Méthodes bayesiennes pour l'estimation de l'histoire démographique et de la pression de sélection à partir de la structure génétique des populations.

Matthieu Foll

Université Grenoble I - Joseph Fourier
Ecole Doctorale Chimie et Sciences du Vivant

Année 2007

Thèse

présentée en vu d'obtenir le grade de

Docteur en Biologie

Spécialité « Biodiversité - Ecologie - Environnement »

Méthodes bayesiennes pour l'estimation de l'histoire démographique et de la pression de sélection à partir de la structure génétique des populations.

Présentée par

Matthieu Foll

Thèse dirigée par Oscar Gaggiotti

Thèse soutenue le 21 décembre 2007 devant le jury composé de :

Laurent Excoffier	Rapporteur	Professeur, Université de Berne, Suisse
Olivier François	Rapporteur	Professeur, Institut National Polytechnique de Grenoble
Michael Blum	Examinateur	CR CNRS, Grenoble
Arnaud Estoup	Examinateur	DR CNRS, Montpellier
Oscar Gaggiotti	Directeur de thèse	Professeur, Université de Grenoble I

Thèse préparée au sein du
Laboratoire d'Ecologie Alpine
UMR CNRS 5553, Université Joseph Fourier, BP 53 38041 Grenoble Cedex 9

Titre : Méthodes bayesiennes pour l'estimation de l'histoire démographique et de la pression de sélection à partir de la structure génétique des populations.

Résumé : Les récents progrès, dans les domaines de la biologie computationnelle et des techniques de biologie moléculaire, ont conduit à l'émergence d'une nouvelle discipline appelée génomique des populations, et dont l'un des objectifs principaux est l'étude de la structure spatiale de la diversité génétique. Cette structure est déterminée à la fois par des forces neutres, comme la migration et la dérive, et des forces adaptatives comme la sélection naturelle, et trouve des applications importantes dans de nombreux domaines comme la génétique médicale ou la biologie de la conservation. Nous développons ici de nouvelles méthodes statistiques pour évaluer le rôle de la sélection naturelle et de l'environnement dans cette structure spatiale. Le modèle bayésien Dirichlet-multinomial de différenciation génétique est utilisé comme base à ces différentes méthodes. Dans un premier temps, nous proposons d'inclure des variables environnementales dans l'estimation de la structure génétique afin d'identifier les facteurs biotiques et abiotiques qui la déterminent. Ensuite, nous étudions la possibilité d'étendre le modèle Dirichlet-multinomial aux marqueurs dominants, devenus très populaires ces dernières années, mais affectés par différents biais de recrutement. Enfin, nous cherchons à séparer les effets neutres des effets de la sélection naturelle, afin, en particulier, d'identifier les régions du génome qui y sont soumis. Trois bases de données ont été analysées pour illustrer l'utilisation de ces nouvelles méthodes : des données humaines, des données de l'arganier du Maroc et des données de littorine. Finalement, nous avons développé trois logiciels implémentant ces différents modèles.

Mots-clés : Génomique des populations, structure génétique, sélection naturelle, histoire démographique, statistiques bayésiennes, marqueurs AFLP, biais de recrutement.

Title: Bayesian methods for infering demographic history and selective pressure from the genetic structure of populations.

Abstract: Recent advances in the fields of computational biology and molecular biology techniques have led to the emerging discipline of population genomics, whose main objective is the study of the spatial structure of genetic diversity. This structure is determined by both neutral forces, like migration and drift, and adaptive forces, like natural selection, and has important applications in many fields like medical genetics or conservation biology. Here, we develop new statistical methods to evaluate the role of natural selection and environment in this spatial structure. All these methods are based on the Bayesian Dirichlet-multinomial model of genetic differentiation. First, we propose to include environmental variables in the estimation process, in order to identify the biotic and abiotic factors that determine the genetic structure. Then, we study the possibility of extending the Dirichlet-multinomial model to dominant markers, which have become very popular in the last few years, but which are affected by various ascertainment biases. Finally, we try to separate neutral effects from adaptive effects on the genetic structure, in order to identify regions of the genome influenced by natural selection. Three databases have been analyzed as illustrations of the use of these new methods : human data, data of argan tree in Morocco, and data of periwinkle. Finally, we developed three softwares implementing these various models.

Keywords: Population genomics, genetic structure, natural selection, demographic history, Bayesian statistics, AFLP markers, ascertainment bias.

Table des matières

Table des matières v

Préface 1

1 Introduction 3
- 1.1 Contexte . . . 5
  - 1.1.1 Enjeux et motivations . . . 5
  - 1.1.2 Objectifs et approche générale . . . 5
  - 1.1.3 Différentes approches statistiques . . . 7
- 1.2 Statistiques bayésiennes . . . 8
  - 1.2.1 Principe . . . 8
  - 1.2.2 Une méthode d'estimation : les MCMC . . . 9
  - 1.2.3 Une méthode approchée : l'ABC . . . 11
  - 1.2.4 La sélection de modèles : les sauts réversibles . . . 12
- 1.3 Les variations génétiques et l'adaptation locale . . . 13
  - 1.3.1 Les forces évolutives impliquées . . . 14
  - 1.3.2 La génétique des populations comme outil . . . 15
  - 1.3.3 Mesurer les variations génétiques . . . 16
- 1.4 Un modèle bayésien . . . 19
  - 1.4.1 La vraisemblance . . . 19
  - 1.4.2 A priori et à postériori . . . 23
  - 1.4.3 Les extensions développées . . . 24

2 Identifier les facteurs environnementaux qui déterminent la structure génétique des populations 31
- 2.1 Problématique et démarche scientifique . . . 33
- 2.2 Contribution scientifique : article A . . . 34
  - 2.2.1 Matériels et méthodes . . . 34
  - 2.2.2 Résultats . . . 37
  - 2.2.3 Discussion . . . 38
- Conclusion . . . 40

Chapitre 2 - Article A 41
Abstract . . . 42
Introduction . . . 43
Methods . . . 44
Simulation study . . . 49
Applications . . . 57
Discussion . . . 62
Acknowledgments . . . 68
Appendix . . . 70
Supporting informations . . . 73

3 La structure génétique et les marqueurs dominants : intégrer le biais de recrutement 77
3.1 Problématique et démarche scientifique . . . 79
3.2 Contribution scientifique : article B . . . 80
3.2.1 Matériels et méthodes . . . 80
3.2.2 Résultats . . . 81
3.2.3 Discussion . . . 81
Conclusion . . . 83

Chapitre 3 - Article B 85
Abstract . . . 86
Introduction . . . 87
The Bayesian model . . . 90
Sources of bias . . . 92
A solution : an ABC approach . . . 101
Discussion . . . 107
Acknowledgments . . . 109

4 Détection de loci sous sélection : différents marqueurs et différents scénarios démographiques 111
4.1 Problématique et démarche scientifique . . . 113
4.2 Contribution scientifique : article C . . . 113
4.2.1 Matériels et méthodes . . . 114
4.2.2 Résultats . . . 116
4.2.3 Discussion . . . 118
Conclusion . . . 120

Chapitre 4 - Article C 121
Abstract . . . 122

Introduction 123
Methods 124
Simulation study 131
Application 140
Discussion 147
Acknowledgments 150

Discussion **151**

Conclusion générale **157**

Bibliographie **159**

Préface

La génomique des populations est une discipline récente qui s'est imposée grâce aux progrès dans les domaines de la biologie computationnelle et des techniques de biologie moléculaire. Elle nécessite d'analyser des données issues de criblage génomique d'individus issus de différentes populations et permet de répondre à des questions encore réservées hier à quelques espèce modèles. En particulier, l'un des objectifs principaux est l'étude de la structure spatiale de la diversité génétique. Cette structure est déterminée à la fois par des forces neutres, comme la migration ou la dérive, et des forces adaptatives comme la sélection naturelle, et trouve des applications importantes dans de nombreux domaines comme la génétique médicale ou la biologie de la conservation. Ce travail s'articule autour de trois chapitres, qui correspondent à trois nouvelles méthodes statistiques que nous proposons pour évaluer le rôle de la sélection naturelle et de l'environnement dans cette structure spatiale. Un chapitre introductif présente le cadre général de la génomique des populations et le modèle bayésien Dirichlet-multinomial qui sera la base des trois modèles proposés ici.

Dans le second chapitre nous proposons une nouvelle méthode bayésienne hiérarchique pour identifier les facteurs environnementaux qui déterminent la structure génétique des populations. Parmi les différentes forces de l'évolution, l'environnement dans lequel évolue une population va particulièrement influencer les migrations (barrières géographiques, ressources disponibles etc.) et la sélection naturelle (adaptation locale). Ainsi des conditions environnementales distinctes entre populations vont conduire à des variations génétiques entre elles. Pour cette raison, identifier les facteurs environnementaux qui permettent d'expliquer la nature de la différenciation génétique observée est un problème fondamental en génétique des populations.

Dans le troisième chapitre, on étudie la possibilité d'utiliser le modèle Dirichlet-multinomial avec des marqueurs dits dominants pour lesquels il est impossible de distinguer les individus hétérozygotes des individus homozygotes pour l'allèle dominant. Cette question présente un enjeu particulièrement important pour les marqueurs AFLP qui sont devenus très populaires ces dernières années. Nous montrons qu'il est nécessaire de tenir compte

du biais de recrutement qui touche ces marqueurs et nous proposons une nouvelle approche originale de type « Approximate Bayesian Computation » (ABC) pour résoudre ce problème.

Enfin le dernier chapitre propose de séparer les effets neutres des effets adaptatifs qui conditionnent la structure génétique. Cela permet en particulier d'identifier les régions du génome qui sont soumises à la sélection naturelle. Cette question est essentielle dans les domaines de la génétique médicale, de l'agronomie et de la biologie de la conservation. En effet, on s'attend à ce que les marqueurs sélectionnés soient liés à une fonction biologique importante, et ils sont donc de bons candidats pour être impliqués dans des maladies génétiques. De plus il existe un effet de confusion entre histoire démographique et sélection naturelle, ce qui fait de cette question un préalable à toute étude de la structure spatiale de la diversité génétique.

Différentes bases de données sont utilisées afin d'illustrer le potentiel d'application des méthodes proposées, mais aussi d'apporter de nouvelles réponses à des questions les concernant : des données humaines, des données de l'arganier du Maroc et des données de littorine. Cette thèse a fait l'objet de plusieurs travaux écrits qui sont intégrés dans chaque chapitre.

Introduction

1

« Il n'y a pas la place dans notre système pour des spéculations concernant la probabilité que le soleil se lève demain. Avant de parler de cela, il faudrait se mettre d'accord sur un modèle (idéalisé) qui tournerait vraisemblablement autour de "parmi un nombre infini de mondes, on en choisit un au hasard…" Un peu d'imagination est nécessaire pour construire un tel modèle, mais il semble à la fois inintéressant et dénué de sens. »

William Feller

1.1 Contexte

1.1.1 Enjeux et motivations

La structure spatiale de la diversité génétique est déterminée à la fois par des forces *neutres*, comme la migration ou la dérive, et des forces *adaptatives* comme la sélection naturelle, et son étude trouve des applications importantes dans de nombreux domaines.

On peut citer tout d'abord la conservation de la biodiversité qui nécessite de connaitre la structure génétique pour identifier les populations qui doivent être la cible de mesure de conservation. Ainsi on cherchera par exemple à connaitre les facteurs qui influencent cette structure pour concevoir une stratégie de gestion. De plus, il est important d'identifier des régions du génome responsables de traits phénotypiques qui améliorent l'adaptation locale ou qui peuvent faciliter l'adaptation à un changement futur (climatique par exemple). On pourra alors conserver en priorité les populations qui ont une fréquence élevée de ces gènes ou de plusieurs d'entre eux.

Ensuite, l'identification de gènes soumis à la sélection naturelle permet l'amélioration d'espèces agricoles, l'identification des gènes impliqués dans des maladies génétiques ou dans une résistance à un pathogène, chez l'Homme et toute espèce ayant une importance économique forte. En effet, les régions du génome soumises à la sélection ont une plus grande probabilité d'avoir un rôle dans une fonction biologique importante, et sont également de bons candidats pour être impliqués dans des maladies ou des dysfonctionnements génétiques. Cependant, ce problème n'est pas indépendant de l'identification de la structure génétique, car il existe un effet de confusion entre histoire démographique et sélection. Ainsi, ignorer la structure génétique conduit à l'identification de faux positifs dans les tests de neutralité.

Enfin, au delà de permettre des tests de neutralité plus fiables, l'identification de la structure spatiale permet de retracer l'histoire démographique et ainsi de comprendre comment la diversité actuellement observée s'est constituée au fil du temps. On peut aussi citer l'intérêt historique dans ces applications au genre *Homo* qui fait de la génétique un outil important de la paléoanthropologie.

1.1.2 Objectifs et approche générale

L'objectif est de développer de nouvelles méthodes statistiques pour évaluer le rôle de la sélection naturelle et de l'environnement dans la structure spa-

tiale de la diversité génétique des populations. Plus précisément, deux problèmes principaux seront traités :

- Quels sont les facteurs environnementaux à l'origine de la structure génétique observée ?
- Quelles sont les parties du génome soumises à la sélection naturelle ?

Dans ce cadre, le but sera de développer des méthodes mais aussi de les implémenter dans des logiciels mis à disposition de la communauté. Un enjeu supplémentaire est de pouvoir adapter ces solutions aux données génétiques disponibles pour des espèces non-modèles, et donc de les rendre accessibles dans des domaines tels que la biologie de la conservation et l'écologie moléculaire.

Plusieurs bases de données seront utilisées afin d'illustrer le potentiel d'application des méthodes proposées, mais aussi d'apporter de nouvelles réponses à des questions les concernant. En particulier, nous analyserons trois bases de données :

- Les données de génétique humaine du HGDP-CEPH publiées par Cann *et al.* (2002) et composées de 1056 individus, répartis dans 51 populations avec 835 marqueurs microsatellites. On cherchera à retracer l'histoire de l'expansion démographique des Hommes et à identifier les gènes qui ont été soumis à la sélection naturelle au cours de ce processus.
- Les données de l'arganier du Maroc publiées par Petit *et al.* (1998) et composées de 12 populations de 20 à 50 individus avec 12 marqueurs polymorphes (isozymes). On cherchera à déterminer si le degré d'isolement génétique est influencé par l'isolement géographique et/ou un autre facteur, ici, l'altitude.
- Les données de littorine publiées par Wilding *et al.* (2001) et composées de 6 populations d'une cinquantaine d'individus, pour lesquelles 290 marqueurs dominants AFLP ont été développés. On cherchera à identifier les régions du génome soumises à la sélection naturelle et qui ont conduit à un isolement reproductif partiel entre deux morphotypes. Cette étude sera menée en prenant soin de séparer les effets de l'histoire démographique de ceux de la sélection.

Le point commun de toutes ces méthodes est l'estimation de paramètres à partir de données de génomique des populations. Ces données sont disponibles en quantité de plus en plus importante, et ce, même pour des espèces non-modèles, grâce au développement de nouvelles techniques de biologie moléculaire. En même temps, les sujets étudiés dans ce domaine rendent souvent long et couteux l'échantillonnage d'un grand nombre d'individus ou de

populations. Ces deux raisons poussent à utiliser des méthodes statistiques basées sur une fonction de vraisemblance, afin d'intégrer toute l'information disponible dans les estimations.

1.1.3 Différentes approches statistiques

Les statistiques sont depuis longtemps un outil incontournable des Sciences du vivant et les récentes avancées dans le domaine de la biologie moléculaire ont crée un engouement supplémentaire autour de ce nouveau type de données constamment renouvelées. L'objectif de toute méthode statistique est d'*interpréter* ou d'*expliquer* des données. Par exemple, nous serons amenés ici à vouloir répondre aux questions suivantes à partir de données génomiques de plusieurs populations :

- Quels niveaux de différenciation génétique ont subi ces populations ?
- Quels sont les facteurs environnementaux responsables de cette différenciation génétique ?
- Quelle est l'histoire démographique des populations ?
- La sélection naturelle a-t-elle influencée ce marqueur moléculaire ?

On peut distinguer deux grandes familles de méthodes : les statistiques descriptives et les statistiques inférentielles. On peut noter que le terme « statistiques » désigne à la fois la discipline (aussi appelée « *la* statistique ») et le produit des analyses qui en découlent (par exemple la moyenne et la variance sont *des* statistiques). Les statistiques descriptives ont pour but de *résumer* et de *représenter* un ensemble de données à l'aide de statistiques (« statistiques sommaires »), de tableaux ou de figures quand celles ci sont nombreuses. Les statistiques inférentielles considèrent les données observées comme un échantillon issu d'un processus ou d'une population plus large dont on cherche à induire certaines caractéristiques. Plus précisément, on est amené à *modéliser* le processus qui a permis d'obtenir l'échantillon observé et on cherche à estimer les paramètres qui définissent ce modèle. De plus, on cherchera souvent à effectuer des tests d'hypothèses sur la validité du modèle choisi. Ces deux familles de méthodes sont complémentaires et s'utilisent généralement de paire.

Dans le domaine des statistiques inférentielles, il existe deux écoles : l'inférence fréquentiste (appelée aussi « statistiques classiques ») et l'inférence bayésienne. L'inférence fréquentiste définit la probabilité d'un évènement comme la limite de sa fréquence lors d'un grand nombre d'épreuves. Au contraire, l'approche bayésienne utilise une vision subjective des probabilités mesurant le degré de certitude d'une hypothèse. Elles doivent simplement

obéir aux axiomes classiques des probabilités et peuvent s'appliquer à tout type d'évènement (évènements passés, non répétables etc.). En statistiques classiques, on considère ainsi qu'il existe une valeur « réelle » pour la probabilité d'un évènement qui lui est inhérente (comme par exemple la probabilité d'être malade quand on a obtenu un résultat positif d'un test) et dont on peut obtenir une estimation à partir d'un échantillon. Au contraire, en statistiques bayésiennes, les probabilités sont définies par rapport aux observations faites (comme par exemple la probabilité que *cette* personne soit malade étant donné que son test est positif, question qui n'a pas de sens en statistiques classiques).

Les statistiques bayésiennes ont fait l'objet de nombreuses controverses depuis leur introduction et une image caricaturale persiste, basée sur l'idée que cette approche reposerait entièrement sur la subjectivité du choix d'une loi à priori qui déterminerait totalement l'inférence obtenue. La solution de facilité consiste à rétorquer que l'on peut tout à fait utiliser un à priori dit « non informatif » pour éviter ce problème. Néanmoins cela conduit le plus souvent à penser que si, dans ce cas, les statistiques bayésiennes ne sont pas arbitraires, elle deviennent simplement inutiles car équivalentes à une approche classique de maximum de vraisemblance. Aujourd'hui le débat entre partisans et opposants de l'approche bayésienne semble s'estomper avec le nombre de problèmes qu'elle a permis de résoudre. En effet, les statistiques bayésiennes se sont imposées dans des domaines très diverses pour des raisons pragmatiques bien plus que pour des raisons philosophiques, et les différents modèles proposés ici en sont une illustration supplémentaire.

1.2 Statistiques bayésiennes

1.2.1 Principe

On se place dans le cas général où l'on a observé des données D que l'on peut modéliser par un modèle M, paramétré par θ. Les statistiques classiques sont basées sur $P(D|\theta)$, la probabilité des données connaissant les paramètres, aussi appelée « vraisemblance » des données et notée $L(\theta)$. En statistiques bayésiennes, on est amené à travailler avec $P(\theta|D)$, la probabilité des paramètres connaissant les données (appelée « à postériori »). En statistiques classiques, θ n'est pas considérée comme une variable aléatoire et on ne peut pas définir $P(\theta|D)$. L'approche bayésienne consiste à regarder θ comme une variable aléatoire et donc à lui attribuer une loi de probabilité $P(\theta)$ (appelée « à priori »).

L'estimation de l'à postériori se fait alors avec la formule de Bayes :

$$P(\theta|D) = \frac{L(\theta)P(\theta)}{P(D)}$$

Plus généralement, quand θ_i forme une partition de l'espace des évènements, on a :

$$P(\theta_i|D) = \frac{L(\theta_i)P(\theta_i)}{\sum_j P(D|\theta_j)P(\theta_j)}$$

Cette formule n'est valable que pour un paramètre θ discret et on peut les généraliser au cas d'une variable continue. Si on note f la densité de probabilité à priori du paramètre θ et π sa distribution à postériori, on a :

$$\pi(\theta|D) = \frac{L(\theta)f(\theta)}{\int P(D|\theta)f(\theta)d\theta}$$

Le dénominateur est une constante, et on note souvent :

$$\pi(\theta|D) \propto L(\theta)f(\theta).$$

Dans la plupart des modèles complexes, il n'est pas possible de calculer directement la loi à postériori à cause de l'intégrale du dénominateur, ni de calculer des estimateurs, comme les moments, basés sur cette distribution. Différentes méthodes existent pour approcher cette loi et nous en présentons deux qui seront utilisées dans les modèles proposés ici. Ces deux méthodes reposent sur la même idée : on cherche à obtenir des échantillons tirés selon la loi à postériori que l'on utilise pour en estimer les caractéristiques (moyenne, « intervalles de crédibilité », densité etc.). Enfin une généralisation de la première méthode sera présentée, qui permet d'estimer en même temps la probabilité de modèles de dimensions différentes.

1.2.2 Une méthode d'estimation : les MCMC

Les méthodes de Monte Carlo classiques permettent d'approcher la loi π si on peut la simuler, ce qui n'est pas le cas ici. L'algorithme de Metropolis-Hasting (Monte Carlo par chaines de Markov , MCMC, Hastings 1970) permet de contourner ce problème en simulant une chaine de Markov[1] ayant la distribution stationnaire π même définie à une constante de normalisation près. Pour cela il convient de trouver une densité de transition q qui vérifie :

[1]On parle ici de chaine de Markov continue où la matrice de transition est remplacée par une densité de transition.

$$\pi(\theta|D)\, q(\theta,\theta') = \pi(\theta'|\, D) q(\theta',\theta) \tag{1.1}$$

Généralement, on ne peut pas trouver facilement la densité de transition q qui convient ; l'idée de l'algorithme est de modifier le processus de simulation de la chaine pour s'adapter au « mauvais choix » fait sur q que l'on laisse libre. Quand le processus est en un point θ, il passe en un point θ' généré par $q(\theta,\cdot)$. A cause du mauvais choix de q on aura par exemple $\pi(\theta|D)\, q(\theta,\theta') > \pi(\theta'|D)\, q(\theta',\theta)$. Dans ce cas, le processus a tendance à aller trop souvent de θ vers θ' et trop rarement de θ' vers θ. Une solution pour réduire le nombre de sauts de θ vers θ' est d'introduire une probabilité de faire effectivement le saut notée $\alpha(\theta,\theta') < 1$. Cela signifie que lorsque le processus s'apprête à passer dans un nouvel état θ', on lui refuse la transition avec la probabilité $1-\alpha(\theta,\theta')$ et il reste dans l'état actuel. Ainsi la densité de transition effective est :

$$p(\theta,\theta') = q(\theta,\theta')\,\alpha(\theta,\theta')$$

Il reste maintenant à déterminer $\alpha(\theta,\theta')$ tel que p vérifie l'équation de réversibilité 1.1. Dans l'exemple précédent, on va évidemment définir $\alpha(\theta',\theta) = 1$. On a alors la condition :

$$\pi(\theta|D)\, q(\theta,\theta')\,\alpha(\theta,\theta') = \pi(\theta'|D)\, q(\theta',\theta)$$

On peut effectuer le même raisonnement si le processus a tendance à aller trop souvent de θ' vers θ. Finalement, la probabilité de saut est définie par :

$$\begin{aligned}\alpha(\theta,\theta') &= \min\left(\frac{\pi(\theta'|D)\, q(\theta',\theta)}{\pi(\theta|D)\, q(\theta,\theta')},1\right) \text{ si } \pi(\theta|D)\, q(\theta,\theta') \neq 0\\ &= 1 \text{ sinon.}\end{aligned}$$

On note qu'il est bien possible d'appliquer la méthode pour la distribution π que l'on ne connait qu'à une constante multiplicative près, car elle va se simplifier dans le quotient de l'équation précédente. Comme n'importe quelle méthode de simulation de chaine de Markov, il convient de déterminer le nombre n_0 à partir duquel on considère que la distribution stationnaire est atteinte et que l'effet de la valeur initiale peut être ignoré. De même, comme toute méthode de Monte Carlo, il convient de déterminer le nombre N d'itérations pour lequel on considère avoir un nombre suffisant de valeurs pour approcher la distribution π. Il faut aussi noter que les simulations produites

par la chaine de Markov ne sont pas indépendantes ; il faudra alors obtenir un échantillon plus important que si elles l'étaient.

1.2.3 Une méthode approchée : l'ABC

L'algorithme MCMC présenté précédemment est *approché* dans le sens où l'on n'obtient la distribution π qu'à travers des échantillons qui en sont issus. La méthode que nous allons présenter maintenant est dite *approchée* dans le sens où l'on va non seulement obtenir une distribution à postériori à travers des échantillons qui en sont issus, mais aussi où la distribution approchée ne sera elle même qu'une approximation de π. Pour cette raison on la nomme « computation bayésienne approchée » (ABC, pour Approximate Bayesian Compuation).

L'algorithme ABC est utile dans des situations où l'on ne peut pas calculer la fonction de vraisemblance $P(D|\theta)$ ou si par exemple son évaluation est très couteuse en temps de calcul. Pour remplacer cette vraisemblance, il suffit de pouvoir simuler des données selon le modèle M. L'idée est basée sur une méthode de rejet, et afin d'augmenter le taux d'acceptation, on choisit de résumer les données D par un ensemble de statistiques sommaires S. Si ces statistiques sont exhaustives, on aura $\pi(\theta|D) = \pi(\theta|S)$. L'algorithme proposé par Pritchard *et al.* (1999) est le suivant :

Algorithme 1.

