuals separated by a very small distance (Rousset 1997), it should be appropriate to remove individuals separated by a distance lower than σ in our analysis. A low negative slope (–0.0017) was obtained when pairs of individuals closer than 0.5 km were removed. The 0.025 and 0.975 threshold values of the 95% confidence interval were S = 67 and infinity. Thus, the estimates of S themselves give no evidence for isolation by distance. A plot of pairwise genetic differentiation between individuals against logarithm of distance as well as the regression lines including or excluding individuals distant by more than 500 m are presented in the Fig. 1.

This paper describes an application with microsatellite markers of Rousset (2000a) regression method based on individual genotypes. The variance of classical F_{ST} estimators (i.e. F_{ST} computed with subpopulation as unit) decreases with increasing genetic diversity of markers (Goudet *et al*. 1996). The same trend is expected for F_{ST}-like statistic between pairs of individuals. Hence, for a given number of loci, the regression method is expected to be more accurate with microsatellites than with less variable markers such as enzymes. However, in the Byron bay region, the level of variability at microsatellite loci was only slightly higher (mean allele number and gene diversity of 2.9 and 0.51, respectively) than that at enzymatic markers in the same area (e.g. 2.1 and 0.32, Easteal 1985). This is likely to be due to a series of bottlenecks which occurred during the introduction of *Bufo marinus* in the Caribbean and Pacific Islands (Easteal 1981), and possibly during the colonization of Australia (Slade & Moritz 1998). Therefore, the potential gain of precision from using microsatellites instead of allozymes may be small in the present study.

As a matter of fact, though 120 individuals and seven microsatellite loci were analysed, the level of precision of the regression method appears to be low as suggested by the large confidence intervals. A total of 12 microsatellite loci have been optimized on cane toad (Tikel *et al*. 2000), but three of them were monomorphic in the Byron bay region and two of them were discarded from analysis due to sex linkage (results not shown). More precision would be expected by genotyping more loci and, to a lesser extent, more individuals (Rousset 2000a), but it is likely that this would not alter the general conclusion of low isolation by distance in continuous populations of *B. marinus*.

Comparison with ecological, historical and demographic data

In contrast to the high level of philopatry observed in most amphibian species (reviewed in Waldman &

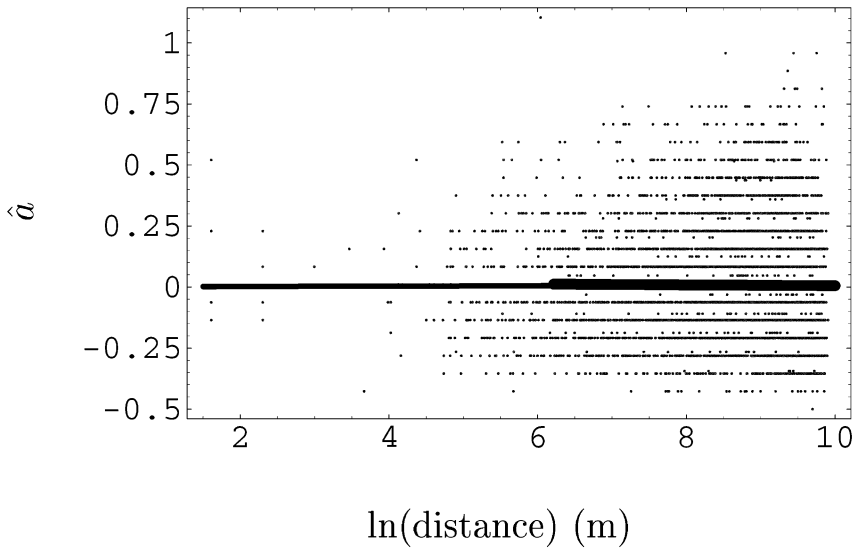

Fig. 1 Genetic differentiation in Cane toads. Pairwise genetic differentiation between individuals are plotted against logarithm of distance (in metres). All estimates for pairs of individuals at nonzero distance are shown, as well as the regression line computed from these estimates (thin line) and from estimates beyond 500 m (thick line).

McKinnon 1993), both mark-recapture and radiotracking studies have shown that cane toads rapidly move away from the location where they were captured (with individual distance per night ranging from 0 m to 1.3 km) and very seldom return (Bayliss 1994; Alford *et al*. 1995; L. Schwartzkopf & R. Alford, personal com.). This behavioural feature, as well as the high capability for rapid colonization of large areas in Australia (Easteal *et al*. 1985), are congruent with the absence of evidence for isolation by distance in the present study (Slatkin 1993).

Ecological, historical and demographic data available on *B. marinus* populations allows a rough estimation of population densities and dispersal rates, and hence that of 'neighbourhood sizes' ($S = 4\pi D\sigma^2$). Cane toad population density is known to vary in relation to time since colonization as well as to environmental features (Easteal & Floyd 1986; Freeland 1986). The population studied here is a relatively long established population located in semi-urban and agricultural areas. For this type of population, mark-recapture methods gave a density of 1500–3000 toads/km^2 in tropical and subtropical populations (Pearse 1979; Easteal & Floyd 1986) with ≈40% of mature individuals (Freeland 1986). *B. marinus* is a highly fecund species (7500–20 000 eggs/female, Alford *et al*. 1995), so it is likely that the reproductive variance is sufficiently large to substantially reduce the proportion of 'effective individuals'. However, no data are available allowing estimation of the female and male reproductive variance in cane toads. Hence, only a correction based on the proportion of mature individuals could be made on density estimates, so that mean dispersal rates of 0.535 km/year and 1.25 km/year (cf materials and methods) correspond to S-values between 2160 and 23 560 mature individuals in the Byron Bay region. Although those estimations are inaccurate and may be lower due to the reproductive variance, such large S-values are in agreement with the absence of evidence for isolation by distance (Fig. 1), and it is to be expected that the confidence intervals computed from microsatellite data would also contain large values.

Implications for the interpretation of genetic structure at different geographical and temporal scales

The absence of evidence for isolation by distance in a continuous population of *B. marinus* favours the hypothesis that the substantial genetic differentiation previously observed among populations of the Moreton Bay region, as well as the finding of significant autocorrelation over various distance classes at most enzyme loci (Easteal 1985), mainly results from discontinuities in the colonization process, with founder effects occurring at the time of the establishment of new populations. Theoretical studies have shown that founding events are likely to increase the divergence among populations (Slatkin 1977; Wade & McCauley 1988). A large increase of differentiation is particularly expected if the new populations are not immediately connected by gene flow to the main range of the species and if the population sampling includes a large proportion of recently founded populations (Slatkin 1993). Many of the populations studied in the Moreton bay region had been in existence for less than 10 years, and some for less than five, when they were sampled (Easteal & Floyd 1986). In contrast, the continuous population studied here has been established for 25–35 years and should be less affected by past demographic fluctuation than would populations closer to the colonization front. Finally, the extent and persistence of genetic differentiation may be locally enhanced by environmental features that are known to limit dispersal in *B. marinus*, especially large rivers, high density forest and mountainous areas (Zug & Zug 1979; Easteal 1985; Easteal & Floyd 1986). Studies at a local geographical scale, as in the present paper, are less prone to such environmental heterogeneity (Rousset 2000b) and, therefore, provide a clearer perspective on intrinsic limits to dispersal. We are currently completing additional population studies using microsatellite markers in order to characterize the colonization process of this invasive amphibian species in Australia.

Acknowledgements

We thank Chloe Schauble and Fiona Manson for assistance in toad sampling, David Paetkau for technical support, and Lin Schwartzkopf for sharing an unpublished manuscript and helpful discussions. This work was supported by a special investigator award from the Australian Research Council and a grant from the Institut National de Recherche Agronomique.

References

Alford RA, Cohen MP, Crossland MR, Hearnden MN, James D, Schwartzkopf L (1995) Population biology of *Bufo marinus* in Northern Australia. In: *Wetland Research in the Wet-Dry Tropics of Australia* (ed. Finlayson M), pp. 173–181. Office of the supervising Scientist report 101, Canberra, Australia.

Bayliss P (1994) *The ecology of post-metamorphic* Bufo Marinus *in Central Amazonian savanna, Brazil*. PhD Thesis. University of Queensland, Australia.

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals (with discussion). *Statistical Science*, **11**, 189–228.

Easteal S (1981) The history of introductions of *Bufo marinus* (Amphibia: anura); a natural experiment in evolution. *Biological Journal of the Linnean Society*, **16**, 93–113.

Easteal S (1985) The genetics of introduced populations of the giant toad *Bufo marinus*: III. Geographical patterns of variation. *Evolution*, **39**, 1065–1075.

Easteal S (1988) Range expansion and its genetic consequences in populations of the giant toad, *Bufo marinus*. *Evolutionary Biology*, **23**, 49–84.

Easteal S, Floyd RB (1986) The ecological genetics of introduced

populations of the giant toad, *Bufo marinus* (Amphibia: anura): dispersal and neighborhood size. *Linnean Society of London*, **27**, 17–45.

Estoup A, Largiader CR, Perrot E, Chourrout D (1996) Rapid one-tube DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Molecular Marine Biology and Biotechnology*, **5**, 295–298.

Freeland WJ (1986) Population of the cane toad, *Bufo marinus*, in relation to time since colonisation. *Australian Wildlife Research*, **13**, 321–330.

Guinand B, Easteal S (1996) Multivariate patterns of genetic differentiation support complex colonisation schemes in *Bufo marinus* populations. *Evolution*, **50**, 944–951.

Goudet J, Raymond M, de Meeus T, Rousset F (1996) Testing differentiation in diploid populations. *Genetics*, **144**, 1933–1940.

Pearse BW (1979) A population and home range study of *Bufo Marinus*. *Report to Australian Environmental studies*. Griffith University Brisbane, Australia.

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

Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Rousset F (2000a) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58–62.

Rousset F (2000b) Genetic approaches to the estimation of dispersal rates. In: *Dispersal: Individual, Population and Community* (eds Clobert J, Nichols JD, Danchin É, Dhondt A), Oxford University Press, Oxford, in press.

Slade RW, Moritz C (1998) Phylogeography of *Bufo marinus* from its natural and introduced ranges. *Proceedings of the Royal Society of London B*, **265**, 769–777.

Sabath MD, Boughton WC, Easteal S (1981) Expansion of the range of the introduced toad *Bufo marinus* in Australia from 1935 to 1974. *Copeia*, **1981**, 676–680.

Skellam JG (1951) Random dispersal in theoretical population. *Biometrica*, **38**, 196–218.

Slatkin M (1977) Gene flow and genetic drift in a species subject to frequent extinctions. *Theoretical Population Biology*, **12**, 253–262.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Tikel D, Peatkau D, Cortinas N, Leblois R, Moritz C, Estoup A (2000) Polymerase chain reaction primers for polymorphic microsatellite loci in the invasive toad species *Bufo marinus*. *Molecular Ecology*, **9**, 1927–1929.

Van Beurden EK, Grigg GC (1980) An isolated and expanding population of the introduced toad *Bufo marinus* in New South Wales. *Australian Wildlife Research*, **7**, 305–310.

Wade MJ, McCauley DE (1988) Extinction and recolonisation: their effects on the genetic differentiation of local populations. *Evolution*, **48**, 1114–1120.

Waldman B, McKinnon JS (1993) Inbreeding and outbreeding in fishes, amphibians and reptiles. In: *The Natural History of Inbreeding and Outbreeding: Theoretical and Empirical Perspectives* (ed. Thornhill NW), pp. 251–282. University of Chicago press, Chicago.

Wolfram S (1999) *The Mathematica Book*. 4th edn. Wolfram Media/Cambridge University Press, Cambridge.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Zug GR, Zug PB (1979) The marine toad, *Bufo marinus*: a natural history résumé of native populations. *Smithsonian Contribution to Zoology*, **284**, 1–58.

B-2

Leblois R., Estoup A., Rousset F. 2003.
Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance.
***Molecular Biology and Evolution*. 20 :491-502.**

Influence of Mutational and Sampling Factors on the Estimation of Demographic Parameters in a "Continuous" Population Under Isolation by Distance

Raphaël Leblois,† Arnaud Estoup,* and François Rousset†*

*Laboratoire Modélisation et Biologie Evolutive, CBGP-INRA, Montferrier sur Lez, France; and †Laboratoire Génétique et Environnement, CNRS-UMR 5554, Montpellier, France

In numerous species, individual dispersal is restricted in space so that "continuous" populations evolve under isolation by distance. A method based on individual genotypes assuming a lattice population model was recently developed to estimate the product $D\sigma^2$, where D is the population density and σ^2 is the average squared parent-offspring distance. We evaluated the influence on this method of both mutation rate and mutation model, with a particular reference to microsatellite markers, as well as that of the spatial scale of sampling. Moreover, we developed and tested a nonparametric bootstrap procedure allowing the construction of confidence intervals for the estimation of $D\sigma^2$. These two objectives prompted us to develop a computer simulation algorithm based on the coalescent theory giving individual genotypes for a continuous population under isolation by distance. Our results show that the characteristics of mutational processes at microsatellite loci, namely the allele size homoplasy generated by stepwise mutations, constraints on allele size, and change of slippage rate with repeat number, have little influence on the estimation of $D\sigma^2$. In contrast, a high genetic diversity (≈0.7–0.8), as is commonly observed for microsatellite markers, substantially increases the precision of the estimation. However, very high levels of genetic diversity (>0.85) were found to bias the estimation. We also show that statistics taking into account allele size differences give unreliable estimations (i.e., high variance of $D\sigma^2$ estimation) even under a strict stepwise mutation model. Finally, although we show that this method is reasonably robust with respect to the sampling scale, sampling individuals at a local geographical scale gives more precise estimations of $D\sigma^2$.

Introduction

Dispersal rates and population sizes or densities are important demographic parameters in evolutionary processes. Many studies have attempted to estimate such parameters using either direct methods (e.g., mark-recapture methods) or indirect methods (e.g., genetic markers). A number of indirect methods for demographic parameter estimation using genetic data at neutral loci or clines of selected markers have been defined (see Slatkin (1994) and Rousset (2001*b*) for reviews). Discrepancies between estimations made with direct and indirect methods have often been attributed to inadequacies of the assumptions of the genetic models made in indirect methods (Hastings and Harrison 1994; Koenig et al. 1996; Slatkin 1994). The kinds of assumptions usually considered to be inadequate are those related to (1) the modalities of dispersal (e.g., the island model), (2) the demographic stability in space and time, (3) the mutation rates and mutation processes of genetic markers, and (4) the selective neutrality of genetic markers.

In numerous species, individual dispersal is restricted in space. This means that there is a higher probability that individuals mate with individuals born in close proximity to themselves than to individuals born far away. Several studies on animals or plants have shown such restricted dispersal (e.g., for plant data, see Crawford 1984; and for animal data, Rousset 1997, 2000; Spong and Creel 2001; Sumner et al. 2001). Isolation by distance models taking into account this biological feature were introduced by Wright (1943 and 1946). Under these models the genetic differentiation at neutral loci is expected to increase with geographical distance (e.g., Malécot 1950, 1967; Sawyer 1977). Empirical data indicate that such a relationship holds for many species (Endler 1977; Slatkin 1993). Recently, a method of analysis was developed based on the increase, at a local scale, of genetic differentiation between individuals with geographical distance in a "continuous" population evolving under isolation by distance (Rousset 2000). The method makes use of the regression of estimators of a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals, and the logarithm of the geographical distance, to estimate the product $D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance. It is expected to perform better than previous methods for several reasons. First, the demographic model on which the method is based makes weak assumptions about the shape of the distribution of dispersal distances. In particular, the method is valid for leptokurtic distributions of dispersal distance (Rousset 2000), a feature commonly observed in natural populations (for review and data, see Endler 1977; Portnoy and Willson 1993; Clark et al. 1999). Second, analysis of genetic differentiation is made at a small (local) geographical scale so that heterogeneity of demographic parameters such as dispersal or density is reduced and hence its influence on genetic differentiation is also reduced (Slatkin 1993; Rousset 2001*b*). In a similar way, influence of non-neutrality of the genetic markers may be less problematic for studies at local scale because selection parameters may be less heterogeneous at a small geographical scale. On the other hand, the theory on which the method is based shows that only estimations from analysis over short distances will be accurate (Rousset 1997). These expectations have been confirmed by several comparisons of direct and indirect estimates of $D\sigma^2$ (Rousset 1997, 2000; Sumner et al. 2001). Although the geographical scale at which the sampling has been done is

Key words: coalescence, dispersal, isolation by distance, microsatellite DNA, nonparametric ABC bootstrap.

E-mail: leblois@isem.univ-montp2.fr.

Mol. Biol. Evol. 20(4):491–502. 2003
DOI: 10.1093/molbev/msg034

expected to influence the quality of the estimation of $D\sigma^2$, very few analytical or simulation studies have formally addressed this question.

Since their discovery in the 1980s, microsatellite loci have been increasingly used as genetic markers. Rapid progress in molecular biology technologies, especially the development of the polymerase chain reaction, and attractive evolutionary features (e.g., high level of polymorphism), explain why this category of markers are progressively replacing, or at least complementing, classical markers such as allozymes for numerous applications in molecular systematics, population genetics, and ecology (reviewed in Estoup and Angers 1998; Estoup, Jarne, and Cornuet. 2002). However, the mutation processes (i.e., the nature of mutations) at microsatellite loci are complex and not yet well understood (e.g., Estoup and Cornuet 1999). The effect of the mutation processes on evolutionary inferences depends in large part on the method, the statistics, and the evolutionary time scale considered (e.g., Estoup, Jarne, and Cornuet 2002). Some authors have discussed the effect of the nature of the mutation on F_{ST} values (Slatkin 1995; Rousset 1996). Because a stepwise mutation process occurs at microsatellite loci, several statistics taking into account the allele size have been proposed (Goldstein et al. 1995; Slatkin 1995; Michalakis and Excoffier 1996). Their utility, however, has often been criticized (e.g., Takezaki and Nei 1996; Gaggiotti et al. 1999). Overall, the potential interest of the different statistics has never been addressed in the context of the estimation of demographic parameters under isolation by distance.

In this study, we developed an original simulation algorithm based on the coalescent theory in order to study the sensitivity of the estimation of $D\sigma^2$ to different factors: (1) the sampling scale of individuals, (2) the mutation model of markers and (3) their mutation rate, with particular reference to microsatellite markers for the two latest points. This algorithm was also used to test a nonparametric ABC bootstrap procedure allowing the construction of confidence intervals on the $D\sigma^2$ estimation. Finally, we draw guidelines that could be useful for empirical investigators using the individual-based method of Rousset (2000).

Models and Methods

Demographic Model and Population Cycle

The model that we considered for "continuous" populations is the lattice model with each lattice node corresponding to one diploid individual. This model without demic structure is viewed as an approximation for truly continuous populations with infinite local competition (Malécot 1975; Rousset 2000). More realistic continuous models would incorporate the feature that individuals could settle in any position in a continuous space. Although such models have been formulated (e.g., Malécot 1967; Sawyer 1977), it is known that they do not follow a well-defined set of biological assumptions (Maruyama 1972; Felsenstein 1975; see Barton et al. 2002 for an alternative approach for continuous populations). Individuals are assumed to be diploids by a model with two independent genes per node. To avoid edge effects, the lattice is represented on a circle for a one-dimensional model or a torus for a two-dimensional model. Edge effects have little influence on local differentiation when the habitat area (i.e., the lattice size) is large when compared to the mean dispersal. Finally, we considered that dispersal occurs through gametes only.

The life cycle is divided into four steps: (1) at each reproductive event, each individual gives birth to a great number of gametes, and then dies; (2) gametes undergo the effect of mutations; (3) gametes disperse; (4) diploid individuals are formed, and (5) competition brings back the number of adults in each deme to one.

Coalescent Algorithm

The genealogical tree of a sample of n genes taken from a panmictic population of constant size N can be modeled using a stochastic process known as the n-coalescent. This process was introduced by Kingman (1982*a*, 1982*b*) as an approximation of a gene genealogy under the "Wright-Fisher" neutral model (see also Hudson 1990, Tajima 1983). More sophisticated models have since been developed for analysis of more complex evolutionary scenarios with recombination, selfing, and variable population size (reviewed in Nordborg 2001).

The n-coalescent approximation can be used in the same context as diffusion equations (Nordborg 2001). It is thus valid for a restricted numbers of models of population structure, e.g., panmictic populations or the infinite island model. In the present work, we focused on isolation by distance. For this category of models, no analytical treatment of coalescence time or coalescence probabilities has been done for more than two genes. Algorithms such as those developed for likelihood estimation by Griffiths and collaborators (see Nath and Griffiths 1996; Bahlo and Griffiths 2000) could in principle deal with continuous models; however, they are not ready for demographic inferences (De Iorio and Griffiths, personal communication). The coalescent algorithm we developed is not based on the n-coalescent theory; rather it is an algorithm for which coalescence and migration events are considered "generation by generation" until the common ancestor of the sample has been found. The idea of tracing lineages back in time generation by generation is fundamental in the coalescence theory, and is well described in Nordborg (2001). At least one study already used this simple concept for simulations (i.e., Pope, Estoup, and Morris 2000). Although such a generation-by-generation algorithm leads to less efficient simulations in terms of computation time than those based on the n-coalescent theory, it is much more flexible when complex demographic and dispersal features are considered. The algorithm described below and the program used in this study were checked at every step during elaboration by comparison with exact analytical results for probabilities of identity in models of isolation by distance on finite lattice (e.g., Malécot 1975 for the lattice model, adapted to different mutation models following Rousset 1996). These comparisons show that estimates of identity probabilities from our program and

analytical expectations differ by less than one per thousand for sufficiently long runs.

Let us consider, at a given time and on a two-dimensional lattice, a sample of $n(0)$ genes numbered 1 to $n(0)$. The position of each gene on this lattice is given by a pair of coordinates (x,y). The set of coordinates of sampled genes is given by the two vectors $X(0) = [x_1(0), \ldots, x_{n(0)}(0)]$, $Y(0) = [y_1(0), \ldots, y_{n(0)}(0)]$, where $x_i(0)$ and $y_i(0)$ are the coordinates of the gene i at $G = 0$, with G corresponding to the number of generations since sampling.

This algorithm goes backward in time, generation by generation (considering discrete generations). At $G = 1$, parents of our $n(0)$ sampled genes have coordinates $x_i(1) = x_i(0) + dx$, $y_i(1) = y_i(0) + dy$, where dx and dy are random variables representing dispersal distance in one dimension, expressed in number of steps on the lattice. Under a two-dimensional model, the density function of the random variable (dx,dy) is given by $b_{dx,dy}$, the "backward" dispersal function. The term *backward* is used because the position of the parental gene is determined knowing the position of its descendant gene. This function is calculated using $f_{dx,dy}$, the forward dispersal density function describing where descendants go. The dispersal functions are detailed in the next section. We assume that dispersal is independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. Considering that density is homogenous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$.