1. Choisir des statistiques S, calculer leur valeur s pour D
2. Choisir une tolérance δ
3. Générer θ selon $f(\cdot)$
4. Simuler D' selon M avec les paramètres θ
5. Calculer s' pour D'
6. Calculer la distance $\|s - s'\|$
7. Accepter θ si $\|s - s'\| < \delta$
8. Retourner à 3.

Cet algorithme est itéré jusqu'à l'obtention d'un nombre m d'échantillons souhaités. Les observations acceptées à l'étape 7 ont pour distribution $\pi(\theta|S)$ quand $\delta \to 0$. Beaumont *et al.* (2002) ont proposé une modification afin de

corriger les estimations obtenues. Avec l'algorithme précédent, on a simulé des paires indépendantes (θ_i, s_i), $i = 1 \dots m$. L'amélioration proposée consiste à pondérer les valeurs θ_i selon $\|s_i - s'\|$ et à ajuster θ_i en utilisant une régression linéaire locale avec cette pondération. L'idée est qu'en moyenne, plus s_i est loin de s, plus θ est loin de la valeur réelle du paramètre à estimer. On accorde un poids plus fort aux θ qui sont associés à des statistiques plus proches des données observées. On pose :

$$\theta_i = \alpha + (s_i - s)^T \beta + \epsilon_i, \ i = 1 \dots m$$

Quand $s_i = s$, $E[\theta|S = s] = \alpha$ et on ajuste les valeurs : $\theta_i^* = \theta_i - (s_i - s)^T \beta$. En pratique on estime α et β par la méthode des moindres carrés, pondérée par $K_\delta(\|s_i - s'\|)$, où K_δ est le noyau d'Epanetchnikov de paramètre δ.

Cet algorithme sera utilisé dans le chapitre 3 pour permettre d'inclure le biais de recrutement des marqueurs AFLP dans les estimations.

1.2.4 La sélection de modèles : les sauts réversibles

On peux généraliser l'approche MCMC pour considérer des modèles alternatifs ayant des dimensions différentes dont on veut estimer la probabilité respective. On note $\boldsymbol{\theta}_M$ le vecteur des paramètres sous le modèle M de dimension n_M. On cherche alors à estimer la distribution à postériori :

$$\pi(\boldsymbol{\theta}_M, M|D) \propto L(\boldsymbol{\theta}_M) f(\boldsymbol{\theta}_M) P(M).$$

Où $P(M)$ désigne la probabilité à priori du modèle M. L'algorithme qui vient d'être présenté permet de construire la chaine pour un modèle donné, mais pas de passer d'un modèle M à un modèle M'. Pour se déplacer entre les différents modèles, on utilise une généralisation de cet algorithme, dit à sauts réversibles (« Reversible Jump MCMC », Green 1995) et noté RJMCMC. On note $p\,(M \to M')$ la probabilité de la transition de M à M'. Par exemple, si on suppose que $n_{M'} > n_M$, on crée un vecteur $\boldsymbol{u}$ composé de $n_u = n_{M'} - n_M$ variables générées selon une densité proposée $q_{M,M'}\,(\boldsymbol{u})$, indépendante de $\boldsymbol{\theta}_M$, et on prend $\boldsymbol{\theta}_{M'} = g_{M,M'}\,(\boldsymbol{\theta}_M, \boldsymbol{u})$. Le saut inverse $(M' \to M)$ peut être fait en prenant la transformation inverse, qui est déterministe dans ce sens une fois M choisi. Le saut de $(\boldsymbol{\theta}_M, M)$ vers $(\boldsymbol{\theta}_{M'}, M')$ est accepté avec la probabilité $\alpha\,(M, M') = min\,[1, A\,(M, M')]$, avec :

$$A\,(M, M') = \frac{\pi\,(\boldsymbol{\theta}_{M'}, M'|D)\, q_{M',M}\,(\boldsymbol{u}')}{\pi\,(\boldsymbol{\theta}_M, M|D)\, q_{M,M'}\,(\boldsymbol{u})} \frac{p\,(M' \to M)}{p\,(M \to M')} \left| \frac{\partial g\,(\boldsymbol{\theta}_M, \boldsymbol{u})}{\partial\,(\boldsymbol{\theta}_M, \boldsymbol{u})} \right|$$

On peut alors estimer les probabilités de chaque modèle en comptant le nombre de fois où la chaine les visite. L'équation précédente se simplifie fortement dans beaucoup de cas pratiques. Par exemple, on supposera souvent que tous les modèles ont la même probabilité à priori et on proposera aussi souvent de passer de M à M' que l'inverse. Dans le cas où le modèle M' consiste à ajouter un nouveau paramètre λ au modèle M et où l'on construit $\boldsymbol{\theta}_{M'} = (\boldsymbol{\theta}_M, \lambda)$ en tirant λ selon une loi q, on aura :

$$A\left(M, M'\right) = \frac{L\left(\boldsymbol{\theta}_{M'}\right) f(\lambda)}{L\left(\boldsymbol{\theta}_M\right) q(\lambda)}$$

Pour considérer le saut de M' à M, il suffit d'accepter de retirer λ du modèle avec la probabilité $min\left[1, 1/A\left(M, M'\right)\right]$. Cette formulation simplifiée sera utilisée dans les chapitres 2 et 4.

1.3 Les variations génétiques et l'adaptation locale

Le processus de l'évolution commun à toutes les espèces peut être résumé en deux étapes :

1. Un nouveau caractère héréditaire apparait chez un, ou plusieurs individus.
2. Ce caractère se repend, soit parce qu'il offre une meilleur adaptation, soit par un effet du hasard.

On peut définir l'adaptation comme l'état optimum d'un organisme par rapport à l'environnement dans lequel il se trouve. Par extension, l'adaptation définie aussi l'ensemble des processus évolutifs qui mènent à cet état. Les exemples d'adaptation sont nombreux dans la nature et les plus spectaculaires sont par exemple le mimétisme qui aide à échapper aux prédateurs où bien sûr le long cou des girafes. A ce titre, l'adaptation est certainement la partie la plus visible de l'évolution des espèces en général.

Les premières études de l'adaptation avaient pour but de comprendre les processus de la spéciation et la traitaient donc à l'échelle des espèces (macro-évolution). Pourtant, les individus d'une même espèce se trouvent souvent dans un environnement hétérogène et qui peut être amené à changer au cours du temps. Ainsi, un processus d'adaptation peut tout à fait agir sur certaines populations d'une même espèce. On parle alors d'adaptation locale ou de micro-évolution.

1.3.1 Les forces évolutives impliquées

L'adaptation locale est le résultat de l'action de plusieurs mécanismes évolutifs inter-dépendants, appelés forces évolutives. Le processus de l'évolution défini précédemment nécessite l'héritabilité des nouveaux caractères, et implique donc des variations au niveau génétique. Ainsi, alors que certaines forces sont d'ordre démographique, d'autres sont purement d'ordre génétique. Nous détaillons ici l'influence des principales forces impliquées.

La sélection naturelle

La sélection naturelle désigne la survie et la reproduction des individus les mieux adaptés à leur environnement, et par conséquence, le fait que les caractères héréditaires responsables de cette adaptation soient mieux transmis et que leur fréquence augmente au cours des générations. On retrouve ici trois principes fondamentaux dans le processus de sélection naturelle : l'existence d'une *variation*, qui entraine une meilleure *adaptation* et qui est *héréditaire*.

La mutation

La mutation est la force évolutive qui permet de donner naissance aux variations héréditaires nécessaires à la sélection naturelle. La plupart des mutations sont sans effet (neutres) et elles peuvent disparaitre ou se fixer par un effet aléatoire. Les mutations délétères (défavorables) sont éliminées par la sélection naturelle, alors qu'au contraire celles responsables d'un caractère avantageux vont voir leur fréquence augmenter.

La dérive génétique

Les variations génétiques liées au hasard évoquées pour la mutation sont le résultat de la dérive génétique. Cette force est la conséquence du processus d'échantillonnage aléatoire des gamètes à chaque génération. La dérive a donc tendance à fixer aléatoirement un allèle quelconque et son effet est d'autant plus important que la population est petite.

La migration

Les migrations permettent le flux de gènes entre plusieurs populations. Elles s'opposent en ce sens à l'adaptation locale, mais sont aussi une source de renouvèlement de la diversité génétique. En effet, la dérive génétique agissant

indépendamment dans chaque population, elle va fixer des allèles différents, et la migration peut réintroduire des allèles perdus dans une population.

1.3.2 La génétique des populations comme outil

La plupart des espèces sont structurées dans l'espace : il est plus probable que des individus se reproduisent avec des individus géographiquement proches, qu'avec des individus éloignés. De plus, des contraintes environnementales ou géographiques peuvent renforcer cet isolement. De cette façon, les individus peuvent être regroupés en « dèmes[2] ». On peut noter que les dèmes eux même peuvent posséder une structure similaire, en étant par exemple regroupés par région. Selon les espèces, ce principe peut être itéré avec plus ou moins de profondeur et on observe alors une structure hiérarchique des populations ayant un certain niveau d'autosimilarité.

Les différentes forces évolutives présentées précédemment et impliquées dans le processus d'adaptation, vont laisser une signature génétique. Certaines vont agir sur tout le génome comme la migration ou la dérive alors que d'autres, comme la sélection, vont agir sur les déterminants génétiques des caractères qu'elles influencent. Pour ces deux raisons, l'adaptation locale peut être étudiée à travers les variations génétiques observées entre différentes populations d'une même espèce. Ainsi, la génétique des populations est un outil de premier choix pour l'étude de cette variation dite « intra-spécifique ».

Encore aujourd'hui, le séquençage systématique d'un grand nombre d'individus ne peut être réalisé que pour quelques espèces modèles, principalement à cause de son cout financier. Pour cette raison, on a recours à des marqueurs moléculaires pour déterminer et comparer des génotypes. Les marqueurs peuvent détecter des mutations ponctuelles de la séquence d'ADN ou bien des modifications du nombre de copies de motifs répétés. Certains permettent de distinguer les individus possédant deux copies du même allèle et sont dits codominants, alors que d'autres ne permettent pas de distinguer les individus hétérozygotes des individus homozygotes pour l'allèle dominant, et sont alors dits dominants. Un grand nombre de marqueurs différents existent, et nous en présentons ici brièvement trois, qui sont particulièrement utilisés depuis quelques années :

- Les SNP (« Single Nucleotide Polymorphism ») représentent simplement le changement d'une base en une autre. La majorité du polymorphisme du génome est constitué de SNP. Ces marqueurs présentent en général deux allèles différents et sont codominants.

[2]Les dèmes sont aussi appelés sous-populations ou, simplement, populations

- Les microsatellites ou SSR (« Single Sequence Repeats ») sont des motifs des quelques bases répétées plusieurs fois dans le génome. Le nombre de répétitions observées définit alors les différents allèles du marqueur. Ces marqueurs sont très polymorphes et codominants.
- Les AFLP (« Amplified Fragment Length Polymorphism ») sont de courtes séquences d'ADN basées sur la combinaison d'une enzyme de restriction et d'une amorce. Ces marqueurs sont dominants et on ne détecte que la présence ou l'absence d'une séquence de longueur donnée chez un individu.

1.3.3 Mesurer les variations génétiques

Les différentes forces évolutives présentées précédemment vont faire varier les fréquences alléliques au cours des générations et les différents marqueurs présentés vont permettre de mesurer ces variations génétiques. Wright (1931) a montré que ces fréquences allaient alors atteindre un équilibre dans différents modèles démographiques simples. On considère deux allèles A et a de fréquences respectives p et $1-p$ dans une population panmictique de taille N, où, à chaque génération, une force évolutive fait varier la fréquence p de Δp. Si la distribution des fréquences alléliques atteint un équilibre, la densité de probabilité de la fréquence p_c vérifie l'équation suivante :

$$\varphi(p_c) = \frac{\Gamma(2N)}{p_c(1-p_c)\Gamma(2Np_c)\Gamma(2N(1-p_c))} \int_0^1 (p+\Delta p)^{2Np_c}(1-p-\Delta p)^{2N(1-p_c)}\varphi(p)dp \tag{1.2}$$

Cette formulation obtenue par Wright (1931) est très générale et permet de calculer la distribution des fréquences alléliques sous différents scénarios en adoptant différentes formes de Δp. On peut résoudre cette équation dans un modèle simple de migration : on considère une ile dans laquelle une proportion m d'individus est remplacée à chaque génération par des immigrants d'une grande population où la fréquence des allèles A et a est constante et notée p_m et $1-p_m$. Dans ce cas, on a $\Delta p = -m(p-p_m)$ et si on a $4Nm \ll 1$ la solution de l'équation 1.2 se simplifie :

$$\varphi(p) = \frac{\Gamma(4Nm)}{\Gamma(4Nmp_m)\Gamma(4Nm(1-p_m))} p^{4Nmp_m-1}(1-p)^{4Nm(1-p_m)-1}$$

Cette distribution est une distribution beta avec les paramètres $4Nmp_m$ et $4Nm(1-p_m)$ (voir figure 1.1). On peut généraliser ce résultat dans le cas de

k allèles (Wright 1969) en utilisant une distribution de Dirichlet. La densité de probabilité du vecteur $\boldsymbol{p} = (p_1, \ldots, p_k)$ est alors :

$$\varphi(\boldsymbol{p}) = \Gamma(\theta) \prod_{i=1}^{k} \frac{p_i^{\theta p_{mi} - 1}}{\Gamma(\theta p_i)} \tag{1.3}$$

où $\boldsymbol{p_m} = (p_{m1}, \ldots, p_{mk})$ est le vecteur des fréquences alléliques parmi les immigrants.

Le modèle de Wright (1931) peut être étendu pour considérer plusieurs iles qui reçoivent toutes et indépendamment des immigrants d'une source avec une fréquence allélique constante (« Island-Mainland metapopulation »). On rencontre aussi cette situation si on suppose que les immigrants entrant dans une ile sont choisis aléatoirement dans un grand nombre d'iles (appelé « Infinite Island Model »).

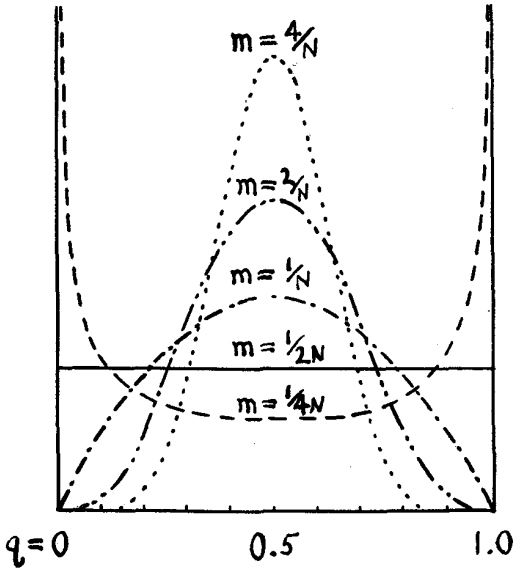

FIGURE 6.—Distribution of frequencies of a gene among subdivisions of a population in which $q_m = 1/2$ (or probability array of gene within a subdivision) under various amounts of intermigration. $y = Cq^{4Nmq_m - 1}(1-q)^{4Nm(1-q_m)-1}$.

FIG. 1.1 – *Distribution de la fréquence d'un allèle A dans une ile recevant des immigrants d'une grande population avec $p_m = 1/2$, d'après* Wright *(1931)*

Comme expliqué précédemment, on peut mesurer la différenciation génétique d'une population en utilisant le coefficient F_{ST}. Crow et Kimura (1970) ont défini ce coefficient comme étant la probabilité que deux gènes tirés aléatoirement dans un dème aient un ancêtre commun dans celui ci. Dans ce modèle en iles, on aura ainsi à l'équilibre :

$$F_{ST} = (1-m)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{ST} \right]$$

Et si on a $m \ll 1$, on obtient :

$$F_{ST} \approx \frac{1}{1+4Nm}$$

Propriété 1.1 *On a montré ici que le vecteur des fréquences alléliques $\boldsymbol{p}$ d'une population de taille N, dans laquelle une proportion $m \ll 1$ d'individus sont remplacés à chaque génération par des immigrants tirés aléatoirement parmi un grand nombre de populations, suivait une distribution de Dirichlet de paramètre $\theta \boldsymbol{p_m}$, avec $F_{ST} = 1/(1+\theta)$, $\theta = 4Nm$ et $\boldsymbol{p_m}$ le vecteur des fréquences alléliques parmi les immigrants.*

Plus récemment, Nicholson *et al.* (2002) ont proposé un modèle de « pure fission » avec deux allèles dans lequel les dèmes ont évolué en isolement après s'être tous séparés en même temps d'une population ancestrale. L'avantage de ce modèle est qu'il ne considère pas une situation d'équilibre comme celui de Wright (1931). Leur formulation ne conduit pas exactement à la même distribution, mais à une loi normale tronquée de moyenne p_m et de variance $cp_m(1-p_m)$, avec p_m la fréquence dans la population ancestrale. Le paramètre c est analogue au coefficient F_{ST} et ils sont égaux quand $c \to 0$. Quand $p_m \approx 0.5$ et F_{ST} n'est pas trop élevé, les deux distributions sont similaires. De plus, les fréquences alléliques dans les deux modèles ont les mêmes moments d'ordre 1 et 2. (Marchini et Cardon 2002) ont comparé l'adéquation des deux modèles à des jeux de données d'humains et ont observé que les deux paramétrisations donnaient des résultats comparables, excepté en Europe où le modèle de Nicholson *et al.* (2002) était meilleur. Ainsi, Falush *et al.* (2003) ont proposé d'utiliser la distribution de Dirichlet de l'équation 1.4 dans un modèle de pure fission.

Dans le cadre de l'étude de la structure génétique, on est amené à utiliser des données avec de multiples marqueurs dans plusieurs populations. Dans toute la suite, on considère ce modèle dans le cas de J populations et de I loci, avec K_i allèles au locus i. On peux noter que le modèle de Wright (1931) ainsi généralisé n'impose pas que le taux de migration m ou le nombre d'individus N soit le même dans chaque population. Ainsi, on note F_{ST}^{ij} le coefficient de différenciation génétique de la population j au locus i[3]. On note $\boldsymbol{p_i} = \{p_{ik}\}$ le vecteur des fréquences alléliques de la population ancestrale au locus i, où p_{ik} est la fréquence de l'allèle k ($\sum_k p_{ik} = 1$) et de la même façon, $\widetilde{\boldsymbol{p_{ij}}} = \{\widetilde{p_{ijk}}\}$ le vecteur des fréquences alléliques de la population j au locus i. On a alors avec le modèle présenté ci-dessus :

[3]Le paramètre F_{ST}^{ij} étant directement lié au nombres d'immigrants arrivant dans la population j, il devrait être constant à travers les loci i. On conserve ici la notation la plus générale, ce problème sera discuté ensuite.

$$\widetilde{\boldsymbol{p_{ij}}} \sim \mathcal{D}irichlet(\theta_{ij}\boldsymbol{p_i}) \tag{1.4}$$

et :

$$\pi\left(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) = \frac{1}{Z\left(\theta_{ij}\boldsymbol{p_i}\right)} \prod_{k=1}^{K_i} \widetilde{p_{ijk}}^{\theta_{ij}p_{ik}-1}$$

Avec $\theta_{ij} = 1/F_{ST}^{ij} - 1$. Si $\boldsymbol{v} = (v_1, \ldots, v_K)$, $Z\left(\boldsymbol{v}\right)$ est une constante définie par :

$$Z\left(\boldsymbol{v}\right) = \int \cdots \int t_1^{v_1-1} \ldots t_K^{v_K-1} dt_1 \ldots dt_K \tag{1.5}$$

$$= \frac{\prod_{k=1}^{K} \Gamma\left(v_k\right)}{\Gamma\left(\sum_{k=1}^{K} v_k\right)} \tag{1.6}$$

1.4 Un modèle bayésien

Plusieurs arguments plaident en la faveur d'une modélisation bayésienne de la différenciation génétique. Tout d'abord, la structure génétique des populations semble se prêter naturellement à un modèle hiérarchique. En effet, comme expliqué précédemment, la plupart des populations se subdivisent géographiquement de manière hiérarchique. Ensuite, on a montré que l'on pouvait calculer dans des modèles simples la distribution attendue des fréquences alléliques en fonction de la différenciation génétique subie. De plus, il se trouve que cette distribution est l'à priori conjugué de la vraisemblance obtenue en échantillonnant des génotypes dans les populations, ce qui permet de simplifier les estimations des distributions à postériori. Enfin, un modèle bayésien est suffisamment flexible pour inclure d'autres variables ou tester des modèles alternatifs en tenant compte de toute l'information disponible dans les données et de leur incertitude (voir ci-dessous).

1.4.1 La vraisemblance

On a montré que l'on pouvait obtenir la distribution des fréquences alléliques dans chaque population. Pour pouvoir l'intégrer dans un modèle bayésien, on doit obtenir la vraisemblance en fonction de ces fréquences. De cette façon, la distribution de Dirichlet de l'équation 1.4 pourra être utilisée comme l'à priori

sur les fréquences alléliques. Les données considérées sont différentes selon que les marqueurs utilisés sont codominants ou dominants et conduisent à deux fonctions de vraisemblances distinctes.

Marqueurs codominants

On peut directement obtenir la vraisemblance de données multilocus à partir de marqueurs codominants. On considère que dans chaque population j, on a prélevé n_{ij} allèles au locus i (par exemple provenant de $n_{ij}/2$ individus diploïdes). Le nombre d'allèles de type k compté parmi les K_i différents allèles observés est noté a_{ijk} et on a $n_{ij} = \sum_k a_{ijk}$. L'ensemble des données peut être représenté par une matrice $\boldsymbol{A} = \{\boldsymbol{a_{ij}}\}$, avec $\boldsymbol{a_{ij}} = \{a_{ij1}, a_{ij2}, \ldots, a_{ijK_i}\}$ le vecteur des allèles comptés au locus i dans la population j. On peux considérer le vecteur $\boldsymbol{a_{ij}}$ comme un tirage aléatoire dans la population de fréquence $\widetilde{\boldsymbol{p_{ij}}}$ et ainsi il peut être décrit par une loi multinomiale (Holsinger 1999) :

$$\boldsymbol{a_{ij}} \sim \mathcal{M}ultinomial(n_{ij}; \widetilde{\boldsymbol{p_{ij}}}) \tag{1.7}$$

$$P\left(\boldsymbol{a_{ij}}|\widetilde{\boldsymbol{p_{ij}}}\right) = \binom{n_{ij}}{a_{ij1}a_{ij2}\cdots a_{ijK_i}} \prod_{k=1}^{K_i} \widetilde{p_{ijk}}^{a_{ijk}}$$

La loi multinomiale est une généralisation de la loi binomiale que l'on retrouve dans le cas de marqueurs bi-alléliques. Il est intéressant de noter ici que la distribution de Dirichlet de l'équation 1.4 est l'à priori conjugué de la distribution multinomiale de l'équation 1.7. Cela signifie que l'on peut calculer de manière exacte la distribution marginale de $\boldsymbol{a_{ij}}$ et ainsi éliminer du modèle les paramètres $\widetilde{\boldsymbol{p_{ij}}}$:

$$P\left(\boldsymbol{a_{ij}}|\boldsymbol{p_i}, \theta_{ij}\right) = \int \cdots \int \pi\left(\boldsymbol{a_{ij}}, \widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) d\widetilde{p_{ij1}} \cdots d\widetilde{p_{ijK_i}}$$

On a aussi :

$$\begin{aligned} \pi\left(\boldsymbol{a_{ij}}, \widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) &= P\left(\boldsymbol{a_{ij}}|\widetilde{\boldsymbol{p_{ij}}}, \boldsymbol{p_i}, \theta_{ij}\right) \pi\left(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) \\ &= P\left(\boldsymbol{a_{ij}}|\widetilde{\boldsymbol{p_{ij}}}\right) \pi\left(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) \end{aligned}$$

Avec les densités des distributions de Dirichlet et multinomiale, on obtient :

$$
\begin{aligned}
P\left(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_{ij}\right) &= \frac{n_{ij}!\Gamma\left(\sum_{k=1}^{K_i}\theta_{ij}p_{ik}\right)}{\prod_{k=1}^{K_i}a_{ijk}!\prod_{k=1}^{K_i}\Gamma\left(\theta_{ij}p_{ik}\right)}\int\cdots\int\prod_{k=1}^{K_i}\widetilde{p_{ijk}}^{\theta_{ij}p_{ik}+a_{ijk}-1}d\widetilde{p_{ij1}}\cdots d\widetilde{p_{ijK_i}} \\
&= \frac{n_{ij}!\Gamma\left(\theta_{ij}\right)}{\prod_{k=1}^{K_i}a_{ijk}!\Gamma\left(\theta_{ij}p_{ik}\right)}\frac{\prod_{k=1}^{K_i}\Gamma\left(\theta_{ij}p_{ik}+a_{ijk}\right)}{\Gamma\left(\sum_{k=1}^{K_i}\theta_{ij}p_{ik}+a_{ijk}\right)} \text{ avec l'équation 1.5}
\end{aligned}
$$

En utilisant finalement $n_{ij} = \sum_k a_{ijk}$ et $\sum_k p_{ik} = 1$ on obtient l'équation 1.8 (Balding 2003) qui correspond à la densité de la distribution Dirichlet-multinomiale.

Propriété 1.2 *La vraisemblance du vecteur du nombre d'allèles comptés $\boldsymbol{a_{ij}}$ dans une population j à un locus i s'exprime directement en fonction du vecteur des fréquences alléliques $\boldsymbol{p_i}$ des immigrants tirés aléatoirement parmi un grand nombre de populations et du coefficient $F_{ST}^{ij} = 1/(1+\theta_{ij})$:*

$$
P\left(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_{ij}\right) = \frac{n_{ij}!\Gamma\left(\theta_{ij}\right)}{\Gamma\left(n_{ij}+\theta_{ij}\right)}\prod_{k=1}^{K_i}\frac{\Gamma\left(a_{ijk}+\theta_{ij}p_{ik}\right)}{a_{ijk}!\,\Gamma\left(\theta_{ij}p_{ik}\right)} \tag{1.8}
$$

Cette distribution est aussi la généralisation de la distribution beta-binomiale obtenue dans le cas d'un marqueur bi-allélique. Si on suppose que les allèles sont échantillonnés indépendamment dans chaque population et que l'on ignore le déséquilibre d'association gamétique, on peut obtenir la vraisemblance totale :

$$
L\left(\boldsymbol{p},\boldsymbol{\theta}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J}P\left(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_{ij}\right)
$$

Nous utiliserons cette formulation dans les chapitres 2 et 4.

Marqueurs dominants

Les marqueurs dominants ne permettent pas d'estimer directement les fréquences alléliques à travers une simple distribution multinomiale comme pour les marqueurs codominants. Ainsi les AFLP par exemple fournissent des données binaires : pour chaque individu l'information est la « présence de bande » ou l'« absence de bande » que l'on peut interpréter comme un phénotype. Une solution simple souvent utilisée est de supposer l'équilibre

de Hardy-Weinberg mais cela implique de fortes hypothèses et en particulier l'absence de consanguinité (Lynch et Milligan 1994, Zhivotovsky 1999, Hill et Weir 2004). Une seule méthode permettant d'inclure l'estimation du coefficient de consanguinité F_{IS} a été proposée par Holsinger *et al.* (2002). Dans leur modèle, Holsinger *et al.* (2002) ont considéré les coefficients F_{IS} et F_{ST} comme communs à tous les loci et à toutes les populations. Ici nous présentons une généralisation de cette approche.

Avec des marqueurs dominants, les données $\mathbf{N}$ représentent le nombre de phénotypes observés plutôt que le nombre d'allèles. On note $n_{[A1],ij}$ et $n_{[A2],ij}$ le nombre de phénotypes $[A1]$ et $[A2]$ au locus i pour la population j. L'ensemble des données est alors la matrice $\mathbf{N} = \left\{n_{[A1],ij}, n_{[A2],ij}\right\}$ et la taille de l'échantillon au locus i pour la population j est $n_{ij} = n_{[A1],ij} + n_{[A2],ij}$. On peut considérer que le nombre de phénotypes $n_{[A1],ij}$ suit une loi binomiale de paramètres $g_{[A1],ij}$ et n_{ij}, où $g_{[A1],ij}$ est la fréquence du phénotype $[A1]$ au locus i dans la population j :

$$n_{[A1],ij} \sim \mathcal{B}inomial\left(g_{[A1],ij}, n_{ij}\right) \tag{1.9}$$

Même si la distribution beta est l'à priori conjugué de la loi binomiale, il n'est pas possible ici de calculer exactement la distribution marginale comme dans le cas des marqueurs codominants. En effet, le paramètre de la loi binomiale est ici la fréquence phénotypique au lieu d'être la fréquence allélique dans le cas de marqueurs codominants. Ainsi, le modèle bayésien ne peut pas se simplifier et l'équation 1.4 devra être intégrée à travers un à priori dans un second niveau hiérarchique. Comme précédemment si on suppose les échantillons indépendants entre les loci et les populations, on obtient la vraisemblance totale :

$$L\left(\widetilde{\mathbf{p}}, \mathbf{F_{IS}}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} P\left(n_{[A1],ij} | g_{[A1],ij}\right)$$

La fréquence phénotypique $g_{[A1],ij}$ peut être reliée à la fréquence correspondante p_{ij} de l'allèle $A1$ et au taux de consanguinité F_{IS}^{j} de la population j en utilisant les équations suivantes :

$$\begin{aligned} g_{[A1],ij} &= \widetilde{p_{ij}}^2(1 - F_{IS}^{j}) + F_{IS}^{j}\widetilde{p_{ij}} + \left(1 - F_{IS}^{j}\right) 2\widetilde{p_{ij}}\left(1 - \widetilde{p_{ij}}\right) \\ g_{[A2],ij} &= \left(1 - F_{IS}^{j}\right)\left(1 - \widetilde{p_{ij}}\right)^2 + F_{IS}^{j}\left(1 - \widetilde{p_{ij}}\right) \\ &= 1 - g_{[A1],ij} \end{aligned}$$

De la même façon, l'à priori complet des fréquences alléliques sera donné par :

$$\pi\left(\widetilde{\boldsymbol{p}}|\boldsymbol{p},\boldsymbol{\theta}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\pi\left(\widetilde{p_{ij}}|p_i,\theta_{ij}\right)$$

Avec p_i la fréquence de l'allèle $A1$ parmi les immigrants. Nous utiliserons ce modèle dans les chapitres 3 et 4 en lui apportant quelques modifications pour tenir compte du biais de recrutement qui affecte les marqueurs AFLP.

1.4.2 A priori et à postériori

Dans le cas des marqueurs codominants, la distribution à postériori est donnée par l'équation suivante :

$$\pi(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{N})\propto L(\boldsymbol{p},\boldsymbol{\theta})\pi(\boldsymbol{p})\pi(\boldsymbol{\theta})$$

Pour les marqueurs dominants, comme expliqué précédemment, nous avons un niveau hiérarchique supplémentaire et les coefficients de consanguinité $\boldsymbol{F_{IS}}$ de chaque population en plus :

$$\pi\left(\boldsymbol{p},\boldsymbol{F_{IS}},\widetilde{\boldsymbol{p}},\boldsymbol{\theta}|\boldsymbol{N}\right)\propto L\left(\widetilde{\boldsymbol{p}},\boldsymbol{F_{IS}}\right)\pi\left(\widetilde{\boldsymbol{p}}|\boldsymbol{p},\boldsymbol{\theta}\right)\pi\left(\boldsymbol{F_{IS}}\right)\pi\left(\boldsymbol{p}\right)\pi\left(\boldsymbol{\theta}\right)$$

Il nous reste à choisir les à priori des fréquences alléliques $\boldsymbol{p}$ de la population d'immigrants, des coefficients de différenciation génétique $\boldsymbol{\theta}$ et des coefficients de consanguinité $\boldsymbol{F_{IS}}$ pour les marqueurs dominants. Pour les coefficients F_{IS} et F_{ST}, nous utilisons un à priori uniforme dans les cas où aucune information supplémentaire n'est disponible. Pour les coefficients F_{ST}, les chapitres 2 et 4 proposent deux formes différentes pour cet à priori qui répondent à deux questions biologiques différentes (voir ci-dessous).

Pour les fréquences alléliques de la population d'immigrants $\boldsymbol{p}$, la solution généralement choisie est de supposer un à priori dit « non-informatif » avec une distribution de Dirichlet pour chaque locus i dont les K_i paramètres sont égaux à 1[4]. Dans le cas de marqueurs codominants, nous avons observé que les données contenaient suffisamment d'information sur ces fréquences pour que cet à priori n'ait qu'une influence très limitée sur les estimations obtenues (chapitre 4). Au contraire, nous avons montré (chapitre 3) que cet à priori influençait fortement les estimations pour les marqueurs dominants lorsque la distribution des fréquences $\boldsymbol{p}$ n'était pas uniforme.

[4]Cette distribution est une distribution uniforme dans le cas de marqueurs bi-alléliques.

Il n'y a aucune raison de penser que cette distribution est uniforme. A titre d'exemple, on peut obtenir la distribution à l'équilibre des fréquences alléliques d'une population isolée soumise à de la mutation dans des cas simples. L'équation 1.2 permet d'obtenir cette distribution dans le cas d'une mutation réversible à deux allèles. Si on note u le taux $A \rightarrow a$ et v le taux $a \rightarrow A$, on a $\Delta p = -up + v(1-p)$ et on obtient la distribution suivante si $4Nu \ll 1$ et $4Nv \ll 1$:

$$\varphi(p) = \frac{\Gamma(4Nu + 4Nv)}{\Gamma(4Nu)\Gamma(4Nv)} p^{4Nv-1}(1-p)^{4Nu-1}$$

Cette distribution est une distribution beta avec les paramètres $4Nv$ et $4Nu$. Si $u = v$ cette distribution est symétrique (voir figure 1.2). On notera qu'il n'est pas possible d'obtenir le même type de résultat pour des cas plus complexes comme par exemple le modèle de mutation « Stepwise Mutation Model » (SMM) utilsé pour les marqueurs microsatellites (Moran 1975, Graham *et al.* 2000).

Ainsi, dans le cas des marqueurs dominants, nous proposons d'ajouter un à priori plus général qu'une distribution uniforme avec une distribution beta symétrique $beta(a, a)$ où le paramètre a est estimé dans le modèle. Le modèle comprend alors un niveau hiérarchique supplémentaire et la distribution à postériori s'écrit :

$$\pi\left(\boldsymbol{p}, \boldsymbol{F_{IS}}, \widetilde{\boldsymbol{p}}, \boldsymbol{\theta}, a | \boldsymbol{N}\right) \propto L\left(\widetilde{\boldsymbol{p}}, \boldsymbol{F_{IS}}\right) \pi\left(\widetilde{\boldsymbol{p}} | \boldsymbol{p}, \boldsymbol{\theta}\right) \pi\left(\boldsymbol{F_{IS}}\right) \pi\left(\boldsymbol{p} | a\right) \pi\left(\boldsymbol{\theta}\right) \pi\left(a\right)$$

Le paramètre a est estimé en utilisant un à priori positif log-normal : $a \sim log\mathcal{N}ormal(0, 1)$. Ce modèle est utilisé dans les chapitres 3 et 4.

1.4.3 Les extensions développées

Le modèle Dirichlet-multinomial présenté ici est très général et nous proposons maintenant trois extensions de ce modèle, qui seront développées dans les chapitres qui suivent. La première permet d'intégrer les facteurs environnementaux dans les estimations et en même temps de tester différentes hypothèses sur leur influence dans la différenciation génétique. La deuxième permet de l'appliquer aux marqueurs dominants en prenant en compte le biais de recrutement affectant les AFLP en particulier. Enfin la troisième propose de séparer les effets neutres, comme la dérive ou la migration, de l'effet de la sélection naturelle qui n'affecte que certains loci et ainsi de les identifier.

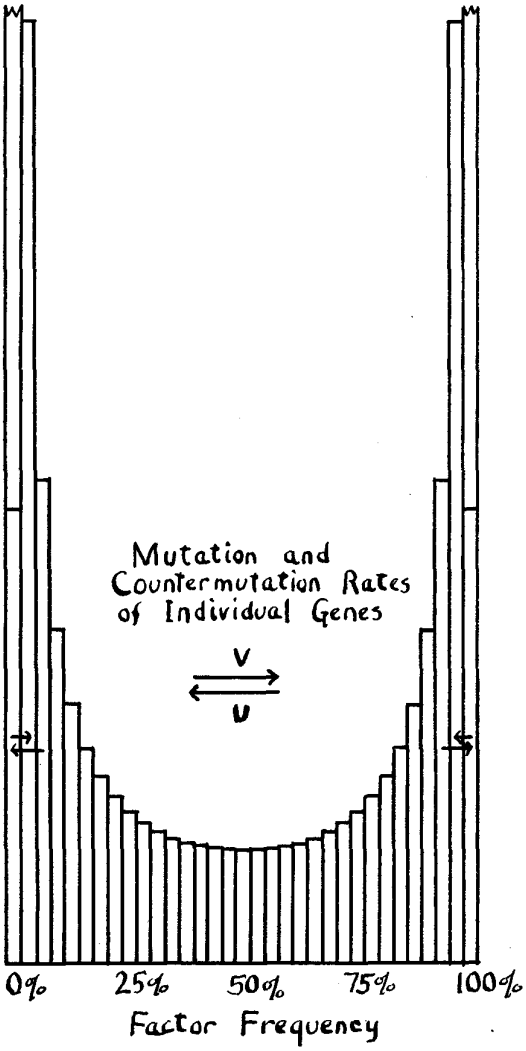

FIGURE 5.—Distribution of gene frequencies (or probability array of gene) where equilibrium with mutation has been attained. Population so small that the terms 4Nu and 4Nv are both much smaller than 1. $y = Cq^{4Nv-1}(1-q)^{4Nu-1}$, approximately $\frac{q^{-1}(1-q)^{-1}}{2 \log 3.6N}$.

FIG. 1.2 – *Distribution des fréquences alléliques dans une population isolée soumise à une mutation réversible symétrique, d'après* WRIGHT *(1931)*

Les facteurs environnementaux

On se place tout d'abord dans le cas classique où l'on souhaite décrire la structure génétique d'un ensemble de populations. On est donc amené à estimer un unique coefficient F_{ST}^{j} dans chaque population j en combinant l'information disponible à travers les différents loci. Un des problèmes fondamental en génétique des populations est alors d'identifier les facteurs environnementaux responsables de cette structure génétique. La méthode la plus couramment utilisée pour répondre à cette question est le test multivarié de Mantel (SMOUSE *et al.* 1986) qui est basé sur le calcul de distances génétiques et environnementales entre paires de populations. Cette méthode présente différents inconvénients développés au chapitre 2 et en particulier le fait de ne pouvoir

prendre en compte que des variables qui s'expriment comme une distance entre chaque paire de populations. Le modèle proposé ici permet au contraire d'estimer un coefficient F_{ST} spécifique à chaque population. Ainsi, si on veut tester l'effet de F facteurs environnementaux $G^{(f)} = (G_1^{(f)}, \ldots, G_J^{(f)})$ mesurés dans chaque population, les coefficients F_{ST} étant bornés entre 0 et 1, on peut utiliser le modèle de régression logistique :

$$\ln\left(\frac{F_{ST}^j}{1 - F_{ST}^j}\right) = \alpha_0 + \sum_{f=1}^{F} \alpha_f G_j^{(f)} \tag{1.10}$$

La valeur absolue et le signe de chaque paramètre α_f indiquent l'intensité et la direction de l'influence du facteur environnemental f. Cependant, pour effectuer cette régression, il faut choisir une estimation ponctuelle pour les coefficients F_{ST}^j à partir de leur distribution à postériori. L'inconvénient principal de cette approche est de résumer toute l'information génétique dans ces seules statistiques sommaires. De cette façon la régression ne tiendra pas compte de l'incertitude sur les estimations des coefficients F_{ST}^j. De même, cette incertitude ne sera pas non plus prise en compte si l'on utilise la régression pour effectuer un choix de modèle afin de déterminer si chacun des facteurs participe à la différenciation génétique. Ainsi, il est intéressant de pouvoir intégrer directement cette possibilité dans le modèle bayésien.

Gaggiotti *et al.* (2004) ont proposé une méthode que l'on peut facilement généraliser pour permettre d'effectuer ce type de régression au sein d'un modèle bayésien. Pour cela, on considère un à priori log-Normal sur $F_{ST}^j/(1 - F_{ST}^j)$, où la moyenne est une combinaison linéaire des variables avec une variance commune :

$$\ln\left(\frac{F_{ST}^j}{1 - F_{ST}^j}\right) \sim \mathcal{N}(\mu_j, \sigma^2) \tag{1.11}$$

avec

$$\mu_j = \alpha_0 + \sum_{f=1}^{F} \alpha_f G_j^{(f)} \tag{1.12}$$

Cette formulation permet d'estimer dans le même modèle les coefficients F_{ST}, les paramètres de la régression $\boldsymbol{\alpha}$, et l'adéquation entre les valeurs estimées F_{ST} et les valeurs prédîtes par la régression à travers le paramètre σ. De plus, il est possible de déterminer en même temps si la différenciation génétique dépend de chaque facteur proposé. On considère tous les modèles alternatifs en adoptant différentes formes pour l'équation 1.12. Par exemple,

le modèle le plus simple est $\mu_j = \alpha_0$ où aucun facteur n'a d'effet sur la différenciation génétique. Le cas $\mu_j = \alpha_0 + \alpha_2 G_j^{(2)}$ correspond au modèle où seul le facteur $G^{(2)}$ a un effet. La probabilité de chacun des différents modèles possibles est estimée dans le modèle en utilisant un algorithme de Monte Carlo par Chaine de Markov à sauts réversibles (RJMCMC, Green 1995). Le chapitre 2 est consacré à l'étude de cette nouvelle méthode.

Le biais de recrutement des AFLP

Comme expliqué précédemment, un grand nombre de marqueurs moléculaires différents existent pour étudier la structure génétique. Parmi ceux là, les marqueurs dominants sont devenus très populaires et en particulier les AFLP dans le domaine de l'écologie. Comme expliqué précédemment, ces marqueurs ne permettent pas d'estimer directement les fréquences alléliques à partir des données. Le modèle Dirichlet-multinomial a été généralisé aux marqueurs dominants par Holsinger *et al.* (2002) et cela nécessite en particulier d'estimer en même temps le coefficient de consanguinité F_{IS}^j de chaque population j. Cependant, les auteurs ont remarqué par la suite que la méthode fournissait sur certains jeux de données des estimations invraisemblables pour F_{IS} (voir le manuel de Hickory, Holsinger et Lewis 2002).

Nous montrons dans le chapitre 3 que le biais de recrutement des marqueurs AFLP peut être à l'origine des mauvaises estimations observées. Le processus de découverte des marqueurs AFLP est complexe : tout d'abord si tous les individus sont homozygotes pour l'allèle récessif, aucune bande n'est observée et le marqueur ne sera pas identifiable. De plus, les marqueurs ne sont généralement pas choisis aléatoirement et on préfère sélectionner les marqueurs « polymorphes », c'est à dire pour lesquels certains individus possèdent la bande et d'autres pas. Enfin, il est souvent difficile de distinguer certaines bandes d'artéfacts, et on choisit alors de ne retenir que les marqueurs pour lesquels la fréquence de la bande est comprise entre deux bornes (par exemple 1% et 99%). Ces différentes raisons font que la distribution binomiale de l'équation 1.9 utilisée dans chaque population n'est pas correcte.

Nicholson *et al.* (2002) et Nielsen *et al.* (2004) ont proposé de modifier la fonction de vraisemblance pour prendre en compte le biais de recrutement des marqueurs SNP. Cependant cette approche ne correspond pas au processus de découverte des AFLP décrit ici, et il n'est pas possible d'obtenir une fonction de vraisemblance simple. Nous proposons une méthode alternative basée sur un algorithme de type « Approximate Bayesian Computation » (ABC). Cet algorithme permet d'estimer les distributions à postériori sans

avoir besoin de calculer la fonction de vraisemblance. A la place, il est nécessaire de pouvoir simuler des données selon le modèle. Ici on peut simuler des données en incluant le biais de recrutement avec l'algorithme suivant :

Algorithme 2.

1. Simuler $a \sim logNormal(0, 1)$.
2. Pour chaque population j dans $1..J$:
   (a) Simuler $F_{IS}^{j} \sim \mathcal{U}[0, 1]$.
   (b) Simuler $F_{ST}^{j} \sim \mathcal{U}[0, 1]$ et calculer $\theta_j = 1/F_{ST}^{j} - 1$.
3. Pour chaque locus i dans $1..I$:
   (a) Simuler les fréquences alléliques de la population ancestrale : $p_i \sim Beta(a, a)$.
   (b) Pour chaque population j dans $1..J$:
      i. Simuler les fréquences alléliques j : $\widetilde{p_{ij}} \sim Beta(\theta p_i, \theta(1 - p_i))$.
      ii. Calculer les fréquences phénotypiques $g_{[A1],ij} = \widetilde{p_{ij}}^2(1 - F_{IS}^{j}) + F_{IS}^{j}\widetilde{p_{ij}} + \left(1 - F_{IS}^{j}\right) 2\widetilde{p_{ij}}\left(1 - \widetilde{p_{ij}}\right)$.
   (c) Pour chaque population j dans $1..J$:
      i. Simuler les nombre de bandes $n_{[A1],ij} \sim Binomial(n_{[A1],ij}, g_{[A1],ij})$.
   (d) Si la condition de recrutement n'est pas vérifiée, retourner à 3a.

La méthode proposée par Nicholson *et al.* (2002) était basée sur l'utilisation d'une loi binomiale tronquée et cela revenait à considérer que le saut conditionnel de l'étape 3d s'effectuait vers l'étape 3c. Cela implique un processus de découverte particulier qui n'est pas réaliste : si un locus ne vérifie pas la condition de recrutement à l'étape 3d, il sera éliminé et le marqueur suivant aura *exactement les même* paramètres (fréquence allélique ancestrale et dans chaque population) que celui qui vient d'être éliminé.

Nous avons montré que l'utilisation de l'algorithme ABC dans le modèle Dirichlet-multinomial permet de tenir compte du biais de recrutement des marqueurs AFLP et d'obtenir des estimations non biaisées. Ce modèle sera développé au chapitre 3.

La sélection

Beaucoup de méthodes différentes utilisent des données de criblage génomique (grand nombre de loci) pour détecter des loci influencés par la sélection naturelle. La forte baisse des couts et l'augmentation de la vitesse pour obtenir ce type de données, y compris pour des organismes non modèles, a fait resurgir une ancienne idée de Lewontin et Krakauer (1973), et émerger de nombreuses méthodes qui en sont dérivées (Beaumont 2005). L'idée générale est d'identifier les loci qui présentent un coefficient F_{ST} avec une valeur différente de ce que l'on attend dans un modèle neutre. Plus précisément, un locus sous l'influence de sélection directionnelle présentera une différenciation génétique plus élevée que les loci neutres, et au contraire, un locus sous l'influence d'une sélection balancée présentera une différenciation génétique plus faible.

Dans un modèle en iles neutre, θ_{ij} correspond au nombre d'immigrants entrant dans la population j et doit donc être constant pour tous les loci i, comme proposé dans les deux méthodes précédentes. Ainsi une valeur singulière de θ_{ij} peut indiquer que le locus i a subi un effet de la sélection. On peut ici aussi généraliser la méthode et intégrer cette possibilité à travers $\boldsymbol{\theta}$. Beaumont et Balding (2004) ont proposé de séparer les effets affectant tous les loci d'une population (dérive, migration etc.) de ceux spécifiques à un locus (sélection) en remplaçant θ_{ij} par :

$$\ln\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right) = \alpha_i + \beta_j$$

Nous proposons comme pour les facteurs environnementaux de tester différents modèles alternatifs. En particulier, on considère pour chaque locus i l'hypothèse de neutralité ($\alpha_i = 0$) dans laquelle seul le coefficient β_j est présent dans l'équation précédente. La méthode estime alors la probabilité à postériori que chaque locus soit soumis à la sélection en plus de la distribution à postériori des α_i. Ce modèle sera développé en détail dans le chapitre 4.

2

Identifier les facteurs environnementaux qui déterminent la structure génétique des populations

Parmi les différentes forces évolutives exposées en introduction, l'environnement dans lequel évolue une population va particulièrement influencer les migrations (barrières géographiques, ressources disponibles etc.) et la sélection (adaptation locale). Ainsi des conditions environnementales distinctes entre populations vont conduire à différents niveaux de variation génétique. Identifier les facteurs environnementaux qui permettent d'expliquer la nature de la différenciation génétique observée est un problème fondamental en génétique des populations. Il trouve de nombreuses applications comme, entre autres, retracer l'histoire des populations, conserver la biodiversité, identifier des gènes associés à des maladies ou à une résistance à une maladie chez l'homme ou d'autres espèces ayant une valeur économique importante. Dans ce chapitre, nous proposons une nouvelle méthode bayésienne hiérarchique pour identifier les facteurs environnementaux qui déterminent la structure génétique des populations.

« *La variation, quelle qu'en soit la cause, est le phénomène essentiel de l'évolution.* »

William Bateson

2.1 Problématique et démarche scientifique

La méthode la plus couramment utilisée pour répondre à cette question est le test multivarié de Mantel (Smouse *et al.* 1986) qui est basé sur le calcul de distances génétiques et environnementales entre paires de populations. Plus précisément, la méthode consiste d'abord à calculer une matrice de distance génétique entre les populations avec la statistique sommaire F_{ST} présentée en introduction. Ensuite, une technique de randomisation permet d'estimer si la corrélation observée entre les matrices de distances génétiques et environnementales est uniquement due au hasard ou si elle est significative. Cependant, différentes critiques ont été formulées sur ce test. D'abord, la validité de la méthode permettant d'estimer des intervalles de confiance (jackknife) lors de l'analyse de plus d'un facteur environnemental a été remise en cause (Raufaste et Rousset 2001, Rousset 2002). De même, le test a pu être généralisé en utilisant différentes techniques de randomisation, mais le choix de la technique la plus appropriée doit être fait à la main (Legendre 2000). Ensuite, toute l'information sur la distance génétique entre deux populations est résumée par la statistique sommaire F_{ST}, qui ne prend pas en compte toute l'information disponible. Comme expliqué en introduction, cela est particulièrement problématique pour les petits jeux de données. Enfin le principe même du test fait qu'il est uniquement possible d'étudier l'effet de variables qui s'expriment sous forme de distances entre paires de populations. Par exemple, l'effet de la taille locale des populations ne peut pas être représenté par paire puisque son influence sur la dérive génétique est déterminée par la taille efficace N_e. Ainsi, deux populations ayant une même faible taille efficace pourront être séparées par une distance génétique élevée.

La structure génétique est la conséquence de la dérive génétique que chaque population locale a subi. Elle peut être due à sa taille efficace, à son niveau d'isolement géographique ou écologique, ou à toute condition locale ayant entrainé de la sélection. Pour cette raison, nous proposons de baser les estimations sur des paramètres et des variables qui sont spécifiques à chaque population locale, plutôt qu'exprimés par paires comme dans le test de Mantel. Pour cela, nous proposons d'utiliser le modèle Dirichlet-multinomial détaillé en introduction (Balding 2003) pour estimer un coefficient de différenciation génétique (F_{ST}) dans chaque population locale. A partir de ce modèle, nous avons développé une nouvelle méthode bayésienne hiérarchique complémentaire au test de Mantel. L'innovation principale est d'introduire les données non génétiques (environnementales) via la distribution à priori des coefficients F_{ST} en utilisant un modèle linéaire généralisé (Gaggiotti *et al.*

2002). Nous considérons différents modèles alternatifs, chacun incluant différentes combinaisons des variables environnementales, et estimons leur probabilité à postériori.