Once the position of the parents on the lattice is known, the coalescence events occurring at $G = 1$ are assessed. In other words, we determine whether some genes share a common parent at $G = 1$. This step corresponds to the idea of "individuals picking their parents at random from the previous generation" (Nordborg 2001). A coalescence event occurs if genes are both on the same lattice node and if they originate from the same parental gene. Multiple coalescences are allowed. The probability for a coalescence of k genes in a given parental gene is $1/2^{k-1}$ under the model with one individual per lattice node. In this case, the remaining j genes from the same lattice node coalesce in the other parental gene. For convenience, we keep the numbering ($i \in [1, \ldots, n(0)]$) of descendant genes for their parents when these genes do not coalesce and attribute new numbers ($i \in [n(0) + 1, \ldots, n(1)]$) for the parents of the coalesced genes. A gene i at $G = 0$ and its parent at $G = 1$ have the same number if there was no coalescence event between the gene i and another gene at $G = 0$. Thus our numbering refers more to the branches of the coalescent tree than to the genes themselves. This particular numbering of branches, nodes, and genes is illustrated in figure 1. At G = 1, we have $X(1) = (x_1(1), ., x_{n(1)}(1))$, $Y(1) = (y_1(1), ., y_{n(1)}(1))$, the $n(1)$ geographic coordinates at $G = 1$ for each branch corresponding to a lineage of our sample. We keep in memory the ages of the tree "nodes" (corresponding to coalescence events) and the labels of the branches descending from this "node." The entire process is repeated over generations until the most recent common ancestor of our entire gene sample has been found.

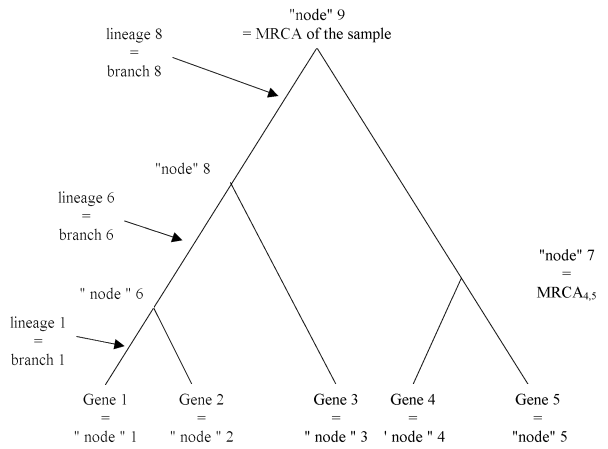

FIG. 1.—Numbering of branches, genes, and nodes of a genealogical tree for a sample of five genes as described by our coalescence algorithm.

Dispersal Functions

Biologically realistic dispersal functions often have a high kurtosis (Endler 1977; Kot, Lewis, and van den Driessche 1996). Forward dispersal distributions for which the probability of moving k steps (for $0 < k \leq K_{max}$) in one direction is of the form $f_k = f_{-k} = M/k^n$ were considered, with parameters M and n controlling the total dispersal rate and the kurtosis, respectively.

By suitable choice of the two parameter values, large kurtosis can be obtained with high migration rates (Rousset 2000). For all of our simulations, we used a dispersal distribution with a moderate σ^2 value ($\sigma^2 = 4$), corresponding to a dispersal distribution with parameters:

$$f_1 = f_{-1} = 0.06, \qquad f_2 = f_{-2} = 0.03 \quad \text{and for}$$
$$2 < k < 49, \qquad M = 0.802 \quad \text{and} \quad n = 2.518. \quad (1)$$

With such a dispersal distribution the product $4\pi D\sigma^2$ is 50.26. This value corresponds to a relatively strong isolation by distance, which appears biologically reasonable for many species (see references cited in the *Introduction*).

Mutation Processes

One interesting feature of the coalescent-based approach is that, for neutral loci, genealogical and mutation processes are totally independent, so that the effects of mutation are simply superimposed on the genealogical tree obtained for the gene sample.

Two theoretical mutation models, the infinite allele model (IAM: Kimura and Crow 1964) and the K-allele model (KAM: Crow and Kimura 1970), have sometimes been used for microsatellite loci. However, the most widely adopted model for microsatellite mutation is the stepwise mutation model (SMM: Ohta and Kimura 1973) in which the mutant allele differs from its parent by one repeat. Direct and indirect studies have shown that mutations of several repeats also occurred, indicating that a strict one-step model is inappropriate (Estoup and Angers 1998; Gonser et al. 2000; Ellegren 2000). In

practice, modeling assumptions are commonly limited to the SMM (e.g., Reich and Goldstein 1998; Wilson and Balding 1998), and sensitivity of the final inferences to this assumption may be substantial, although this is rarely investigated. In several studies (e.g., Pritchard et al. 1999), a generalization of the SMM was adopted in which the change in the number of repeat units forms a geometric random variable. This generalization was named the GSM (generalized stepwise mutation) model. The geometric distribution in our GSM model refers to a change expressed in an (absolute) number of repeat units subsequently added or withdrawn to the mutating allele with equal probability. Under this model, the large data set of microsatellite mutations of Dib et al. (1996) in humans suggests an estimate of the variance of the geometric distribution near 0.36 (Estoup et al. 2001). The GSM does not capture all the complexity of the mutation process at microsatellite loci. In particular, constraints on allele size occur at some microsatellite loci (reviewed in Amos 1999; Estoup and Cornuet 1999; Ellegren 2000) and potentially affect various statistics in population genetics (Estoup et al. 2002). This evolutionary feature, particular to microsatellite loci, was thus tested on our method. Allele size constraints were included in our simulations by imposing reflecting boundaries to the allele size range (e.g., Feldman et al. 1997; Estoup et al. 1999). Another outstanding feature of the microsatellite mutation process is that within-loci mutation rate increases with allele length (Ellegren 2000; Huang et al. 2002). Whether this increase is linear with the number of repeats remains subject to further investigation (Schlötterer 2000; Stumpf and Goldstein 2001; Brohede et al. 2002). In our simulations, we considered a linear model in which (1) the mutation rate was fixed to 5×10^{-4} for the allelic state of the root of the tree (fixed at 100 repeats units and considered the "middle size allele"); (2) a decrease in mutation rate with allele size of 0.1% or 1% per repeat unit for a weak or a strong variation, respectively is simulated for alleles shorter than 100 repeat units; (3) a similar increase is simulated for alleles longer than 100 repeat. In other words, this leads to the linear form: $\mu(L) = \mu_0 + s*L$, where $\mu(L)$ is the mutation rate for an allele of size L, μ_0 the mutation rate for the smallest allele, and s the increase per repeats unit. We set $s = 0.1\%$ or 1% for a weak or a strong variation, respectively, to be close to the value given in Brohede et al. (2002).

Interlocus variability in the mutation rate potentially decreases the precision of parameter estimation in population genetics (Takezaki and Nei 1996; Gonser et al. 2000). The effect of variable mutation rate was thus tested as well. Little information is available on the interlocus variance of the mutation rate at microsatellite loci. Several pedigree studies show that the mutation rates can differ across loci in important respects (reviewed in Schlötterer 2000). Without more information, we modeled variable mutation rates at microsatellite loci by drawing single locus mutation rate values in a gamma distribution with parameters (shape, scale) being (2, 2.5 10^{-4}). This distribution has a mean equal to 5×10^{-4}, a value considered as the average mutation rate in many species (reviewed in Estoup and Angers 1998), and 2.5% and 97.5% quantiles equal to 6×10^{-5} and 1.4×10^{-3}, respectively. These values are similar to the mean and 95% confidence interval values typically considered for autosomal microsatellites in humans (Weber and Wong 1993).

The following step-by-step procedure was used to add mutations to the genealogical tree. Take at random two genes i, j and their most recent common ancestor, the gene l, and let $state_i$, $state_j$, $state_l$ be their respective allelic states. The number of mutations that occurred in lineage i is proportional to the length L_i (expressed in number of generations) of branch i (from l to i) and is given by a binomial distribution with parameters (μ, L_i), which can be approximated by a Poisson process with parameter μL_i. Let m_i be the number of mutations that occurred on branch i. One can easily deduce $state_i$ from $state_l$ through m_i successive steps, each step corresponding to a mutation event under the chosen mutation model. The allelic states of the various genes of the sample were obtained starting from a given state for the common ancestor of the sample (root of the genealogical tree) and going forward in time on each branch.

Method of Analysis

Each simulation iteration gave the genotypes at l polymorphic loci for ($n \times n$) individuals denoted by their coordinates on the lattice. l independent coalescent trees were used to simulate multi-locus genotypes. This process was repeated 1,000 times giving 1,000 multilocus samples sharing the same demographic conditions. We computed estimates of the parameter

$$a_r \equiv \frac{Q_w - Q_r}{1 - Q_w}$$

for each pair of individuals, where Q_w is the probability of identity in state for two genes taken from the same individual, and Q_r the probability of identity in state for two genes at geographical distance r (Rousset 2000). The statistic a_r is a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals (and not between populations, as in Rousset 1997). An estimator of a_r for a pair π of individuals taken from the P different possible pairs is:

$$\hat{a} \equiv \frac{SS_{b(\pi)}P}{\sum_{k=1}^{P} SS_{w(k)}} - \frac{1}{2}$$

with

$$SS_{b[etween](\pi)} \equiv \sum_{i,u} (X_{i.:u} - X_{..:u})^2$$

and

$$SS_{w[ithin](\pi)} \equiv \sum_{i,j,u} (X_{ij:u} - X_{i.:u})^2,$$

where $X_{ij:u}$ is an indicator variable taking the value 1 if gene i of individual j is of allelic type u and the value 0 otherwise (Rousset 2000).

To test the effect of using a statistic that takes into account the allele length differences (and hence the stepwise mutational process occurring at microsatellite

loci), we defined another parameter b_r, equivalent to a_r, except that it is defined in terms of squared differences in microsatellite allele lengths (SD) instead of probabilities of non- identity in state ($1 - Q$). Thus, we have

$$b_r \equiv \frac{SD_r - SD_w}{SD_w},$$

where SD_r is the expectation of the squared length differences between two genes at geographical distance r and SD_w is the expectation of the squared length differences between two genes taken in the same individual. b_r was estimated for a pair π of individuals taken from the P different possible pairs in a way similar to a_r:

$$\hat{b} \equiv \frac{SSD_{b(\pi)}P}{\sum_{k=1}^{P} SSD_{w(k)}} - \frac{1}{2}$$

with

$$SSD_{b[etween](\pi)} \equiv \sum_i (S_{i.} - S_{..})^2$$

and

$$SSD_{w[ithin](\pi)} \equiv \sum_{i,j} (S_{ij} - S_{i.})^2,$$

where S_{ij} is a variable representing the size of gene i of individual j, expressed in number of repeat units.

For each of the 1,000 repetitions, the value of the slope of the regression line between $\hat{a}$ (or $\hat{b}$) and the logarithm of geographical distance was computed. In the limit of low mutation rates, the inverse of the slope is an estimate of the product $4\pi D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance (Rousset 1997). It is worth noting that high mutation rates should not result in an asymptotic bias as long as the focus is on local processes involving distances between sampled individuals

$$r \ll \frac{\sigma}{\sqrt{2\mu}}.$$

Beyond this limit, the linear relationship between a_r (or b_r) and the logarithm of the distance holds less well (for details, see Rousset 1997). Thus, if the analysis is done at a small geographical scale, the use of highly variable loci such as microsatellite loci should not bias the estimation. However, the effect of mutation on small sample properties of the estimator needs to be tested. The quality of an estimator is usually assessed through the computation of its bias and its mean square error (MSE). These measures are suitable when estimates have approximately a normal distribution but not when the estimate is sometimes infinite. In the present case, a negative slope should be interpreted as an infinite estimate of $D\sigma^2$. Therefore we chose to work on the slope values and not on $D\sigma^2$ estimates. The following statistics were estimated over all repetitions: (1) the mean relative bias between the value of the slope and the expected value $1/(4\pi D\sigma^2)$; (2) the standard error on this relative bias; and (3) the mean square error (MSE = Bias^2 + var). The bias and the MSE are relative values, as they are computed from the ratio of the estimate to the value to be estimated, $1/(4\pi D\sigma^2)$. We also computed the proportion of negative slopes found and the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$. Note that the latest measure is strictly equivalent to the probability that the $D\sigma^2$ estimate was within a factor of 2 from the expected $D\sigma^2$ value.

An accurate estimate of the uncertainty associated with parameter estimates is important to avoid misleading inferences. The nonparametric ABC bootstrap procedure described in DiCiccio and Efron (1996) was adapted to compute 95% confidence intervals around the regression slope. ABC bootstrap is a procedure that generates approximated bootstrap confidence intervals without real resampling. It is useful for estimation methods with high computation time needs. In this procedure, we considered genotypic data at each locus as independent replicates of the genealogical process. Tests of this procedure were performed using the same simulation program described above by calculating probability coverage of the confidence intervals for 1,000 simulated data sets. We choose arbitrarily a dispersal distribution with $\sigma^2 = 4$ [parameters given in equation (1)]. For each repetition, 100 individuals were sampled every two lattice nodes within an area of $(10\sigma \times 10\sigma)$ on a (100×100) lattice. Estimates of a_r and 95% confidence intervals were calculated for 7, 13, or 25 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}.

Table 1
Coverage Probability of 95% Confidence Intervals Around the Regression Slope Using an ABC Bootstrap Procedure

	Bootstrap Sample Size		
	7 loci	13 loci	25 loci
Coverage probability	0.842	0.885	0.90
Proportion of intervals below the slope value	0.020	0.030	0.030
Proportion of intervals above the slope value	0.138	0.085	0.070

Results

ABC Bootstrap

Table 1 shows that the non parametric ABC bootstrap procedure gives inaccurate 95% confidence intervals in terms of coverage probability even for large number of loci (e.g., coverage probability is 0.90 instead of 0.95 for 25 loci). The inaccuracy mostly concerns the lower bound of the confidence intervals for the regression slope (i.e., the proportion of intervals above the slope value is 0.07 instead of 0.025 for 25 loci; table 1). This may reflect the asymmetrical shape of the distribution with a long tail for small values (i.e., large $D\sigma^2$, data not shown). The effect of asymmetrical distribution on ABC bootstrap was tested on a simpler statistical model. ABC confidence intervals were computed for the mean of a random sample drawn in a bivariate student distribution with density

$$\Pr(r) = 2\pi r \frac{\Gamma[1+p]}{\pi u \Gamma[p]} (1 + r^2/u)^{-1-p}$$

and parameters (p,u) being (1,1). This distribution is asymmetrical with an infinite kurtosis and an infinite skewness. Even for very large sample sizes (5000

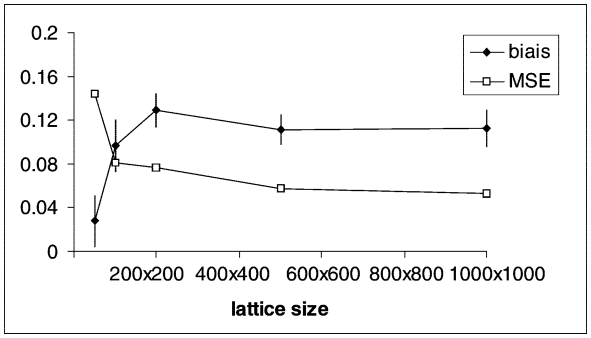

FIG. 2.—Influence of the lattice size on the estimation of the product $1/4\pi D\sigma^2$. NOTE—Only 500 iteration were done for each case. Vertical bars represent standard errors on the bias.

replicates, results not shown), the ABC procedure gives an inaccurate upper bound, resulting in underestimated confidence intervals (results not shown). In the case of the regression slope, the inaccuracy increases for small sample size (e.g., 0.842 instead of 0.95 for seven loci; table 1).

Because of the important computation time needed to construct ABC confidence intervals, this procedure was not used for evaluating the influence of the sampling scale and mutational factors on the estimation of $D\sigma^2$ (see *Models and Methods*).

Influence of the Sampling Scale

Previous simulations with two-allele loci suggested that the regression method would be efficient if one can sample all individuals within an area of about $10\sigma \times 10\sigma$, giving a sample size of $100D\sigma^2$ individuals (Rousset 2000). It is worth noting that if $D\sigma^2$ is greater than say 5, it becomes difficult in practice to sample and genotype all individuals (>500 individuals). Hence, since the number of individuals to sample is necessarily limited, the method should be less efficient when $D\sigma^2$ increases. In practice, biologists collect samples of a reasonably large number of individuals (say 100) within an area larger or smaller than the recommended ($10\sigma \times 10\sigma$) area when $D\sigma^2$ is small or large respectively. In order to assess the effect of such practical "non-scaled sampling," we simulated a distribution of dispersal with $\sigma^2 = 4$ [parameters given in expression (1)] and four different sampling schemes. One hundred individuals were taken: (1) every lattice node within an area of ($5\sigma \times 5\sigma$), for the first sampling scheme; (2) every two lattice nodes within an area of ($10\sigma \times 10\sigma$), for the second one; (3) every five lattice nodes within an area of ($25\sigma \times 25\sigma$) for the third one; and (4) every ten lattice nodes within an area of ($50\sigma \times 50\sigma$) for the last one. For each repetition the parameter estimated is a_r for 13 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}. We considered that a set of 13 loci represents a reasonable number of loci in empirical studies using microsatellites. A two dimensional lattice of (200×200) individuals was considered for the first three sampling schemes and of (500×500) individuals for the last one, to avoid edge effects on the estimations when considering samples larger than half the length of the lattice. Figure 2 shows that lattice size has no major effect on the estimation, except if it is less than ten times the mean dispersal distance (simulation parameters are those used in this paragraph). Unless the lattice size is very small (50*50), the bias and the MSE do not differ notably from those for a very large lattice size (1000*1000).

The sampling scale seems to have only a limited effect on the MSE of the $D\sigma^2$ estimation (table 2). Whatever sampling scale is considered (i.e., smaller or larger than the recommended area) the MSE is low (values between 5% and 12% in the studied cases). In contrast, the sampling scale has a great effect on the bias. A sample taken from an area two times smaller than the recommended area (first column of table 2) gave a large and positive bias (22%). The bias decreases when the sampling area increases and becomes negative when the sampling area is larger than the recommended area, reaching high values (e.g., −21%, fifth column of table 2). However, it is worth noting that even for extreme sampling situations, estimates of $D\sigma^2$ are not very different from the expected value, as shown by the large proportion of estimated values falling within a factor of two from $D\sigma^2$ (>93%).

Influence of the Mutation Model

The following mutation models were considered: (1) the infinite allele model (IAM); (2) the *K*-allele model (KAM) with an arbitrary choice of $K = 10$ possible allelic states; (3) the stepwise mutation model (SMM); (4) the generalized stepwise model (GSM) with variance of the geometric distribution equal to 0.36; and (5) the GSM with constraints on allele size (bounded GSM). In the bounded GSM, the number of possible allelic states was equal to 10 or 20, each allelic state being separated by a single repeat unit.

Simulations were run considering a sample of 100 individuals for 13 loci evolving in a two-dimensional lattice of (100×100) individuals. For each repetition of the simulation process the parameter estimated is a_r. As it is often not easy in practice to sample most individuals from a small area, we considered a sample of (10×10) individuals taken every two nodes from an area of (20×20) nodes in the lattice. By doing so, we approximated the sampling scheme typically used in empirical studies. We also chose a dispersal distribution with a relatively large σ^2 value [i.e., $\sigma^2 = 4$, parameters given in equation (1)]. The logic underlying this choice is that the method may be inaccurate in this case and that it is more relevant to distinguish differences in efficiency when the method does not perform extremely well, than when it performs well, whatever the mutation model.

The mutation rate was first fixed at 5×10^{-4} for all loci for each mutation model. Our results show that the nature of the mutation model has little influence on the estimation of the product $D\sigma^2$ (table 3). Whatever mutation model is considered, the bias is positive and around 10%. Although the precision of the method is maximum under the IAM (MSE of 6%) and minimum under the GSM with strong constraints ($K = 10$, MSE = 0.11), these differences are small. For all mutation models more

Table 2
Influence of Sampling Scale on the Estimation of $1/4\pi D\sigma^2$

	Sampling Scale (Sampling Area)			
	1 (10 × 10)	2 (20 × 20)	5 (50 × 50)	10 (100 × 100)
Bias	0.219	0.130	−0.056	−0.205
(standard error)	(0.0077)	(0.0077)	(0.0072)	(0.0064)
MSE	0.106	0.0763	0.0554	0.082
2× coverage	0.999	0.996	0.967	0.93
Negative slope	0	0	0	0

NOTE—Sampling area is expressed in lattice node unit (see text for details). 2× coverages correspond to the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$.

than 97% of the estimations are within a factor 2 from the expected $D\sigma^2$ value.

For a given mutation rate, level of genetic diversity varies according to the mutation model considered. Because the level of genetic diversity is likely to have an important effect on the estimation of the product $D\sigma^2$, we studied the influence of different mutational models for the same level of diversity. The genetic diversity can be expressed in terms of probability of identity by $(1 - Q_w)$, where Q_w is the probability of identity in state of two genes taken in the same individual. This corresponds to the fraction of heterozygous individuals in the population. The influence of mutation models was thus studied with the same Q_w value for all mutation models. The conclusions are similar to those obtained with a mutation rate fixed at the same value for all mutation models (table 3). For a given value of genetic diversity, the bias and the MSE of $D\sigma^2$ estimates shows little variation among mutational models.

Influence of the Mutation Rate

The influence of the mutation rate (or the genetic diversity) has been studied for the GSM, a mutation model considered as more realistic for microsatellite loci than the SMM, the KAM, or the IAM (e.g., Estoup and Cornuet 1999). All other simulation parameters are those used for evaluating the influence of the mutation model. Our simulations showed that the mutation rate has a substantial effect on the bias and the MSE (fig. 3 and table 4). The MSE is more strongly influenced by the mutation rate than the bias. For "low" genetic diversities (i.e., H = 0.5), the observed bias is positive and never greater than 12%. In contrast, for genetic diversity lower than 0.6, the MSE is greater than 20% and increases relatively rapidly when the genetic diversity decreases. However, even for a genetic diversity lower than the mean genetic diversity observed in most microsatellite studies (e.g., about 0.5), 85% of the estimations are within a factor of two from $D\sigma^2$, but 15 negative slopes were found (table 4).

It is worth mentioning that the observed bias may be of two types: (1) the bias, inherent in the method, that is due to the effect of high mutation rate on the parameter value (we will name it the "parametric bias"); and (2) the bias due to the deviation of the estimates in relation to the parameter value considering a finite sample of individuals and loci (which we will name "small sample bias"). The method is expected to perform poorly for very high mutation rates because distances between some pairs of sampled individuals are then larger than

$$\frac{\sigma}{\sqrt{2\mu}}$$

(Rousset 1997). In such a case, the parametric bias is expected to be negative because the slope of the regression line will be underestimated (for details, see Rousset 1997). In our simulations, we have $\sigma = 2$ and the maximal distance between individuals equals $20\sqrt{2}$ lattice units, which is within

$$\frac{\sigma}{\sqrt{2\mu}}$$

for mutation rates lower than 0.001. However, our results show that for a genetic diversity of 0.8 (corresponding to a mutation rate of c. 0.005 in our model) the bias and the MSE are very low. The low values of the bias and the MSE in this case are likely to result from some compensatory effects between a positive "small sample bias" and a negative "parametric bias." When higher genetic diversity is considered (i.e., H = 0.85 corresponding to mutation rates of c. 0.05 in our model), the bias becomes large and negative and the MSE rapidly increases (table 4). This result is in agreement with the above prediction: for very high mutation rates the "parametric bias" becomes more important than the "small sample bias," so that the global bias observed for high mutation rates is negative.

It is sometimes considered that the large variation between loci of the mutation rate decreases the precision of parameter estimation in population genetics (e.g., Takezaki and Nei 1996; Gonser et al. 2000). To address this question, we considered 13 loci evolving under the GSM with mutation rates drawn for each locus in a gamma distribution of mean 5×10^{-4} (see earlier under *Models and Methods: Mutation Model*), all other simulation parameter values being the same as those used in the previous section. Our simulation results show that variable mutation rates for microsatellite loci have little effect on the estimation of $D\sigma^2$ (table 4). The bias and the MSE values are 11% and 11%, respectively, which does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. More than 98% of the estimations are within a factor of 2 from $D\sigma^2$ and no negative estimates were found. Finally, our simulation results show that a linear increase in mutation rates with allele length has little effect on the estimation of $D\sigma^2$ (table 4). Strong or weak

Table 3
Influence of Mutational Processes on the Estimation of $1/4\pi D\sigma^2$ with Constant Mutation Rate or Constant Genetic Diversity for All Mutation Models

	Mutation Model											
	Constant Mutation Rate						Constant Genetic Diversity					
	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)
Genetic diversity	0.787	0.711	0.703	0.772	0.679	0.720	0.68	0.68	0.68	0.68	0.68	0.68
Mutation rate	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0001	0.000218	0.000342	0.00012	0.0005	0.0002
Bias	0.109	0.0919	0.0917	0.104	0.0997	0.128	0.111	0.104	0.118	0.121	0.0997	0.108
(standard error)	(0.0067)	(0.0088)	(0.0093)	(0.00863)	(0.0101)	(0.009)	(0.01)	(0.01)	(0.015)	(0.0120)	(0.0101)	(0.010)
MSE	0.057	0.0853	0.0953	0.0852	0.112	0.098	0.119	0.109	0.119	0.159	0.112	0.121
$2\times$ coverage	0.998	0.982	0.975	0.987	0.976	0.984	0.96	0.97	0.96	0.938	0.976	0.962
Negative slope	0	0	0	0	0	0	0.001	0.001	0	0.001	0	0.002

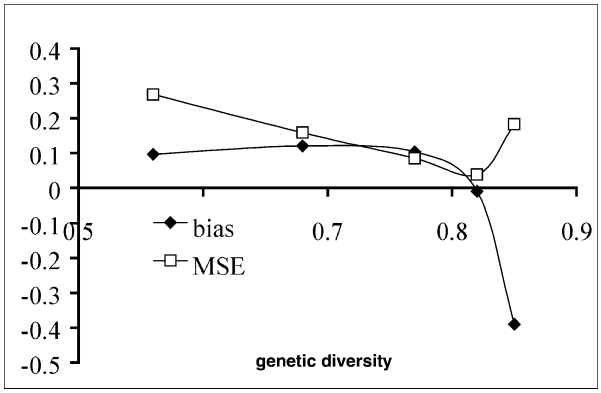

FIG. 3.—Influence of the mutation rate on the estimation of the product $1/4\pi D\sigma^2$. The mutation model is a GSM.

variations give similar results. The bias and the MSE values are about 10%–11% and 8%, respectively, which again does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. No negative estimates were found, and more than 99% of the estimations are within a factor of 2 from $D\sigma^2$.

Test for a Statistic Taking into Account Allele Size Differences

The behavior of the statistic b_r, an equivalent of a_r based on allele sizes, has been studied under both the SMM (i.e., the mutation model under which this statistic is expected to perform optimally) and the GSM with a mutation rate fixed at 5×10^{-4}. All other simulation parameters values are those used in the two previous sections. Table 5 shows that the method of estimation of $D\sigma^2$ performs poorly when b_r is used. Under both the SMM and GSM, the increase in MSE as well as the number of negative slopes is spectacular. For instance the MSE goes from about 10% when using the classical measure a_r to values greater than 100% when using b_r. In contrast, the bias is only slightly increased compared to estimations using a_r. Although slight, the bias increase appears higher under the GSM than the SMM (+ 9% versus + 4%).

Discussion

A first general conclusion of this study is that the mutation model of the markers has little influence on the efficiency of the method of estimation of $D\sigma^2$ based on individual genotypes and allelic identity. Hence, the allele size homoplasy typically produced under stepwise mutation models (SMM and GSM), and specifically of microsatellite markers (reviewed in Estoup, Jarne, and Cornuet 2002 for different population genetics statistics), is not a feature prejudicial for the method described in this article. Our results dealing with constraints on allele sizes, an evolutionary feature also specific to microsatellite markers and known for substantially increasing size homoplasy, show that even extremely strong constraints (e.g., $K = 10$) have little effect on the estimation of $D\sigma^2$. These results can be interpreted in the context of

Table 4
Influence of the Mutation Rate on the Estimation of the Product $1/4\pi D\sigma^2$

	Mutation Rate					Interloci Variability (*)	Intraloci Variability (**)	
	0.00005	0.00012	0.0005	0.005	0.05		Weak	Strong
Genetic diversity	0.56	0.68	0.77	0.82	0.85	0.77	0.77	0.77
Bias	0.0972	0.121	0.104	0.00946	−0.390	0.114	0.0965	0.111
(standard error)	(0.01609)	(0.0120)	(0.00863)	(0.00616)	(0.0055)	(0.0096)	(0.00846)	(0.0081)
MSE	0.268	0.159	0.0852	0.0380	0.182	0.105	0.0808	0.0778
2× coverage	0.844	0.938	0.987	0.996	0.761	0.983	0.991	0.993
Negative slope	0.015	0.001	0	0	0	0	0	0

NOTE.—The mutation model is a GSM. (*) Mutation rate drawn in a gamma (2, 2.5 10^{-4}) distribution. (**) Variation in mutation rate with allele length is 0.1% and 1% per repeat unit for weak and strong variation, respectively (see text under *Influence of Mutation Rate* for details).

coalescent theory. Values of F-statistics, under the assumption of low mutation rate, can be deduced from the comparison between the distributions of coalescence probability for different pairs of genes (e.g., pairs from the same deme and pairs from different demes) (Rousset 1996, 2002). These distributions differ essentially by an "excess" of coalescence probability for the most related genes, this excess being concentrated in a brief period in the recent past. Under isolation by distance, the more distant the demes are, the more the "recent past" is extended to the distant past, permitting more mutations to act and thus to increase the sensitivity to variation in the mutation process. By contrast, sensitivity to range constraints has been observed for statistics that are not related to differences of distribution of coalescence times (e.g., genetic distances, Nauta and Weissing 1996) or for F-statistics when the excess probability of coalescence is not concentrated in a recent enough past (large subpopulation sizes and low dispersal rates, Gaggiotti et al. 1999). Because the method of Rousset (2000) focuses on local differentiation and thus on recent evolutionary processes corresponding to a narrow recent past zone, it is no surprise that mutation processes (including allele size constraints) have little influence on the estimation of $D\sigma^2$.

A second major conclusion of this study is that the mutation rate, or the genetic diversity (the latest being largely dependent on the mutation rate), has a strong influence on the estimation of $D\sigma^2$. This is in agreement with previous studies demonstrating that mutation rate is a more important feature than mutation processes for the estimation of demographic parameters through F-statistics (reviewed in Rousset 2001*a*; Estoup, Jarne, and Cornuet 2002). Interestingly, the heterozygosities at microsatellite loci are typically between 0.5 and 0.8 (reviewed in Estoup and Angers 1998), a range of values corresponding to the level of genetic diversity that was found to maximize the efficiency of the estimation of $D\sigma^2$. Moreover, the potential effect on the estimation of interlocus and intralocus variability in the mutation rate seems to be weak. Therefore microsatellites are more appropriate to estimate the product $D\sigma^2$ than less polymorphic markers such as allozymes. The importance of the level of variability of the loci used to estimate population parameters has been illustrated by several theoretical and empirical studies. For example, Robertson and Hill (1984) showed that precision in estimates of heterozygote deficiency (F_{is}) increases with the level of variability of the markers. Goudet et al. (1996) also showed that the power of statistical tests of differentiation increases with the number of alleles. In practice, although precise information on mutation rate is difficult to obtain, it is straightforward to calculate a genetic diversity index for a set of markers from which a level of efficiency can be inferred for the estimation of $D\sigma^2$. Our simulations also indicate that future studies should avoid loci with a very high level of genetic diversity (higher than, say, 0.85), because those loci were found to strongly bias negatively the estimations of $D\sigma^2$.

Many studies emphasize that traditional F_{ST} does not make use of the additional information provided by the difference in the number of repeat units at microsatellite loci. However, statistics developed for this purpose often have higher variance than statistics based on allele frequencies (e.g., Gaggiotti et al. 1999). In agreement with this finding, estimates computed using a statistic taking into account allele size differences increases by at least a factor of 10 the MSE compared to a statistic based on identity in state. This result parallels those of Gaggiotti et al. (1999), which showed that in many cases, especially when sample size and number of loci are "small" (i.e., under the conditions of most empirical studies), population structure measures based on allele frequencies alone are more reliable than measures specifically designed for microsatellite loci. Takezaki and Nei (1996) also showed that even for loci evolving under a strict SMM, genetic distances taking into account allele size differences are less efficient for phylogenetic inference than those based on identity in state, especially for short to moderate divergence times. The poor efficiency of this category of statistics appears to be a general feature of studies of evolutionary events, especially those referring to fine geographical and temporal scales.

The effects of the mutation processes and high mutation rates on the estimation of $D\sigma^2$ are expected to be more important at large geographical scales (Rousset 1997). In agreement with this expectation, our results showed that sampling at large distance leads to an underestimation of the regression slope and thus to an overestimation of $D\sigma^2$. Therefore sampling at large distance makes it less likely to detect a pattern of isolation by distance. In contrast, sampling from too small an area leads to an overestimation of the regression slope and thus

Table 5
$D\sigma^2$ Estimation Using a Statistic Taking into Account the Differences in Allele Length (B_r)

	Mutation Model[a]			
	SMM	SMM	GSM	GSM
Parameter estimated	a_r	b_r	a_r	b_r
Bias	0.0917	0.128	0.104	0.19
(standard error)	(0.0093)	(0.036)	(0.00863)	(0.034)
MSE	0.0953	1.13	0.0852	1.25
2× coverage	0.975	0.518	0.987	0.497
Negative slope	0	0.154	0	0.141

NOTE.—Mutation rate is 5.10^{-4}.

[a] SMM: stepwise mutation model; GSM: generalized stepwise mutation model.

to an underestimation of the product $D\sigma^2$. A possible explanation for this overestimation is that the linear relationship between estimates of a_r and the logarithm of the geographical distance is expected to hold less well over very short distances (Rousset 1997). However, using a sample not exactly appropriate to the biological case studied [i.e., a few times larger or smaller than the recommended area of $(10\sigma \times 10\sigma)$] still gives reasonably robust estimations because, in most cases, the estimated $D\sigma^2$ fell within a factor of 2 from the expected $D\sigma^2$ value.

Given our result on bootstrap confidence intervals, we alert biologists using this method on a standard-sized data set (10 loci and 150 individuals, e.g., Sumner et al. 2001) that ABC confidence intervals overestimate the lower bound for the regression slope and thus underestimate the upper bound for $D\sigma^2$. Construction of reliable confidence intervals based on the bootstrap is an ongoing problem for which a satisfactory solution has not yet been found, especially when the number of replications is limited computationally (DiCiccio and Efron 1996). Nevertheless, the ABC bootstrap procedure evaluated here should give an idea of the uncertainty of the $D\sigma^2$ estimate, namely a correct lower bound for $D\sigma^2$ and a minimal value for the upper bound. This procedure will be implemented in the next version of the population genetics package Genepop (Raymond and Rousset 1995).

Conclusion

Three conclusions inferred from our simulation study have important consequences for empirical investigations. First, we recommended using loci with high levels of polymorphism (genetic diversity around 0.7), although loci with too high genetic diversity, e.g., more than 0.85, should be avoided. Because the mutational processes, specifically size homoplasy and allele size constraints, have little influence on $D\sigma^2$ estimations, microsatellite markers seem to be the best choice at the present time. Second, using statistics based on allele size differences at microsatellite loci gives unreliable estimations of $D\sigma^2$ because of the very high variance of those estimations. Third, it is important to restrict the sampling design to a relatively small geographical area in order to work at a local geographical scale; however, it is necessary to sample on a relatively large scale when σ is high. Optimizing the method studied here requires a previous knowledge of σ, and we therefore recommended using a preliminary estimate of σ to allow subsequent design of an appropriate sampling scheme. In the absence of a preliminary estimate of σ, a rough estimate of this parameter deduced from consideration of known dispersal mechanisms should be useful to define the minimal scale of the study (e.g., Leblois et al. 2000). If these aspects are approximately satisfied, the method should give estimates of the product $D\sigma^2$ with low bias and low mean square error. Finally, the ABC bootstrap procedure, as implemented in the package Genepop (Raymond and Rousset 1995), should be useful to estimate a 95% confidence interval on $D\sigma^2$, although the upper bound of this interval is likely to be underestimated.

Acknowledgments

We thank R. Streiff, B. Danforth, and three anonymous reviewers for constructive comments on an earlier version of the manuscript. This work was supported financially by the AIP no. 00202 "biodiversité" from the Institut Français de Biodiversité. This is paper 2003-002 of the Institut des Sciences de l'Evolution.

Literature Cited

Amos, W. 1999. A comparative approach to the study of microsatellite evolution. Pp. 66–79 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Bahlo, M., and R. C. Griffiths. 2000. Inference from GeneTree in a subdivided population. Theor. Pop. Biol. **57**:79–95.

Barton, N. H., F. Depaulis, and A. M. Etheridge. 2002. Neutral evolution in spatially continuous populations. Theor. Popul. Biol. **61**:31–48.

Brohede, J., C. Primmer, A. Møller, and H. Ellegren. 2002. Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. **30**:1997–2003.

Clark, J. S., M. Silman, R. Kern, E. Macklin, and J. HilleRis-Lambers. 1999. Seed dispersal near and far: patterns across temperate and tropical forests. Ecology **80**:1475–1494.

Crawford, T. J. 1984. The estimation of neighborhood parameters for plant populations. Heredity **52**:273–283.

Crow, J. F., and M. Kimura, 1970. An introduction to population genetics theory. Harper & Row, New York.

Dib, C., S. Faure, C. Fizames et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152–154.

DiCiccio, T. J., and B. Efron. 1996. Bootstrap confidence intervals (with discussion). Stat. Sci. **11**:189–228.

Ellegren, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**:400–402.

Endler, J. A. 1977. Geographical variation, speciation, and clines. Princeton University Press, Princeton, N.J.

Estoup, A., and B. Angers. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. Pp. 55–86 *in* G. Carvalho, ed. Advances in molecular ecology. NATO ASI series. IOS Press, Amsterdam.

Estoup, A., and J.-M. Cornuet. 1999. Microsatellite evolution: inferences from population data. Pp. 49–65 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasy at microsatellite loci and its consequences for population genetics analysis. Mol. Ecol. **11**:1591–1604.

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

Felsenstein, J. 1975. A pain in the torus: some difficulties with models of isolation by distance. Am. Nat. **109**:359–368.

Gaggiotti, O. E., O. Lange, K. Rassmann, and C. Gliddon. 1999. A comparison of two methods for estimating average levels of gene flow using microsatellites data. Mol. Ecol. **8**:1513–1520.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**:6723–6727.

Gonser, R., P. Donnelly, G. Nicholson, and A. Di Rienzo. 2000. Microsatellite mutations and inferences about human demography. Genetics **154**:1793–1807.

Goudet, J., M. Raymond, T. de Meeüs, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics **144**:1931–1938.

Hastings, A., and S. Harrison. 1994. Metapopulation dynamics and genetics. Annu. Rev. Ecol. Syst. **25**:167–188.

Huang, Q.-Y., F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu, Y.-Z. Liu, J.-L. Li, R. R. Becker, and H.-W. Deng. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyama and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics **49**:725–738.

Kingman, J. F. C. 1982*a*. The coalescent. Stochast. Proc. Appl. **13**:235–248.

———. 1982*b*. On the genealogy of large populations. J. Appl. Prob. **19A**:27–43.

Koenig, W. D., D. Van Vuren, and P. N. Hooge. 1996. Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11**:514–517.

Kot, M., M. A. Lewis, and P. van den Driessche. 1996. Dispersal data and the spread of invading organisms. Ecology **77**:2027–2042.

Leblois, R., F. Rousset, D. Tikel, C. Moritz, and A. Estoup. 2000. Absence of evidence for isolation by distance in expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes. Mol. Ecol. **9**:1905–1909.

Malécot, G. 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon A **13**:37–60.

———. 1967. Identical loci and relationship. Pp. 317–332 *in* L. M. Lecam and J. Neyman, eds. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4. California University Press, Berkeley.

———. 1975. Heterozygoty and relationship in regularly subdivided populations. Theor. Popul. Biol. **8**:212–241.

Maruyama, T. 1972. Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics **70**:639–651.

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. Genetics **142**: 1061–1064.

Nath, H. B., and R. C. Griffiths. 1996. Estimation in an island model using simulation. Theor. Pop. Biol. **50**:227–253.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143**:1021–1032.

Nordborg, M. 2001. Coalescent theory. Pp. 179–208 *in* D.A. Balding, M. Bishop and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

Pope, L. C., A. Estoup, and C. Moritz. 2000. Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropica*, determined using mtDNA and microsatellites. Mol. Ecol. **9**:2041–2053.

Portnoy, S., and M. F. Willson. 1993. Seed dispersal curves: behavior of the tail of the distribution. Evol. Ecol. **7**:25–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

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

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

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.

———. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

———. 2000. Genetic differentiation between individuals. J. Evol. Biol. **13**:58–62.

———. 2001*a*. Genetic approaches to the estimation of dispersal rates. Pp. 18–28 *in* J. Clobert, E. Danchin, A. A. Dhondt, and J. D. Nichols, eds. Dispersal: individual, population and community. Oxford University Press, Oxford.

———. 2001*b*. Inferences from spatial population genetics. Pp. 239–265 *in* D. A. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

Sawyer, S. 1977. Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Prob. **9**:268–282.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma **109**:365–371.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47**:264–279.

———. 1994. Gene flow and population structure. Pp. 3–17 *in* L. A. Real, ed. Ecological genetics. Princeton University Press, Princeton, N.J.

———. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

Spong, G., and S. Creel. 2001. Deriving dispersal distances from genetic data. Proc. R. Soc. Lond. Ser. B **268**:2571–2574.

Stumpf, M. P. H., and D. B. Goldstein. 2001. Genealogical and evolutionary inference with the human Y chromosome. Science **291**:1738–1742.

Sumner, J., F. Rousset, A. Estoup, and C. Moritz. 2001. "Neighborhood" size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol. Ecol. **10**:1917–1927.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Takezaki, N., and M. Nei. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellites DNA. Genetics **144**:389–399.
Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.
Wilson, I. J., and D. J. Balding. 1998. Genealogical inference from microsatellite data. Genetics **150**:499–510.
Wright, S. 1943. Isolation by distance. Genetics **28**:114–138.
———. 1946. Isolation by distance under diverse systems of mating. Genetics **31**:39–59.

Pierre Capy, Associate Editor

Accepted October 11, 2002

B-3

Leblois R., Rousset F., Estoup A. 2004.
Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population from microsatellite data.
***Genetics* 166 : 1081-1092.**

Influence of Spatial and Temporal Heterogeneities on the Estimation of Demographic Parameters in a Continuous Population Using Individual Microsatellite Data

Raphael Leblois,*,†,[1] François Rousset† and Arnaud Estoup*

**Centre de Biologie et de Gestion des Populations, Campus International de Baillarguet CS 30 016, 34988 Montferrier sur Lez, France and* †*Laboratoire Génétique et Environnement, Centre National de la Recherche Scientifique-UMR 5554, 34095 Montpellier, France*

Manuscript received July 4, 2003
Accepted for publication October 18, 2003

ABSTRACT

Drift and migration disequilibrium are very common in animal and plant populations. Yet their impact on methods of estimation of demographic parameters was rarely evaluated especially in complex realistic population models. The effect of such disequilibria on the estimation of demographic parameters depends on the population model, the statistics, and the genetic markers used. Here we considered the estimation of the product $D\sigma^2$ from individual microsatellite data, where D is the density of adults and σ^2 the average squared axial parent-offspring distance in a continuous population evolving under isolation by distance. A coalescence-based simulation algorithm was used to study the effect on $D\sigma^2$ estimation of temporal and spatial fluctuations of demographic parameters. Estimation of present-time $D\sigma^2$ values was found to be robust to temporal changes in dispersal, to density reduction, and to spatial expansions with constant density, even for relatively recent changes (*i.e.*, a few tens of generations ago). By contrast, density increase in the recent past gave $D\sigma^2$ estimations biased largely toward past demographic parameters values. The method was also robust to spatial heterogeneity in density and estimated local demographic parameters when the density is homogenous around the sampling area (*e.g.*, on a surface that equals four times the sampling area). Hence, in the limit of the situations studied in this article, and with the exception of the case of density increase, temporal and spatial fluctuations of demographic parameters appear to have a limited influence on the estimation of local and present-time demographic parameters with the method studied.

DISPERSAL rates and population sizes or densities are important demographic parameters in evolutionary processes. Many studies have attempted to estimate those parameters, using direct methods (*e.g.*, mark-recapture methods) or indirect methods (genetic markers). Discrepancies between estimations based on direct and indirect methods have often been attributed to inadequacies of the assumptions of the genetic models in indirect methods (Hastings and Harrison 1994; Slatkin 1994; Koenig *et al.* 1996). The assumptions that have usually been considered inadequate are those related to the modalities of dispersal (*e.g.*, the island model), the mutation rates and processes of genetic markers, the selective neutrality of genetic markers, and the demographic stability in time and space. The latter assumption raises the question of the exact meaning of demographic parameter estimations in biological systems for which temporal and/or spatial fluctuations of demographic parameters have occurred. With a few exceptions (*e.g.*, Stone and Sunnucks 1993; Beebee and Rowe 2001; Spong and Hellborg 2002), population geneticists usually consider that contemporary spatial patterns of diversity reflect the past more than the present-time population dynamics of a species. Whitlock and McCauley (1999) recently concluded that estimates of the number of migrants between subpopulations from *F*-statistics under the assumption of an island model at equilibrium were "likely to be correct within a few orders of magnitude" only because assumptions of the genetic model (*i.e.*, equal migration, no selection, and demographic stability) are often violated in biological systems. This degree of precision is of little value for understanding the present-time demographic processes of populations. This is particularly worrying in a practical context since reliable estimates of present or at least recent migration rates, dispersal distances, or densities are increasingly demanded as integral elements of applied management and conservation decisions.