2.2 Contribution scientifique : article A

Foll, M., et O. Gaggiotti, 2006 Identifying the environmental factors that determine the genetic structure of populations. Genetics **174**: 875–891

2.2.1 Matériels et méthodes

Modèle

Le point de départ de la méthode est le modèle Dirichlet-multinomial présenté en introduction permettant d'estimer un coefficient de différenciation génétique (F_{ST}^j) dans chaque population locale j. Nous voulons déterminer l'influence de facteurs biotiques ou abiotiques sur les coefficients F_{ST}. On considère ici deux facteurs [1] avec $G_j^{(1)}$ et $G_j^{(2)}$ représentant respectivement les valeurs observées du premier et du second facteur environnemental pour la population j. Gaggiotti *et al.* (2004) ont proposé un cadre général présenté en introduction pour inclure cette régression directement dans le modèle bayesien. Dans notre cas, on considère un à priori log-Normal sur $F_{ST}^j/(1-F_{ST}^j)$, où la moyenne est une combinaison linéaire des variables avec une variance commune :

$$\ln\left(\frac{F_{ST}^j}{1-F_{ST}^j}\right) \sim \mathcal{N}(\mu_j, \sigma^2) \tag{2.1}$$

avec

$$\mu_j = \alpha_0 + \alpha_1 G_j^{(1)} + \alpha_2 G_j^{(2)} + \alpha_3 G_j^{(1)} G_j^{(2)} \tag{2.2}$$

Cette formulation permet d'estimer dans le même modèle les coefficients F_{ST}, les paramètres de la régression $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$, et l'adéquation entre les valeurs estimées F_{ST} et les valeurs prédîtes par la régression à travers le paramètre σ. La valeur absolue et le signe de chaque paramètre α_i indiquent l'intensité et la direction de l'influence du facteur environnemental correspondant. Le paramètre α_3 correspond à l'interaction des deux facteurs. La distribution à postériori de tous les paramètres est estimée à l'aide d'un algorithme de Monte Carlo par Chaine de Markov (MCMC, Hastings 1970).

[1] La méthode a été généralisée par la suite à un nombre quelconque de facteurs.

Au delà de la simple estimation des paramètres de la régression, nous souhaitons déterminer si la différenciation génétique dépend vraiment de chaque facteur proposé. Pour cela, on considère tous les modèles alternatifs en adoptant différentes formes pour l'équation 2.2. Par exemple, le modèle le plus simple est $\mu_j = \alpha_0$ où aucun facteur n'a d'effet sur la différenciation génétique. Le cas $\mu_j = \alpha_0 + \alpha_1 G_j^{(2)}$ correspond au modèle où seul le facteur $G^{(2)}$ a un effet. La probabilité de chacun des différents modèles possibles est estimée dans le modèle en utilisant une méthode de Monte Carlo par Chaine de Markov à sauts réversibles (RJMCMC, Green 1995).

Méthode de validation

Nous avons évalué les performances de la méthode en utilisant trois différentes approches pour générer des données simulées. La première approche utilise le même modèle statistique que la méthode proposée et permet d'étudier l'effet de la qualité des données (nombre de marqueurs, taille de l'échantillon etc.) et différents scénarios biologiques (nombre de populations, facteurs environnementaux considérés etc.) sur la précision des résultats obtenus. Les deux autres approches ont été choisies pour permettre d'explorer les performances de la méthode pour des scénarios qui s'écartent du modèle en ile. Nous avons utilisé EasyPop (Balloux 2001) pour simuler un scénario d'isolement par la distance, et SPLATCHE (Currat *et al.* 2004) pour un scénario d'expansion spatiale des populations.

Jeux de données analysés

La méthode peut être appliquée pour tester l'effet de différents facteurs biotiques ou abiotiques et pour étudier les processus d'expansion des populations. Deux jeux de données ont été analysés pour illustrer les différentes possibilités d'application : les données publiées par Petit *et al.* (1998) de l'arganier du Maroc (*Argania spinosa*) et les données de génétique humaine du « Human Genome Diversity Cell Line Panel – Centre d'Etude du Polymorphisme Humain » (HGDP-CEPH) présentées par Cann *et al.* (2002) et disponibles sur `http://www.cephb.fr/HGDP-CEPH-Panel/`.

– Arganier (*Argania spinosa*) : Un des objectifs de la génétique de la conservation est d'identifier les populations cibles à conserver. Cela nécessite de caractériser chaque unité en termes de diversité génétique et de différenciation par rapport aux autres, et de comprendre les facteurs qui en sont responsables. Les données de l'arganier du Maroc étudiées par Petit *et al.* (1998) sont un bon exemple de ce type d'analyse. Nous avons

utilisé ce jeu de données pour déterminer si le degré d'isolement génétique est influencé par l'isolement géographique et/ou un autre facteur, ici, l'altitude. Les données sont constituées de 12 populations (voir figure 2.1) de 20 à 50 individus, pour lesquelles 12 loci polymorphes ont été développés. Le degré d'isolement géographique de chaque population j est mesuré par leur connectivité S_j (Moilanen et Nieminen 2002) : $S_j = \sum_{\substack{i=1 \\ i \neq j}}^{J} exp\left(-\beta d_{ij}\right)$, avec d_{ij} la distance entre les populations i et j, et β mesurant l'effet de la distance sur la probabilité de migration.

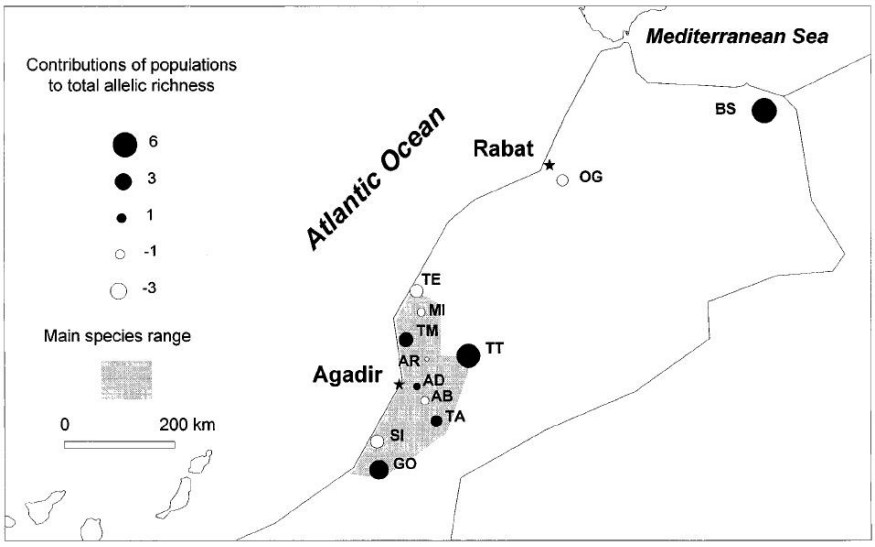

Fig. 2.1 – *Carte de la répartition des 12 populations d'arganier du Maroc, d'après* Petit et al. *(1998).*

– Humain : Une des théories les plus répandues sur l'origine des hommes modernes est le modèle dit de l'origine africaine récente (RAO). Ce modèle postule que notre espèce a migré depuis une population d'Afrique de l'Est qui a colonisé tout le globe. Ici nous utilisons les données du HGDP-CEPH pour illustrer comment notre méthode peut être utilisée pour retracer l'expansion des populations. Le jeu de données contient 1056 individus, répartis dans 51 populations (voir figure 2.2) avec 377 marqueurs microsatellites développés [2]. Deux analyses ont été effectuées. La première considère l'effet de la distance géographique depuis l'Afrique de l'Est le long des routes de colonisation (Prugnolle *et al.* 2005), la seconde considère un modèle avec deux facteurs : la latitude et la longitude.

[2] 835 marqueurs microsatellites sont maintenant disponibles (voir chapitre 4). L'étude de sensibilité a montré que 377 marqueurs étaient déjà très largement suffisants pour obtenir des résultats précis.

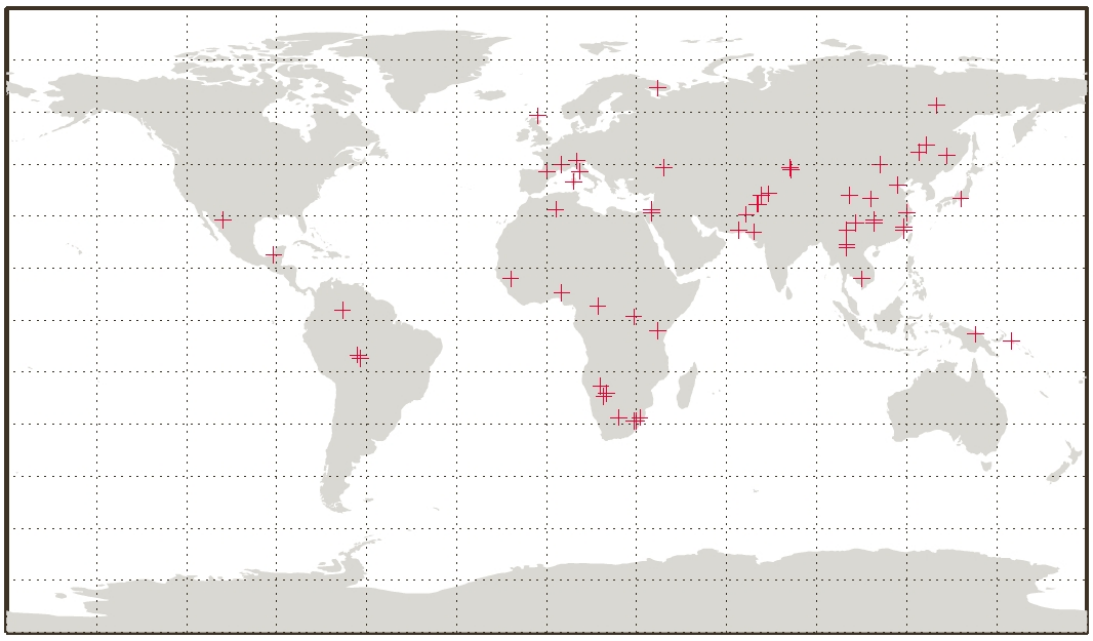

FIG. 2.2 – *Carte de la répartition des 51 populations humaines du HGDP-CEPH, d'après* CANN et al. *(2002).*

2.2.2 Résultats

Les résultats de l'étude de sensibilité montrent que la méthode peut fournir des estimations précises des paramètres et identifier le bon modèle, même si les données sont de qualité moyenne. En particulier, on obtient de bons résultats à partir d'une vingtaine d'individus par population et d'une dizaine de loci. Les différents scénarios considérés montrent aussi que le modèle simple utilisé est robuste à des situations réalistes s'écartant de celui-ci.

Pour les données de l'arganier, nous avons estimé que le modèle ayant la probabilité à postériori la plus élevée n'incluait que la connectivité (P=0.60). Ce modèle domine largement tous les autres et il apparait que l'altitude n'affecte pas la différenciation génétique. L'estimation à postériori du paramètre $\sigma^2 = 0.27$ indique que la distance explique assez bien la différenciation observée et que, même si d'autres facteurs peuvent jouer un rôle, leur effet est probablement assez faible. Le coefficient de régression de l'équation 2.2 correspondant à la connectivité a une distribution à postériori centrée autour de -0.55, ce qui indique que la différenciation génétique augmente avec l'isolement géographique.

Si les humains se sont propagés depuis l'Afrique de l'Est et si ce processus a engendré des goulots d'étranglement successifs de faible amplitude (PRUGNOLLE *et al.* 2005), on attend une augmentation graduelle des coefficients F_{ST} au fur et à mesure que la distance depuis l'origine augmente. Ici la probabilité à postériori du modèle qui inclut ce facteur est de 1. Seules les

populations africaines semblent en dehors de la régression et plusieurs hypothèses sont formulées pour expliquer ce résultat dans la discussion. L'analyse comprenant les deux facteurs latitude et longitude conduit à une probabilité très élevée pour le modèle les incluant tous les deux ($P = 0.97$). La valeur absolue du coefficient de régression se rapportant à la longitude est plus importante que pour la latitude, confirmant que les mouvements de population ont été plus importants le long de cet axe (Piazza *et al.* 1981). L'influence de la latitude peut être expliquée par les effets combinés de mouvements vers le sud (principalement en Amérique mais aussi en Afrique), du climat, et de tailles efficaces plus faibles des populations méridionales dans ce jeu de données. Ces deux facteurs expliquent bien la structure génétique observée au vu de l'estimation à postériori $\sigma^2 = 0.19$. Le facteur unique de la distance à l'Afrique de l'Est conduit à $\sigma^2 = 0.27$, et $\sigma^2 = 0.18$ si on ne considère pas les population africaines.

La méthode a été implémentée dans un logiciel écrit en langage C++ avec une interface graphique conviviale, le rendant simple d'utilisation. L'interface permet aussi de visualiser les résultats sous forme graphique. Le logiciel est diffusé librement sur le site internet du Laboratoire d'Ecologie Alpine à l'adresse `http://www-leca.ujf-grenoble.fr/logiciels.htm` sous le nom de GESTE pour « GEnetic STructure inference based on genetic and Environmental data ».

2.2.3 Discussion

Contrairement au test de Mantel qui résume l'information génétique sous forme d'une statistique sommaire (F_{ST}), notre méthode est basée sur un modèle et utilise toute l'information disponible. De plus, elle évite les problèmes rencontrés par les approches fréquentistes utilisant des techniques de randomisation. Elle permet aussi d'inclure des variables qui ne peuvent pas s'exprimer sous forme de distance deux à deux. Enfin, le fait que la méthode soit robuste à différents scénarios démographiques rend la méthode assez générale pour être appliquée à des données très diverses. Dans tous les cas, l'estimation d'un coefficient F_{ST} pour chaque population permet de prendre en compte le fait que l'intensité de la dérive génétique diffère pour chaque population.

Différentes explications ont été envisagées pour expliquer le comportement atypique des populations africaines par rapport à la régression en fonction de la distance depuis l'Afrique de l'Est. Les simulations effectuées avec le modèle d'expansion de SPLATCHE (Currat *et al.* 2004) écartent l'hypo-

thèse d'un effet dû à l'écart trop important du modèle RAO supposé dans la méthode. Une première explication possible est la faible taille efficace des populations africaines de ce jeu de données, qui sont des groupes tribaux assez petits et isolés. L'autre raison invoquée est le biais de recrutement des marqueurs qui ont été choisis parmi un panel européen du CEPH (Ray *et al.* 2005). Des simulations supplémentaires ont été effectuées à l'aide du logiciel SPLATCHE (Currat *et al.* 2004) et ont permis de montrer que ce biais produisait le même effet sur les populations africaines que celui observé avec les données réelles.

Les résultats de l'application de notre méthode sur les données humaines suggèrent une expansion rapide au Moyen-Orient et en Europe, suivie d'une expansion plus lente en Asie de L'Est, en Australasie et en Amérique. Contrairement aux méthodes fréquentistes multivariées complexes (voir par exemple Piazza *et al.* 1981) qui amènent à des résultats difficiles à interpréter, notre méthode bayésienne fournit des conclusions simples. Elle peut être utilisée pour identifier les facteurs environnementaux responsables de la structure génétique spatiale observée, mais aussi pour étudier les processus d'expansion, en introduisant simplement les coordonnées géographiques comme variables explicatives.

Conclusion

Notre méthode fournit de nombreuses informations qui peuvent aider à mieux comprendre l'histoire démographique et évolutive des espèces étudiées. Elle peut être appliquée dans un grand nombre de problèmes, allant de la génétique humaine, à la biologie de la conservation et à l'agronomie. Dans le domaine de la génétique humaine, le modèle peut être utilisé pour tester l'importance de facteurs culturels pour décrire les variations génétiques et ainsi fournir des informations importantes pour l'identification de gènes impliqués dans des maladies génétiques. Dans le domaine de la biologie de la conservation, identifier les facteurs qui déterminent la différenciation peut aider à concevoir des stratégies de management pour mieux préserver la diversité génétique. Enfin l'analyse des données du HGDP-CEPH illustre bien la possibilité d'utiliser la méthode pour reconstruire l'histoire démographique des populations.

Chapitre 2 - Article A

Foll, M., et O. Gaggiotti, 2006 Identifying the environmental factors that determine the genetic structure of populations. Genetics **174**: 875–891

Abstract

The study of population genetic structure is a fundamental problem in population biology because it helps us obtain a deeper understanding of the evolutionary process. One of the issues most assiduously studied in this context is the assessment of the relative importance of environmental factors (geographic distance, language, temperature, altitude, etc.) on the genetic structure of populations. The most widely used method to address this question is the multivariate Mantel test, a non-parametric method that calculates a correlation coefficient between a dependent matrix of pairwise population genetic distances and one or more independent matrices of environmental differences. Here we present a hierarchical Bayesian method that estimates F_{ST} values for each local population and relates them to environmental factors using a generalized linear model. The method is demonstrated by applying it to two data sets, a data set for a population of the argan tree and a human data set comprising 51 populations distributed worldwide. We also carry out a simulation study to investigate the performance of the method and find that it can correctly identify the factors that play a role in the structuring of genetic diversity under a wide range of scenarios.

Key words : genetic structure, environmental factors, demographic history, human expansion, Bayesian statistics, MCMC.

Running title : Environmental determinants of genetic structure.

Introduction

One of the fundamental problems in population genetics is the study of the nature of genetic differentiation that is found in real populations and, if possible, to identify the factors that are responsible for the observed spatial structuring of genetic diversity. A clear understanding of these issues is of fundamental importance for a wide range of applications that include, among others, the inference of population histories, biodiversity conservation and the identification of disease genes and/or disease-resistant genes in humans and economically important species. There are many methods that estimate different measures of genetic differentiation among populations (Excoffier 2001, Rousset 2001) but there is a paucity of methods that allow us to identify factors that influence genetic structuring. The most commonly used method, the multivariate Mantel test (Smouse *et al.* 1986), is based on the calculation of genetic and environmental distance measures between every pair of populations.

Genetic structuring of neutral markers is a consequence of the amount of genetic drift to which each local population has been subjected, either due to its local effective size and/or due to its overall degree of geographic/ecologic isolation. Thus, it seems appropriate to base the estimation on parameters and variables that are specific to each local population. However, the study of population genetic structuring is traditionally done using global measures such as F_{ST} or G_{ST}, which ignore differences in the strength of genetic drift across populations. Over a decade ago, Balding et Nichols (1995) proposed the use of population specific F_{ST}'s in the context of a migration-drift equilibrium model. They considered bi-allelic markers and modeled allele frequencies using a beta distribution with expectation p and variance $p(1-p)/(1+\theta)$ so that $F_{ST} = 1/(1+\theta)$. Such a formulation enabled them to use a likelihood-based approach to estimate population specific F_{ST}'s. More recently, Nicholson *et al.* (2002) used a truncated normal distribution with mean p and variance $cp(1-p)$ instead of a beta because they were interested in a non-equilibrium fission model where subpopulations evolve in isolation after splitting from an ancestral population in which the allele frequency is p. F_{ST} and c are the same in the limit as $c \downarrow 0$ but differ when p is close to 0 or 1, or c is large (Balding 2003). Marchini et Cardon (2002) compared the two formulations by fitting them to two human datasets and concluded that although Nicholson *et al.* (2002) parameterisation fitted better the European dataset, both formulations performed equally well with the global dataset. Recently, both formulations were extended to consider multi-allelic loci; Balding (2003) considered the

migration-drift equilibrium model while Falush *et al.* (2003) considered the non-equilibrium fission model. Interestingly, both formulations lead to the same multinomial-Dirichlet distribution for the subpopulation allele frequencies, $\widetilde{\boldsymbol{p_{ij}}} \sim \mathcal{D}ir(\theta_j p_{i1}, \ldots, \theta_j p_{iK_i})$. The main difference between the two models resides in the interpretation given to F_{ST}. In the case of the migration-drift model, the F_{ST}s measure how divergent each local population is from the metapopulation as a whole, while in the case of the fission model they measure the degree of genetic differentiation between each descendant population and the ancestral population.

Using as starting point the methodological advances described above, we have developed a novel hierarchical Bayesian method that represents a more informative complement to the Mantel test. The most important innovation of our method is that it introduces non-genetic (environmental) data via the prior distribution of the population specific F_{ST}s, which are modeled using a generalised linear model. We can thus consider different alternative models, each including a different set of environmental variables, and estimate their posterior probabilities. Using these posterior probabilities we can make inferences about the factors that influence genetic structure. The method is easy to employ, intuitive and of wide applicability. It can be used to test for the effect of specific biotic/abiotic factors and also to study population expansion processes. In this latter case, the geographic coordinates of the samples are used to define the prior distribution of F_{ST}s. We carry out a detailed simulation study that demonstrates the method can correctly identify the biotic/factors that influence the genetic structure of populations and we present examples that illustrate how the method can be applied to study a wide range of scenarios.

Methods

As stated in the introduction, the model we use is based on ideas put forward by Balding et Nichols (1995) and later extended by Nicholson *et al.* (2002), Falush *et al.* (2003) and Balding (2003). Balding et Nichols (1995) and Balding (2003) considered a migration-drift model while Nicholson *et al.* (2002) and Falush *et al.* (2003) considered a fission model. For the sake of simplicity we will describe the details of our approach using the terminology of the fission model but it should be kept in mind that it also applies to island models. We consider a collection of J subpopulations that evolved in isolation after splitting from an ancestral population. The derived subpopulations

may have been subject to different amounts of genetic drift and, therefore, their allele frequencies will show different degrees of differentiation from the ancestral allele frequency. The extent of differentiation between subpopulation j and the ancestral population is measured by F_{ST}^{j} and is the result of its demographic history.

We consider a set of I loci and let K_i be the number of alleles at the i^{th} locus. Let $\boldsymbol{p_i} = \{p_{ik}\}$ denote the allele frequencies of the ancestral population at locus i, where p_{ik} is the frequency of the allele k at locus i ($\sum_k p_{ik} = 1$). We use $\boldsymbol{p} = \{\boldsymbol{p_i}\}$ to denote the entire set of allele frequencies of the ancestral population and $\widetilde{\boldsymbol{p_{ij}}} = \{\widetilde{p_{ijk}}\}$ to denote the current allele frequencies at locus i for subpopulation j. Under these assumptions, the allele frequencies at locus i in subpopulation j follows a Dirichlet distribution with parameters $\theta_j \boldsymbol{p_i}$,

$$\widetilde{\boldsymbol{p_{ij}}} \sim \mathcal{D}ir(\theta_j p_{i1}, \ldots, \theta_j p_{iK_i}) \tag{2.3}$$

where $\theta_j = 1/F_{ST}^{j} - 1$. The parameters F_{ST}^{j}s are very closely related to Wright's F_{ST} (Wright 1951) parameter and are interpreted as measures of the shared ancestry within each of the subpopulations (see Balding 2003, for a more detailed explanation).

Model Parameters

Our objective is to estimate the F_{ST}s by combining genetic and environmental data. For simplicity we will focus on the derived parameters θ_js but the same results would be obtained by focusing directly on the F_{ST}s. We use a hierarchical Bayesian approach that introduces the genetic data through the likelihood function and the environmental data through the prior distribution of the θ_js. We will also estimate the allele frequencies of each subpopulation and that of the ancestral population since they are also unknowns. The data consist of allele counts obtained from samples of size n_{ij} (where the subscript i refers to locus and the subscript j to population). We use a_{ijk} to denote the number of alleles k observed at locus i in the sample from subpopulation j. Thus, $n_{ij} = \sum_k a_{ijk}$. The full data set can be presented as a matrix $\boldsymbol{A} = \{\boldsymbol{a_{ij}}\}$, where $\boldsymbol{a_{ij}} = \{a_{ij1}, a_{ij2}, \ldots, a_{ijK_i}\}$ is the allele count at locus i for subpopulation j. The observed allele frequencies, $\boldsymbol{a_{ij}}$, can be considered as sampled from the true alleles frequencies $\widetilde{\boldsymbol{p_{ij}}}$ and, therefore, can be described by the multinomial distribution (Holsinger 1999) :

$$\boldsymbol{a_{ij}} \sim \mathcal{M}ultinomial\{n_{ij}; \widetilde{p_{ij1}}, \widetilde{p_{ij2}}, \ldots, \widetilde{p_{ijK_i}}\} \tag{2.4}$$

Let us first construct the likelihood function for the allele count $\boldsymbol{a_{ij}}$ at locus i for subpopulation j. In principle, we could use as likelihood the multinomial distribution (equation 4.5) and consider equation 4.2 as a Bayesian prior. However, in our case, we can calculate exactly the marginal distribution of $\boldsymbol{a_{ij}}$ because the Dirichlet distribution is the conjugate prior of the multinomial. This allows us to eliminate the nuisance parameters $\widetilde{\boldsymbol{p_{ij}}}$ which are not of immediate interest but are needed by the model. The marginal distribution is obtained by integrating out $\widetilde{\boldsymbol{p_{ij}}}$:

$$P(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_j) = \int \cdots \int \pi(\boldsymbol{a_{ij}},\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i},\theta_j) d\widetilde{p_{ij1}} \cdots d\widetilde{p_{ijK_i}}$$

where, $\pi(\boldsymbol{a_{ij}},\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i},\theta_j) = P(\boldsymbol{a_{ij}}|\widetilde{\boldsymbol{p_{ij}}})\pi(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i},\theta_j)$. The right-hand terms $\pi(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i},\theta_j)$ and $P(\boldsymbol{a_{ij}}|\widetilde{\boldsymbol{p_{ij}}})$ are given by equations 4.2 and 4.5, respectively. Thus, we obtain the multinomial-Dirichlet distribution :

$$P(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_j) = \frac{n_{ij}!\Gamma(\theta_j)}{\Gamma(n_{ij}+\theta_j)} \prod_{k=1}^{K_i} \frac{\Gamma(a_{ijk}+\theta_j p_{ik})}{a_{ijk}!\,\Gamma(\theta_j p_{ik})}$$

This equation is basically the same as eq. 13 in Balding (2003). The only difference is that here we express it in more general terms by using subindexes to identify loci and populations. It is also equivalent to eq. 5 in Falush *et al.* (2003) and differs from the parameterisations of Balding et Nichols (1995) and Nicholson *et al.* (2002) in that they use respectively the Beta or a truncated normal distribution instead of the Dirichlet because they are concerned with biallelic markers while we deal with multiallelic loci.