The effect of temporal and spatial fluctuations on the estimation of demographic parameters strongly depends on the type and intensity of the fluctuation encountered. However, it also strongly depends on the population models assumed, the statistics computed, and the genetic markers used. Most studies dealing with disequilibrium situations referred to the classical island model or to the Wright-Fisher population model and

[1]*Corresponding author:* Laboratoire Génétique et Environnement, Institut des Sciences de l'Evolution, UMR 5554-CC065, Université des Sciences et Techniques du Languedoc, Pl. E. Bataillon, 34095 Montpellier, France. E-mail: leblois@isem.univ-montp2.fr

only a few of them have considered more sophisticated and realistic models (but see Slatkin 1993). In numerous species, individual dispersal is restricted in space (see references in Leblois *et al.* 2003). A method of analysis adapted to a "continuous" population evolving under isolation by distance was developed to estimate the product $D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance (Rousset 2000). This method uses a regression of estimators of a parameter a_r to the geographical distances or the logarithm of the geographical distances in one or two dimensions, respectively. The parameter a_r, defined in Rousset (2000), is analogous to the parameter $F_{ST}/(1 - F_{ST})$ but is calculated between individuals (see *Method of analysis* for details about this parameter and its estimator). The inverse of the slope of the regression line gives an estimate of $4\pi D\sigma^2$ (Rousset 1997). The method is valid for leptokurtic distributions of dispersal distance (Rousset 2000; Leblois *et al.* 2003), a feature commonly observed in natural populations (review and data in Endler 1977; Portnoy and Willson 1993). Because analysis of genetic differentiation is made at a small (local) geographical scale, heterogeneity of demographic parameters such as dispersal or density is reduced and hence its influence on genetic differentiation is also reduced (Slatkin 1993; Rousset 2001). The good properties of this method have been confirmed by comparisons of direct and indirect estimates of $D\sigma^2$ (Rousset 2000; Sumner *et al.* 2001).

As for any population genetics method of demographic parameter estimation, the quality of the estimation of $D\sigma^2$ using this method may be affected by local and temporal spatial heterogeneities in demographic parameters. In this study, we adapted the coalescence-based simulation algorithm of Leblois *et al.* (2003) to study the effect of temporal and spatial fluctuations of demographic parameters on the estimation of present-time $D\sigma^2$. Although one can imagine many scenarios dealing with demographic heterogeneities in space and time, we have chosen to focus our study on demographic scenarios often met in empirical surveys in conservation biology and in the study of introduced invading species. In this context, we assessed the effect on the estimation of the present-time $D\sigma^2$ of (i) a temporal change of the dispersal feature, (ii) a density reduction (bottleneck) or increase (flush) in time, (iii) a spatial expansion with constant density, and (iv) a sample of individuals taken from a high-density zone within a lower-density area.

MODELS AND METHODS

Spatial model and population cycle: The model that we considered for "continuous" populations is the lattice model with each lattice node corresponding to one diploid individual. This model without demic structure is viewed as an approximation for truly continuous populations with infinitely strong density regulation (Malécot 1975; Rousset 2000). More realistic continuous models would incorporate the feature that individuals could settle in any position in a continuous space. Although such models have been formulated (*e.g.*, Malécot 1967; Sawyer 1977), it is known that they do not follow a well-defined set of biological assumptions (Maruyama 1972; Felsenstein 1975; see Barton *et al.* 2002 for an alternative approach for continuous populations). To avoid edge effects, a two-dimensional lattice is represented on a torus. Edges and lattice size have little effect on local differentiation when the habitat area (*i.e.*, the lattice size) is large compared to the mean dispersal (Leblois *et al.* 2003). Finally, we considered diploid individuals with dispersal through gametes only. The life cycle is divided into five steps: (i) at each reproductive event, each individual gives birth to a great number of gametes and dies; (ii) gametes undergo the effect of mutations; (iii) gametes disperse; (iv) diploid individuals are formed; and (v) competition brings back the number of adults in each deme to N (usually $N = 1$ but see *Spatial and temporal heterogeneities*). We assume here random assortment of gametes present after dispersal at a given node. This is akin to random selfing in a population of N diploids without spatial structure, by which selfing occurs with frequency $1/N$. How alternative assumptions would affect the analysis is discussed below.

Coalescent algorithm: In this work, we focused on isolation by distance. For this category of models, no analytical treatment of coalescence time or coalescence probabilities has been done for more than two genes. The coalescent algorithm used in this study is thus not based on the large-N approximation of the n-coalescent theory; rather it is an exact algorithm for which coalescence and migration events are considered *generation by generation* until the common ancestor of the sample has been found. The idea of tracing lineages back in time generation by generation is fundamental in the coalescence theory, and is well described in Nordborg (2001). Such a *generation-by-generation* algorithm leads to less efficient simulations in terms of computation time than do those based on the n-coalescent theory (Kingman 1982a,b; Nordborg 2001). However, this algorithm is much more flexible when complex demographic and dispersal features are considered. Note that, since multiple coalescent events are taken into account by considering the probability of a coalescence event of k genes in a given parental node ($= 1/2^{k-1}$ under the model with one individual per lattice node), it allows us to build an exact coalescent tree under very small population size. The entire *generation-by-generation* algorithm that gives the coalescent tree for a sample of n genes evolving under isolation by distance, with density and dispersal homogenous in space and time, is detailed in Leblois *et al.* (2003). The algorithm and the program used in this study were checked at every step during its elaboration by comparing simulated values of probabilities of

identity of two genes under models of isolation by distance on finite lattices with their exact analytically computed values (*e.g.*, Malécot 1975 for the lattice model) with adaptation to different mutation models following general methods valid for any assumption about dispersal and density (Rousset 1996). These comparisons show that estimates of identity probabilities from our program and analytical expectations differ by less than one per thousand for sufficiently long runs.

Dispersal functions: Let (dx, dy) be the parent-offspring axial distance, backward in time, expressed in number of steps on the lattice. Under a two-dimensional model, the probability distribution of the random variable (dx, dy) is given by $b_{dx,dy}$, the "backward" dispersal function. The term backward is used because the position of the parental gene is determined knowing the position of its descendant gene. This function is calculated using $f_{dx,dy}$, the forward dispersal density function describing where descendants go. Biologically realistic dispersal functions often have a high kurtosis (Endler 1977; Kot *et al.* 1996). As previously explained (Rousset 2000), the commonly used discrete probability distributions for dispersal are not appropriate here because high kurtosis can be achieved only by assuming a low dispersal probability, *i.e.*, that most offspring reproduce exactly where their parents reproduced. Thus we used forward dispersal distributions for which the probability of moving k steps (for $0 < k \leq K_{\max}$) in one direction is of the form

$$f_k = f_{-k} = M/k^n, \tag{1}$$

with parameters M and n controlling the total dispersal rate and the kurtosis, respectively. This distribution corresponds to a truncated variant of the discrete Pareto, or ζ, distribution (see, *e.g.*, Patil and Joshi 1968). By suitable choice of the two parameter values, large kurtosis can be obtained with high migration rates (Rousset 2000). For some distributions, the first p terms were arbitrarily fixed:

$$f_1 = f_{-1} = M_1, \quad f_2 = f_{-2} = M_2, \ldots, \quad f_p = f_{-p} = M_p,$$
$$\text{and for } p < k \leq K_{\max}, \quad f_k = f_{-k} = M/k^n. \tag{2}$$

Dispersal was assumed to be independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. When density is homogenous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$.

Mutation processes: The number of mutations on each branch of the coalescent tree follows a binomial distribution with parameter (μ, L), where μ is the mutation rate and L the length of the branch. The allelic states of each gene of the sample were obtained starting from the common ancestor of the sample (root of the genealogical tree) from an allelic state determined according to a probability distribution determined by the mutation model and then going forward in time adding mutations one by one on each branch of the tree. The study of Leblois *et al.* (2003) stressed the interest in using loci with high levels of polymorphism for $D\sigma^2$ estimation. Therefore, microsatellite markers were simulated in the present study. On the basis of direct observations of mutations at human microsatellite loci (Dib *et al.* 1996; Ellegren 2000), the generalized stepwise model (GSM) in which the change in the number of repeat units forms a geometric random variable was adopted (Pritchard *et al.* 1999; Estoup *et al.* 2001). The variance of the geometric distribution was fixed at 0.36 (Estoup *et al.* 2001), a value computed from the mutation data in Dib *et al.* (1996). The mutation rate was equal to 5×10^{-4}, a value considered as the average mutation rate in many species (reviewed in Estoup and Angers 1998). The GSM does not capture all the complexity of the mutation process at microsatellite loci (reviewed in Ellegren 2000; Schlötterer 2000). However, Leblois *et al.* (2003) have shown that exact mutation processes, and in particular the occurrence of constraints on allele size and increase of mutation rate with allele length, have little influence on $D\sigma^2$ estimations.

Method of analysis: Each simulation iteration gives the genotypes at 10 polymorphic loci of 100 (*i.e.*, 10×10) individuals characterized by their coordinates on the lattice. Ten loci and 100 individuals were considered as representative of the number of loci and individuals commonly analyzed in empirical studies based on microsatellites. Independent coalescent trees were used to simulate multilocus genotypes at independent loci. In practice it is difficult to sample all individuals in a small area. Simulations were run for a sample of (10×10) individuals taken every two nodes from an area of (20×20) nodes in the lattice. In this we aimed to roughly mimic a sampling scheme commonly achieved in empirical studies. This process was repeated 1000 times giving 1000 multilocus samples of 100 individuals sharing the same demographic history.

For each simulated multilocus sample, estimates of the parameter $a_r = (Q_w - Q_r)/(1 - Q_w)$ were computed for each pair of individuals, with Q_w the probability of identity in state for two genes taken from the same individual and Q_r the probability of identity in state for two genes at geographical distance r (Rousset 2000). The parameter a_r is a parameter analogous to $F_{ST}/(1 - F_{ST})$ calculated between individuals (not between populations as in Rousset 1997). An estimator of a_r for a pair ξ of individuals taken from the P different possible pairs is

$$\hat{a} \equiv \frac{SS_{b(\xi)}P}{\sum_{k=1}^{P} SS_{w(k)}} - \frac{1}{2}, \tag{3}$$

where $SS_{b[\text{etween}](\xi)} \equiv \Sigma_{i,j}(X_{i:u} - X_{i:u})^2$ measures divergence between genes taken from two different individuals and $SS_{w[\text{ithin}](\xi)} \equiv \Sigma_{i,j,u}(X_{ij:u} - X_{i.:u})^2$ measures divergence between genes within the same individual ($X_{ij:u}$ is an indica-

tor variable taking the value 1 if gene i of individual j is of allelic type u and the value 0 otherwise; Rousset 2000). Thus, $\hat{a}$ compares the genetic divergence of individuals at distance r (numerator) to the divergence of the two-gene copy within the individual (denominator), which is essentially what the parameter a_r does. Because stepwise mutations occur at microsatellite loci, a statistic taking into account the allele size might appear to be attractive. However, Leblois *et al.* (2003) have shown that incorporation of allele size into the estimate of a_r gives unreliable results due to the high variance of the estimates. Therefore, only the parameter a_r described in Equation 3 was used in this study.

The generalized random selfing assumption made in this article implies that the identity within individuals is identical to the identity between juveniles competing for a site. More generally, $D\sigma^2$ is related to the parameter

$$\frac{\rho_r}{1-\rho_r} = \frac{Q_0 - Q_r}{((1-Q_w)/2) - Q_0}, \quad (4)$$

where Q_w is the probability of identity of genes within individuals, Q_r is the probability of identity of two genes in different individuals at distance r, and Q_0 is the probability of identity of two genes in different individuals in the same node (Rousset 2004, Equation 8.12). Without random selfing, $\hat{a}_r$ is not the most relevant statistic. Rather one should estimate not only Q_w but also Q_0. Since there is only one adult per node of the lattice, Q_0 cannot be estimated directly from adults: it must be approximated as the identity between close adults or (better) between close juveniles before competition (see Rousset 2004, Chap. 8, for further discussion). In this way, it is easy to adapt the methods considered in this article, but this is not considered further.

For each simulated data set, the value of the slope of the regression line between $\hat{a}$ and the logarithm of geographical distance was computed. In the limit of low mutation rates, the inverse of the slope is an estimate of the product $4\pi D\sigma^2$ (Rousset 1997). High mutation rates should not result in a large sample bias as long as one focuses on local processes involving distances between sampled individuals, $r \ll \sigma/\sqrt{2\mu}$. Beyond this limit, the linear relationship between a_r and the logarithm of the distance holds less well (see Rousset 1997 for theoretical details). Thus, if the analysis is done on a small geographical scale, the use of loci with high mutation rates such as microsatellites does not bias the estimation. This is illustrated by Leblois *et al.* (2003), using simulations.

The quality of an estimator is usually assessed through the computation of its bias and its mean square error (MSE). These measures are suitable when estimates have an approximately normal distribution but not when estimates are sometimes infinite. In the present case, a negative slope should be interpreted as an infinite estimate of $D\sigma^2$. Therefore, we present the bias and the MSE for the slope values of the regression lines and not for $D\sigma^2$ estimates. Thus, the following statistics were estimated over all repetitions: (i) the mean relative bias between the value of the slope and the expected value, $1/(4\pi D\sigma^2)$ [*i.e.*, (observed slope − expected slope)/expected slope]; (ii) the standard error on this relative bias; and (iii) the mean square error [*i.e.*, MSE = ((observed slope − expected slope)/expected slope)^2]. The bias and the MSE are relative values since they are computed from the ratio of the observed to the expected value. We also computed the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

Spatial and temporal heterogeneities: One important advantage of the generation-by-generation algorithm is that virtually any demographic model including those with variations in time and space of demographic parameters can be easily implemented.

Temporal change in dispersal: We first studied the effect of a simple decrease of dispersal capabilities in time. Decrease in dispersal under isolation-by-distance models can be modeled in various ways (*i.e.*, changing various parameters in the dispersal distributions). Here we considered a decrease over time of the average squared axial parent-offspring distance (σ^2). Two different dispersal distributions with different σ^2 values were used, while all other parameters of the distribution (*i.e.*, the global shape of the distribution) remained unchanged. This situation corresponds to a change in a landscape (*e.g.*, a fragmentation) resulting in modifying the ability of a species to move within this landscape (*e.g.*, Brooker and Brooker 2002). Simulations were run with a two-dimensional lattice of (500 × 500) nodes with one individual per node. A first dispersal distribution, given in expression (2) with parameters

$$M = 0.555 \text{ and } n = 2.744 \quad \text{for } 0 < k \leq 48, \quad (5)$$

has a moderate σ^2 value ($\sigma^2 = 4$ in lattice units) and is the dispersal distribution from the present until the time of change, G_c. A second dispersal distribution, with parameters $M = 0.187$ and $n = 1.246$ for $0 < k \leq 48$ corresponds to a very high σ^2 value ($\sigma^2 = 100$) and is the dispersal distribution from the time of change G_c until the time of the most recent common ancestor (TMRCA). Four simulations were run with $G_c = 10$, $G_c = 20$, $G_c = 100$ generations (going backward in time), and G_c infinite as baseline (*i.e.*, no change in dispersal features over time).

Temporal change in density: A second category of fluctuations is temporal variations in density of individuals. We studied two simple situations: (i) a decrease in density from past to present (population bottleneck) and (ii) an increase in density from past to present (population flush). Such bottleneck or flush events are expected to occur in endangered or invasive populations, respectively. These situations were implemented in our simula-

TABLE 1

Models used to study the effects of density variation in time on the estimation of $1/(4\pi D\sigma^2)$

Demographic change	Density (no. of individuals per lattice node): From sampling time to G_c	Density (no. of individuals per lattice node): From G_c to the TMRCA	Factor
Bottleneck			
Weak decrease	1	10	10
Strong decrease	1/9	10	90
Flush			
Weak increase	1	1/9	9
Strong increase	1	1/100	100

The number of generations, G_c, indicates the moment in the past when the density variation occurred. TMRCA corresponds to the time of the most recent common ancestor of the sampled genes.

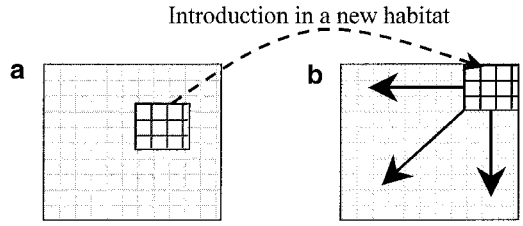

FIGURE 1.—Schema of a demographic expansion with constant density as modeled in this study. (a) The source population from which a subpopulation (dark gray grid) is introduced in an empty habitat (dotted arrow). (b) The empty habitat on which the introduced population spreads within a few generations (solid arrows). In our simulations, two-dimensional habitats are represented on a torus and not on a plane square as in this figure.

tions by changing the number of individuals per lattice node over time. Four different lattice models were used: one with 1 individual per node, one with 10 individuals per node, one with 1 individual every 3 nodes in each direction, and one with 1 individual every 10 nodes in each direction. These models correspond to densities of 1, 10, 1/9, and 1/100, respectively. Having less than 1 individual per node avoids the consideration of models with a too high number of individuals per node (*i.e.* >10) before or after a change in density, which would strongly deviate from the concept of continuous population to which the method of estimation applies. For easier coding, we modeled densities lower than 1 individual per node, considering that a given proportion of nodes of the lattice are always "empty" (*e.g.*, for a density of 1/9, 8/9 of the nodes are empty). This is equivalent to a model with a larger lattice unit (*e.g.*, a lattice unit three times larger in each dimension for a density of 1/9 compared to the lattice unit for a density of 1). A summary of the different density changes studied is presented in Table 1.

For the model with 1 individual every 9 nodes, we adapted the dispersal distribution to keep a constant $\sigma^2 = 4$. Since dispersal may occur only between "non-empty" nodes, the dispersal distribution parameters are then $M = 0.299$ and $n = 4.159$ for $0 < k \leq 48$. For the model with 1/100, 1, or 10 individuals per node, the dispersal distribution parameters are those used in the previous section [*cf.* expression (5)]. We have not adapted the dispersal distribution to keep a constant $\sigma^2 = 4$ for the model with 1 individual every 100 nodes because it was mathematically impossible to adjust this distribution with a too small number of points in the distribution (*i.e.*, in this case, there are only five possible moves in each direction between "suitable" nodes, which are located at 0, 10, 20, 30, and 40 lattice units). However, additional simulations with a 90-fold density increase (from 1/9 to 10 individuals per node) and a dispersal distribution adapted to keep a constant σ^2 gave similar results (results not shown).

For each case of density change considered, four simulations were run, using a two-dimensional habitat of (500 × 500) nodes with $G_c = 10$, $G_c = 20$, $G_c = 100$ generations, and G_c infinite as baseline. For each bottleneck and flush case, we simulated a weak density variation (10 and 9 times density change, respectively) and a strong density variation (90 and 100 times density change, respectively). In the case of bottleneck, the low-density models (1 and 1/9 individuals per node for weak and strong variations, respectively) were implemented from sampling time to G_c and the high-density models (10 individuals per node) from G_c to the TMRCA. In the case of density flush, the high-density models (1 individual per node) were implemented from sampling time to G_c and the low-density models (1/9 and 1/100 individuals per node for weak and strong variations, respectively) from G_c to the TMRCA (Table 1).

Spatial expansion with constant density: The third type of studied situation is a population expansion in space with constant density of individuals (Figure 1). The population introduced into an empty habitat is composed of individuals that have evolved in a source population at equilibrium with some demographic features (*i.e.*, density and dispersal distribution). The introduced population spreads within a few generations on an empty two-dimensional habitat with the same demographic features as the source population. This situation corresponds to the case of an introduced species that colonizes a new territory with similar ecological features to that of its native territory. Before expansion (*i.e.*, at generation G_c), the introduced population is composed of 100 individuals located on a (10 × 10) area, which were sampled from a (10 × 10) area in the source population, which itself evolved on a (160 × 160) lattice. From generation G_c to present, the introduced population spreads over a lattice of (160 × 160) nodes. The

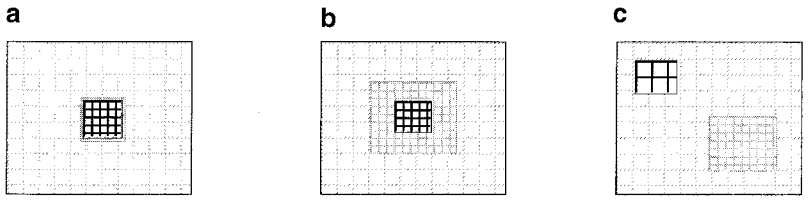

FIGURE 2.—Schema of the spatial density heterogeneities as modeled in this study. (a) A small high-density zone (dark gray grid) strictly corresponds to the sampling area (black grid) on a two-dimensional habitat with a lower density (light gray grid). (b) A large high-density zone (dark gray grid) includes the sampling area (black-grid) on a two-dimensional habitat with a lower density (light gray grid). (c) A large high-density zone (dark gray grid) is present on a two-dimensional habitat with a lower density (light gray grid); the sampling area (black grid) is located outside the high-density zone. In our simulations, two-dimensional habitats are represented on a torus and not on a plane square as in this figure.

entire (160 × 160) matrix is potentially occupied in two generations. At sampling time, as in the previous sections, 100 individuals were taken from an area of (20 × 20) nodes located outside the area of introduction, the distance between the introduction area and the sampling area being equal to 50 nodes. The forward dispersal distribution parameters are those given in expression (5) and correspond to a $\sigma^2 = 4$. Four simulations were run with $G_c = 10$, $G_c = 20$, $G_c = 100$, and G_c infinite as baseline.

Spatial density heterogeneities: The situations we choose to study reflect the fact that biologists usually collect individual samples in localities where they are easy to collect, that is, in high-density areas. Hence, we considered a lattice model with homogenous density except on a squared area where the density of individuals is higher (Figure 2). In such models with density heterogeneities in space, backward and forward dispersal differ. Each lattice node has a backward distribution that depends on the density of each surrounding node (*e.g.*, each node being at a distance less or equal to the K_{max} step). Those surrounding nodes correspond to all locations from which genes could have come in one generation (forward in time). Since those nodes are occupied by different numbers of individuals and because nodes occupied by more individuals contribute potentially more to the number of immigrants that reach a given node, we have to weight each term of the backward dispersal distribution by the number of individuals of the node from where immigrants have come. Let $N_{x,y,G}$ be the number of individuals at node (x, y) at generation G. Then for any node (x, y) the probability $b_{dx,dy}$ for a gene to move backward dx steps in one direction and dy in the other is equal to

$$b_{dx,dy} = \frac{N_{(x+dx),(y+dy),G} \cdot f_{dx,dy}}{\sum_{dx,dy \le K_{max}} N_{(x+dx),(y+dy),G} \cdot f_{dx,dy}}. \quad (6)$$

Simulations were run for a sample of 100 individuals taken every two nodes from an area of (20 × 20) nodes evolving in a (160 × 160) lattice. Density is one individual per node, except on a $(n \times n)$ zone including the sample area where density is 10 individuals per node. Two cases were considered: (i) a small high-density zone of (20 × 20) nodes, which strictly corresponds to the sample area (Figure 2a), and (ii) a larger high-density zone of (40 × 40) nodes, which includes the (20 × 20) nodes sample area (Figure 2b). We were particularly interested in assessing whether the estimated density corresponds to the density on the sampling area (*i.e.*, the local density) or whether the estimation is influenced largely by the density surrounding the sampling area (*i.e.*, the neighboring density). This was performed by alternatively considering that the expected $D\sigma^2$ value corresponded to a density of 10 (local density) and 1 (surrounding density) individuals per node. An additional simulation was run with a single large high-density zone of (40 × 40) nodes located outside the sampling area, the distance between the high-density and sampling zones being equal to 50 nodes (Figure 2c).

RESULTS

Interpretation of observed bias: Observed bias in our simulations might be attributable to (i) a bias, inherent to the method, due to the effect of a high mutation rate on the parameter value (this we call "mutational bias"), (ii) a bias due to the deviation of the estimates relative to the parameter value considering a finite sample of individuals and loci (this we name "small sample bias"), and (iii) a bias introduced by the demographic fluctuations studied. Additional details on the small sample and mutational biases can be found in LEBLOIS *et al.* (2003). All results in the present study should be interpreted taking into account the small sample and mutational biases that can be observed in the simulations without demographic fluctuations that were included in all situations studied as baseline (G_c infinite). For example, in the case of a reduction of density (bottleneck, Table 3), the mutational and small sample bias is large when considering an intermediate-density model (baseline simulation for a weak reduction) and much lower when considering a low-density model (baseline simulation for a stronger reduction). This difference is due partly to the different densities of individuals in the two baseline simulations, which influence the global level of genetic diversity in the sample. LEBLOIS *et al.* (2003) indeed showed that differences in genetic diversity have a substantial effect on the estimation of $D\sigma^2$.

TABLE 2

Effect of a temporal reduction of dispersal on the estimation of $1/(4\pi D\sigma^2)$

G_c	Infinite	100	20	10
Bias (standard error)	0.444 (0.0062)	0.0923 (0.0081)	−0.0795 (0.0076)	−0.234 (0.0074)
MSE	0.228	0.0743	0.0642	0.109
2× coverage	0.995	0.989	0.965	0.876

The number of generations, G_c, indicates the moment in the past when the dispersal reduction occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

Temporal change in dispersal: Simulation results show that the bias due to a reduction of dispersal is negative (Table 2) and thus corresponds to an overestimation of the present time $D\sigma^2$. This result is in agreement with a transition from a high $D\sigma^2$ value ($\sigma^2 = 100$) during the past generations (*i.e.*, before G_c) to a much lower value after G_c ($\sigma^2 = 4$). In other words, the method of $D\sigma^2$ estimation has a memory of temporal changes in dispersal. However, this memory is short term since a reduction of dispersal 100 generations ago gave only a slight negative bias compensated by the positive small sample and mutational biases (*cf.* first column of Table 2). Moreover, even for a recent reduction of dispersal ($G_c = 10$), the bias is <25% (*i.e.*, <0.25), a relatively low value compared to the high amplitude of the dispersal change. Standard error of the estimation also remains low for all G_c values, and for changes older than 20 generations, >95% of the estimations are within a factor of two of the present-time $D\sigma^2$. Hence, our simulations generally show that the precision of the present-time $D\sigma^2$ estimation is relatively robust to temporal changes in dispersal.

Temporal reduction of density (bottleneck): The negative bias observed in Table 3 (*i.e.*, overestimation of $D\sigma^2$) reflects the higher population density from generation G_c until the TMRCA. For a 10 times reduction of density, the method is quite robust when the density change occurred 20 or more generations ago. The bias and the MSE are low (<10%) and almost 99% of the estimations are within a factor of two of the present-time $D\sigma^2$ value. For very recent density change (*e.g.*, $G_c = 10$) the bias is substantial. However, the MSE remains low and >90% of the estimations are still within a factor of two of the present-time $D\sigma^2$ value.

The effect of reduction of density is more marked for a stronger change in density (*i.e.*, 90 times density reduction). For a very recent density reduction (*i.e.*, 10 generations ago), the negative bias reaches 50% and only 24% of the estimations are within a factor of two of the present-time $D\sigma^2$ value. For $G_c = 100$, the bias and the MSE become similar to the baseline. Note that all estimations are within a factor of two of the present-time $D\sigma^2$ for $G_c \geq 20$. Therefore, even for large recent density reductions, the method appears to be relatively robust.

Temporal increase in density (demographic flush): The positive bias observed in Table 4, which corresponds to an underestimation of the present-time $D\sigma^2$, reflects the lower population density from generation G_c until the TMRCA. For a small increase in density (10

TABLE 3

Effect of a weak (10 times density reduction) and strong (90 times density reduction) bottleneck on the estimation of $1/(4\pi D\sigma^2)$

Intensity	G_c	Infinite	100	20	10
Weak	Bias (standard error)	0.444 (0.0062)	0.0990 (0.0070)	−0.0625 (0.0064)	−0.222 (0.0061)
	MSE	0.228	0.0588	0.0449	0.0868
	2× coverage	0.995	0.997	0.989	0.915
Strong	Bias (standard error)	−0.0138 (0.0042)	−0.0743 (0.0027)	−0.330 (0.0017)	−0.526 (0.0012)
	MSE	0.0175	0.0128	0.115	0.278
	2× coverage	1	1	1	0.238

The number of generations, G_c, indicates the moment in the past when the density reduction occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

TABLE 4

Effect of a weak (9 times density increase) and strong (100 times density increase) density flush on the estimation of $1/(4\pi D\sigma^2)$

Intensity	G_c	Infinite	100	20	10
Weak	Bias (standard error)	0.444 (0.0062)	0.315 (0.040)	0.685 (0.043)	1.4 (0.046)
	MSE	0.228	1.72	2.33	4.07
	2× coverage	0.995	0.45	0.381	0.238
Strong	Bias (standard error)	0.432 (0.00644)	0.648 (0.0094)	2.24 (0.015)	3.91 (0.0193)
	MSE	0.228	0.508	5.27	15.8
	2× coverage	0.999	0.89	0.00262	0

The number of generations, G_c, indicates the moment in the past when the density increase occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

times), the bias and the MSE are high even for a relatively ancient flush (*e.g.*, $G_c = 100$). The proportion of estimations being within a factor of two of $D\sigma^2$ remains small (<50%) even for $G_c = 100$. The effect of the flush also increases substantially with the intensity of the density change. For a 100-fold density change and for $G_c = 10$, the bias reaches 391% and none of the estimations are within a factor of two of $D\sigma^2$ (Table 4). Hence, although the bias and the MSE decrease when G_c increases, the estimation remains unreliable for both 100- and 10-fold density change. These results contrast sharply with those pertaining to bottlenecks and dispersal changes.

Spatial increase in population size with constant density (demographic expansion): All measures (bias, MSE, and proportion of estimates within a factor of two) indicate that the estimation of the present-time $D\sigma^2$ is good when the spatial expansion occurred 20 or more generations ago (Table 5). For $G_c = 10$ only, an 8% negative bias is observed, which corresponds to an overestimation of the present-time $D\sigma^2$ (Table 5). However, the MSE is very small (10%) and 97% of the estimations are within a factor of two of the expected $D\sigma^2$ value. Hence, a spatial expansion as modeled here has only a short-term and limited influence on the present-time $D\sigma^2$ estimation; the method is precise even for very recent expansions.

Spatial heterogeneity in density (sampling within a high-density zone): Table 6 shows that $D\sigma^2$ estimation is not robust when the high-density zone is small and strictly corresponds to the sampling area. The bias and MSE values indicate that in this case the low-density area surrounding the sampling area strongly influences the $D\sigma^2$ estimation, which becomes a bad measure of both local density (*i.e.*, the density on the sampling area) and surrounding density (*i.e.*, the density surrounding the sampling area). It can be seen, however, that two times coverage probabilities, although globally low, are higher when referring to the local rather than to the surrounding area density as expected ($D\sigma^2$ value 0.018 *vs.* 0.001). This suggests that there is a tendency for the method to measure the local rather than the surrounding density. This trend becomes obvious when looking at results for a larger high-density zone (Table

TABLE 5

Effect of a spatial expansion

G_c	Infinite	100	20	10
Bias (standard error)	0.430 (0.0076)	0.387 (0.0126)	0.133 (0.0111)	−0.0824 (0.0101)
MSE	0.243	0.23	0.08	0.0581
2× coverage	0.989	0.98	0.996	0.972

The number of generations, G_c, indicates the moment in the past when the spatial expansion occurred. The expansion occurred without density and dispersal changes. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

TABLE 6

Effect of spatial heterogeneities in density

	Local density		Surrounding density	
Spatial heterogeneity	Estimation	Control	Estimation	Control
Small high-density zone				
Bias (standard error)	2.11 (0.017)	0.45 (0.025)	−0.689 (0.0017)	0.430 (0.0076)
MSE	4.76	0.83	0.477	0.243
2× coverage	0.018	0.65	0.001	0.989
Large high-density zone				
Bias (standard error)	0.393 (0.013)	0.45 (0.025)	−0.861 (0.0013)	0.43 (0.0076)
MSE	0.330	0.83	0.743	0.243
2× coverage	0.9	0.65	0	0.989
Large high-density zone outside sampling area				
Bias (standard error)	0.447 (0.00752)	0.43 (0.0076)	13.5 (0.0752)	0.45 (0.025)
MSE	0.256	0.243	187	0.83
2× coverage	0.99	0.989	0	0.65

Sampling was done on a small or large high-density zone of (20 × 20) and (40 × 40) nodes, respectively. Local density, the expected density is the local density (*i.e.*, density in the sampling area); surrounding density, the expected density is the surrounding density (*i.e.*, around the sampling area). Controls correspond to a homogenous lattice with density being the local or the surrounding density for the local and surrounding estimation cases, respectively. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

6). In this case, the bias and the MSE are much lower when considering the local rather than the surrounding zone for the $D\sigma^2$ value. About 90% of the estimates are within a factor of two of the local $D\sigma^2$ value, while none of them are within a factor of two of the surrounding $D\sigma^2$ value. The third case of a large high-density zone located outside the sampling area (*i.e.*, 50 nodes away) confirms this result (Table 6). Hence, our simulations generally show that the method estimates local demographic parameters and is robust for such measurement when the density is relatively homogenous around the sampling area (*e.g.*, over an area equal to four times the sampling area).

DISCUSSION

This work is the first one focusing on the study of evolutionary disequilibrium situations in the complex but realistic population model of a continuous population evolving under isolation by distance. Within the limits of the situations studied in this article, and with the exception of the case of a density flush, we found that temporal and spatial fluctuations of demographic parameters, if not too strong and not too recent (*i.e.*, more than, say, 20–50 generation in the past), have a limited influence on the estimation of local and present-time demographic parameters with the method of Rousset (2000). It is worth noting that we are talking about changes on timescales of a few tens of generations in the past, which may be very recent by standards in population genetics, but not for lots of species undergoing demographic changes due to ongoing human impact. Moreover, the numbers of generations defining the time of demographic change in this study should be considered as indicative of only the length of the effect of the demographic changes studied rather than as absolute reference numbers. As a matter of fact, the persistence in time of the effect of demographic fluctuations strongly depends on various features of the demographic model (*e.g.*, σ^2 values) and disequilibrium situations. It is thus preferable to consider general trends rather than precise numbers for each situation. For clarity, those trends have been summarized in Table 7.

The robustness of the method of Rousset (2000) to several temporal and spatial demographic fluctuations somewhat contradicts previous studies dealing with the study of evolutionary disequilibrium. In their review, Koenig *et al.* (1996) concluded that estimations of dispersal parameters from genetic data give ideas about past rather than present dispersal and gene flow, so that direct methods, such as mark-recapture methods, should give a better estimation of actual dispersal parameters. Boileau *et al.* (1992) similarly showed that hundreds or thousands of generations are required to erase the effects of colonization processes on "F_{ST}-like estimates" from allozyme data in large populations, con-

TABLE 7

Qualitative summary of the effects of different temporal and spatial heterogeneities

		Effect on $D\sigma^2$ estimation			
		Bias			
Demographic change		Sign	Intensity	2× coverage	Duration
Temporal	Dispersal increase (25 times)	Positive	Medium	Good	Short
	Density decrease (10–90 times)	Positive	Low to medium	Good to poor	Short
	Density increase (9–100 times)	Negative	High	Poor	Medium
Spatial	Local high-density zone (10 times)	Negative	Low (local) to high (surrounding)	Good (local) to poor (surrounding)	NA
Temporal and spatial	Spatial expansion	Negative	Low	Good	Short

Low intensity, mean relative bias <50%; high intensity, mean relative bias >100%; good, 2× coverage >85%; poor, 2× coverage <85%; short duration, few (10–20) generations; medium duration, >100 generations; NA, not appropriate.

cluding that estimates of gene flow from genetic data should be taken with care. We fully agree that temporal demographic fluctuations in a population are likely to have a strong and persistent effect on some population genetics statistics and methods. However, the present study shows that some indirect methods and genetic markers give accurate estimations of present-time density and dispersal features even when the demographic history includes relatively recent demographic changes.

The general robustness to spatial and temporal heterogeneities of the present *F*-statistic-based method can be interpreted using arguments from the coalescence theory and analytical treatment available in this field. Values of *F*-statistics, under the assumption of low mutation rate, can be deduced by comparing the distributions of coalescence probability for different pairs of genes (*e.g.*, pairs from the same deme and pairs from different demes; *e.g.*, Rousset 2002). These distributions differ essentially by an excess of coalescence probability for the most related genes, this excess being concentrated in a brief period τ in the recent past. *F*-statistics thus depend mainly on differences between the distributions of coalescence probability for different pairs of genes in recent generations. As the sensitivity of *F*-statistics values to past demographic fluctuations is also related to this recent time period, past demographic fluctuations have less effect when the time period τ is short. This recent time period τ is shorter when high dispersal rates and/or low deme size are considered (Rousset 2004). Hence, if models with small deme size and high migration rates, such as isolation by distance between individuals where each deme is of size two genes, are considered the influence of past demographic fluctuations on the estimation of demographic parameters from *F*-statistics is limited. By contrast, under the classical island model with large deme size and low migration rates, the effect of past demographic fluctuations is expected to be more problematic. Moreover, under isolation-by-distance models, the more distant the demes are on the lattice, the more the period τ is expanding to the past, increasing the effect of past demographic parameter fluctuations (Slatkin 1994; Rousset 2004). Because the present method focuses on local differentiation and thus on recent evolutionary processes corresponding to a narrow recent past zone, it is again logical that past demographic fluctuations have limited effects on the estimation of the present-time and local $D\sigma^2$ with this method. The same reasoning can be used to understand why the method gives estimates of the local demographic parameter values rather than estimates of the surrounding demographic parameter values. As the period τ is short in the models considered, *F*-statistics depend mainly on genetic events (migration, coalescence, mutation) that occurred in a recent past and, because dispersal is localized, at a local geographical scale. Therefore, the estimate of $D\sigma^2$ by the present method should correspond to the local demographic parameter values on the sampling area and should not be much influenced by demographic features of zones that are far away from the sampling area.

Close examination of our results brings up several issues. Our simulations showed that, for the study of invading species, the present method should give precise estimates of the present-time $D\sigma^2$ provided that no demographic flush occurred during the expansion process. This is an interesting feature of the method, which makes it appropriate to study invasive organisms for which demographic features are similar in the newly founded population and in the original source population. Our simulations further showed that if a change in dispersal occurred during the invasion process, this new dispersal feature should translate quickly in the estimation of the present-time $D\sigma^2$. On the other hand, density flushes (and to a much lower extent population

bottlenecks) may strongly affect present-time $D\sigma^2$ estimation. Invading species populations often experience complex demographic fluctuations that may include both bottlenecks (*i.e.*, founder events) and density flushes during their spreading (*e.g.*, Williamson 1996; Estoup *et al.* 2001). Therefore, it seems necessary to run additional simulations adapted to those complex demographic scenarios to thoroughly evaluate the robustness of the estimation of the present-time $D\sigma^2$.

Our simulations also show that for conservation biology studies dealing with bottlenecked populations the estimation of $D\sigma^2$ is potentially biased toward past demographic parameter values. However, the memory of past demographic parameter values is short so that this bias is important for only a strong and recent decrease in density. A major genetic consequence of a population bottleneck is that the number of alleles decreases much faster than the heterozygosity (Nei *et al.* 1975; Luikart and Cornuet 1998). One might have expected the precision of the method to be reduced due to the lower number of alleles in the bottlenecked population. However, standard error on the bias was weak whatever the strength of the bottleneck. One possible explanation for this result is that the method's precision depends more on the mean heterozygosity level than on the average number of alleles.

Our simulations indicate that surrounding densities considerably influence the estimation of local $D\sigma^2$ when the sample is taken on a small high-density zone. In this case, the estimates correspond neither to the $D\sigma^2$ values on the sampling area nor to the surrounding $D\sigma^2$ values. However, if sampling is done in a sufficiently large high-density zone (*e.g.*, on a surface equals to four times the sampling area), the estimates correspond more to the local density (*i.e.*, the density in the sampling area). Our simulations allowed us to study the case of a high-density zone in the middle of a large homogenous zone with low density. This situation is realistic for various demographic systems and mimics a classical experimental bias (*i.e.*, the fact that biologists generally collect their samples in high-density areas). However, many biological situations with spatial density heterogeneities would correspond rather to random density fluctuations on each lattice node. It is expected that differentiation in such scenarios will be a function of some "effective" density and dispersal rate. The lack of analytical formulas for these effective parameters limits the interpretation of a simulation study of the performance of estimators. Nevertheless, there is no obvious reason to believe that the estimation of the effective $D\sigma^2$ would be affected more by such random fluctuations than by previously studied spatial heterogeneities.

We thank Thomas Lenormand and Franck Shaw for constructive comments on the manuscript. This work was financially supported by the Action Incitative Programmée no. 00202 "biodiversité" from the Institut Français de la Biodiversité and grant no. D4E/SRP/01118 "biological invasion" from the Ministère de l'Ecologie et du Développement Durable. This is paper ISEM 2004-007.

LITERATURE CITED

Barton, N. H., F. Depaulis and A. M. Etheridge, 2002 Neutral evolution in spatially continuous populations. Theor. Popul. Biol. **61:** 31–48.

Beebee, T., and G. Rowe, 2001 Application of genetic bottleneck testing to the investigation of amphibian declines: a case study with natterjack toads. Conserv. Biol. **15:** 266–270.

Boileau, M. G., P. D. N. Hebert and S. S. Schwartz, 1992 Nonequilibrium gene frequency divergence: persistent founder effects in natural populations. J. Evol. Biol. **5:** 25–39.

Brooker, L., and M. Brooker, 2002 Dispersal and population dynamics of the blue-breasted fairy-wren, *Malurus pulcherrimus*, in fragmented habitat in the Western Australian wheatbelt. Wildlife Res. **29:** 225–233.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Ellegren, H., 2000 Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24:** 400–402.

Endler, J. A., 1977 Geographical variation, speciation, and clines. Princeton University Press, Princeton, NJ.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology NATO ASI Series*, edited by G. Carvalho. IOS Press, Amsterdam.

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

Felsenstein, J., 1975 A pain in the torus: some difficulties with models of isolation by distance. Am. Nat. **109:** 359–368.

Hastings, A., and S. Harrison, 1994 Metapopulation dynamics and genetics. Annu. Rev. Ecol. Syst. **25:** 167–188.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Koenig, W. D., D. Van Vuren and P. N. Hooge, 1996 Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11:** 514–517.

Kot, M., M. A. Lewis and P. van den Driessche, 1996 Dispersal data and the spread of invading organisms. Ecology **77:** 2027–2042.

Leblois, R., A. Estoup and F. Rousset, 2003 Influence of mutational and sampling factors on the estimation of demographic parameters in a 'continuous' population under isolation by distance. Mol. Biol. Evol. **20:** 491–502.

Luikart, G., and J.-M. Cornuet, 1998 Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conserv. Biol. **12:** 228–237.

Malécot, G., 1967 Identical loci and relationship, pp. 317–332 in *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and Probability*, Vol. 4, edited by L. M. Lecam and J. Neyman. University of California Press, Berkeley, CA.

Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **8:** 212–241.

Maruyama, T., 1972 Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics **70:** 639–651.

Nei, M., T. Maruyama and R. Chakraborty, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–10.

Nordborg, M., 2001 Coalescent theory, pp. 179–208 in *Handbook of Statistical Genetics*, edited by D. A. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Patil, G. P., and S. W. Joshi, 1968 *A Dictionary and Bibliography of Discrete Distribution.* Oliver & Boyd, Edinburgh.

Portnoy, S., and M. F. Willson, 1993 Seed dispersal curves: behavior of the tail of the distribution. Evol. Ecol. **7:** 25–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

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., 2000 Genetic differentiation between individuals. J. Evol. Biol. **13:** 58–62.

Rousset, F., 2001 Inferences from spatial population genetics, pp. 239–265 in *Handbook of Statistical Genetics*, edited by D. A. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Rousset, F., 2002 Inbreeding and relatedness coefficients: What do they measure? Heredity **88:** 371–380.

Rousset, F., 2004 *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Sawyer, S., 1977 Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Probab. **9:** 268–282.

Schlötterer, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109:** 365–371.

Slatkin, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264–279.

Slatkin, M., 1994 Gene flow and population structure, pp. 3–17 in *Ecological Genetics*, edited by L. A. Real. Princeton University Press, Princeton, NJ.

Spong, G., and L. Hellborg, 2002 A near-extinction event in lynx: Do microsatellite data tell the tale? Conserv. Ecol. **6** (1): 15.

Stone, G. N., and P. Sunnucks, 1993 Genetic consequences of an invasion through a patchy environment: the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Mol. Ecol. **2:** 251–268.

Sumner, J., F. Rousset, A. Estoup and C. Moritz, 2001 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol. Ecol. **10:** 1917–1927.

Williamson, M., 1996 *Biological Invasions*. Chapman & Hall, London.

Whitlock, M. C., and D. E. McCauley, 1999 Indirect measure of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity **82:** 117–125.

Communicating editor: L. Excoffier

B-4

De Iorio M., Griffiths R., Leblois R., Rousset F. 2004.
Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models.
Soumis à *Theoretical Population biology*

Version 18, 16th September, 2004
Compiled with LaTeX September 16, 2004
Revised for *Theoretical Population Biology*
Corresponding author: R. C. Griffiths
email: griff@stats.ox.ac.uk

Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models

Maria De Iorio, [1] **Robert C. Griffiths,** [2] **Raphael Leblois,** [3] **François Rousset.** [3]

Department of Mathematics, Imperial College, [1]
Department of Statistics, University of Oxford, [2]
Laboratoire Génétique et Environnement. [3]

ABSTRACT

An importance sampling algorithm for computing the likelihood of a sample of genes at loci under a stepwise mutation model in a subdivided population is developed. This allows maximum likelihood estimation of migration rates between subpopulations. The time to the most recent common ancestor of the sample can also be computed. The technique is illustrated by an analysis of a data set of Australian red fox populations.

Keywords. Coalescent process, Importance sampling, Migration, Stepwise mutation
AMS 2000 Subject Classification: 60G40 93E25 92D25.