The likelihood is obtained by multiplying across all loci and populations :

$$L(\boldsymbol{p},\boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(\boldsymbol{a_{ij}}|\boldsymbol{p_i},\theta_j) \tag{2.5}$$

Since the allele frequencies in the ancestral population are unknown, we have to estimate them by introducing a non-informative Dirichlet prior, $\boldsymbol{p_i} \sim \mathcal{D}ir(1,\ldots,1)$, into our Bayesian model.

Incorporation of non-genetic data

We want to determine the influence of biotic/abiotic factors on the F_{ST}s coefficients or equivalently on the θ_js. For example, we often want to ascertain if geographic isolation (usually measured in terms of pairwise geographic distances but best described by a connectivity measure; see Moilanen et

Nieminen (2002) for a review) and/or linguistic affinity have influenced the genetic structure of a species (Sokal *et al.* 1992, Cavalli-Sforza et Feldman 2003, Geffen *et al.* 2004, Hunley et Long 2005). Given that F_{ST} values are bounded between 0 and 1, we could use the familiar logistic regression model

$$\begin{aligned} ln\left(\frac{F_{ST}^{j}}{1-F_{ST}^{j}}\right) &= -ln\ \theta_j \\ &= \alpha_0 + \alpha_1 G_j^{(1)} + \alpha_2 G_j^{(2)} + \alpha_3 G_j^{(1)} G_j^{(2)} \end{aligned} \tag{2.6}$$

where $G_j^{(1)}$ and $G_j^{(2)}$ are respectively the observed values of the first and the second environmental factor for population j and can be presented as elements of a matrix $G = \{G_1^{(1)}, \ldots, G_J^{(1)}, G_1^{(2)}, \ldots, G_J^{(2)}\}$. Here we consider only two factors but the method can be easily extended to include many more factors. Although this is in principle a valid approach, in practice it would impose a very strong prior on the θ_js. Thus, we chose instead to use a logNormal prior where the mean is a function of environmental variables (as in Gaggiotti *et al.* 2004) :

$$ln\ \theta_j \sim \mathcal{N}(\mu_j, \sigma^2)$$

with :

$$\mu_j = \alpha_0 + \alpha_1 G_j^{(1)} + \alpha_2 G_j^{(2)} + \alpha_3 G_j^{(1)} G_j^{(2)} \tag{2.7}$$

where $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ denotes a vector of model parameters to be estimated. The magnitude and sign of these parameters indicate the strength and direction of the effect of the corresponding biotic/abiotic factors. α_3 corresponds to the interaction of the two factors. In order to improve computational efficiency and facilitate the posterior interpretation process, the factors $\boldsymbol{\alpha}$ should be normalized so as to have mean 0 and variance 1. The variance of the logNormal distribution, σ^2, is a measure of model fit and has to be estimated too. In order to estimate this and all other parameters related to the effect of environmental factors, we use vague priors. More specifically, we take $\alpha_l \sim \mathcal{N}(0, 10)$ and $1/\sigma^2 \sim \mathcal{G}amma(1, 1)$.

With these likelihood and priors, the full posterior distribution represented by the Directed Acyclic Graph (DAG) in Figure 2.3 is given by :

$$\pi(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{A}) \propto L(\boldsymbol{p}, \boldsymbol{\theta}) \pi(\boldsymbol{p}) \pi(\boldsymbol{\theta} | \boldsymbol{\alpha}, \sigma^2) \pi(\boldsymbol{\alpha}) \pi(\sigma^2) \tag{2.8}$$

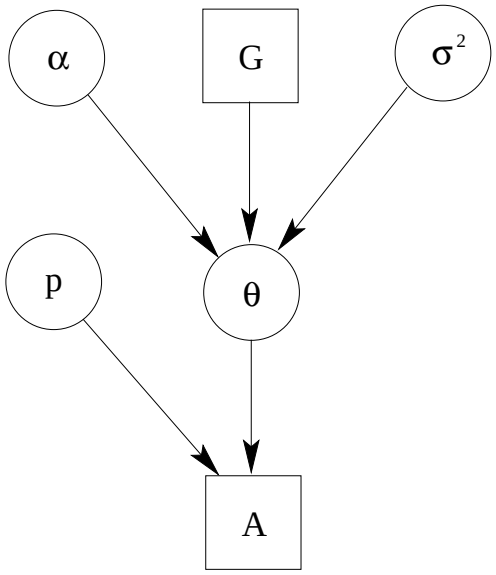

FIG. 2.3 – *The DAG of the model given in equation 3.8. Square nodes denote known quantities (i.e. data) and circles represent parameters to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text. A is the genetic data and G the environmental data used to explain the genetic differentiation.* $\boldsymbol{p}$ *is the matrix of ancestral allele frequencies,* $\boldsymbol{\theta}$ *is the vector of genetic differentiation coefficient for each local population,* $\boldsymbol{\alpha}$ *is the vector of regression coefficients that correspond to factors in G and* σ^2 *measures the deviation from the regression. The multinomial-Dirichlet distribution allows the exact calculation of the marginal likelihood to eliminate the nuisance parameters* $\widetilde{\boldsymbol{p}}$.

Posterior model probabilities

As well as parameters estimates, we are interested in determining the dependence (or otherwise) of the factors upon the genetic differentiation. Alternative models are obtained by adopting different forms for μ_j given in equation 2.7. The simplest model might set $\mu_j = \alpha_0$, corresponding to the assumption that neither factor has an effect upon the genetic differentiation. A suitable alternative might be $\mu_j = \alpha_0 + \alpha_2 G_j^{(2)}$, in which we assume that factor $G^{(2)}$ has some influence. Clearly these models can be expressed by setting the appropriate $\alpha_l = 0$ in equation 2.6 with the obvious restriction that the interaction term can only be included if both $G^{(1)}$ and $G^{(2)}$ are included in the model. Thus with two environmental factors there are nine alternative models as shown in Table 2.1 for the human example (see below).

In order to discriminate between the different models we let $\boldsymbol{\alpha}_M$ denote the vector of nonzero α-values under model M and then derive the following posterior distribution over both parameter and model space :

$$\pi(\boldsymbol{p}, \boldsymbol{\alpha}_M, \boldsymbol{\theta}, \sigma^2, M|\boldsymbol{A}) \propto L(\boldsymbol{p}, \boldsymbol{\theta})\pi(\boldsymbol{p})\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}_M, \sigma^2)\pi(\boldsymbol{\alpha}_M)\pi(\sigma^2)p(M) \qquad (2.9)$$

Model	Factors included	Probability
1	Constant	0
2	Latitude	0
3	Constant and Latitude	0
4	Longitude	0
5	Constant and Longitude	0
6	Latitude and Longitude	0
7	Constant, Latitude and Longitude	0.97
8	Latitude, Longitude and interaction	0
9	All	0.03

TAB. 2.1 – *Posterior probabilities of the nine possible models to explain the genetic differentiation in humans as a function of latitude and longitude. We used an uniform prior for model probabilities.*

where $p(M)$ denotes the prior model probability. In most cases, it makes sense to use a uniform prior where all models are equally likely, so this is the strategy we adopted for all the analyses that follow.

Implementation

The estimation of model parameters is carried out using a combination of MCMC and RJMCMC (GREEN 1995) techniques that are described in Appendix 1. We evaluated the convergence of the method using the diagnostic tests implemented in the R BOA package (Bayesian Output Analysis Program version 1.1.5 SMITH 2005). The tests indicated that a burn-in of 2000 iterations was enough to attain convergence and it has been implemented as part of the pilot-tuning process (see Appendix 1). We used a sample size of 10000 and a thinning interval of 10 as suggested by an autocorrelation analysis. With these parameter values, the total length of the chain was 100 000 iterations. Posterior model probabilities are estimated from the number of times the algorithm visited each model. F_{ST}s are estimated using model-averaged posterior means and, therefore, take into account model uncertainty. The parameters of the linear model, α and σ^2, are specific to each model and, therefore, we report the estimates obtained for the model with the highest posterior probability. Additionally, instead of using the posterior means of σ^2 we use its mode because its posterior density is highly asymmetric.

SIMULATION STUDY

We evaluate the performance of the method using three different approaches to generate simulated data. The first approach uses the same statistical mo-

del assumed by our method (the inference model) and allows us to study the effect of the quality of the samples (number of loci, sample sizes) and different biological scenarios (number of populations, environmental factors considered, etc.) on the accuracy of the estimates obtained. The two other approaches allow us to explore the performance of the method under scenarios that deviate from the assumptions of the inference model. We used EasyPop (Balloux 2001) to generate data under a migration-drift-mutation model that incorporates the effect of isolation by distance, and SPLATCHE (Currat *et al.* 2004) to generate data under a population fission where new populations are formed progressively from older populations.

Inference model

The inference model used to derive our method applies to two different scenarios, an island model at migration-drift equilibrium, and a simple population fission model. The same algorithm can be used to generate synthetic samples under these two scenarios and is described in Appendix 2. We chose a set of default values for the key parameters of the inference model (see Table 2.2) and first evaluated the accuracy of the method using the Relative Mean Square Error (RMSE) :

$$RMSE(p) = \frac{1}{10}\sum_{k=1}^{10}\left(\frac{p_k - \tilde{p}}{\tilde{p}}\right)^2$$

where $\tilde{p}$ is its true value of the parameter being estimated and p_k is its estimated value for replicate k. We analysed 10 replicate datasets generated using the default values (Table 2.2). We considered samples of good quality (50 individuals and 20 loci) but within the range of those commonly seen in the literature. We also studied the effect of changing the value of one parameter at a time. We used data simulated under model 7 in which includes two factors and a constant term. The degree of uncertainty of the estimations is measured by the 95% Highest Probability Density Interval (HPDI, the smallest interval which contains 95% of the values).

Accuracy : In general, all estimates have a small RMSE and narrow HPDI with the exception of σ^2 (Table 2.3). Lower values of F_{ST} lead to slightly higher RMSEs. The poor estimation of σ^2 is due to the uncertainty in the estimation of the α's, which cannot be separated from the deviation from the regression included in the model. As it will be shown below, the quality of the estimation for this parameter can be greatly improved by increasing the number of populations sampled.

Parameter	Value
Populations	10
Loci	20
Sample size per population	50
Mean number of alleles	10
Mean F_{ST}	0.1
Model	7
α_0	-2.4
α_1	-0.6
α_2	0.5
σ^2	0.1

TAB. 2.2 – *Default parameters used in data simulated under the inference model discussed in text.*

Parameter	True value	Estimated value	RMSE	95% HPDI
F_{ST}^1	0.101	0.111	2.8%	[0.087 ; 0.136]
F_{ST}^2	0.100	0.107	2.1%	[0.083 ; 0.131]
F_{ST}^3	0.146	0.151	0.8%	[0.121 ; 0.184]
F_{ST}^4	0.072	0.076	1.0%	[0.058 ; 0.095]
F_{ST}^5	0.029	0.031	4.3%	[0.022 ; 0.042]
F_{ST}^6	0.036	0.041	5.3%	[0.029 ; 0.053]
F_{ST}^7	0.236	0.247	0.7%	[0.203 ; 0.295]
F_{ST}^8	0.112	0.111	0.4%	[0.086 ; 0.136]
F_{ST}^9	0.029	0.034	7.9%	[0.024 ; 0.045]
F_{ST}^{10}	0.180	0.183	1.3%	[0.149 ; 0.224]
σ^2	0.10	0.245	167.3%	[0.095 ; 0.678]
α_0	-2.40	-2.310	0.2%	[-2.683 ; -1.943]
α_1	-0.60	-0.553	1.1%	[-0.916 ; -0.200]
α_2	0.50	0.495	0.5%	[0.142 ; 0.843]

TAB. 2.3 – *Estimations, RMSE and 95% HPDIs based on 10 replicates simulated under the inference model. Estimated values correspond to the posterior means with the exception of σ^2 which is estimated by the mode. See Table 2.2 for parameter values used.*

Sample sizes and number of loci : The quality of the estimates is good even with the lowest sample size and number of loci (Tables 2.4 and 2.5) and increases as they increase, specially for estimates of F_{ST}. The posterior probability of the true model (7) was always the highest, independently of sample size and number of loci but model determination did increase substantially as they increased. The second most probable model always had a probability at least three times lower than that of the true model. Note that it is easier to improve estimates (i.e. decrease HPDIs and increase posterior probability of true model) by increasing the number of loci studied than by increasing

sample sizes. In particular, increasing sample sizes beyond 50 does not improve estimates.

		Sample size per population			
Parameter	True value	10	20	50	100
F_{ST}^1	0.101	0.106	0.133	0.109	0.098
		[0.069 ; 0.144]	[0.101 ; 0.171]	[0.085 ; 0.134]	[0.078 ; 0.120]
F_{ST}^2	0.100	0.105	0.085	0.110	0.111
		[0.068 ; 0.141]	[0.061 ; 0.111]	[0.085 ; 0.135]	[0.089 ; 0.134]
F_{ST}^3	0.146	0.156	0.160	0.170	0.128
		[0.109 ; 0.201]	[0.122 ; 0.200]	[0.137 ; 0.206]	[0.103 ; 0.154]
F_{ST}^4	0.072	0.100	0.085	0.084	0.080
		[0.065 ; 0.138]	[0.060 ; 0.110]	[0.066 ; 0.105]	[0.063 ; 0.097]
F_{ST}^5	0.029	0.033	0.031	0.025	0.032
		[0.014 ; 0.054]	[0.019 ; 0.044]	[0.017 ; 0.034]	[0.024 ; 0.041]
F_{ST}^6	0.036	0.044	0.036	0.044	0.032
		[0.022 ; 0.067]	[0.022 ; 0.051]	[0.031 ; 0.057]	[0.024 ; 0.041]
F_{ST}^7	0.236	0.207	0.281	0.251	0.231
		[0.149 ; 0.265]	[0.221 ; 0.342]	[0.206 ; 0.299]	[0.191 ; 0.273]
F_{ST}^8	0.112	0.117	0.137	0.121	0.095
		[0.079 ; 0.157]	[0.102 ; 0.173]	[0.094 ; 0.148]	[0.076 ; 0.115]
F_{ST}^9	0.029	0.037	0.035	0.045	0.031
		[0.017 ; 0.059]	[0.021 ; 0.051]	[0.033 ; 0.057]	[0.024 ; 0.040]
F_{ST}^{10}	0.180	0.167	0.139	0.212	0.196
		[0.120 ; 0.218]	[0.104 ; 0.175]	[0.171 ; 0.254]	[0.161 ; 0.232]
σ^2	0.10	0.229	0.243	0.246	0.233
		[0.095 ; 0.673]	[0.096 ; 0.704]	[0.096 ; 0.708]	[0.094 ; 0.655]
α_0	-2.40	-2.31	-2.30	-2.24	-2.38
		[-2.70 ; -1.91]	[-2.67 ; -1.90]	[-2.63 ; -1.88]	[-2.73 ; -2.03]
α_1	-0.60	-0.507	-0.607	-0.576	-0.525
		[-0.876 ; -0.126]	[-0.976 ; -0.228]	[-0.941 ; -0.190]	[-0.887 ; -0.201]
α_2	0.50	0.470	0.473	0.487	0.533
		[0.109 ; 0.817]	[0.107 ; 0.817]	[0.144 ; 0.843]	[0.192 ; 0.887]
$p(M = 7)$		0.59	0.72	0.80	0.80

TAB. 2.4 – *True values, estimated values and 95% HPDIs from simulated data for various sample sizes. Estimated values correspond to the posterior means with the exception of σ^2 which is estimated by the mode. The last line presents the posterior probability of the true model, $p(M = 7)$. See Table 2.2 for parameter values used.*

Number of populations : Varying the number of populations changes the number of F_{ST} values considered in the model, therefore, we only evaluate its effect on model determination and on estimates of regression parameters. With only 5 or 7 populations the method fails to identify the true model and provides low quality estimates of regression parameters. However, with 10 or more populations model determination greatly improves (the posterior probability for the true model is at least 0.80) and bias and uncertainty of all

		Number of loci		
Parameter	True value	10	20	50
F_{ST}^1	0.101	0.089	0.109	0.096
		[0.062 ; 0.119]	[0.085 ; 0.134]	[0.083 ; 0.111]
F_{ST}^2	0.100	0.085	0.110	0.109
		[0.057 ; 0.114]	[0.085 ; 0.135]	[0.094 ; 0.126]
F_{ST}^3	0.146	0.143	0.170	0.150
		[0.104 ; 0.184]	[0.137 ; 0.206]	[0.1305 ; 0.1711]
F_{ST}^4	0.072	0.076	0.084	0.070
		[0.052 ; 0.102]	[0.066 ; 0.105]	[0.059 ; 0.081]
F_{ST}^5	0.029	0.041	0.025	0.041
		[0.026 ; 0.059]	[0.017 ; 0.034]	[0.033 ; 0.048]
F_{ST}^6	0.036	0.043	0.044	0.041
		[0.027 ; 0.061]	[0.031 ; 0.057]	[0.034 ; 0.048]
F_{ST}^7	0.236	0.186	0.251	0.243
		[0.136 ; 0.236]	[0.206 ; 0.299]	[0.214 ; 0.273]
F_{ST}^8	0.112	0.118	0.121	0.120
		[0.083 ; 0.153]	[0.094 ; 0.148]	[0.103 ; 0.137]
F_{ST}^9	0.029	0.020	0.045	0.034
		[0.010 ; 0.030]	[0.033 ; 0.057]	[0.027 ; 0.041]
F_{ST}^{10}	0.180	0.150	0.212	0.175
		[0.107 ; 0.194]	[0.171 ; 0.254]	[0.152 ; 0.199]
σ^2	0.10	0.252	0.246	0.230
		[0.105 ; 0.737]	[0.096 ; 0.708]	[0.100 ; 0.613]
α_0	-2.40	-2.449	-2.24	-2.31
		[-2.831 ; -2.038]	[-2.63 ; -1.88]	[-2.65 ; -1.96]
α_1	-0.60	-0.529	-0.576	-0.509
		[-0.892 ; -0.151]	[-0.941 ; -0.190]	[-0.857 ; -0.198]
α_2	0.50	0.458	0.487	0.468
		[0.100 ; 0.843]	[0.144 ; 0.843]	[0.154 ; 0.799]
$p(M = 7)$		0.48	0.80	0.80

TAB. 2.5 – *True values, estimated values and 95% HPDIs from simulated datasets with different number of loci. Estimated values correspond to the posterior means with the exception of σ^2 which is estimated by the mode. The last line presents the posterior probability of the true model, $p(M = 7)$. See Table 2.2 for parameter values used.*

regression parameter estimates decreases pronouncedly. This is particularly the case for σ^2 estimates, which bias almost disappears with 20 populations. The results obtained depend largely on the values chosen for the coefficients of the regression ($\alpha_0 = -2.4$, $\alpha_1 = -0.6$ and $\alpha_2 = 0.5$). The effect of a factor on population genetic structure is proportional to the absolute value of the corresponding regression coefficient. Factors with a strong effect are more easily identified than factors with a small effect. Thus, with only five populations, the highest posterior probability correspond to the model with the constant term only. With seven populations, the method can also identify

the effect of the first factor but not that of the second and, therefore, the posterior probability is highest for model 5. The effect of the second factor is identified only with 10 or more populations. Additionally, models that do not include the constant term are readily excluded because its effect is very strong and easily identified by the method. It is important to note that when the method fails to identify the true model, the results obtained give hints that they are not reliable. For instance, the posterior probability of models with either factor is non negligible (over 15%) and the estimates of σ^2 are very high (upper bound of the HPDI well over 1). Thus alert users of the method will not be mislead to accept a false model ; instead they will conclude that the results are inconclusive.

		Number of populations			
Parameter	True value	5	7	10	20
α_0	-2.4	-2.67	-2.15	-2.23	-2.33
		[-3.23 ; -1.40]	[-2.80 ; -1.50]	[-2.63 ; -1.88]	[-2.52 ; -2.15]
α_1	-0.6	*	*	-0.57	-0.60
		*	*	[-0.941 ; -0.190]	[-0.783 ; -0.423]
α_2	0.5	*	0.67	0.46	0.49
		*	[0.059 ; 1.25]	[0.144 ; 0.843]	[0.311 ; 0.670]
σ^2	0.1	0.67	0.51	0.25	0.12
		[0.203 ; 2.84]	[0.152 ; 1.71]	[0.096 ; 0.708]	[0.065 ; 0.261]
Model 1		0.53	0.17	0.01	0
Model 2		0	0	0	0
Model 3		0.22	0.22	0.11	0
Model 4		0	0	0	0
Model 5		0.16	0.32	0.02	0
Model 6		0	0	0	0
Model 7		0.08	0.26	0.80	0.97
Model 8		0	0	0	0
Model 9		0.01	0.03	0.06	0.03

TAB. 2.6 – *True values, estimated values, 95% HPDIs and posterior model probabilities from simulated datasets with different number of populations. Estimated values correspond to the posterior means with the exception of σ^2 which is estimated by the mode. Posterior model probabilities are estimated from the number of times the algorithm visited each model. Model not chown are never visited by the chain. See Table 2.2 for parameter values used. Symbol * correponds to datasets where the highest probability model does not include the corresponding parameter.*

Model determination : We simulated data under all models except equivalent ones (e.g. model 3 and model 5). The true model has always the highest probability. We note that this probability seems to increase with the number of parameters included in the true model. The probability of the true models

are respectively 0.86, 0.82, 0.91, 0.70, 0.80, 0.93 and 1.00 for model 1, 2, 3, 6, 7, 8 and 9.

Overall the results of these simulations indicate that the method can accurately estimate the parameters and identify the true model if samples are of at least average quality.

Subdivided population model

The first interpretation of our inference model (see above), considers a subdivided population at migration-drift equilibrium where the proportion of migrants can differ among local population but where all migrant groups are drawn from the same migrant pool (see Balding 2003). Here we consider instead an isolation by distance scenario where the composition of the migrant group arriving at any local population depends on the distance between the focal population and each one of the other local populations. Synthetic data for this scenario can be generated using EASYPOP, a software that implements an individual based model for the simulation of genetic data under different scenarios of population subdivision. We considered a scenario with 14 local populations, each consisting of 50 individuals (see Figure 2.4a). The probability of being a migrant was fixed to 0.05 and the migration rate between populations followed a negative exponential kernel with parameter $r = 1/\alpha$. Thus, when $\alpha \to 0$ the populations are totally isolated and as α increases, the effect of distance on migration rates decreases ; here we take $\alpha = 0.75$. We simulate 15 loci, each with 6 allelic states and with a mutation rate of 0.0001. Using these setting we generated 10 replicate dataset consistent with a scenario where population genetic structure is influenced only by the degree of geographic isolation of local populations, which is measured by their connectivity. This measure is typically used in metapopulation studies to describe the effect of the habitat matrix properties on the degree of isolation of local populations (e.g. Moilanen et Nieminen 2002). There are many connectivity measures to choose from, some include geographic distance and habitat patch size and shape. Since we only consider the degree of geographic isolation we chose to use the connectivity of population j, S_j :

$$S_j = \sum_{\substack{i=1 \\ i \neq j}}^{J} exp\left(-\beta d_{ij}\right)$$

where d_{ij} is the distance between populations j and i, and β measures the effect of distance on migration probability.

The analyses of these datasets consider three alternative models, a null model that includes the constant regression term (model 1), another that includes only connectivity (model 2) and finally a model that includes both the constant term and the distance (model 3).

Our method correctly identifies model 3 as the most probable with a posterior probability of 0.90. The regression coefficients α_0 and α_1 are estimated respectively as -1.12 and -0.50. More isolated populations (e.g. 1, 2 or 8) have a higher F_{ST} than well connected populations (e.g. 3, 4 or 5). The connectivity explains relatively well the genetic differentiation because the posterior mode of σ^2 is 0.24 and there is little scattering of points around the regression line (Figure 2.4b). This results indicate that the method provides reliable results under an isolation-by-distance model applicable to a wide range of species.