---

[1]Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK. [2]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [3]Laboratoire Génétique et Environnement, CNRS-UMR 5554, 34095 Montpellier, France. MDI, RCG were supported by BBSRC Bioinformatics grant 43/BIO14435; RL, FR were supported by grants AIP Biodiversité INRA; and a ACI Jeune chercheur grant supported FR.

1 Introduction

Recently there has been much research into computation methods of ancestral inference from samples of genes conditional on their observed type configuration using Importance sampling (IS), MCMC and Bayesian techniques. In the simplest stepwise mutation model inference involves finding a maximum likelihood estimate of θ, the mutation rate. Wilson and Balding (1998) use a Bayesian MCMC scheme implemented in `Micsat` for microsatellite data, Beerli *et al.* (1999) use an MCMC scheme implemented in `Migrate` to estimate migration rates from data which could have a stepwise mutation mechanism. Nielsen (1997) uses an importance sampling algorithm based on an algorithm in Griffiths and Tavaré (1994). Stephens and Donnelly (2000) develop a sequential importance sampling technique, improving the algorithm of Griffiths and Tavaré (1994) to find the likelihood of samples of genes under a general mutation model which includes the stepwise model. Stephens (2001) gives an algorithm based on importance sampling to simulate genealogies of selected alleles in a population of variable size. The paper also describes the use of importance sampling methods in population genetics in a more general framework. Chen *et al.* (2004) improve sequential importance sampling by running multiple processes and resampling at sequential steps in the algorithm. A different approach based on F-statistics and analysis of variance analogues is reviewed in Excoffier (2001) and Rousset (2001). De Iorio and Griffiths, (2004a,b) develop a technique to construct sequential importance sampling proposal distributions on coalescent histories in population genetic models based on the diffusion process generator that describes the distribution of population gene frequencies. The technique extends proposal distributions of Stephens and Donnelly (2000) to a wider class of models, with a focus on subdivided population models of the island model type. Apart from likelihood calculations, ancestral inference questions involving time, such as the time to the most recent common ancestor (TMRCA) of the genes in a data set, can be answered by including the time between events in the underlying coalescent process.

The ancestry of a sample of n genes is described by a coalescent tree, (Kingman, 1982), where pairs of ancestral lines coalesce at unit rate forming a tree back in time to the ancestor of the sample genes. Mutations occur at rate $\theta/2$ along the edges of the coalescent tree according to a Poisson process.

An accessible introduction to the coalescent process is Nordborg (2001), with subdivided coalescent studies by Notohara (1990) and Herbots (1997). Two papers on ancestral inference are Griffiths (2001) and Stephens (2001). This current paper derives a detailed algorithm from the techniques in De Iorio and Griffiths (2004b) for likelihood calculations under a stepwise mutation model in a subdivided population.

In the stepwise mutation model the allele type space is the set of integers $\{\ldots -2, -1, 0, 1, \ldots\}$. Transitions of allele type when a mutation occurs are made according to a random walk from state j to a state $j + Z$, where Z is an integer-valued random variable. In the simplest case studied by Ohta and Kimura (1973), Moran (1975), Moran (1976), $Z = \pm 1$ with probability 1/2.

There is a distribution of the configuration of types in a sample with positions measured relative to the most recent common ancestor of the sample, which could be taken without loss of generality to be 0. The distribution is shift invariant, only depending on the relative positions of the types on a line.

In this paper an importance sampling algorithm for computing the likelihood of a sample of genes at loci under a stepwise mutation model in a subdivided population is developed. This allows maximum likelihood estimation of migration rates between subpopulations. The algorithm and program code used were thoroughly checked. Intermediate calculations also agree with exact analytic results for probabilities of identity of type under a finite island model. See Nagylaki (1983) for the finite island model, and Rousset (1996) for an adaption to different mutation models.

The precision of migration and mutation rate maximum likelihood estimates is checked by an analysis of multilocus simulated data from a two population model.

To illustrate the algorithm with a real data set, an analysis is performed on data from two Australian red fox *(Vulpes vulpes)* populations.

2 Coalescent histories and importance sampling

Let E be the set of possible types of a gene. Denote the sample configuration of the numbers of different types as $\boldsymbol{n} = (n_j, j \in E)$, and $p(\boldsymbol{n})$ the probability of obtaining a sample $\boldsymbol{n}$. $\boldsymbol{e}_j$ will denote the jth unit vector. For $j \in E$, let $\pi(j \mid \boldsymbol{n})$ be the probability that an additional gene chosen from the population is of type j, given that we have an observed configuration $\boldsymbol{n}$. These conditional distributions are important in the sampling distribution and coalescent history process. If $j_1, j_2, \ldots, j_n$ are the types of genes sequentially sampled such that there are $n_j, j \in E$ genes of type j in the sample, then

$$p(\boldsymbol{n}) = \frac{n!}{\prod_{j \in E} n_j!} \prod_{l=1}^{n} \pi(j_l \mid \boldsymbol{e}_{j_1} + \cdots + \boldsymbol{e}_{j_{l-1}}).$$

The distribution $p(\boldsymbol{n})$ is invariant under sequential sampling order.

A coalescent history $\{H_k,\ k = 0, -1, \ldots, -m\}$ is defined as the set of ancestral configurations at the embedded events in the Markov process where coalescence, mutation or other events take place. H_0 denotes the current state, and H_{-m} the state when a singleton ancestor is reached. The Markov nature of the process implies that

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}). \tag{2.1}$$

$p(H_k)$ and $\{p(H_{k-1})\}$ are unknown, whereas the probabilities $p(H_k \mid H_{k-1})$ are easily derived from the distribution of the coalescent tree. A coalescent history is illustrated in Figure 1. In (2.1) history probabilities are evaluated in

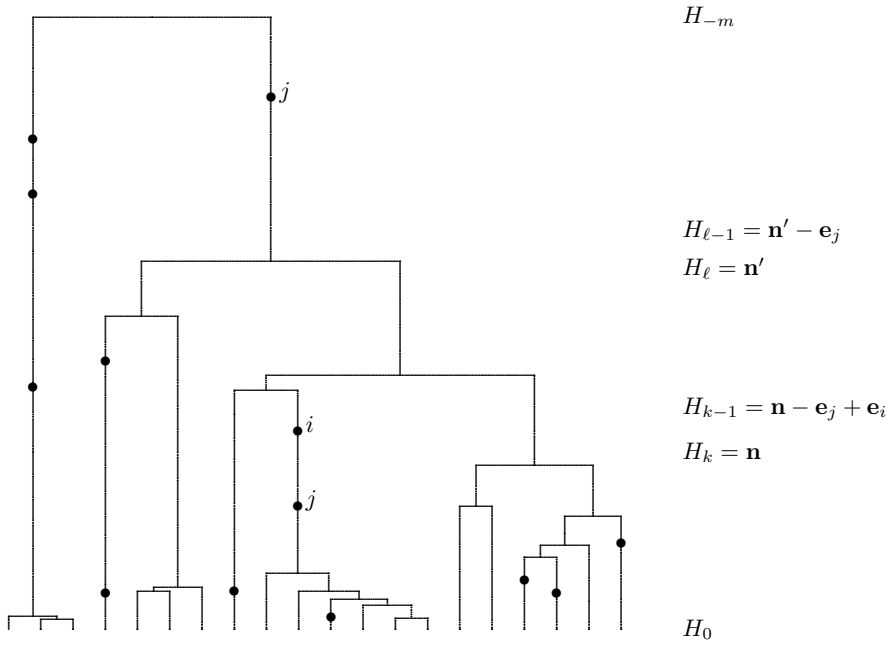

Figure 1: Coalescent tree.

the direction from the ancestor type to the sample data configuration. However a reverse history process from the sample data to the ancestor is required to efficiently evaluate the sample likelihood. A sequential importance sampling representation is based on an approximation $\widehat{p}(H_{k-1} \mid H_k)$ to the unknown reverse probabilities $p(H_{k-1} \mid H_k)$. In one step

$$\begin{aligned} p(H_k) &= \sum_{\{H_{k-1}\}} \frac{p(H_k \mid H_{k-1})}{\widehat{p}(H_{k-1} \mid H_k)} p(H_{k-1}) \widehat{p}(H_{k-1} \mid H_k) \\ &= \mathbb{E}_{\widehat{p}} \sum_{\{H_{k-1}\}} \Big[\frac{p(H_k \mid H_{k-1})}{\widehat{p}(H_{k-1} \mid H_k)} \mid H_k \Big] \end{aligned} \tag{2.2}$$

The full sequential importance sampling representation from continuing (2.2) over states $H_0, H_{-1}, \ldots, H_{-m}$ is

$$p(H_0) = \mathbb{E}_{\widehat{p}} \Big[\frac{p(H_0 \mid H_{-1})}{\widehat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\widehat{p}(H_{-m} \mid H_{-m+1})} p(H_{-m}) \Big] \tag{2.3}$$

where $\mathbb{E}_{\widehat{p}}$ is taken over histories $H_{-1}, \ldots, H_{-m}$ with $\widehat{p}(\cdot \mid \cdot)$ being the reverse chain transition probabilities. Probabilities of a history sample path $\mathcal{H}$ are evaluated in forward and reverse directions in the numerator and denominator of (2.3). The likelihood of the data can be evaluated by repeated simulation of sample histories in a reverse direction from the current sample configuration H_0

to H_{-m} under $\widehat{p}(\cdot)$ with transition probabilities $\widehat{p}(H_{k-1} \mid H_k)$, then averaging the sequential importance sampling weights

$$\frac{p(H_0 \mid H_{-1})}{\widehat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\widehat{p}(H_{-m} \mid H_{-m+1})} p(H_{-m})$$

obtained on each run to obtain an estimate of the likelihood.

In a panmictic model if a historical configuration is $H_k = \boldsymbol{n}$ then either $H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j$ for some $j \in E$ corresponding to coalescence of two type j genes, or $H_{k-1} = \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j$ for some $i, j \in E$ corresponding to mutation forward in time from i to j chosen with transition probability matrix P. (See Figure 1 for an illustration.) In detail (2.1) becomes

$$\begin{aligned} p(\boldsymbol{n}) &= \frac{\theta}{n+\theta-1} \sum_{i,j \in E, n_j > 0} \frac{n_i + 1 - \delta_{ij}}{n} P_{ij} p(\boldsymbol{n} - \boldsymbol{e}_j + \boldsymbol{e}_i) \\ &\quad + \frac{n-1}{n+\theta-1} \sum_{j \in E, n_j > 0} \frac{n_j - 1}{n-1} p(\boldsymbol{n} - \boldsymbol{e}_j). \end{aligned} \tag{2.4}$$

In (2.4) $\delta_{ij} = 1$ if $i = j$ or $\delta_{ij} = 0$ otherwise. Griffiths and Tavaré (1994) derive (2.4) from a coalescent argument and also by considering a sample from the population frequencies in a diffusion process model.

It is possible to express the reverse chain probabilities by using Bayes' rule as

$$P(H_{k-1} \mid H_k) = \begin{cases} \dfrac{n_j - 1}{n+\theta-1} \cdot \dfrac{n_j}{n} \cdot \dfrac{1}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)} & \text{if } H_{k-1} = \boldsymbol{n} - \boldsymbol{e}_j \\ \dfrac{\theta}{n+\theta-1} \cdot \dfrac{n_j}{n} \cdot \dfrac{\pi(i \mid \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j)}{\pi(j \mid \boldsymbol{n} - \boldsymbol{e}_j)} & \text{if } H_{k-1} = \boldsymbol{n} + \boldsymbol{e}_i - \boldsymbol{e}_j. \end{cases} \tag{2.5}$$

Importance sampling distributions are found by substituting an approximation $\widehat{\pi}$ to obtain $\widehat{P}(H_{k-1} \mid H_k)$. $p(\boldsymbol{n})$ can then be calculated by simulation. Stephens and Donnelly (2000) construct an importance sampling proposal distribution on coalescent histories approximating $\pi(\cdot \mid \boldsymbol{n})$, by $\widehat{\pi}(\cdot \mid \boldsymbol{n})$, the stationary distribution in a Markov chain with transition probability matrix

$$\frac{\theta P + \boldsymbol{n}}{n+\theta}. \tag{2.6}$$

That is, for $j \in E$,

$$\widehat{\pi}(j \mid \boldsymbol{n}) = \sum_{i \in E} \widehat{\pi}(i \mid \boldsymbol{n}) \frac{\theta P_{ij} + n_j}{n+\theta}. \tag{2.7}$$

An explicit solution to (2.7) is that

$$\begin{aligned} \widehat{\pi}(j \mid \boldsymbol{n}) &= \sum_{i \in E} \frac{n_i}{n} \sum_{k=0}^{\infty} \rho^k (1-\rho) P_{ij}^{(k)} \\ &= \sum_{i \in E} \frac{n_i}{n} Q_{ij}\,, \end{aligned} \tag{2.8}$$

where $\rho = \theta/(n+\theta)$ and the transition matrix $Q = (1-\rho)(I-\rho P)^{-1}$. De Iorio and Griffiths (2004a, 2004b) provide three ways of justifying the approximation $\widehat{\pi}$; from an approximation to the generator of the diffusion process describing the distribution of the population gene frequencies; from the recursive equations for the sampling distribution; and from a coalescent argument.

A stochastic interpretation of (2.8) is to choose a type $i \in E$ gene with probability n_i/n, then obtain a type $j \in E$ gene from a Geometric $\Big(\theta/(n+\theta)\Big)$ number of mutations according to the transion matrix P. In a stepwise mutation model where $P_{ij} = 1/2$ if $|i-j| = 1$, or zero otherwise, it is shown in this paper that

$$Q_{ij} = \frac{1-\rho}{\sqrt{1-\rho^2}} \cdot \Big[\frac{\rho}{1+\sqrt{1-\rho^2}}\Big]^{|j-i|}. \tag{2.9}$$

In a subdivided population model with g subpopulations let S be the subpopulation type space. A gene's type is then indexed by $S \times E$, the subpopulation it is in, and its allele type. Possible transitions back in time to a sample of genes at a prior history event to a gene type (α, j) are: coalescence of a pair of genes of type (α, j); mutation forward in time from type (α, i) to (α, j) with rate $\theta/2$ and transition probability P_{ij}; and migration back in time of a type j gene from subpopulation α to β at rate $m_{\alpha\beta}/2$. Let $m_\alpha = \sum_{\beta \neq \alpha} m_{\alpha\beta}$, and denote $(q_\alpha, \alpha \in S)$ as relative subpopulation sizes. A Wright-Fisher model in discrete time gives rise to this model as subpopulation sizes tend to infinity. Let $(N_\alpha)_{\alpha \in S}$ be the subpopulation sizes, $N = \sum_{\alpha \in S} N_\alpha$, $q_\alpha = N_\alpha/N$, $\alpha \in S$, and $v_{\alpha\beta}$, $\alpha, \beta \in S$ be the probability that the parent of an offspring in subpopulation α is from subpopulation β in the previous generation. The backward migration rates are defined as $m_{\alpha\beta} = 2Nv_{\alpha\beta}$, $\alpha, \beta \in \Gamma$, $\alpha \neq \beta$ with the overall rate $m_\alpha = \sum_{\beta \neq \alpha} m_{\alpha\beta}$. If $(\widetilde{m}_{\beta\alpha})$ are the forward migration rates then $\widetilde{m}_{\beta\alpha} = N_\alpha m_{\alpha\beta}/N_\beta$. The model considered here is the usual coalescent time scaled model where time is measured in units of N generations, and $N \to \infty$ while migration and mutation rates are kept constant. A careful treatment of the limit is in Herbots (1997). The analogue of (2.4) for a subdivided population model, derived in De Iorio and Griffiths (2004b), is

$$\begin{aligned}
&\left(\sum_{\alpha=1}^{g} n_\alpha(n_\alpha - 1)q_\alpha^{-1} + \sum_{\alpha=1}^{g} n_\alpha m_\alpha + n\theta\right)p(\boldsymbol{n}) = \\
&\qquad \sum_{\alpha=1}^{g}\sum_{j\in E} n_\alpha(n_{\alpha j} - 1)q_\alpha^{-1} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j}) \\
&\quad +\theta \sum_{\alpha=1}^{g}\sum_{i,j\in E} (n_{\alpha i} + 1 - \delta_{ij})P_{ij}\, p(\boldsymbol{n} + \boldsymbol{e}_{\alpha i} - \boldsymbol{e}_{\alpha j}) \\
&+\sum_{\alpha=1}^{g}\sum_{j\in E}\sum_{\beta\neq\alpha} m_{\alpha\beta}\frac{n_\alpha}{n_\beta + 1}(n_{\beta j} + 1)p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j} + \boldsymbol{e}_{\beta j}).
\end{aligned} \tag{2.10}$$

Bahlo and Griffiths (2001) obtain a general solution for $p(\boldsymbol{n})$ when the sample size is $n = 2$, though the form of solution is not simple. For $(\alpha, j) \in S \times E$,

Table 1: Importance sampling proposal distribution and importance weights for a coalescent model with migration

H_{k-1}	Proposal distribution	Importance weight
$\boldsymbol{n}-\boldsymbol{e}_{\alpha j}$	$\dfrac{n_{\alpha j}(n_{\alpha j}-1)q_\alpha^{-1}}{\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})D(\boldsymbol{n})}$	$\dfrac{n_\alpha}{n_{\alpha j}}\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})$
$\boldsymbol{n}-\boldsymbol{e}_{\alpha j}+\boldsymbol{e}_{\alpha i}$	$\dfrac{n_{\alpha j}\theta P_{ij}\widehat{\pi}(i \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}{\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})D(\boldsymbol{n})}$	$\dfrac{(n_{\alpha i}+1-\delta_{ij})}{n_{\alpha j}}\dfrac{\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}{\widehat{\pi}(i \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}$
$\boldsymbol{n}-\boldsymbol{e}_{\alpha j}+\boldsymbol{e}_{\beta j}$	$\dfrac{n_{\alpha j}m_{\alpha\beta}\widehat{\pi}(j \mid \beta, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}{\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})D(\boldsymbol{n})}$	$\dfrac{(n_{\beta j}+1)}{n_{\alpha j}}\dfrac{n_\alpha}{(n_\beta+1)}\dfrac{\widehat{\pi}(j \mid \alpha, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}{\widehat{\pi}(j \mid \beta, \boldsymbol{n}-\boldsymbol{e}_{\alpha j})}$

let $\pi(j \mid \alpha, \boldsymbol{n})$ be the probability that an additional gene taken from subpopulation α is of type j, given that we have an observed configuration $\boldsymbol{n} = (n_{\alpha j})$. The scaling is such that $\sum_{j\in E}\pi(j \mid \alpha, \boldsymbol{n}) = 1$. The reverse chain probabilities $p(H_{k-1} \mid H_k)$ can be expressed in terms of $\pi(\cdot \mid \alpha, \boldsymbol{n})$. The proposal distribution $\widehat{p}(H_{k-1} \mid H_k)$ and one-step importance sampling weights, $p(H_k \mid H_{k-1})/\widehat{p}(H_{k-1} \mid H_k)$ based on approximate distributions $\widehat{\pi}(\cdot \mid \alpha, \boldsymbol{n})$, derived in De Iorio and Griffiths (2004b), are shown in Table 1. The probability distributions $\widehat{\pi}(\cdot \mid \alpha, \boldsymbol{n})$ are defined by a system of equations

$$\left(n_\alpha q_\alpha^{-1} + m_\alpha + \theta\right)\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) = n_{\alpha j}q_\alpha^{-1} + \theta\sum_{i\in E}P_{ij}\widehat{\pi}(i \mid \alpha, \boldsymbol{n}) + \sum_{\beta\neq\alpha}m_{\alpha\beta}\widehat{\pi}(j \mid \beta, \boldsymbol{n}). \tag{2.11}$$

The overall event rate in subpopulation α is $d_\alpha/2$, where $d_\alpha = n_\alpha(n_\alpha-1)q_\alpha^{-1} + n_\alpha m_\alpha + n_\alpha\theta$, and the total event rate is $D(\boldsymbol{n})/2$ where $D(\boldsymbol{n}) = \sum_{\alpha=1}^{g} d_\alpha$. In practice it is easiest to choose a gene of type (α, j) to change with probability

$$\frac{n_{\alpha j}((n_\alpha-1)q_\alpha^{-1} + m_\alpha + \theta)}{D(\boldsymbol{n})}, \tag{2.12}$$

then select an associated event from the conditional proposal distribution found by dividing the proposal distribution in Table 1 by (2.12).

Here we describe briefly how a coalescent approximation gives rise to $\widehat{\pi}$. Let $B_{\alpha j}$ be the event that a gene from subpopulation α of type j is involved in the first event back in the coalescent history of the process and $\boldsymbol{Y}$ a random vector

describing the configuration of types. Then

$$\begin{aligned} p(B_{\alpha j} \cap \{\boldsymbol{Y} = \boldsymbol{n}\}) &= p(\boldsymbol{n})\mathbb{P}(B_{\alpha j} \mid \boldsymbol{Y} = \boldsymbol{n}) \\ &= \frac{n_\alpha(n_\alpha - 1)q_\alpha^{-1}}{D(\boldsymbol{n})} \sum_{j \in E} \frac{n_{\alpha j} - 1}{n_\alpha - 1} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j}) \\ &\quad + \frac{n_\alpha \theta}{D(\boldsymbol{n})} \sum_{i,j \in E} \frac{n_{\alpha i} + 1 - \delta_{ij}}{n_\alpha} P_{ij} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j} + \boldsymbol{e}_{\alpha i}) \\ &\quad + \frac{n_\alpha m_\alpha}{D(\boldsymbol{n})} \sum_{j \in E} \sum_{\beta \neq \alpha} \frac{m_{\alpha\beta}}{m_\alpha} \cdot \frac{n_{\beta j} + 1}{n_\beta + 1} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j} + \boldsymbol{e}_{\beta j}). \end{aligned} \tag{2.13}$$

Exchangeability in the order of sampled genes implies that

$$\begin{aligned} \pi(j \mid \alpha, \boldsymbol{n} - \boldsymbol{e}_{\alpha j}) p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j}) &= \frac{n_{\alpha j}}{n_\alpha} p(\boldsymbol{n}) \\ \pi(i \mid \alpha, \boldsymbol{n} - \boldsymbol{e}_{\alpha j}) p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j}) &= \frac{n_{\alpha i} + 1 - \delta_{ij}}{n_\alpha} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j} + \boldsymbol{e}_{\alpha i}) \\ \pi(j \mid \beta, \boldsymbol{n} - \boldsymbol{e}_{\alpha j}) p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j}) &= \frac{n_{\beta j} + 1}{n_\beta + 1} p(\boldsymbol{n} - \boldsymbol{e}_{\alpha j} + \boldsymbol{e}_{\beta j}). \end{aligned} \tag{2.14}$$

Substituting from (2.14) into (2.13)

$$\begin{aligned} &\pi(j \mid \alpha, \boldsymbol{n} - \boldsymbol{e}_j)\mathbb{P}(B_{\alpha j} \mid \boldsymbol{Y} = \boldsymbol{n}) D(\boldsymbol{n})/n_\alpha \\ &\quad = (n_{\alpha j} - 1)q_\alpha^{-1} + \theta \sum_{i \in E} P_{ij} \pi(i \mid \alpha, \boldsymbol{n} - \boldsymbol{e}_j) + \sum_{\beta \neq \alpha} m_{\alpha\beta} \pi(j \mid \beta, \boldsymbol{n} - \boldsymbol{e}_j). \end{aligned} \tag{2.15}$$

The system (2.15) is exact, rather than approximate. To obtain the approximate system (2.11) assume that

$$\mathbb{P}(B_{\alpha j} \mid \boldsymbol{Y} = \boldsymbol{n}) = \frac{n_\alpha(n_\alpha - 1)q_\alpha^{-1} + n_\alpha m_\alpha + n_\alpha \theta}{D(\boldsymbol{n})} \cdot \frac{n_{\alpha j}}{n_\alpha},$$

the probability of the last history event being in subpopulation α, times the approximate probability $n_{\alpha j}/n_\alpha$. Then substituting in (2.15) and setting $\boldsymbol{n} - \boldsymbol{e}_j \to \boldsymbol{n}$ yields (2.11).