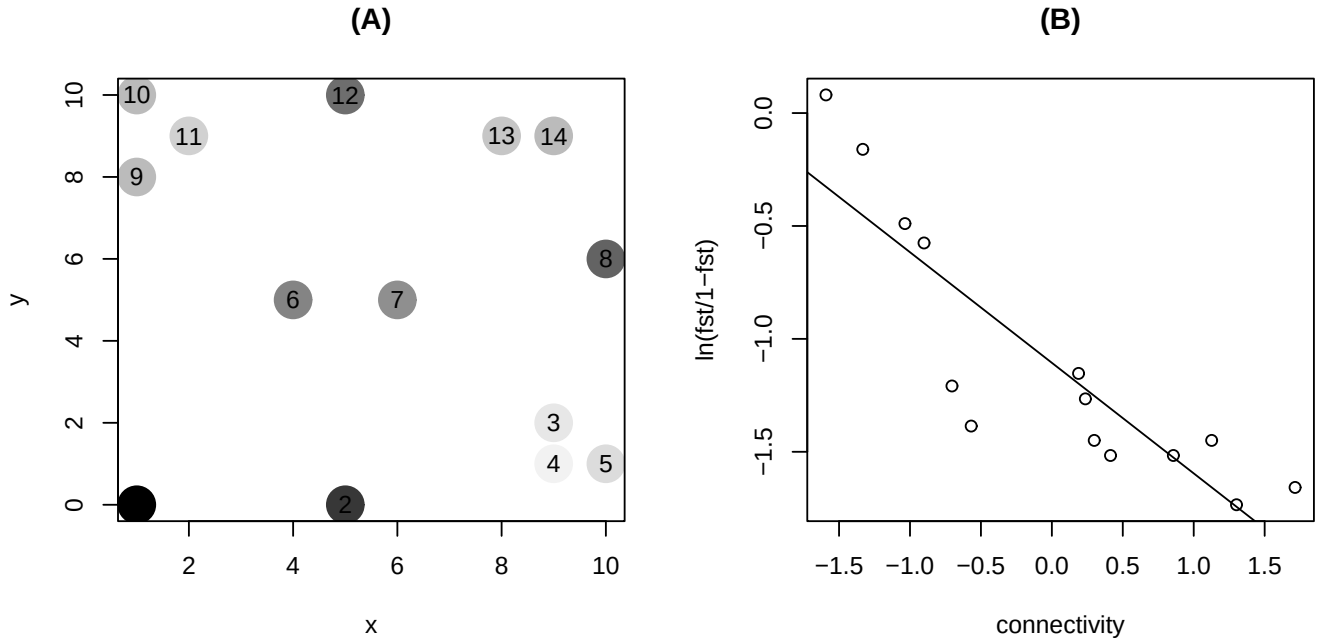

FIG. 2.4 – *(A) Map of the 14 populations used for the isolation by distance model simulated with EASYPOP. The connectivity of each population is described by the gray scale. Dark populations are isolated while lightly colored ones are well connected to others. (B) Plot of* $ln\ F_{ST}/(1-F_{ST})$ *against the connectivity for the isolation by distance simulated data. The equation of the regression line is* $ln\ F_{ST}/(1-F_{ST}) = -1.12 - 0.50S$*, where S denotes the connectivity measure defined in text.*

Splatche

The second interpretation of our inference model considers a simple fission scenario where local populations evolve in isolation after splitting from an ancestral population. To evaluate the performance of the method when the true evolutionary model differs radically from this scenario we used SPLATCHE (CURRAT *et al.* 2004). More specifically, SPLATCHE simulates a population

expansion from a single origin in a two-dimensional habitat (strict two-dimensional stepping-stone model) and generates genetic samples for geographic locations chosen by the user. We used the example of the human population expansion with origin in East Africa (see below) as a template for our simulated scenarios. We used a growth rate of 0.10, a carrying capacity of 100 for all demes, and a migration rate of 0.20. With these settings, the whole world is colonized after around 4000 generations. We simulated ten replicates of this scenario and for each of them we "collected" genetic data for 36 populations chosen uniformly on the map (Figure 2.5 shows the map of the world with the sampling locations). We calculated the distance from East Africa to each of the sampling locations using the shortest possible land-based route. The objective of this part of the study is to determine if our method can detect the genetic signal left by a recent spatial population expansion so we use distance from East Africa as a factor in our analysis. However, we also include a second factor, "insularity", to illustrate the flexibility of our method. Clearly, genetic structure is influenced by physical barriers to migration such as water masses. Therefore, the second factor considered by the logistic regression model (2.6) takes values of either 0 for continental populations (e.g. those in mainland Europe, Africa, Asia and America), or 1 for insular populations (those in England, Japan, Indonesia, New Zealand, Australia, and Greenland)

The results of this analysis clearly show that our method can accurately detect the effect of not only distance but also other physical barriers despite the fact that the synthetic data was generated using a demographic model that differs radically from the simple fission model. Figure 2.6 shows the plots of $ln[F_{ST}/(1-F_{ST})]$ versus distance and "insularity". Data points correspond to mean values over the 10 replicated scenarios. The posterior probability of the model with the constant factor and both distance from East Africa and insularity is 0.97. Figure 2.6 shows the strong effect of distance on F_{ST} and it also indicates that insular populations (identified with solid squares) would not fit very well the regression line without including the insularity factor. Here we estimate $\sigma^2 = 0.15$ while in another analysis without the insularity factor we obtained $\sigma^2 = 0.25$.

Applications

Argan tree

One of the goals of conservation genetic is to identify populations for on site conservation. This requires the characterisation of the status of each local po-

FIG. 2.5 – *Map of the 36 populations (gray dots) used in the simulation of the colonization of the world with SPLATCHE. The solid square indicates the origin of the expansion.*

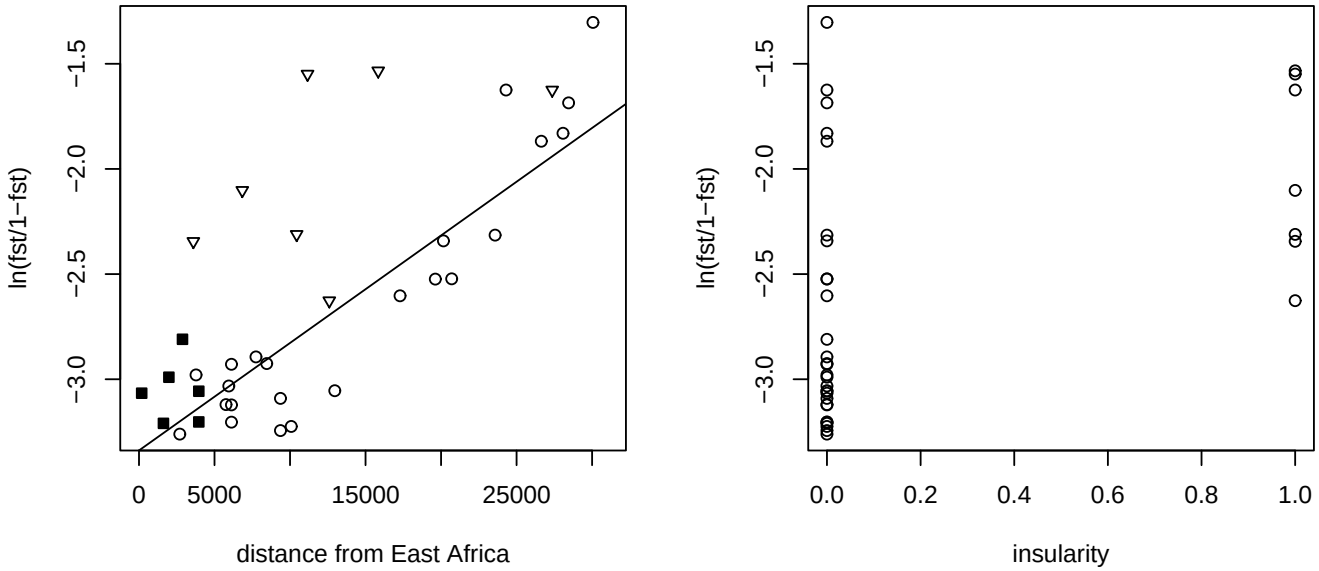

FIG. 2.6 – *Plots of* $ln\ F_{ST}/(1-F_{ST})$ *against the two factors, distance to East Africa and insularity, for the human simulated data. Triangles indicate insular populations and solid squares indicate African populations. The equation of the regression line is* $ln\ F_{ST}/(1-F_{ST}) = -2.59 + 0.46d$*, where d denotes the distance to East Africa.*

pulation in terms of their genetic diversity and their degree of differentiation with respect to all other populations, and an understanding of the factors that are responsible for the observed genetic structuring. A good example of this type of studies is presented by PETIT *et al.* (1998), who analysed the genetic structure of the argan tree of Morocco. Here we use their data to illustrate how our approach can be applied to study metapopulation scenarios with the objective of determining if the degree of isolation is influenced by geographic location and/or one other factor, in this case altitude. We note that this latter variable was not considered by PETIT *et al.* (1998) and is only included here for

the sake of illustration. **Data :** The Argan tree data consists of samples from 12 populations that were analyzed for 12 polymorphic loci. Sample size varied between 20 and 50 individuals per population. Geographic isolation of each local population is modeled using two factors, connectivity and elevation. Connectivity was calculated using the geographic distance between populations (kindly provided by R. Petit) and the same connectivity measure, S_j as in the simulation study (see above). The elevation data was obtained from Table 1 in Petit *et al.* (1998).

Results : The highest probability model (p = 0.60) includes a constant and connectivity. This model dominates well all the others, and it seems clear that elevation doesn't influence genetic differentiation since models that include this factor have low posterior probability (between 0 and 0.07). The estimated σ^2 is 0.27 (Highest Probability Density Interval HPDI = [0.13 ; 0.82]), which indicates that distance explains well the genetic differentiation and that, although some other unknown environmental factor may also play a role, its effect is unlikely to be very strong. The estimate of $\alpha_2 = -0.55$, which is negative indicating that genetic differentiation does indeed increase with increasing geographic isolation (see Figure 2.7).

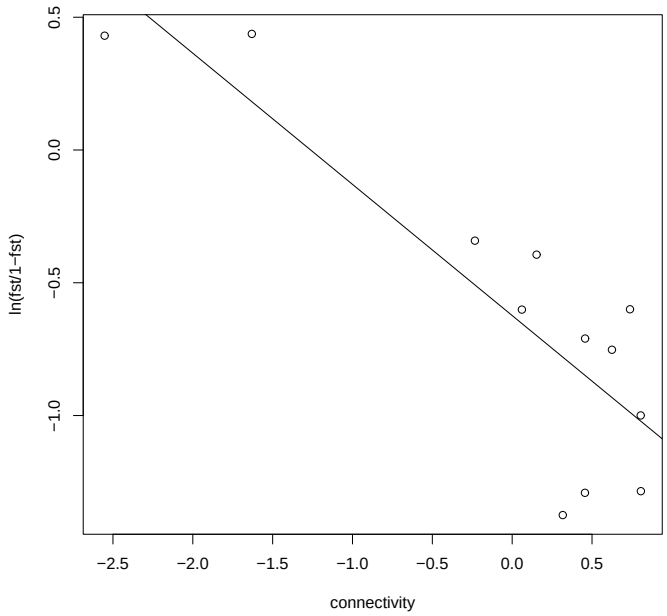

Fig. 2.7 – *Plots of* $ln\ F_{ST}/(1-F_{ST})$ *against the connectivity for the Argan Tree data. The equation of the regression line is* $ln\ F_{ST}/(1-F_{ST}) = -0.66 - 0.55S$*, where S denotes the connectivity measure defined in text.*

Humans

One of the most widely accepted theories on the origin of modern humans, the recent African origin (RAO) model, postulates that our species has evol-

ved from a small East African population that had subsequently colonized the whole world. Numerous studies provide evidence supporting this model but less is known about the specific details of the demographic history of humans (see Excoffier 2002). For example, some studies have found that an important component of human diversity seems to have evolved outside Africa (Zhao *et al.* 2000, Yu *et al.* 2001), which has been interpreted as implying a migration back to Africa. Also, the strength of the bottlenecks that may have taken place in the colonization process and the magnitude of the subsequent population expansions remain controversial (see Excoffier 2002, for a review). Here we use the HGDP-CEPH Human Genome Diversity Cell Line Panel presented by Cann *et al.* (2002) in order to illustrate how our method can be used to make inferences about population expansion events such as those underwent by humans.

Data : The data consist of 1056 individuals from 51 subpopulations, which were scored for 377 microsatellite. We carried out two analyzes. The first one considers the effect of one factor, geographic distance from East Africa along colonization routes (as in Prugnolle *et al.* 2005) while the second one considers a model with two factors, longitude and latitude.

Results : Estimated F_{ST} values are presented in Appendix 3. If humans spread from East Africa and in the process underwent successive bottlenecks of small amplitude (as proposed by Prugnolle *et al.* 2005) we should expect a gradual increase in F_{ST} values as the distance from the center of origin increases. Thus, we followed Prugnolle *et al.* (2005) and calculated the distance from East Africa along likely colonization routes for each one of the 51 populations and estimated the posterior probability of the three alternative models, with and without distance. The estimated posterior probability of the model with the constant factor and distance is 1. Figure 2.8 shows the strong effect of distance on θ but it also indicates that some of the African populations (identified by solid squares) do not fit very well the regression line. Indeed repeating the estimation process without these populations improves the fit from $\sigma^2 = 0.27$ to $\sigma^2 = 0.18$. Several factors can explain the relative lack of fit of African populations (see discussion).

Another way of uncovering the genetic signature left by the demographic history of humans is to carry out a second analysis that uses as factors latitude and longitude. The rationale for this is that under the RAO model there should be a clear effect of longitude on F_{ST} since the most important population movements have occurred along this axis (e.g. Piazza *et al.* 1981). We also expect an effect of latitude due to the combined effect of southward population movements (mainly in the Americas but also in Africa), climate, and

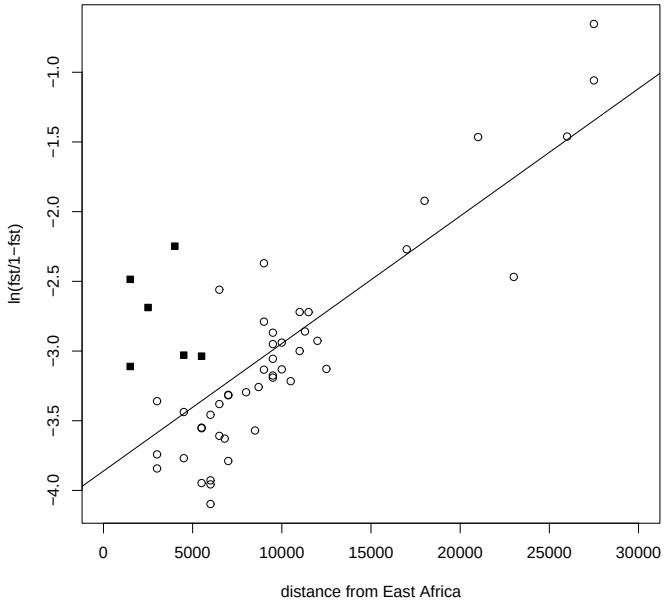

FIG. 2.8 – *Plots of* $ln\ F_{ST}/(1-F_{ST})$ *against the distance to East Africa for the human data. Black squares indicate African populations. The equation of the regression line is* $ln\ F_{ST}/(1-F_{ST}) = -3.01 + 0.56d$*, where d denotes the distance to East Africa.*

smaller effective sizes of the southern populations included in the data set (the African and Amerindian populations are much smaller than the European and most Asian populations). Thus, latitude would take into account the increased effect of genetic drift due to all this factors.

In order to take into account the fact that the Americas where colonized from east Asia, we measured longitude on a scale between 0 and 360 ° from West to East and starting at the Greenwich meridian. With the two factors considered, there are nine alternative models but almost all the probability mass was allocated to model seven (Table 2.1) that includes the constant term and both latitude and longitude. The only other model visited is model nine with probability 0.02 which also includes the two factors. So it is clear that they both play a major role in explaining the observed genetic structure. In addition, the estimate of σ^2 is 0.19 with a HPDI of $[0.13; 0.29]$, yet another indication that these two factors explain well the observed genetic pattern. Note also that this estimate of σ^2 is lower than that obtained above for the analysis that included the African populations and is almost identical to that obtained excluding these populations. Thus, latitude and longitude explain more of the variation than distance alone. The stronger effect of longitude is indicated by the larger absolute value of its regression coefficient (see Table 2.7) and the less pronounced scattering of points in the plot of $ln\ [F_{ST}/(1-F_{ST})]$ against longitude than against latitude (Figure 2.9). The negative value of α_2 indicates that colonization occurred mainly from West to East while the

negative value of α_1 indicates that the effect of genetic drift was stronger in the southern populations.

Parameter	Factor	Mean	95% HPDI
α_0	Constant	-3.01	[-3.13 ; -2.88]
α_1	Latitude	-0.30	[-0.44 ; -0.18]
α_2	Longitude	-0.43	[-0.56 ; -0.30]

TAB. 2.7 – *Estimation of α_l regression coefficients with 95% HPDIs.*

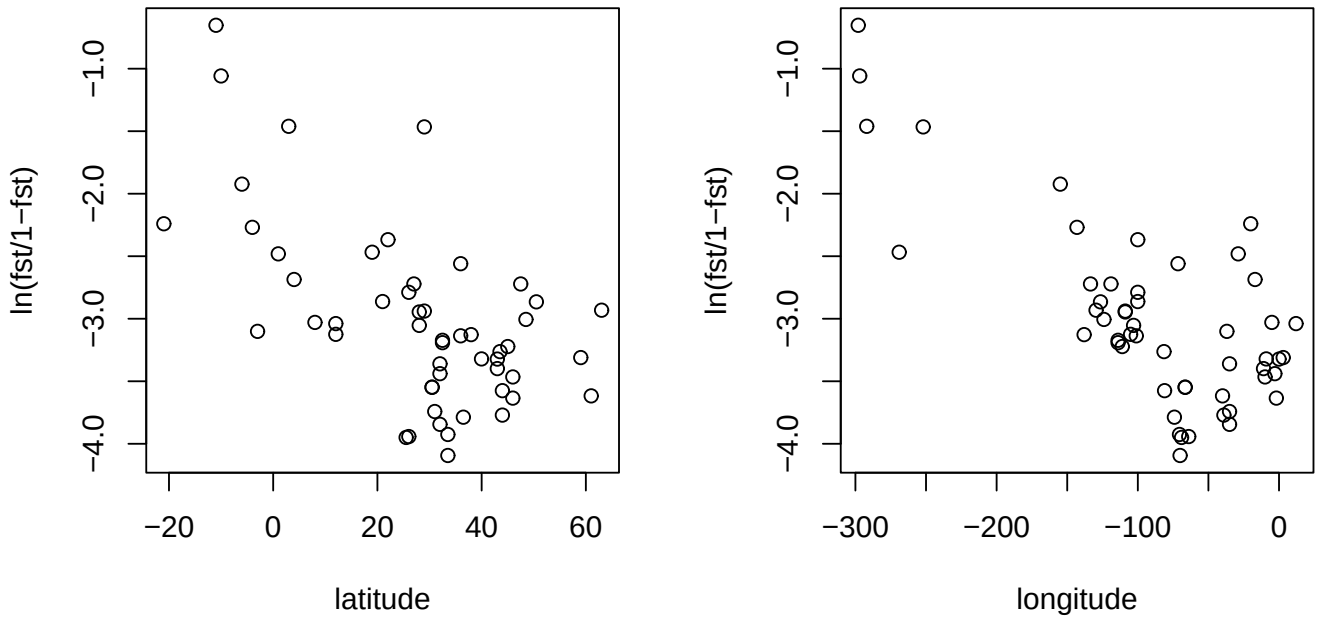

FIG. 2.9 – *Plots of $ln\ F_{ST}/(1-F_{ST})$ against the two factors : latitude and longitude for the human data.*

DISCUSSION

In this article we present a new Bayesian method to evaluate the effect that biotic and abiotic environmental factors (geographic distance, language, temperature, altitude, local population sizes, etc.) have on the genetic structure of populations. It estimates F_{ST} values for each local population and relates them to environmental factors using a generalized linear model. The method requires genetic data from codominant markers (e.g. allozymes, microsatellites, or SNPs) and environmental data specific to each local population. The results of our simulation study indicate that the method can correctly identify the environmental factors that influence the structuring of genetic variability under a wide range of conditions. In particular, it can perform very well for moderate values of sample sizes (20) and number of loci (10) and it is very

robust to deviations from the inference model assumed by our method. Thus, it should be applicable to a wide variety of species.

Our method provides many pieces of information that can help us better understand the demographic and evolutionary history of the species under study. Population specific F_{ST}s allow us to establish how distinct a local population is in terms of its genetic composition; a high F_{ST} indicates that the allele frequency distribution of a local populations is fairly different from that of the metapopulation as a whole. The posterior model probabilities allow us to identify the biotic/abiotic factors controlling genetic structure and also explain why a given local population is or is not genetically distinct. The signs of the regression coefficient estimates indicate if the effect of a factor increases or decreases genetic differentiation and, when the most probable model includes more than one factor, their absolute value tell us which factor has a stronger effect.

The method can be used to study a wide range of problems, including human genetics, conservation biology and species demographic history. In the field of human genetics it can be used to test for the importance of cultural factors as descriptors of the distribution of genetic variation in humans. This information is of fundamental importance for efforts to identify disease genes by association with marker loci. In the field of conservation biology, the identification of the factors that are important determinants of the genetic structure of natural population can help design management strategies for best preserving genetic biodiversity. It can also help gather valuable information for designing and parameterising simulation models aimed at predicting the effect of global climate change on natural populations. Its application to the study of population demographic history is well illustrated by the application of our method to the analysis of the HGDP-CEPH Human Genome Diversity Cell Line Panel (see below).

As opposed to the multivariate Mantel test, which uses a summary statistic (pairwise genetic distance), our method is model based and makes full use of the data (allele frequencies). This is a very important consideration when data sets are not very large, as is usually the case. Additionally, it avoids the problems faced by many frequentist approaches, most of which use randomisation techniques to test for the significance of the relationships. In the particular case of the multivariate Mantel test, Raufaste et Rousset (2001) and Rousset (2002) have recently called into question the validity of the jacknife procedure when two or more environmental factors are included in the analysis. The limitations of the sequential testing procedures were discussed previously by Cheverud *et al.* (1989) and particularly by Legendre (2000)

who described two additional randomisation methods for carrying out the multivariate Mantel test. Legendre (2000) also provided recommendations on ways of deciding which randomisation technique is more appropriate for the data at hand. Thus, the Multivariate Mantel test can still be used as a data exploration technique provided that all these considerations are taken into account. On the other hand, a more information-rich method, such as ours, is more suitable when the aim is to test specific hypothesis about the effect of environmental factors. It is also more appropriate for studying the influence of variables whose effects cannot be measured in terms of pairwise distance measures. For example the effect of local population sizes on genetic structuring cannot be modeled using pairwise distances since the effect of genetic drift is determined by the N_e of each local population and not by their difference. Our approach can readily account for this effect if estimates of local population sizes are available. Also, as shown in our simulation study, it can be used to study the effect of physical barriers such as water masses by considering a binary factor to discriminate between for example continental and insular populations.

The generality of our method stems from the fact that the likelihood function we use arises as an approximation to two different demographic models. In the case of the fission model considered by Nicholson *et al.* (2002) and Falush *et al.* (2003), the population specific F_{ST}s measure the degree of genetic differentiation between each descendant population and the ancestral population. In the case of the island model considered by Balding et Nichols (1995) and Balding (2003), the F_{ST}s measure how divergent each local population is from the metapopulation as a whole. In both cases, the estimation of individual F_{ST}s for each one of the descendant/local populations takes into account the fact that the strength of genetic drift differs among them, something that is overlooked by the traditional procedure of estimating a single F_{ST} value for the whole set of demes.

We illustrate the application of our method using published data sets. The Argan tree example approaches a migration-drift model and uses connectivity and altitude as factors that could explain the genetic structuring. There is no a priori reason to believe that altitude plays a role and our method confirms this expectation. Connectivity on the other hand has a strong effect. Although our example only considers abiotic factors, it is also possible to include biotic factors directly related to dispersal, for example the abundance of a pollinator species.

The human example illustrates the application of our method to a scenario that approaches imperfectly that of a fission model. Our analysis of the effect

of distance to East Africa on F_{ST} values (Figure 2.8) are not entirely consistent with those obtained by Prugnolle *et al.* (2005) who considered its effect on heterozygosity. In our case, although the overall fit of the model is very good, three African populations (Biaka Pigmies, Mbuti Pygmies and San) seem to deviate more from the F_{ST} values predicted by distance. This distinct behavior of African populations has also been observed by Ramachandran *et al.* (2005) using regression methods that did not explicitly assume a specific evolutionary model (but see below). Several factors can explain this lack of fit. The first one to consider is a mismatch between the fission model assumed by the method and the real demographic history of humans. Indeed, the fission model assumes that all subpopulations evolved from an ancestral population while the RAO model considers a series of bottlenecks with new populations being formed progressively from the ones that left the ancestral population. In order to determine if this is the source of the problem, we used SPLATCHE (Currat *et al.* 2004) to generate synthetic data under the RAO model (see Results). The simulations considered a population expansion from a single origin in East Africa in a two-dimensional habitat (strict two-dimensional stepping-stone model) and "collected" genetic data for 36 populations chosen uniformly on the map (c.f. Figure 2.5). As we did in the analysis of the real data set, we use distance from East Africa as a factor but also added a second factor, "insularity", to account for the barrier effect of water masses, which was expected to be very strong due to the way migration across water masses have to be modeled with SPLATCHE. The results (Figure 2.6) show that samples "collected" in Africa are no longer outliers, indicating that the mismatch between the inference model and the RAO model is unlikely to be responsible for the observed lack of fit.

Another potential explanation for the lack of fit of African populations is that they may have smaller effective sizes, a consequence of being rather small and isolated tribal groups. In their analysis, Prugnolle *et al.* (2005), used heterozygosity, a summary statistic that does not really take into account the information provided by very low frequency alleles. Thus, it is less affected by genetic drift due to small population sizes and provides a better fit. The fit of the model could have been improved by including estimates of local population sizes but no such estimates were available. Finally a third possible explanation is a potential ascertainment bias in the choice of the STRs of the database we used. Indeed, these markers were primarily ascertained on an European CEPH panel and may be biased towards loci that are more variable in European than African populations (Ray *et al.* 2005). This would introduce a bias in the estimation of the ancestral allele frequency distribution, which

will be closer to European populations. Prugnolle *et al.* (2005) used heterozygosity which, as mentioned before, is less sensitive to the effect of low frequency alleles and therefore is less influenced by ascertainment bias. We carried out simulations to determine to what extent ascertainment bias can influence the results given by our method. We considered the same scenario described above, i.e. 36 populations chosen uniformly on the world map, and used SPLATCHE to generate a dataset with 100 loci. From these we generated an unbiased dataset by selecting at random 20 loci, and a biased dataset by selecting 20 loci with the highest variability in a French sample. The effect of ascertainment bias can be studied using the averaged squared residuals (average squared difference between observed and fitted values) of the African samples, which in the case of the biased dataset is almost twice as high as that of the unbiased one (0.073 versus 0.042). This result suggest that ascertainment bias could explain at least part of the lack of fit of African samples. Another important effect of ascertainment bias is to skew the estimated allele frequencies of the ancestral population towards those of the population used as reference for choosing the markers. Thus, the lowest F_{ST} for the biased dataset value correspond to the French sample, which in the unbiased dataset ranked 16th. This can have a large effect on studies that attempt to narrow down on the region of Africa that was the epicenter of the human population expansion. This problem was investigated by Ray *et al.* (2005). They carried out realistic simulations of the genetic diversity expected after an expansion of modern humans into the Old World from different possible areas and then compared the results to the same human data set we have analyzed. Their results point towards a North African origin for modern humans. In order to uncover the effect of ascertainment bias they generated 20,000 data sets under the assumption of an East African origin. They then selected 1000 simulations with the highest variability in the subset of European samples (a biased data set) and 1000 additional simulations selected at random (an unbiased dataset). The results obtained with the biased samples indicated a North African origin while the unbiased samples correctly pointed towards an East African origin. In order to take into account the effect of ascertainment bias they used synthetic data sets that presented the same bias as the real human database. With this correction, the best fit between observed and simulated data were obtained for the scenario that considered an East African origin for modern humans. A better understanding of the demographic history of humans requires a more even geographic coverage of sampling efforts and less biased choice in the selection of the loci.

A statistical problem that arises when assessing the effect of geographic

distance from East Africa on human genetic diversity/differentiation is that human populations have a shared history of drift and cannot be considered as independent. For example, Amerindian populations have all been issued from the same bottleneck that took place at the time of colonisation of the American continent and they all bear its genetic signal. This is akin to the problem that Felsenstein (1985) pointed out for studies that use the comparative method and is common to all the recent studies that have used a regression approach to evaluate the validity of the RAO model (e.g. Prugnolle *et al.* 2005, Amos et Manica 2006, Ramachandran *et al.* 2005). More precisely, human populations can be considered as independent for the purpose of statistical analysis only under a strict fission model, in which case the history of genetic drift in the different human populations will be independent. Otherwise, populations in the different continents or regions of the globe will all bear the signal of the bottleneck at the time of colonisation of that particular continent/region. Thus, samples coming from the same geographical area can be considered, to at least some degree, as replicates and may inflate the statistical significance of the regression. Although we use a Bayesian approach, our method does not avoid this problem so we carried out simulations to try to evaluate to what extent this lack of independence can affect the results. The Amerindian populations are the most likely to influence the estimates of σ^2, the variance of θ's, and the posterior model probabilities. Thus, we compared the results of three different scenarios that differ in the number of Amerindian samples. In the first scenario samples were "collected" from 36 populations chosen uniformly on the world map (the unbiased scenario described above) ; they comprised four populations from North America, one from Central America and four others from South America. The second scenario considered 29 populations in total, with only one each in North, Central and South America. Finally, we considered a third scenario with 44 sampled populations of which four were from North America, one was from Central America and twelve were from South America. The results correspond to averages across 10 independent simulations of each scenario. They indicate that the number of Amerindian samples included has little effect on posterior model probabilities (0.97 for scenario 1, 0.93 for scenario 2, and 0.99 for scenario 3) and the estimates of σ^2 (0.15, 0.18, and 0.13 respectively). Duplicated populations can also have an effect similar to that of ascertainment bias (i.e. skew ancestral allele frequencies towards those of European populations) if there is a large number of samples from Europe and the Middle East. We explored this issue using simulations of a scenario were we added six European samples to the unbiased dataset. In this case, the estimated posterior proba-

bility of model 7 is 1 and the estimate of σ^2 is 0.17, indicating that the effect is rather weak. Additionally, the average F_{ST} of European populations for the dataset with more European samples ($F_{ST} = 0.059$) is almost the same as that obtained for the unbiased one ($F_{ST} = 0.061$). Ideally one would like to eliminate all non-independent samples as we did for the American continent but it is not clear how to identify the duplicated samples for large areas such as Asia, without knowing the history of the bottlenecks. In any case, the results of the simulations suggest that despite this problem our method seems capable of correctly identifying the effect of distance or other factors on genetic differentiation when the true model radically differs from the fission model.

The results of the application of our method to the study of human demographic history seem to suggest a rapid population expansion into the middle east and Europe as indicated by the low F_{ST} values obtained for all the populations in these regions. Indeed under a rapid population expansion, the effect of genetic drift is greatly reduced leading to low F_{ST} values. This rapid expansion seems to have been followed by a slower expansion into east Asia, Australasia and the Americas. There are clear longitudinal and latitudinal gradients in the degree of genetic differentiation of subpopulations (Figure 2.9) with values of F_{ST} increasing both toward the east and south, but with a stronger effect of longitude. This is expected because most population movements occurred along a west-east axis. Equivalent results were obtained by the seminal work of Piazza *et al.* (1981) in a much earlier study that considered only aboriginal populations.

As opposed to the complexity of frequentist multivariate techniques such as PCA or DCA, which provide results that are difficult to interpret, our Bayesian method provides results that are easy to interpret. It can be applied to the identification of the environmental factors that underlie observed spatial patterns of genetic diversity but it can also be used to study spatial population processes, such as range expansions, by simply introducing longitude and latitude as the explanatory variables. All these characteristics should make it a very valuable tool for addressing a wide range of problems.

Acknowledgments

We thank Peter Smouse, Robin Waples and Ilkka Hanski for very useful comments on the manuscript and Remy Petit for the Argan tree data. Daniel Falush and an anonymous reviewer provided detailed comments that helped to greatly improve the manuscript. This work was supported by the Fond Na-

tional de la Science (grant ACI-IMPBio-2004-42-PGDA). MF holds a PhD studentship from the Ministère de la Recherche. The software implementing the method is available at http://www-leca.ujf-grenoble.fr/logiciels.htm.

Appendix

Human genetic structure

Estimates of F_{ST} for the 51 human populations with 95% HPDIs. Italics identify populations with a F_{ST} significatively different from others in the same geographic area.

Population	F_{ST}	95% HPDI
America	**0.22**	
Karitiana (Brazil)	0.31	[0.27,0.35]
Surui (Brazil)	0.40	[0.36,0.45]
Piapoco/others (Colombia)	0.26	[0.22,0.30]
Maya (Mexico)	0.11	[0.09,0.12]
Pima (Mexico)	0.24	[0.20,0.27]
Oceania	**0.15**	
Melanesian (Bougainville)	0.16	[0.14,0.18]
Papuan (New Guinea)	0.14	[0.12,0.16]
East Asia	**0.07**	
Cambodian (Cambodia)	0.077	[0.061,0.094]
Dai (China)	0.072	[0.055,0.089]
Daur (China)	0.072	[0.055,0.089]
Han (China)	0.076	[0.061,0.092]
Hezhen (China)	0.091	[0.071,0.11]
Lahu (China)	*0.11*	[0.092,0.14]
Miaozu (China)	0.079	[0.060,0.097]
Mongola (China)	0.057	[0.042,0.071]
Naxi (China)	0.080	[0.062,0.099]
Oroqen (China)	0.076	[0.059,0.094]
She (China)	*0.10*	[0.081,0.13]
Tu (China)	0.061	[0.046,0.077]
Tujia (China)	0.086	[0.067,0.11]
Uygur (China)	*0.040*	[0.028,0.052]
Xibo (China)	0.056	[0.041,0.072]
Yizu (China)	0.069	[0.053,0.086]
Japanese (Japan)	0.069	[0.059,0.080]

Population	F_{ST}	95% HPDI
Eurasia	**0.03**	
Balochi (Pakistan)	0.034	[0.027,0.041]
Brahui (Pakistan)	0.032	[0.026,0.040]
Burusho (Pakistan)	0.027	[0.021,0.033]
Hazara (Pakistan)	0.023	[0.018,0.030]
Kalash (Pakistan)	*0.089*	[0.075,0.10]
Makrani (Pakistan)	0.023	[0.018,0.029]
Pathan (Pakistan)	0.026	[0.020,0.032]
Sindhi (Pakistan)	0.025	[0.019,0.031]
Yakut (Pakistan)	0.061	[0.051,0.073]
Basque (France)	0.040	[0.032,0.048]
French (France various regions)	0.034	[0.027,0.041]
Bergamo (Italy)	0.035	[0.025,0.045]
Sardinian (Italy)	*0.047*	[0.038,0.055]
Tuscan (Italy)	0.036	[0.022,0.050]
Orcadian (Orkney Islands)	0.037	[0.027,0.047]
Russian (Russia)	0.035	[0.027,0.042]
Adygei (Russia Caucasus)	0.030	[0.022,0.038]
Druze (Israel Carmel)	0.041	[0.034,0.047]
Palestinian (Israel Central)	0.026	[0.022,0.031]
Bedouin (Israel Negev)	0.030	[0.025,0.035]
Mozabite (Algeria Mzab)	0.033	[0.027,0.039]
Subsaharan Africa	**0.08**	
Biaka Pygmies (Central Africa)	0.083	[0.073,0.093]
Mbuti Pygmies (Congo)	*0.099*	[0.085,0.11]
Bantu Speakers (Kenya)	0.063	[0.050,0.074]
San (Namibia)	*0.12*	[0.10,0.15]
Yoruba (Nigeria)	0.069	[0.059,0.078]
Mandenka (Senegal)	0.058	[0.049,0.066]

Description of simulation study

The algorithm used in the simulations based on the inference model is given below in pseudo-code :

choose factors $G^{(1)}$ and $G^{(1)}$
choose the influence of factors α and the deviation from the regression σ^2
for each locus i in 1:I
    choose the number K_i of possible allelic states at locus i
    choose the alleles frequencies p_i of the ancestral population at locus i
endfor
for each population j in 1:J
    calculate $\mu_j = \alpha_0 + \alpha_1 G_j^{(1)} + \alpha_2 G_j^{(2)} + \alpha_3 G_j^{(1)} G_j^{(2)}$
    draw θ_j from $LogNormal(\mu_j, \sigma^2)$
    for each locus i in 1:I
        draw the alleles frequencies at locus i in population j $\widetilde{p_{ij}}$ from $\mathcal{D}ir(\theta_j p_{i1}, \ldots, \theta_j p_{iK_i})$
        choose the total number n_{ij} of alleles at locus i in population j
        draw the allele count a_{ij} at locus i in population j from $\mathcal{M}ultinomial\{n_{ij}; \widetilde{p_{ij1}}, \widetilde{p_{ij2}}, \ldots, \widetilde{p_{ijK_i}}\}$
    endfor
endfor

Supporting informations

MCMC methods

The Reversible Jump MCMC (RJMCMC) algorithm is used to estimate the posterior probability of the alternative models and their parameters. RJMCMC updates (Green 1995) provide a natural extension to the basic Metropolis Hastings algorithm to allow moves that change the dimension of the state vector. The estimates of model probabilities are obtained by simply observing the number of times that the chain visits each distinct model. A very thorough sensitivity study has been made to study the convergence of the method and the accuracy of the estimates under different data-set scenarios (sample sizes, number of loci and populations etc.).

Updating p

The frequencies are updated one locus at a time. In each case, the move is accepted with probability $\alpha(\boldsymbol{p}, \boldsymbol{p'}) = min(1, A)$, where :

$$A = \frac{L(\boldsymbol{p'}, \boldsymbol{\theta})\pi(\boldsymbol{p'})q(\boldsymbol{p'}, \boldsymbol{p})}{L(\boldsymbol{p}, \boldsymbol{\theta})\pi(\boldsymbol{p})q(\boldsymbol{p}, \boldsymbol{p'})}$$

The elements of the vector $\boldsymbol{p_i}$ are updated pairwise because they must sum to one. Thus, we select at random two different alleles m and n by choosing m uniformly from $\{1, \ldots, K_i\}$ and then choosing n uniformly from the remaining integers $\{1, \ldots, K_i\} \setminus \{m\}$. $\boldsymbol{p'}$ is the vector $\boldsymbol{p}$ with p_{im} and p_{in} replaced by p'_{im} and p'_{in}. Note that for the elements to sum up to one, $p_{im} + p_{in} = p'_{im} + p'_{in}$. Thus, we first take $p'_{im} = u$, where : $u \sim \mathcal{U}\left[max(0, p_{im} - e), min(p_{im} + p_{in}, p_{im} + e)\right]$ and e is some increment value which can be pilot-tuned to obtain reasonable acceptance rates. Finally, we set $p'_{in} = p_{in} + (p_{im} - p'_{im})$.

Updating α

The $\boldsymbol{\alpha}$ coefficients are updated one at a time. $\boldsymbol{\alpha'}$ is the vector $\boldsymbol{\alpha}$ with element α_l replaced by α'_l. The move is accepted with probability $\alpha(\boldsymbol{\alpha}, \boldsymbol{\alpha'}) = min(1, A)$, where :

$$A = \frac{\pi(\boldsymbol{\alpha'})\pi(\boldsymbol{\theta}|\boldsymbol{\alpha'}, \sigma^2)q(\boldsymbol{\alpha'}, \boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \sigma^2)q(\boldsymbol{\alpha}, \boldsymbol{\alpha'})}$$

We take the proposal value $\alpha_l' = \alpha_l + e$, where $e \sim \mathcal{N}(0, \sigma_\alpha^2)$. Since the proposal normal density is symmetric, $q(\boldsymbol{\alpha'}, \boldsymbol{\alpha}) = q(\boldsymbol{\alpha}, \boldsymbol{\alpha'})$ and thus, both terms drop from the expression for A.

Updating $\boldsymbol{\theta}$

The θ_js are updated one at a time. $\boldsymbol{\theta'}$ is the vector $\boldsymbol{\theta}$ with element θ_j replaced by θ_j'. The move is accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta'}) = min(1, A)$, where :

$$A = \frac{L(\boldsymbol{p}, \boldsymbol{\theta'})\pi(\boldsymbol{\theta'}|\boldsymbol{\alpha}, \sigma^2)q(\boldsymbol{\theta'}, \boldsymbol{\theta})}{L(\boldsymbol{p}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \sigma^2)q(\boldsymbol{\theta}, \boldsymbol{\theta'})}$$

We take the proposal value $ln\ \theta_j' = ln\ \theta_j + e$, where $e \sim \mathcal{N}(0, \sigma_\theta^2)$. So we have $ln\ \theta_j' \sim \mathcal{N}(ln\ \theta_j, \sigma_\theta^2)$ and here the move is not symmetric because of the log transformation : $q(\boldsymbol{\theta'}, \boldsymbol{\theta})/q(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \theta_j'/\theta_j$.

Updating σ^2

We take the proposal value $ln\ \sigma'^2 = ln\ \sigma^2 + e$, where $e \sim \mathcal{N}(0, \sigma_\sigma^2)$ and the move is accepted with probability $\alpha(\sigma^2, \sigma'^2) = min(1, A)$, where :

$$A = \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \sigma'^2)\pi(\sigma'^2)q(\sigma'^2, \sigma)}{\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \sigma^2)\pi(\sigma^2)q(\sigma, \sigma'^2)}$$

Here just like for $\boldsymbol{\theta}$ we have $q(\sigma'^2, \sigma^2)/q(\sigma^2, \sigma'^2) = \sigma'^2/\sigma^2$

Between-model moves

When moving between models, we add or delete one element to the $\boldsymbol{\alpha}$ vector. We randomly select one of these parameters and if it is already in the model, we propose to delete it. If it is not in the current model, we propose to add it. We use the canonical jump function so that when moving to a space of higher dimension (model 5 to model 7 for example), we keep the current values and we propose the new value from a distribution q. The reverse move is achieved by just deleting the parameter no longer included. Suppose we propose to add α_2 to the model and that we draw α_2 from q, then we accept this move with probability $min(1, A)$, where :

$$A = \frac{\pi(\boldsymbol{\theta}|\alpha_0, \alpha_1, \alpha_2, \sigma)\pi(\alpha_2)}{\pi(\boldsymbol{\theta}|\alpha_0, \alpha_1, \alpha_2 = 0, \sigma)q(\alpha_2)}$$

Because we choose the new model uniformly, the ratio of prior model probability simplifies to 1. The Jacobian is one because of the canonical jump

function used. To consider the reverse move, we simply accept the move deleting α_2 with probability $min(1, 1/A)$. All the other reversible jump moves are similar. The problem with this proposal scheme is that we might propose to add α_1 to a model without α_3. To get around this problem, we simply reject any proposal which doesn't make sense (for example α_0, α_1, α_3). The proposal distribution q is a normal distribution with its mean and variance pilot-tuned to improve convergence (see below).

Tuning proposal distribution

Proposal distribution have to be adjusted in order to have acceptance rates between 0.25 and 0.45. Thus, we have to adjust e for the uniform proposal of allele frequencies, and the variances σ_α, σ_θ and σ_σ for the normal proposals of α, θ and σ. If we propose values in a very wide interval, most moves will be rejected because we will often propose values in areas of low posterior probability. On the other hand, if we propose values very close to the current value, the move will be almost always accepted. These values are tuned on the basis of short pilot runs : we run 200 iterations, and for each parameter, the proposal is adjusted in order to reduce or increase the acceptance rate. We make 10 such pilot runs before starting the sampling, which also play the role of burn-in period. At the same time, we can choose the proposal distribution q for the reversible jump. Brooks *et al.* (2003) show that the best choice is to take $q(\boldsymbol{\alpha})$ to be the full conditional distribution of $\boldsymbol{\alpha}$ in model M'. Because we don't know this distribution, we use the pilot run to have an estimation of the mean m_i and variance v_i for all α_l under the saturated model (which corresponds to model 9). Then we propose new value for α_l from $\mathcal{N}(m_i, v_i)$ which is generally close to the full conditional distribution.

3 La structure génétique et les marqueurs dominants : intégrer le biais de recrutement

Parmi les nombreux marqueurs moléculaires existants, les AFLP sont devenus très populaires ces dernières années. En effet, cette technique permet d'obtenir rapidement et à moindre cout un grand nombre de marqueurs pour beaucoup d'organismes différents. Cependant, les AFLP ont le défaut d'être des marqueurs dominants, et il est ainsi impossible de distinguer les individus hétérozygotes des individus homozygotes pour l'allèle dominant. Pour cette raison, on ne peut pas les utiliser directement dans les méthodes basées sur l'estimation des fréquences alléliques, comme celle du chapitre précédent, sans faire d'hypothèse sur le taux de consanguinité. Holsinger *et al.* (2002) ont proposé une première approche pour résoudre ce problème, mais ils ont observé par la suite des estimations invraisemblables sur certains jeux de données, en particulier sur le taux de consanguinité. Dans ce chapitre nous montrons qu'une cause possible de ces mauvaises estimations est le biais de recrutement qui touche les AFLP. Nous montrons aussi que les techniques classiques pour prendre en compte ce biais ne peuvent pas s'appliquer dans un modèle hiérarchique tel que le modèle Dirichlet-multinomial. Enfin nous proposons une nouvelle approche de type « Approximate Bayesian Computation » (ABC) pour résoudre ce problème et estimer les coefficients F_{IS} et F_{ST} dans chaque population en tenant compte du biais de recrutement.

« *Un mathématicien est un aveugle qui, dans une pièce sombre, cherche un chat noir qui n'y est pas.* »

Charles Darwin

3.1 Problématique et démarche scientifique

De nombreux types de marqueurs moléculaires différents existent pour étudier la structure génétique. Les marqueurs codominants comme les allozymes, les microsatellites ou les SNP permettent d'accéder directement aux génotypes et peuvent ainsi être utilisés dans de nombreux logiciels (voir Excoffier et Heckel 2006, pour une revue détaillée). Par ailleurs, l'utilisation de marqueurs dominants conduit à de sérieuses difficultés à cause de l'impossibilité de distinguer les individus hétérozygotes des individus homozygotes pour l'allèle dominant. Néanmoins, ils sont devenus très populaires ces dernières années, avec en particulier l'émergence des marqueurs AFLP. En effet, ils représentent un moyen rapide et peu onéreux d'obtenir un grand nombre de marqueurs pour beaucoup d'organismes différents (Bensch et Akesson 2005, Meudt et Clarke 2007). Ils ont permis par exemple le premier criblage génomique pour une espèce non modèle (Wilding *et al.* 2001). Pour cette raison, il existe une forte demande pour pouvoir utiliser ces marqueurs dans les méthodes existantes.

Comme expliqué au chapitre 1, les marqueurs dominants ne permettent pas d'estimer directement les fréquences alléliques et cela nécessite d'inclure l'estimation du coefficient de consanguinité F_{IS}. Ce modèle bayésien proposé par Holsinger *et al.* (2002) a été implémenté dans le logiciel Hickory, mais les auteurs ont remarqué par la suite que la méthode fournissait sur certains jeux de données des estimations invraisemblables pour F_{IS} (voir le manuel de Hickory, Holsinger et Lewis 2002).

Dans ce chapitre nous proposons une méthode permettant de généraliser le modèle Dirichlet-multinomial présenté dans le chapitre 1 aux marqueurs dominants. En particulier, nous avons établi que le biais de recrutement qui affecte les AFLP permettait d'expliquer les estimations irréalistes de F_{IS} fournies par Hickory. De plus, l'à priori uniforme utilisé dans le modèle de Holsinger *et al.* (2002) sur les fréquences alléliques de la population ancestrale conduit à un biais sur F_{IS}. Nous montrons aussi que les méthodes généralement utilisées (par exemple Nielsen *et al.* 2004) pour prendre en compte le biais de recrutement ne peuvent pas s'appliquer dans le modèle hiérarchique considéré ici. Enfin nous proposons une méthode alternative de type « Approximate Bayesian Computation » (ABC) pour résoudre ce problème et estimer les coefficients F_{IS} et F_{ST} dans chaque population en tenant compte du biais de recrutement.

3.2 Contribution scientifique : article B

Foll, M., M. Beaumont, et O. Gaggiotti, 2007 An approximate bayesian computation approach to overcome biases that arise when using aflp markers to study population structure. Submitted to Genetics

3.2.1 Matériels et méthodes

La méthode proposée est basée sur l'extension aux marqueurs dominants du modèle Dirichlet-multinomial (Holsinger *et al.* 2002) détaillé au chapitre 1. Les AFLP sont des marqueurs bialléliques et la loi Dirichlet se simplifie alors en la loi beta, comme la loi multinomiale se simplifie en la loi binomiale. La différence principale est l'observation du nombre de phénotypes (de bandes) au lieu du nombre d'allèles. Ainsi le paramètre de la loi binomiale est la fréquence phénotypique au lieu de la fréquence allélique. On peut relier ces deux fréquences par une équation simple en introduisant le coefficient de consanguinité de chaque population qu'il faut alors estimer. En combinant l'information partagée par différents marqueurs, on peut en principe obtenir une telle estimation (Holsinger *et al.* 2002). Ici nous généralisons cette méthode en considérant des coefficients F_{ST} et F_{IS} différents dans chaque population.

Holsinger *et al.* (2002) ont utilisé un à priori uniforme, souvent considéré comme « non informatif » sur les fréquences alléliques de la population ancestrale. Pourtant, Wright (1931) a montré que cette fréquence allélique pouvait être approchée par une loi beta, dont les paramètres dépendent de la taille efficace et du taux de mutation (voir chapitre 1). Ici nous montrons qu'imposer un à priori uniforme conduit à des estimations biaisées des coefficients F_{IS}. Nous utilisons à la place un à priori plus général sous la forme d'une loi $Beta(a, a)$ symétrique où le paramètre a est estimé dans la méthode.

Différentes propriétés des marqueurs AFLP font que la distribution binomiale utilisée dans chaque population n'est pas une bonne modélisation du processus de découverte des marqueurs. Tout d'abord, si tous les individus sont homozygotes pour l'allèle récessif, aucune bande n'est observée et le marqueur ne sera pas identifiable. De plus, les marqueurs ne sont généralement pas choisis aléatoirement et on préfère seléctionner les marqueurs « polymorphes », c'est à dire pour lesquels certains individus possèdent la bande et d'autres pas. Enfin, il est souvent difficile de distinguer certaines bandes d'artéfacts, et on choisit alors de ne retenir que les marqueurs pour lesquels la fréquence de la bande est comprise entre deux bornes (par exemple 1%

et 99%). Nous montrons que ces différents biais de recrutement conduisent à des estimations fortement biaisées des coefficients F_{IS} si on les ignore, et dans une moindre mesure des coefficients F_{ST}.

Malheureusement, il n'est pas possible d'obtenir une fonction de vraisemblance simple avec ce biais de recrutement comme l'avait proposé Nicholson *et al.* (2002) pour les marqueurs SNP. Nous proposons alors une méthode alternative basée sur un algorithme de type « Approximate Bayesian Computation » (ABC). Cet algorithme permet d'estimer les distributions à postériori sans avoir besoin de calculer la fonction de vraisemblance. A la place, il est nécessaire de pouvoir simuler des données selon le modèle, ce qui est le cas ici. Cet algorithme a aussi l'avantage d'être fortement parallélisable et donc d'offrir une grande rapidité d'exécution sur des machines à plusieurs processeurs. Une étude de sensibilité a été conduite pour illustrer les possibilités de cette nouvelle méthode.

3.2.2 Résultats

La méthode ABC proposée permet d'obtenir des estimations non biaisées des coefficients F_{IS} et F_{ST} dans chaque population. L'étude de sensibilité montre qu'un grand nombre de loci sont nécessaires pour obtenir une distribution à postériori avec une faible variance en comparaison de ce que l'on obtient généralement avec les méthodes MCMC. L'algorithme ABC a été parallélisé, ce qui rend la méthode très rapide sur des machines multiprocesseurs.

La méthode a été implémentée dans un logiciel écrit en langage C++ avec une interface graphique conviviale, le rendant simple d'utilisation. L'interface permet aussi de visualiser les résultats sous forme graphique. L'algorithme ABC parallélisé tire partie des machines récentes comprenant plusieurs processeurs ou plusieurs cœurs. Le logiciel est diffusé librement sur le site internet du Laboratoire d'Ecologie Alpine à l'adresse `http://www-leca.ujf-grenoble.fr/logiciels.htm`

3.2.3 Discussion

Cette analyse a montré que l'estimation des coefficients F_{IS} à partir de marqueurs dominants était extrêmement sensible au choix non aléatoire des marqueurs. Plus précisément, le fait de ne pas retenir les marqueurs monomorphes, ou de ne retenir que les marqueurs dont le nombre d'individus possédant une bande se situe entre deux bornes, a un effet important. Différentes raisons techniques peuvent imposer ces choix et il est alors nécessaire

de les prendre en compte dans les estimations. De la même façon, il est important, une fois fixé le protocole de choix des marqueurs, de s'y astreindre en choisissant les marqueurs aléatoirement.