A stochastic interpretation of the distribution $\widehat{\pi}$ is shown in De Iorio and Griffiths (2004b) to be the following. Let $M^\circ = (m_{\alpha\beta}/m_\alpha)$ be a transition probability matrix with diagonal elements zero constructed from the migration rate matrix M, so that the rows of M° each add to 1. Denote, for $\alpha \in S$,

$$\phi_\alpha = \frac{m_\alpha}{n_\alpha q_\alpha^{-1} + m_\alpha},\ \rho_\alpha = \frac{\theta}{n_\alpha q_\alpha^{-1} + m_\alpha + \theta},$$

and the transition probability matrix $P_\alpha = (1 - \rho_\alpha)(I - \rho_\alpha P)^{-1}$. A mechanism for choosing a gene of type $j \in E$ from the distribution $\widehat{\pi}(j \mid \alpha, \boldsymbol{n})$ is the

following. Choose a sequence of subpopulations $\alpha_0, \alpha_1, \ldots, \alpha_\tau$, for the migration path of a gene, starting with $\alpha_0 = \alpha$ and stopping at step τ in subpopulation α_τ, with probability

$$\phi_{\alpha_0}\phi_{\alpha_1}\cdots\phi_{\alpha_{\tau-1}}(1-\phi_{\alpha_\tau})\cdot m^\circ_{\alpha_0\alpha_1}m^\circ_{\alpha_1\alpha_2}\cdots m^\circ_{\alpha_{\tau-1}\alpha_\tau}.$$

ϕ_α can be interpreted as the probability of moving from subpopulation α to another subpopulation, while $1-\phi_\alpha$ is the probability of stopping in subpopulation α. Next choose a type at random from subpopulation α_τ, such that the probability of choosing a gene of type i is $n_{\alpha_\tau i}/n_{\alpha_\tau}$. Mutate back along the migration path to α_0, so that a sample path probability of a sequence of mutations which start with type $i_{\alpha_\tau} = i$ and end with a type $i_0 = j$ gene is

$$\frac{n_{\alpha_\tau i_{\alpha_\tau}}}{n_{\alpha_\tau}}P_{\alpha_\tau;i_\tau i_{\tau-1}}\cdots P_{\alpha_1;i_2 i_1}P_{\alpha_0;i_1 i_0}.$$

An interpretation of P_α is that there are a geometrically distributed number of mutations with parameter ρ_α, the probability of a mutation, and a transition matrix P for type changes, in each subpopulation $\alpha \in \{\alpha_\tau, \ldots, \alpha_0\}$ visited in the migration path. The stochastic structure described above can be seen by rewriting (2.11) as

$$\begin{aligned}\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) &= (1-\rho_\alpha)(1-\phi_\alpha)\frac{n_{\alpha j}}{n_\alpha} + (1-\rho_\alpha)\phi_\alpha\sum_{\beta\neq\alpha} m^\circ_{\alpha\beta}\widehat{\pi}(j \mid \beta, \boldsymbol{n}) \\ &\quad +\rho_\alpha\sum_{i\in E} P_{ij}\widehat{\pi}(i \mid \alpha, \boldsymbol{n}). \qquad (2.16)\end{aligned}$$

In the simple stepwise mutation model $P_\alpha = Q$, in (2.8) with parameter $\rho = \rho_\alpha$.

3 Stepwise mutation model

The population genetics model considered in this paper is a subdivided population with stepwise mutations on the line. The gene type space is then $E = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ with a mutation transition matrix of the form

$$P_{ij} = u_{j-i}, \ \ , i, j = 0, \pm 1, \pm 2, \ldots$$

A sample of $(n_\alpha; \alpha \in S)$ genes is taken from the subpopulations. $\boldsymbol{n} = (n_{\alpha j}; (\alpha, j) \in S \times E)$ is the collection of the number of genes in subpopulation α of type j. Backwards migration rates from subpopulation α to β are denoted by $(m_{\alpha\beta}; \alpha, \beta \in S)$. $p(\boldsymbol{n})$ will denote the probability of a configuration $\boldsymbol{n}$ under the model.

Fourier transform methods will be used in the following sections to solve recursive equations obtained. The Fourier transform of an absolutely convergent series $\{a_j; j = 0, \pm 1, \pm 2 \ldots\}$ will be denoted by

$$a^*(\xi) = \sum_{j=-\infty}^{\infty} e^{\iota\xi j}a_j,$$

where $\iota = \sqrt{-1}$.

3.0.1 A sample of two genes

It is possible to obtain a formula for $p(\boldsymbol{n})$ when $n = 2$ by using Fourier transform methods. We provide details of the simplest case with two subpopulations labelled α and β. Let $p(\alpha, \beta; \Delta)$ be the probability that two genes chosen from subpopulations α and β are separated by a signed distance $\Delta = 0, \pm 1, \pm 2, \ldots$. $p(\alpha, \alpha; \Delta)$ is interpreted as the probability that the signed distance of the second gene from the first gene chosen from subpopulation α is Δ. If $\boldsymbol{n}_{\alpha\beta}$ is a sample of two genes from subpopulations α and β with respective positions i and j such that $i - j = \Delta$, then for $\alpha \neq \beta$, $p(\alpha, \beta; d) = p(\boldsymbol{n}_{\alpha\beta})$, while for $\alpha = \beta$, $p(\alpha, \alpha; \Delta) = p(\boldsymbol{n}_{\alpha\alpha})/(2 - \delta_{\Delta,0})$, where $\delta_{\Delta,0} = 1$ if $\Delta = 0$ or $\delta_{\Delta,0} = 0$ if $\Delta \neq 0$. The scaling is such that $\sum_{\Delta=-\infty}^{\infty} p(\alpha, \beta; d) = 1$ for α and β equal or unequal. Without loss of generality choosing $j = 0$, substituting in (2.10), (or a derivation from first principles for a sample of two genes) and using translation invariance gives the following equations for $\alpha \neq \beta$,

$$\begin{aligned}
(m_\alpha + m_\beta + 2\theta)p(\alpha, \beta; \Delta) &= 2\theta \sum_{k=-\infty}^{\infty} p(\alpha, \beta; \Delta - k)u_k \\
&\quad + m_\alpha p(\beta, \beta; \Delta) + m_\beta p(\alpha, \alpha; \Delta), \\
(q_\alpha^{-1} + m_\alpha + \theta)p(\alpha, \alpha; \Delta) &= q_\alpha^{-1}\delta_{\Delta,0} + \theta \sum_{k=-\infty}^{\infty} p(\alpha, \alpha; \Delta - k)u_k \\
&\quad + m_\alpha p(\alpha, \beta; \Delta).
\end{aligned} \tag{3.1}$$

Denote $p^*_{\alpha\beta}(\xi)$ as the Fourier transform of $p(\alpha, \beta; d)$, and $\theta^* = \theta\big(1 - u^*(\xi)\big)$. Then from (3.1), omitting the argument ξ for ease of notation,

$$\begin{aligned}
(m_\alpha + m_\beta + 2\theta^*)p^*_{\alpha\beta} &= m_\alpha p^*_{\beta\beta} + m_\beta p^*_{\alpha\alpha} \\
(q_\alpha^{-1} + m_\alpha + \theta^*)p^*_{\alpha\alpha} &= q_\alpha^{-1} + m_\alpha p^*_{\alpha\beta} \\
(q_\beta^{-1} + m_\beta + \theta^*)p^*_{\beta\beta} &= q_\beta^{-1} + m_\beta p^*_{\alpha\beta}.
\end{aligned}$$

The solution of (3.2) is

$$p^*_{\alpha\beta} = \frac{A}{B}, \tag{3.2}$$

where

$$\begin{aligned}
A &= q_\beta^{-1} m_\alpha (q_\alpha^{-1} + m_\alpha + \theta^*) + q_\alpha^{-1} m_\beta (q_\beta^{-1} + m_\beta + \theta^*), \\
B &= (m_\alpha + m_\beta + 2\theta^*)(q_\alpha^{-1} + m_\alpha + \theta^*)(q_\beta^{-1} + m_\beta + \theta^*) \\
&\quad - m_\alpha m_\beta (q_\alpha^{-1} + m_\alpha + q_\beta^{-1} + m_\beta + 2\theta^*).
\end{aligned}$$

$p^*_{\alpha,\alpha}$ and $p^*_{\beta,\beta}$ can then be found from the second equation of (3.2). Probabilities are found by inversion of the corresponding transform, with

$$p(\alpha, \beta; d) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\iota d\xi} p^*_{\alpha\beta}(\xi) d\xi. \tag{3.3}$$

The general case of more than two subpopulations is similar, with a system of equations to solve for $\alpha \neq \beta$, $\alpha, \beta = 1, \ldots, g$ of

$$\begin{aligned}(m_\alpha + m_\beta + 2\theta^*)p^*_{\alpha\beta} &= \sum_{\gamma\neq\alpha} m_{\alpha,\gamma}p^*_{\gamma\beta} + \sum_{\gamma\neq\beta} m_{\beta,\gamma}p^*_{\gamma\alpha} \\ (q_\alpha^{-1} + m_\alpha + \theta^*)p^*_{\alpha\alpha} &= q_\alpha^{-1} + \sum_{\gamma\neq\alpha} m_{\alpha\gamma}p^*_{\alpha\gamma}.\end{aligned} \tag{3.4}$$

In a symmetric one-step mutation model where $u_{-1} = u_{+1} = 1/2$, $u_j = 0$ if $j \neq \pm 1$, $u^*(\xi) = \cos(\xi)$, and $\theta^* = \theta\big(1 - \cos(\xi)\big)$.

3.1 Importance sampling distribution π

3.1.1 Single population

The system of equations (2.7) in a stepwise mutation model becomes

$$\widehat{\pi}(j \mid \boldsymbol{n}) = \frac{n_j}{n+\theta} + \frac{\theta}{n+\theta}\sum_{i=-\infty}^{\infty} u_{j-i}\widehat{\pi}(i \mid \boldsymbol{n}). \tag{3.5}$$

The Fourier transform equation corresponding to (3.5) is

$$\widehat{\pi}^*(\xi \mid \boldsymbol{n}) = \frac{n^*(\xi)}{n+\theta} + \frac{\theta}{n+\theta}u^*(\xi)\widehat{\pi}^*(\xi \mid \boldsymbol{n}). \tag{3.6}$$

Thus

$$\widehat{\pi}^*(\xi \mid \boldsymbol{n}) = \frac{n^*(\xi)}{n+\theta(1-u^*(\xi))}. \tag{3.7}$$

Inverting the transform

$$\widehat{\pi}(k \mid \boldsymbol{n}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{-k\xi\iota}\cdot n^*(\xi)}{n+\theta(1-u^*(\xi))}d\xi. \tag{3.8}$$

In a symmetric one-step mutation model there is an explicit solution

$$\begin{aligned}\widehat{\pi}(k \mid \boldsymbol{n}) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{-k\xi\iota}\cdot n^*(\xi)}{n+\theta(1-\cos(\xi))}d\xi \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sum_{j=-\infty}^{\infty} n_j\cos((j-k)\xi)}{n+\theta(1-\cos(\xi))}d\xi \\ &= \sum_{j=-\infty}^{\infty}\frac{n_j}{n+\theta}c_{j-k}(\rho) \\ &= \frac{n_k}{n+\theta}c_0(\rho) + \sum_{\ell=1}^{\infty}\frac{n_{k+\ell}+n_{k-\ell}}{n+\theta}c_\ell(\rho),\end{aligned} \tag{3.9}$$

where $\rho = \theta/(n+\theta)$, and

$$c_\ell(\rho) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\cos(\ell\xi)}{1-\rho\cos(\xi)} d\xi = \frac{1}{\sqrt{1-\rho^2}} \cdot \Big[\frac{\rho}{1+\sqrt{1-\rho^2}}\Big]^{|\ell|}. \tag{3.10}$$

In the derivation of (3.9) if $j \neq k$,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sum_{j=-\infty}^{\infty} n_j \sin((j-k)\xi)}{n+\theta(1-\cos(\xi))} d\xi = 0$$

because the integrand is an odd function of ξ about zero.

3.1.2 Subdivided populations

Taking the Fourier transform of (2.11) produces a system of equations for $\alpha \in S$ of

$$(n_\alpha^{-1} q_\alpha^{-1} + m_\alpha + \theta(1-u^*(\xi)))\widehat{\pi}^*(\xi \mid \alpha, \boldsymbol{n}) = n_\alpha^* q_\alpha^{-1} + \sum_{\beta \neq \alpha} m_{\alpha\beta}\widehat{\pi}^*(\xi \mid \beta, \boldsymbol{n}). \tag{3.11}$$

Let $A(\xi) = \mathrm{Diag}\Big(n_\alpha q_\alpha^{-1} + m_\alpha + \theta(1-u^*(\xi))\Big) - M$, $b(\xi) = (n_\alpha^* q_\alpha^{-1})$ and $\widehat{\pi}^*(\xi \mid \boldsymbol{n}) = (\widehat{\pi}(\xi \mid \alpha, \boldsymbol{n}))$. The system (3.11) can be written in the matrix form

$$\widehat{\pi}(\xi \mid \boldsymbol{n}) = A(\xi)^{-1} b(\xi),$$

with a solution of

$$\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-j\xi\iota} A(\xi)^{-1} b(\xi) d\xi. \tag{3.12}$$

An approximate solution to these equations can be found by simple numerical integration on equally spaced points in $[-\pi, \pi]$.

3.1.3 Two subpopulations

The general stochastic representation in Section 2 gives an expression in a model with two subpopulations labelled α and β of

$$\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) = \sum_{k=-\infty}^{\infty} \sum_{v_\alpha=1}^{\infty} \sum_{v_\beta=0}^{\infty} \Big[\frac{n_{\alpha k}}{n_\alpha}(1-\phi_\alpha) + \frac{n_{\beta k}}{n_\beta}(1-\phi_\beta)\Big] \phi_\alpha^{v_\alpha - 1} \phi_\beta^{v_\beta} \Big[P_\alpha^{v_\alpha} P_\beta^{v_\beta}\Big]_{kj}. \tag{3.13}$$

The probability of v_α, v_β visits to α, β starting at α and ending at α is $\phi_\alpha^{v_\alpha - 1}(1-\phi_\alpha)\phi_\beta^{v_\beta}$, and similarly $\phi_\alpha^{v_\alpha - 1}\phi_\beta^{v_\beta}(1-\phi_\beta)$, for ending in β. The transition matrices P_α, P_β commute, so the term containing them in (3.13) is well defined. Denote the mean quantities

$$\widehat{\mu}(\alpha, \boldsymbol{n}) = \sum_{j=-\infty}^{\infty} j\widehat{\pi}(j \mid \alpha, \boldsymbol{n}), \qquad \bar{n}_\alpha = \sum_{j=-\infty}^{\infty} j\frac{n_{\alpha j}}{n_\alpha}$$

$$\widehat{\mu}(\beta, \boldsymbol{n}) = \sum_{j=-\infty}^{\infty} j\widehat{\pi}(j \mid \beta, \boldsymbol{n}), \qquad \bar{n}_\beta = \sum_{j=-\infty}^{\infty} j\frac{n_{\beta j}}{n_\beta}.$$

If the mean mutation distance $\sum_{j=-\infty}^{\infty} jP_{ij}$ is always zero from any position i then by considering the first step in the migration walk

$$\begin{aligned} \widehat{\mu}(\alpha, \boldsymbol{n}) &= \bar{n}_\alpha(1-\phi_\alpha) + \phi_\alpha \widehat{\mu}(\beta, \boldsymbol{n}) \\ \widehat{\mu}(\beta, \boldsymbol{n}) &= \bar{n}_\beta(1-\phi_\beta) + \phi_\beta \widehat{\mu}(\alpha, \boldsymbol{n}). \end{aligned} \tag{3.14}$$

The solution of (3.14) is

$$\begin{aligned} \widehat{\mu}(\alpha, \boldsymbol{n}) &= \frac{\bar{n}_\alpha(1-\phi_\alpha) + \phi_\alpha(1-\phi_\beta)\bar{n}_\beta}{1-\phi_\alpha\phi_\beta} \\ \widehat{\mu}(\beta, \boldsymbol{n}) &= \frac{\bar{n}_\beta(1-\phi_\beta) + \phi_\beta(1-\phi_\alpha)\bar{n}_\alpha}{1-\phi_\alpha\phi_\beta}. \end{aligned} \tag{3.15}$$

The coefficients of $\bar{n}_\alpha, \bar{n}_\beta$ in (3.15) are the probabilities that the parent gene of the gene chosen comes from subpopulations α, β.

The representation (3.13) is transparent to understand, however it is easier to find a detailed solution directly from (3.12). We assume a symmetric one-step mutation model. Then the determinant of $A(\xi)$, $|A(\xi)| = (n_\alpha q_\alpha^{-1} + m_\alpha + \phi)(n_\beta q_\beta^{-1} + m_\beta + \phi) - m_\alpha m_\beta$, where $\phi = \theta(1-\cos(\xi))$. $|A(\xi)| = (\phi - \lambda_1)(\phi - \lambda_2)$, where λ_1, λ_2 are the roots of $\phi^2 + \phi(x_\alpha + x_\beta) + x_\alpha x_\beta - m_\alpha m_\beta = 0$, with $x_\alpha = n_\alpha q_\alpha^{-1} + m_\alpha$, and similarly for x_β. Both roots are real, less than zero, and unequal. Consider the expression

$$\begin{aligned} |A(\xi)|^{-1} &= \frac{1}{\lambda_1 - \lambda_2} \cdot \Big[\frac{1}{\phi - \lambda_1} - \frac{1}{\phi - \lambda_2}\Big] \\ &= \frac{1}{\lambda_1 - \lambda_2}\Big[\frac{1}{-\lambda_1 + \theta} \cdot \frac{1}{1 - \rho_1\cos(\xi)} - \frac{1}{-\lambda_2 + \theta} \cdot \frac{1}{1 - \rho_2\cos(\xi)}\Big], \end{aligned} \tag{3.16}$$

where $\rho_i = \theta/(-\lambda_i + \theta), i = 1, 2$. From the matrix equation for the inversion of the transform and the form of the inverse of $A(\xi)$,

$$\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) = \int_{-\pi}^{\pi} e^{-\iota\xi j}|A(\xi)|^{-1}\Big[\big(n_\beta q_\beta^{-1} + m_\beta + \theta(1-\cos(\xi))q_\alpha^{-1}n_\alpha^* + m_\alpha q_\beta^{-1} n_\beta^*\Big] d\xi \tag{3.17}$$

From equations (3.17) and (3.16)

$$\widehat{\pi}(j \mid \alpha, \boldsymbol{n}) = \frac{1}{\lambda_1 - \lambda_2} \cdot \sum_{k=-\infty}^{\infty}\Big[\frac{a_1(k,j)}{-\lambda_1 + \theta} - \frac{a_2(k,j)}{-\lambda_2 + \theta}\Big], \tag{3.18}$$

where for $i = 1, 2$

$$\begin{aligned} a_i(k,j) &= q_\alpha^{-1} n_{\alpha k}\Big[(n_\beta q_\beta^{-1} + m_\beta + \theta)c_{k-j}(\rho_i) - \frac{\theta}{2}(c_{k-j+1}(\rho_i) + c_{k-j-1}(\rho_i))\Big] \\ &+ \quad q_\beta^{-1} n_{\beta k} m_\alpha c_{k-j}(\rho_i). \end{aligned} \tag{3.19}$$

The summation in (3.18) is finite because only a finite number of $\{n_{\alpha k}, n_{\beta k}\}$ are non-zero.

3.1.4 Computer implementation of the IS algorithm

An implementation of the IS algorithm for two subpopulations (labelled 1 and 2) was based on the formula (3.18) for $\widehat{\pi}$. Using the example of locus DB4 in Table 2 with $\theta = 2.0$, $m_{12} = 5$, and $m_{21} = 3$, and relative subpopulation sizes of (0.25, 0.75) five duplicate computations of the likelihood of the sample and TMRCA were made, each with one million runs and different starting seeds. The accuracy is quite good, with likelihood values 4.2, 4.4, 4.5, 4.2, 4.3 times 10^{-18} and mean TMRCA values in coalescent units of 3.219, 3.218, 3.209, 3.217, 3.217. The time taken for a million runs is approximately 25 minutes on a 2.4 Ghz Pentium 4 computer.

3.1.5 Simulation study

A simulation study was undertaken using the IS algorithm for maximum likelihood estimation of mutation and migration rates. The model has two subpopulations with symmetric migration and a one-step mutation model. Ten data sets of 5 independent loci and 10 data sets of 20 independent loci were used to check the accuracy of estimates, and to see the effect of the number of loci on parameter estimation. Because of computational constraints the study was limited. Likelihood estimation was done on computers of the Centre Informatique National de l'Enseignement Supérieur (CINES, France), occupying roughly 10000 hours on 500 MHz processors. Simulated data sets were generated using a discrete generation coalescent process in a two population model with equal sizes $N_1 = N_2 = 1,000$ genes, mutation rate $\theta = 4.0$, and symmetric migration rates $\gamma = m_{12} = m_{21} = 4.0$. Leblois *et al.* (2003) has details about the discrete generation simulation technique. Likelihoods were computed for each locus using the IS algorithm described in this paper. Likelihoods for each locus were then multiplied to find the overall likelihood for the multilocus data set at each parameter point. Finally, maximum likelihood estimates were computed from the likelihood surface obtained from the likelihood at the different parameter points by kriging as described in the Appendix. A computation based on 5,000 runs at 50 values of the parameter vector yielded less accurate estimates of likelihood than a computation based on 500,000 runs at the same 50 values, but the maximum likelihood estimates were identical (details not shown). Thus, accurate estimation of the likelihood for each point is not essential. Further computations were based on 5,000 runs at 5,000 points. For the two cases of 5 and 20 loci, the relative mean bias and relative mean square error (MSE) were calculated for estimates of the migration rate and mutation rate. Since migration was symmetric in the model, the two migration rate estimates were pooled.

The relative bias (bias/expectation) and relative MSE (MSE/squared expectation) are presented in Table 2. They show good performance of the algorithm, in particular in estimating θ. As expected, the precision increases with the number of loci for estimation of both mutation and migration parameters. In both cases of 5 and 20 loci the bias and the MSE are very small (below 10%). On the

Table 2: Bias and precision of estimates from simulated data sets using the algorithm described in this paper.

		5 loci	20 loci
θ	Bias	0.16	0.03
θ	MSE	0.09	0.01
γ	Bias	0.19	-0.05
γ	MSE	0.55	0.22

Table 3: Bias and precision of estimates from simulated data sets using `Migrate`.

		5 loci	20 loci
θ	Bias	0.23	-0.60
θ	MSE	0.30	0.40
γ	Bias	1.2	0.25
γ	MSE	3.1	0.46

other hand, our results show a slight over-estimation of the migration parameter γ. Nevertheless, the MSE is not very high indicating that estimates are close to parameter values in the model.

Part of the variance of the estimator is due to the variance of the maximum likelihood estimator, and part of it is due to the imprecision in locating the maximum of the likelihood function through sampling points and smoothing the surface. Efficiency of the latter procedure is demonstrated if it contributes only a small part of the total variance. In this case, the correlation between independent applications of the procedure to the same dataset should be high. Thus, we independently applied this procedure (hypercube sampling, likelihood evaluation, and kriging) twice to the ten 5–loci data sets. The correlation for pairs of estimates was > 0.975 for all parameters, demonstrating the efficiency of the procedure.

The discrete generation simulation model differed slightly from the continuous time coalescent model under which estimation was based. The limited number of intervals used in the likelihood grid, or the simulation model, may account for the slight bias of migration estimates.

For comparison `Migrate` (Beerli and Felsenstein, 2001) was used with default settings to analyze the same simulated data sets. Using the default settings, `Migrate` takes about the same computer time as our algorithm to analyze the data. In the study `Migrate` estimates were not precise with a low number of loci. Precision in estimating γ, but not θ, increased with the number of loci (Table 3). Note that estimation with `Migrate` assumed two different parameters θ_1 and θ_2, instead of a single θ parameter for both populations as assumed in the previous estimations.

3.1.6 Microsatellite application

Microsatellite loci are highly variable and the presence of back mutations cannot be ignored. Therefore inference from such data can be challenging and likelihood evaluation extremely computer intensive. The algorithm described in this paper constitutes a significant improvement over previous methods. Nielsen (1997) developed an algorithm based on the pioneering work of Griffiths and Tavaré (1994) to obtain maximum likelihood estimates of the mutation parameter θ at microsatellite loci. The method is computationally inefficient, even for a single locus, and computational time might be large, as many runs through the Markov chain do not contribute anything to the likelihood value. In fact, only a few simulated genealogies will contribute significantly to the likelihood evaluation, while most of the computational effort will be spent on genealogies with very small probability, i.e. not consistent with the observed data. This problem is especially evident in runs in which too many mutation events occur. Nielsen (1997) proposed truncating such runs according to a rule based on the expected number of mutations. The importance sampling proposal described in this paper implicitly solves this problem, concentrating most of the computational effort on trees with high probability, given the observed data. Of course the extension to subdivided populations of the island model type here is new. Stephens and Donnelly (2000) make a comparison on example data sets of their technique with their implementation of the Griffiths-Tavaré technique, and a comparison with `Batwing`, an implementation of a Bayesian technique of Wilson and Balding (1998). The Stephens-Donnelly technique is the most efficient of the three. Their microsatellite state space is truncated, so their distributions $\widehat{\pi}$ are computed from a system of linear equations rather than using (3.9). Chen and Liu (2000) show that in a data example the Griffiths-Tavaré technique combined with path resampling is as efficient as the Stephens-Donnelly technique and produces the same likelihood curves (see also Chen *et al.* (2004)). Path resampling could be used generally in other importance sampling schemes such as in this paper.

3.1.7 Red Fox data example

Lade *et. al.* (1996) collected data on seven microsatellite loci from Australian populations of the red fox *Vulpes vulpes*. As an example the two subpopulations from Phillip Island (PI) and the adjacent mainland at San Remo (SR) separated by a bridge are considered here. The data, coded so the minimum position is 0 and two base pairs are taken as a one unit mutation step, are shown in Table 4. A stepwise mutation model with single steps was used to model the data. This implies that two alleles that differ by one mutation step are more closely related than alleles that differ by many mutation steps. The stepwise model is a possible model for microsatellites when interest is in the relatedness between individuals and in population substructuring. Evolution at the different loci was assumed to be independent with the same mutation rate θ and two migration rates between the populations. Likelihoods for each locus were computed using

the importance sampling algorithm described in this paper, then multiplied to find the overall likelihood for the loci. Maximum likelihood estimates of the mutation and migration rates were found by fitting a likelihood surface using the kriging method described in the Appendix with 1000 design points. Raw likelihood points were based on runs of 10,000 replicates. Parameter estimates and likelihoods are shown in Table 5 for three different population size ratios PI:SR. The ratio with maximum likelihood is (0.25,0.75). There is a large difference between the likelihoods for sizes (0.25,0.75) and (0.75,0.25); the former is consistent with the higher diversity in SR, and the latter is unlikely. A maximum likelihood estimate is $\widehat{\theta} = 2$ with an approximate backward migration rate estimate from PI to SR of $\widehat{m}_{PS} = 5$ and in the other direction $\widehat{m}_{SP} = 3$. Forward migration rates are $\widetilde{m}_{SP} = 1.6$ and $\widetilde{m}_{PS} = 9$. A likelihood surface for the migration rates when $\theta = 2.0$ is shown in Figure 2. Likelihood units are 10^{-89}. Raw likelihood points were based on runs of one million replicates. Pairwise distributions of the number of steps between two genes both from PI, both from SR, or one each from PI and SR are computed from (3.3) and shown in Table 6. The square root of average pairwise difference squares within and between populations is shown in Table 7. There is quite a variation between loci and variation from the theoretical values.

As an illustration the mean and standard deviation of the TMRCA at each locus, shown in Table 8, was computed with the estimated parameters conditional on the stepwise allele configuration observed. If the mutation rate per locus per generation is 10^{-3}, then the effective population size is $N = \theta/(2 \times 10^{-3}) = 1000$, with effective population sizes on PI of 250 and SR of 750. Effective population sizes approximate true population sizes over time by their harmonic means. Supposing a generation time for foxes of five years, (Lade *et. al.*, 1996), then coalescent time units are in units of 5000 years. The TMRCA of the loci are well before the introduction of foxes into Australia around 1870, (Lade *et. al.*, 1996), which is what is expected. Assuming the model holds over the full ancestry may have the effect of inflating migration rates, by explaining allele frequency differences by migration, rather than by drift.

Table 4: Microsatellite allele frequencies from the red fox populations of Phillip Island (PI) and San Remo (SR)

Locus	PI	SR
DB1	n=46	n=42
0		3
1		5
2	32	2
4		10
5		1
7	14	11
9		10
DB3	n=46	n=46
0	38	32
1	8	11
5		3
DB4	n=46	n=44
0	8	20
9	38	14
12		4
14		5
15		1
DB6	n=46	n=42
0	1	2
1	27	27
2	18	13
OB	n=46	n=44
0	28	18
6	8	16
8	10	10
VD10	n=42	n=42
0	5	5
2		13
4	12	12
6	25	12
C213	n=46	n=44
0	4	15
2	42	24
3		5

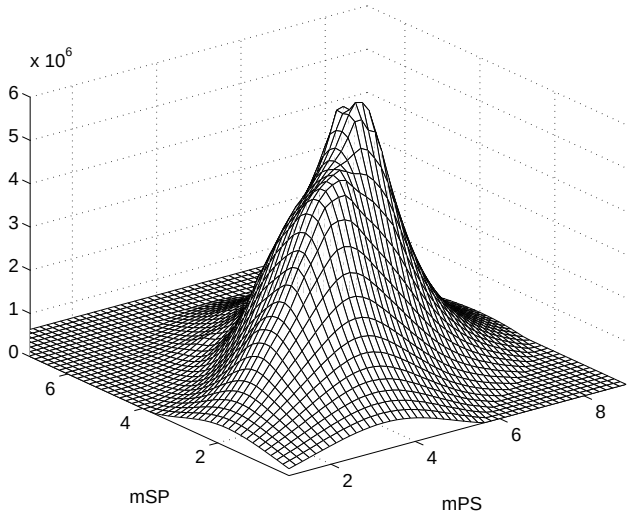

Figure 2: Likelihood surface for migration rates.

Table 5: Likelihood estimates

(q_1, q_2)	Likelihood $\times 10^{-89}$	$\widehat{\theta}$	$\widehat{m}_{PS}$	$\widehat{m}_{SP}$
(0.5, 0.5)	43×10^{3}	1.96	2.07	3.92
(0.25, 0.75)	62×10^{5}	2.00	5.07	3.27
(0.75, 0.25)	64	1.70	1.79	14.6

Table 6: Distance distribution between two genes

$\lvert\Delta\rvert$	PI	SR	PI SR
0	0.565	0.447	0.388
1	0.276	0.333	0.362
2	0.096	0.132	0.150
3	0.038	0.053	0.060
4	0.015	0.021	0.024
5	0.006	0.008	0.010
6	0.002	0.003	0.004
7	0.001	0.001	0.002
8	0.000	0.001	0.000

Table 7: Square root of average square pairwise differences

Locus	PI	SR	PI SR
DB1	3.29	4.33	4.18
DB3	0.54	1.78	1.36
DB4	4.88	8.04	6.76
DB6	0.75	0.77	0.76
OB	5.04	4.88	5.05
VD10	2.78	2.87	3.05
C213	0.81	1.54	1.28
Theory	1.23	1.45	1.54

Table 8: TMRCA in years

Locus	mean	sd
DB1	9100	3700
DB3	8200	3600
DB4	16000	5100
DB6	4600	2500
OB	11200	4200
VD10	7700	3400
C213	5500	2800

A Appendix

A.1 Likelihood surface

The importance sampling algorithm proposed in this paper allows computation of the likelihood with respect to parameters of interest. We have a major interest in estimating migration and mutation parameters and therefore in estimating the likelihood surface. There are two aspects of computing the likelihood surface. The first is pointwise computation of likelihoods; the second is to find a way to interpolate the surface locally about the points already computed taking into account that the surface is a random realization of the true surface. Potentially there are a number of different methods for parameter estimation: importance sampling with respect to a proposal distribution containing the parameter distribution (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000); bridge sampling (Fearnhead and Donnelly, 2001); Markov chain Monte Carlo schemes (Beerli *et al.*, 1999); and Bayesian methods (Wilson and Balding, 1998). Beaumont *et. al.* (2002) propose a method for approximate Bayesian statistical inference based on summary statistics. Their approach replaces the full data with suitable summary statistics and approximates the posterior density of the parameters of interest using kernel density estimation techniques. The use of summary statistics allows for increased computational efficiency, although it does not make use of all the information in the data, as typically in many settings sufficient statistics are not available.

Here we have adopted a different approach based on a local linear predictor for interpolated points on the computed points which takes into account random variation of the surface. The likelihood of the full sample of chromosomes is evaluated at points in a design set and then the likelihood surface is estimated using kriging methods. Let $y(\boldsymbol{\theta})$ denote the likelihood function of a sample of n genes and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ be the vector of parameters of interest (for example, mutation and migration rates), where $\boldsymbol{\theta} \in D \subset \mathbb{R}^d$, $d \geq 1$. Suppose we have evaluated the likelihood at a set of N parameter values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$. Each likelihood evaluation can be extremely computer intensive and can require a long running time. We are interested in estimating the likelihood surface, that is, to predict $y(\boldsymbol{\theta})$ at any $\boldsymbol{\theta} \in D$ given the values $\{y(\boldsymbol{\theta}_1), \ldots, y(\boldsymbol{\theta}_N)\}$. The unknown function $y(\boldsymbol{\theta})$ is assumed to be a realisation of a Gaussian process $Y = \{Y(\boldsymbol{\theta}), \boldsymbol{\theta} \in D\}$ (see Ripley (1981)). A Gaussian process is defined by its mean function and covariance function:

$$\mathbb{E}\left[Y(\boldsymbol{\theta})\right] = \mu\left(\boldsymbol{\theta}\right) \tag{A.1}$$

$$\operatorname{cov}\left[Y(\boldsymbol{\theta}_i), Y(\boldsymbol{\theta}_j)\right] = C\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right). \tag{A.2}$$

Moreover, normality of the finite-dimensional distributions is assumed, that is for every finite subset of points $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\} \subset D$, the joint distribution of $(Y(\boldsymbol{s}_1) \ldots, Y(\boldsymbol{s}_k))$ is a multivariate normal. In a situation where the likelihood is computed from independent sampling realizations, the multivariate normal assumption will hold approximately because of the central limit theorem, but the assumption is not critical because prediction of interpolated points in the local

surface explained below really only depends on best prediction with minimum variance consideration.

Given the value of the likelihood at $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$, we wish to predict the likelihood surface at a new point $\widetilde{\boldsymbol{\theta}}$. Let y_N be the column vector defined as

$$y_N = \begin{pmatrix} y(\boldsymbol{\theta}_1) \\ \vdots \\ y(\boldsymbol{\theta}_N) \end{pmatrix}.$$

Then the minimum mean square error unbiased predictor of $y(\widetilde{\boldsymbol{\theta}})$ is given by

$$\widehat{y}(\widetilde{\boldsymbol{\theta}}) = \mathbb{E}(Y(\widetilde{\boldsymbol{\theta}}) \mid y_N) = \mu(\widetilde{\boldsymbol{\theta}}) + k(\widetilde{\boldsymbol{\theta}})K^{-1}(y_N - \mu_N) \tag{A.3}$$

and its variance is

$$\mathrm{var}(\widehat{y}(\widetilde{\boldsymbol{\theta}})) = \mathrm{var}(Y(\widetilde{\boldsymbol{\theta}}) \mid y_N) = C\left(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}\right) - k(\widetilde{\boldsymbol{\theta}})K^{-1}k(\widetilde{\boldsymbol{\theta}})' \tag{A.4}$$

where

$$\mu_N = \begin{pmatrix} \mu(\boldsymbol{\theta}_1) \\ \vdots \\ \mu(\boldsymbol{\theta}_N) \end{pmatrix}, \qquad k(\widetilde{\boldsymbol{\theta}}) = \begin{pmatrix} C(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_1) \\ \vdots \\ C(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_N) \end{pmatrix}$$

and $K = (K_{ij})$ is $N \times N$ matrix whose elements are given by

$$K_{ij} = C\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right).$$

The linear prediction in (A.3) - (A.4) is known in geostatistics and spatial statistics as kriging (Ripley, 1981; Cressie, 1993; Stein, 1999) and has previously been applied to surface estimation in different contexts, including interpolation, image restoration (Geman and Geman, 1984; Ripley, 1988), prediction of deterministic functions (Currin *et al.*, 1991) and analysis of computer experiments (Currin *et al.*, 1991; Sacks *et al.*, 1989a; Sacks *et al.*, 1989b).

We are looking for a general method to estimate the likelihood surface. The surface is assumed to be smooth, continuous and differentiable. In particular, we require that the mean is constant for all $\boldsymbol{\theta} \in D$, i.e. $\mu(\boldsymbol{\theta}) = \mu$, for all $\boldsymbol{\theta} \in D$. As covariance function we use

$$C(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \sigma^2 \exp\left\{-\lambda \sum_{k=1}^{d} (\theta_{ik} - \theta_{jk})^2\right\}. \tag{A.5}$$

where $\lambda > 0$ determines the correlation structure of Y and σ^2 is a scale factor. Of course, other choices of mean and covariance functions are possible. See, for example, Currin *et al.* (1991), Stein (1999). For example, we could incorporate a linear (or polynomial) model for Y through the mean function. In our experience there is no need for this, especially when d is quite large (greater than 4), as predictions based on a constant mean function are quite good. Determination

of μ, λ and σ^2 is usually achieved by iterative search methods (Ripley,1988; Mardia and Marshall, 1984).

The last issue we need to address is the choice of the N values of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ at which to evaluate the likelihood. Ideally we would like all the areas of the space D to be represented. We have applied *Latin Hypercube Sampling* to determine $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ (McKay *et al.*, 1979). Briefly, *Latin Hypercube Sampling* consists of dividing the range of each θ_k, $k = 1, \ldots, d$, into N intervals of equal marginal probability $1/N$ and then sampling a point at random from each interval. In this way a sample θ_{kj}, $j = 1, \ldots, N$ is obtained, and these sampled values form the k-th component in $\boldsymbol{\theta}_i$, $i = 1, \ldots, N$. The components of the various $\boldsymbol{\theta}_i$ are then matched at random. Therefore, there are N intervals on the range of each element of $\boldsymbol{\theta}_i$ and they combine to form N^d cells which cover the space D. The bigger N is, the better the estimation of the likelihood surface is. Of course, the computational cost increases.

The results presented in this paper have been obtained using the `R` package `fields` (`http://www.cgd.ucar.edu/stats/Software/Fields/index.shtml`). This package contains a collection of functions for curve and function fitting with an emphasis on spatial data. In particular, the function `krig` implements spatial process estimates through kriging. The `R` web site is `http://www.R-project.org` where information and software is available, and the `R` manual is *R: A language and environment for statistical computing* (2004).

References

Bahlo, M and Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* **57** 79–95.

Bahlo, M and Griffiths, R. C. (2001). Coalescence time for two genes from a subdivided population. *J. Math. Biol.*, **43** 397–410.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

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

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

Chen, Y. and Liu, J. S. (2000). Discussion on the paper of Stephens and Donnelly (2000) *J. R. Statist. Soc.* B **62** 644–645.

Chen, Y., Xie, J. and Liu, J. S. (2004) Stopping-Time resampling for sequential Monte Carlo methods. *J. Roy. Statist. Soc.* B In Press.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley, New York.

Currin, C., Mitchell T., Morris, M. and Ylvisaker D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86** 953–963.

De Iorio, M. and Griffiths, R. C. (2004a). Importance sampling on coalescent histories. I *Adv. Appl. Probab.* **36** 417–433.

De Iorio, M. and Griffiths, R. C. (2004b). Importance sampling on coalescent histories. II Subdivided population models. *Adv. Appl. Probab.* **36** 434–454.

Excoffier, L. (2001). Analysis of population subdivision. In *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop M., and Cannings, C.), 271–307. Wiley, Chichester.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetics data. *Genetics* **159** 1299–1318

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

Griffiths, R. C. (2001). Ancestral inference from gene trees. In *Genes, Fossils, and Behaviour: an Integrated Approach to Human Evolution* (eds Donnelly, P and Foley, R), IOS Press, NATO Science Series, Series A: Life Sciences Vol. 310, p137–172.

Griffiths, R. C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46** 131–159.

Herbots H. M. (1997). The structured coalescent. In *Progress in Population Genetics and Human Evolution* (eds Donnelly P. and Tavaré S.), IMA Volumes in Mathematics and its Applications, **87** 231–255. Springer Verlag, Berlin, 330pp.

Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes Appln.* **13** 235–248.

Lade, J. A., Murray, N. D., Marks, C. A. and Robinson, N. A. (1996). Microsatellite differentiation between Philip Island and mainland Australian populations of the red fox *Vulpes vulpes*. *Molecular Ecology* **5** 81–87.

Leblois R., Estoup A., Rousset F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance. *Molecular Biology and Evolution*, **20** 491–502.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146.

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 2390–245.

Moran, P.A.P. (1975). Wandering distributions and the electrophoretic profile, I, *Theor. Popul. Biol.* **8** 318–330.

Moran, P.A.P. (1976) Wandering distributions and the electrophoretic profile, II, *Theor. Popul. Biol.* **10** 145–149.

Nagylaki, T. (1983). The robustness of neutral models of geographical variation. *Theor. Popul. Biol.* **24** 268–294.

Nielsen, R. (1997). A likelihood approach to population samples of microsatellite alleles. *Genetics* **146** 711–716.

Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop M., and Cannings, C.), 179–208. Wiley, Chichester.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29** 59–75.

Ohta, T. and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, *Genet. Res.* **22** 201–204.

R Development Core Team (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley, New York.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge.

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, D. J., Bishop M., and Cannings, C.), 239–269. Wiley, Chichester.

Sacks, J., Schiller, S. B. and Welch, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments, (with comments). *Statistical Science* **4** 409–435.

Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* **78** 49–57.

Stein, M. L. (1999). *Interpolation of Spatial Data.* Springer-Verlag, New York.

Stephens, M. (2001). Inference under the coalescent. In *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop M., and Cannings, C.), 213–238. Wiley, Chichester.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. Roy. Statist. Soc.* B **62** 605–655.

Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150** 499–510.

Estimation de paramètres de dispersion en populations structurées à partir de données génétiques

L'estimation des taux de migration et des distributions de dispersion est déterminante pour la résolution des questions portant sur l'évolution des populations naturelles, notamment dans des contextes de biologie des invasions et de la conservation. Le propos de cette thèse est de montrer comment les modèles de populations subdivisées permettent de telles estimations à partir de données génétiques. Une première étude par simulation montre que l'estimation de paramètres démographiques par F-statistiques en isolement par la distance donne de bons résultats, robustes vis à vis: (i) des processus mutationnels des marqueurs génétiques utilisés, et (ii) d'hétérogénéités temporelles et spatiales des paramètres démographiques. Ce travail montre également que les deux approches d'estimation par maximum de vraisemblance fondées sur la coalescence disponibles à ce jour sont fortement limitées par des temps de calculs extrêmement longs. De plus, nos simulations montrent de mauvaises performances d'un de ces approches pour des modèles d'isolement par la distance. L'autre n'a été appliquée, au stade actuel, qu'à des modèles très simples, tel qu'un modèle à 2 populations avec un processus de mutation par pas, pour lequel les estimations sont relativement précises. Le développement de nouveaux algorithmes devrait réduire les temps de calculs et ainsi permettre la considération de modèles plus réalistes. Enfin, ce travail montre qu'une dispersion limitée dans l'espace a des conséquences importantes sur la détection d'une réduction de taille de l'habitat par des méthodes génétiques.

Mots Clés : populations subdivisées, isolement par la distance, dispersion, paramètres démographiques, coalescence, maximum de vraisemblance, F-statistiques, microsatellites.

Inference of dispersal parameters from genetic data in subdivided populations

Inferences of migration rates and/or dispersal distributions in natural populations are important for the understanding of evolution, especially in the context of bioinvasions and conservation biology. The aim of this thesis is to show how subdivided population models allow the estimation of dispersal parameters from genetic data. First, a simulation study show that demographic parameter inference from F-statistics under isolation by distance gives good estimates and is robust to mutational processes of genetic markers and to temporal and spatial heterogeneities of demographic parameters. This work also show that the two presently available approaches for maximum likelihood estimation based on coalescence are strongly limited because of high computation times. Moreover, our simulations show relatively bad performances of one of those approaches for isolation by distance models. The other approach has presently been applied only to simple models such as a 2 population model with stepwise mutation for which estimations are relatively precise. The development of new algorithms should decrease computational times and thus allow the consideration of more realistic models. Finally, this work shows that geographically limited dispersal has important consequences on the detection of habitat size reduction with genetic methods.

Key words : subdivided populations, isolation by distance, dispersal, demographic parameters, coalescence, maximum likelihood, F-statistics, microsatellite markers.