Nous avons démontré que les approches précédentes pour modéliser l'histoire démographique dans les populations structurées (Nicholson *et al.* 2002, Nielsen *et al.* 2004) ont utilisé des formules pour tenir compte du biais de recrutement qui sont problématiques et qui ne correspondent pas au processus biologique sous-jacent. Ainsi, la solution proposée ici illustre de manière plus générale l'utilité de l'algorithme ABC pour modéliser le biais de recrutement. En effet, il peut être difficile d'obtenir des fonctions de vraisemblance dans des cas complexes (voir par exemple Nielsen *et al.* 2004) alors qu'au contraire il est généralement simple de simuler le processus dans l'algorithme ABC.

Un défi important à relever pour les méthodes ABC est de trouver des statistiques sommaires exhaustives pour les paramètres à estimer. En pratique, de telles statistiques ne sont souvent pas disponibles en génétique des populations, mais des méthodes approchées peuvent fournir des approximations correctes (Tavaré *et al.* 1997). En général les études basées sur une méthode ABC utilisent des statistiques sommaires comme la moyenne ou le mode d'un paramètre donné (F_{ST}, la moyenne du déséquilibre de liaison entre paire de loci, le nombre moyen de différences entre deux séquences ADN etc.). Ici nous avons proposé d'utiliser des quantiles de la distribution d'une statistique sommaire et montré qu'ils fournissaient bien plus d'information qu'une unique estimation ponctuelle comme la moyenne ou le mode.

Conclusion

Nous avons identifié deux sources de biais qui affectent l'estimation des statistiques F_{IS} et F_{ST} avec les marqueurs AFLP dans le modèle Dirichlet-multinomial. En particulier nous avons montré qu'une approche classique du type MCMC ne permet pas de prendre en compte le biais de recrutement et qu'une distribution uniforme ne peut pas être considérée comme « non informative ». De plus notre analyse illustre le fait que les marqueurs monomorphes ne peuvent pas simplement être exclus avec l'idée qu'ils n'apportent aucune information. Nous avons introduit une nouvelle méthode qui fournit des estimations non biaisées des coefficients F_{IS} et F_{ST} dans chaque population en incorporant le biais de recrutement des AFLP. Ainsi notre méthode représente une amélioration importante par rapport à celles qui existaient jusqu'alors et sera très utile pour l'étude de la structure génétique à partir des marqueurs AFLP.

Chapitre 3 - Article B

Foll, M., M. Beaumont, et O. Gaggiotti, 2007 An approximate bayesian computation approach to overcome biases that arise when using aflp markers to study population structure. Submitted to Genetics

Abstract

There is great interest in using Amplified Fragment Length Polymorphism (AFLP) markers because they are inexpensive and easy to produce. It is, therefore, possible to generate a large number of markers that have a wide coverage of species genomes. Several statistical methods have been proposed to study the genetic structure using AFLPs but they assume Hardy-Weinberg equilibrium and do not estimate the inbreeding coefficient, F_{IS}. A Bayesian method has been proposed by Holsinger and colleagues that relaxes these simplifying assumptions but we have identified two sources of bias that can influence estimates based on these markers : (i) the use of a uniform prior on ancestral allele frequencies, and (ii) the ascertainment bias of AFLP markers. We present a new Bayesian method that avoids these biases by using an implementation based on the Approximate Bayesian Computation (ABC) algorithm. This new method estimates population specific F_{IS} and F_{ST} values and offers users the possibility of taking into account the criteria for selecting the markers that are used in the analyzes. The software is available at our website (http://www-leca.ujf-grenoble.fr/logiciels.htm). Finally, we provide advice on how to avoid the effects of ascertainment bias.

Key words : genetic structure, AFLP, Approximate Bayesian Computation, MCMC, ascertainment bias.

Running title : Genetic structure inferred from AFLPs

Introduction

Many if not most natural populations are spatially subdivided. Thus, the genetic diversity of a species is spatially structured into within- and between-components. This so called genetic structure has important implications for the evolution of species and its knowledge is fundamental for applications in the domains of conservation biology and genetic epidemiology. Genetic structuring is typically assessed using the so-called F-statistics first introduced by Wright (1951) who distinguished three statistics, F_{ST}, F_{IT} and F_{IS}. They have been widely used in population genetics but the interpretation of results have been difficult because of ambiguities about their definitions. Balding (2003) proposed a general framework to rigorously define them using the beta-binomial model proposed by Balding et Nichols (1995), which uses population specific estimates of F_{ST}. This new formulation, and in particular its multi-allelic version the multinomial-Dirichlet, has been used recently to address many different problems. Ciofi *et al.* (1999) used it to distinguish between two types of model of population structure and to estimate within-population F_{ST} coefficients, Falush *et al.* (2003) used it for clustering individuals into populations, Beaumont et Balding (2004) used it to identify candidate loci under natural selection and Foll et Gaggiotti (2006) used it to identify biotic/abiotic factors that are responsible for the observed spatial structuring of genetic diversity and to infer population history.

There are a wide variety of molecular markers available for studying genetic structure. The use of codominant markers such as allozymes, microsatellites, or SNPs lead to clearly distinguishable genotypes and, therefore, they can be readily analyzed using existing software (see Excoffier et Heckel 2006). On the other hand, using dominant markers leads to serious difficulties because of the inability to distinguish heterozygous individuals from those that are homozygyous for the dominant allele. Nevertheless, they have became very popular in the last decade, mostly due to the development of the Amplified Fragment Length Polymorphism (AFLP), an inexpensive and easy way of obtaining large number of genetic markers from a wide variety of organisms (Bensch et Akesson 2005, Meudt et Clarke 2007). It is therefore important to clearly understand the potential problems that may arise when dominant markers are used for the study of genetic structure. The main problem is that estimation of F-Statistics requires the allele frequencies to be inferred, which is not straightforward for dominant markers. AFLP's are in fact binary data, for each individual the information is "band-presence" or "band-absence", which can be viewed as a phenotype.

One possible solution is to suppose Hardy-Weinberg equilibrium to estimate allele frequencies but this imposes the strong hypothesis of no inbreeding. Indeed, this is what is assumed by most of the methods available (Lynch et Milligan 1994, Zhivotovsky 1999, Hill et Weir 2004). Simply taking the square root of the frequency of null homozygotes leads to a downward bias in the frequency of the null allele. The method proposed by Lynch et Milligan (1994) for RAPDs is applicable to AFLPs but, as indicated by Zhivotovsky (1999), also leads to a downward bias. Thus, this latter author proposed a Bayesian method that seems to perform better when departures from Hardy-Weinberg equilibrium are not strong. All these methods estimate allele frequencies and use them to subsequently calculate genetic diversity measures such as the heterozygosity. Thus, Hill et Weir (2004) propose a moment based method that simultaneously estimate allele frequencies and diversity measures, but this approach produces estimates with a high variance.

The only method that includes the estimation of the inbreeding coefficient is that of Holsinger *et al.* (2002). The inbreeding coefficient F_{IS} can be defined as the probability that two alleles in an individual are identical by descent. At the population level, we can view F_{IS} as the probability of sampling an individual inbred for a particular locus i. If we denote by $A1$ the dominant allele, with frequency p, and by $A2$ for the recessive allele, with frequency $q = 1 - p$, then, the dominant phenotype frequency $f([A1])$, can be linked to the allele frequency p and the inbreeding coefficient F_{IS} by :

$$f([A1]) = (1 - F_{IS})p^2 + F_{IS}p + (1 - F_{IS})2p(1 - p)$$

We have a similar relation between the phenotype frequency $f([A2])$ and the allele frequency q and F_{IS} :

$$f([A2]) = (1 - F_{IS})q^2 + F_{IS}q \tag{3.1}$$

For simplicity we next focus on this equation without loss of generality because $q = 1 - p$ and $f([A2]) = 1 - f([A1])$. The problem here is that we have only one equation with two unknown parameters and there is an infinite number of different combinations of q and F_{IS} that can give the same observed phenotype frequency $f([A2])$.

Holsinger *et al.* (2002) overcame this problem by considering multiple loci all of which share the same value of F_{IS}. The distribution of $f([A2])$ can be viewed as a mixture of outbred and inbred components, q^2 and q respectively, with respective mixture weights $1 - F_{IS}$ and F_{IS}. So the shape of the

phenotype frequency distribution gives information about F_{IS}. This phenotype distribution can be easily simulated because, as Wright (1931) showed, allele frequency distributions can be modeled using a Beta distribution. Thus, it suffices to choose the value of F_{IS} and draw the allele frequency from a Beta distribution to get the corresponding phenotype frequency from equation 3.1. As an example let us consider a population of 25 individuals with a migration rate of 0.01. This leads to an allele frequency that follows a Beta distribution with both parameters equal to 0.5. Figure 3.1 shows the resulting $[A2]$ phenotype frequency distributions as a function of the value of F_{IS} calculated with equation 3.1. For a given value of F_{IS}, the resulting distribution (figure 3.1b) is a mixture between the case $F_{IS} = 0$ (only outbred individuals, figure 3.1a) and the case $F_{IS} = 1$ (only inbred individuals, figure 3.1c). Note that figure 3.1c is also the distribution of allele frequencies ($Beta(0.5, 0.5)$) because in that case $f([A2]) = q$.

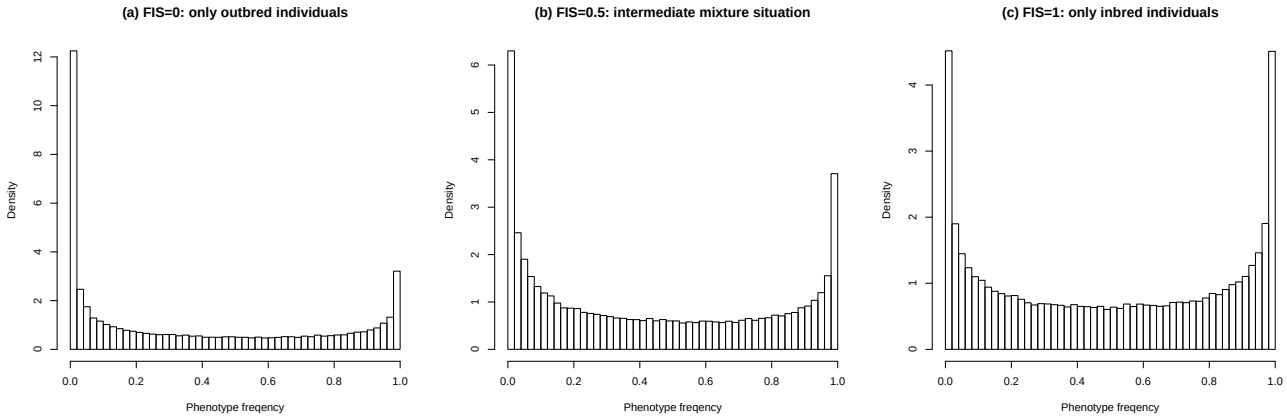

FIG. 3.1 – *Distribution of* $[A2]$ *phenotype frequency depending of the value of* F_{IS} *based on allele frequencies simulated from* $Beta(0.5, 0.5)$. *When* $F_{IS} = 0$ *(left), the distribution corresponds to the Hardy-Weinberg proportions* $f([A2]) = q^2$, *when* $F_{IS} = 1$ *the phenotype distribution is the same as the allele distribution because* $f([A2]) = q$. *An intermediate mixture situation (middle) is presented with* $F_{IS} = 0.5$. *These figures show how multiple dominant loci contain information about* F_{IS} *in the shape of the phenotype distribution.*

Using these principles, HOLSINGER *et al.* (2002) implemented a novel MCMC inference method in the software Hickory that can estimate both F_{IS} and F_{ST}. However, these authors noticed that sometimes the estimates of F_{IS} obtained were implausible based on detailed knowledge of the biology of the studied species (see latest version of the manual of Hickory (1.0.4)). This problem is due to the biases that affect the estimation of F_{IS} from dominant markers, and in particular AFLPs, mostly due to ascertainment in the choice of markers. The objective of this article is to thoroughly describe these problems and propose ways of avoiding them. In doing so we further extend the method to consider population specific F_{IS} and F_{ST} parameters.

In what follows, we first present the Bayesian formulation that we implement in our method and then describe the biases that we identified in the original version of Holsinger *et al.* (2002). We then propose a general solution by the way of an ABC approach and close by giving some suggestions on how to minimize estimation biases when using AFLP data.

The Bayesian model

The model for genetic differentiation used is based on ideas first introduced by Balding et Nichols (1995) (see Foll et Gaggiotti 2006 for a more detailed description of the different formulations leading to that model). Strictly speaking, the approach applies to an island model (Wright 1931) but it has also been used to describe a fission model (Falush *et al.* 2003). For the sake of simplicity we describe the details of our approach using the terminology of this latter model. We consider a collection of J subpopulations that evolved in isolation after splitting from an ancestral population. The extent of differentiation between subpopulation j and the ancestral population is measured by F_{ST}^{j} and is the result of its demographic history. We consider a set of I loci each one with two possible alleles $A1$ and $A2$ and we denote by p_i the frequency of allele $A1$ in the ancestral population at locus i. We denote by $\boldsymbol{p} = \{p_i\}$ the entire set of allele frequencies of the ancestral population and $\widetilde{\boldsymbol{p}} = \left\{\widetilde{p_{ij}}\right\}$ the current frequency of $A1$ at locus i for population j. Under these assumptions, the allele frequencies at locus i in subpopulation j follow a beta distribution with parameters $\theta_j p_i$ and $\theta_j (1 - p_i)$,

$$\widetilde{p_{ij}} \sim \mathcal{B}eta \left(\theta_j p_i, \theta_j (1 - p_i)\right) \tag{3.2}$$

where $\theta_j = 1/F_{ST}^{j} - 1$.

In the context of dominant markers, the data $\boldsymbol{N}$ consists of the sample counts of observed phenotypes instead of alleles counts. They are linked to alleles frequencies by equation 3.1, which includes the inbreeding coefficient F_{IS}^{j} for each population j. Let $n_{[A1],ij}$ and $n_{[A2],ij}$ be the observed number of phenotypes $[A1]$ and $[A2]$ at locus i for population j. The full data set is presented as a matrix $\boldsymbol{N} = \left\{n_{[A1],ij}, n_{[A2],ij}\right\}$ and the sample size at locus i for population j is $n_{ij} = n_{[A1],ij} + n_{[A2],ij}$. We can consider that the number of phenotypes $n_{[A1],ij}$ follows a binomial distribution with parameters $g_{[A1],ij}$ and n_{ij}, where $g_{[A1],ij}$ is the unknown $[A1]$ phenotype frequency at locus i in population j :

$$n_{[A1],ij} \sim \mathcal{B}inomial\left(g_{[A1],ij}, n_{ij}\right) \tag{3.3}$$

And we showed in the previous section that we can write :

$$\begin{aligned} g_{[A1],ij} &= \widetilde{p_{ij}}^2(1-F_{IS}^j) + F_{IS}^j\widetilde{p_{ij}} + \left(1-F_{IS}^j\right)2\widetilde{p_{ij}}\left(1-\widetilde{p_{ij}}\right) && (3.4)\\ g_{[A2],ij} &= \left(1-F_{IS}^j\right)\left(1-\widetilde{p_{ij}}\right)^2 + F_{IS}^j\left(1-\widetilde{p_{ij}}\right) && (3.5)\\ &= 1-g_{[A1],ij} && (3.6) \end{aligned}$$

Note that the binomial distribution is a particular case of the multinomial distribution and the beta distribution a particular case of the Dirichlet distribution both used for models with more than two alleles. If we assume independence we can multiply across loci and populations to obtain the likelihood function,

$$L\left(\widetilde{\boldsymbol{p}}, \boldsymbol{F_{IS}}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} P\left(n_{[A1],ij}|g_{[A1],ij}\right)$$

and the full prior distribution,

$$\pi\left(\widetilde{\boldsymbol{p}}|\boldsymbol{p},\boldsymbol{\theta}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J}\pi\left(\widetilde{p_{ij}}|p_i,\theta_{ij}\right) \tag{3.7}$$

Note that $g_{[A1],ij}$ and $g_{[A2],ij}$ are not parameters of the model because they can be calculated from equations (4.8) and (3.6) ; we use them only to simplify notation.

Up to here, our model only differs from that of Holsinger *et al.* (2002) in that we consider population specific F_{IS}^j and F_{ST}^j parameters. We now introduce an additional modification by assuming a prior for the allele frequency distributions that differs from the uniform used by them. More precisely, we use a $Beta(a,a)$ prior for p, where a is a hyper-parameter to estimate. The justification for this is Wright's (1931) observation that allele frequency distributions for bi-allelic loci can be approached by such a distribution. With these assumptions, the posterior distribution of the full model represented by the Directed Acyclic Graph (DAG) in figure (3.2) is given by :

$$\pi\left(\boldsymbol{p}, a, \boldsymbol{\theta}, \boldsymbol{F_{IS}}, \widetilde{\boldsymbol{p}}|\boldsymbol{N}\right) \propto L\left(\widetilde{\boldsymbol{p}}, \boldsymbol{F_{IS}}\right)\pi\left(\widetilde{\boldsymbol{p}}|\boldsymbol{p},\boldsymbol{\theta}\right)\pi(\boldsymbol{p}|a)\pi\left(\boldsymbol{F_{IS}}\right)\pi\left(\boldsymbol{\theta}\right)\pi(a) \tag{3.8}$$

We take non-informative priors for every F_{IS}^j and F_{ST}^j : $F_{IS}^j \sim \mathcal{U}\left[0,1\right]$ and

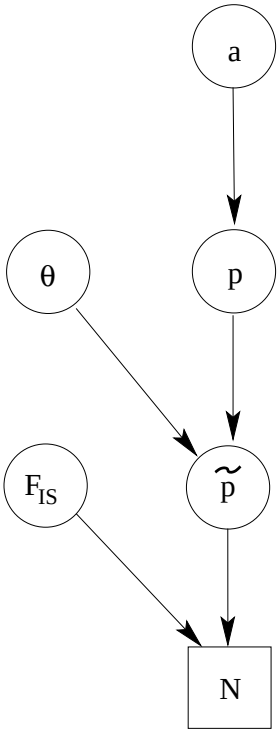

FIG. 3.2 – *DAG of the model given in equation 3.8. Square node denotes know quantity (i.e. data) and circles represents parameters to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text. N is the genetic data, $\boldsymbol{F_{IS}}$ is the vector of inbreeding coefficients, $\widetilde{\boldsymbol{p}}$ and $\boldsymbol{p}$ are respectively the actual and ancestral allele frequencies, θ is the vector of the genetic differentiation coefficient for each local population and a is the hyper-prior determining the shape of the ancestral allele frequencies.*

$F_{ST}^{j} \sim \mathcal{U}[0,1]$. Parameter a is scaled between zero and infinity so we use a log-normal distribution as prior : $a \sim logNormal(0,1)$. This Bayesian formulation was implemented both using a classical MCMC approach and the ABC approach proposed by BEAUMONT *et al.* (2002) and is described in detail below.

SOURCES OF BIAS

In what follows we describe two sources of bias that are introduced when analyzing AFLP data. The first one is due to the use of a supposedly "non-informative" prior for the allele frequency distributions and the other is due to the choice of markers to be included in the analysis (ascertainment bias). Because in what follows we present an ABC approach to solve both biases, in order to compare, we illustrate their influence here using our ABC implementation in which we did not implement the solutions of each bias.

Bias due to "non-informative" priors

Holsinger *et al.* (2002) followed the common practice of using a flat prior on all ancestral allele frequencies p_i. In this model, as we explained above, the information on F_{IS}^j is contained in the shape of the genotype frequency distribution and so, even if a uniform prior is generally called "uninformative", imposing here a flat prior leads to biased F_{IS}^j estimates if data sets (simulated or real) do not match this prior. Even if no information is available individually on frequencies, we have information on the general "shape" that allele frequencies should have in natural populations. As explained above, Wright (1931) showed that they can be approached by a beta distribution. For a single population (with no migration) and assuming low and symmetric mutation rates we obtain a "U-shaped" beta distribution with both parameters equal to $4N\mu < 1$, where N is the effective size and μ is the mutation rate. With migration, and assuming that mutation is negligible, we obtain a uniform distribution if the migration rate $m = 1/2N$, a U-shaped Beta distribution if $m < 1/2N$, and a bell-shaped Beta otherwise. Thus, we use a beta prior for p with both parameters equal to a, which has to be estimated : $p \sim Beta(a, a)$. We suppose that the distribution is symmetric, which is equivalent to assuming symmetric mutation rates and no selection. A more general prior would need a second parameter to estimate, but only little information is available on this hyper-prior and F_{IS}^j so using a second parameter would lead to more uncertainty.

We illustrate the improvement of this new hyper-prior by comparing the results of the full model introduced here where a is estimated and those of a model that use the same uniform prior ($a = 1$) as Holsinger *et al.* (2002). We consider a simple example with 5 populations and 100 loci, $F_{ST} = 0.1$, $F_{IS} = 0.2$ and 50 individuals in each population. We simulate 50 replicates of 3 different data sets : in the first one ancestral allele frequencies were simulated from $Beta(0.5, 0.5)$ ($a = 0.5$), in the second one from a uniform distribution ($a = 1$) and in the last one from a $Beta(2, 2)$ ($a = 2$) as in Holsinger *et al.* (2002). Results are presented in Figure 3.3 and were obtained using the software described below. We present results for only one of the five populations considered in our scenario because results for the remaining ones are very similar. First we observe that all boxplots for F_{IS} where a is estimated by our method are centered around the true value of 0.2. In the case where alleles frequencies are simulated from $Beta(2, 2)$ ($a = 2$) and a uniform prior is imposed, F_{IS}^j's are over-estimated around 0.33. The case where $a = 0.5$ seems to be less influenced by the uniform prior since F_{IS} is only weakly under-

estimated. When $a = 1$ the results are identical whether we estimate a or not. Finally when a is estimated, it appears that the accuracy of the estimates decreases as the parameter a decreases, this will be discussed further below.

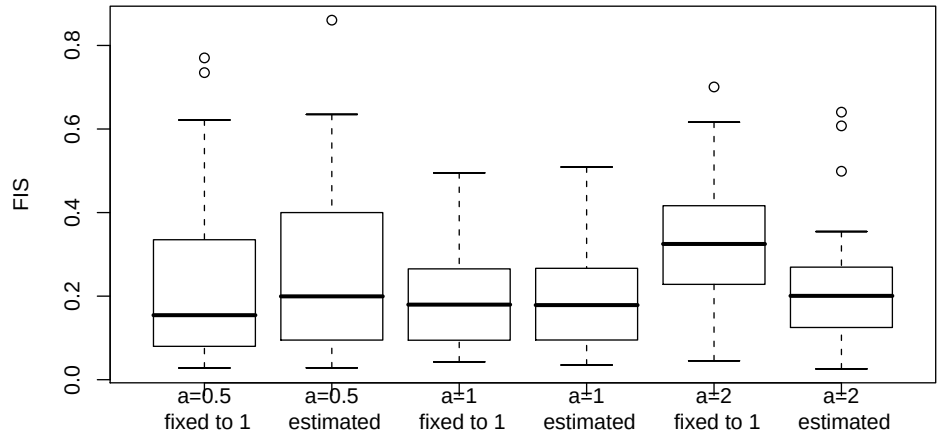

FIG. 3.3 – *Boxplot of the estimates of F_{IS} based on 50 replicates of 3 different data sets depending of the value of the hyper-prior parameter a using the full model presented in figure 3.2 (a estimated) and the original model presented in* HOLSINGER et al. *(2002) with a flat hyper-prior on ancestral allele frequencies (a fixed to 1). Data sets are based on 5 populations and 100 loci, $F_{ST} = 0.1$ and 50 individuals in each population, F_{IS} is fixed to 0.2. Boxes are are constructed using the lower quartile and the upper quartile in order that 50% of the values are in the boxes. The horizontal line in the box gives the median. The vertical lines are called the "whiskers" and indicate the minimum and maximum values but only if they lie within 1.5 times the box height (the interquartile range). Points outside the whiskers are outliers values and are plotted.*

Ascertainment bias

Besides the bias due to the choice of prior, there are also important biases due to intrinsic properties of AFLP markers and to the way these markers are chosen.

An important property of AFLP markers is that if all individuals have the recessive $[A2]$ phenotype, no band will be observed at all and we will not be able to identify this as a locus. As a result, we can never observe a locus i where all individuals have the $[A2]$ phenotype, which corresponds to $\sum_{j=1}^{J} n_{[A1],ij} = 0$ (we call this a hidden locus in the following). This is an intrinsic problem of AFLP markers and can not be avoided. The second is due to the way markers are chosen. In general, markers are not picked up at random and people prefer markers to be polymorphic with the intuition that

they will give more information on genetic diversity. For example Meudt et Clarke (2007) in a review on AFLP markers, suggest that : "a marker must be polymorphic (i.e. show both plus and null alleles) to be informative". It should be noted that what is called a fixed locus is a marker where all individuals have the same dominant $[A1]$ phenotype and $\sum_{j=1}^{J} n_{[A2],ij} = 0$, but this can reflect different genotypes ($A1A1$ or $A1A2$). Excluding non-polymorphic loci can dramatically change the shape of the phenotype distribution. This introduces a strong bias in the estimation of F_{IS} because, as discussed above and illustrated in figure 3.1, the information on F_{IS} is contained in the shape of the phenotype distribution.

There is yet another ascertainment bias that is introduced when choosing the loci that will be used in the analyses. In order to distinguish artifacts from "real" bands, people often fix arbitrary minimum and maximum numbers of individuals with the dominant phenotype $[A1]$ and only choose those loci for which the frequency of $A1$ lies within this interval. For example, some people exclude loci for which the frequency of the band is below 1% or above 99%. This procedure worsens the bias intrinsic to AFLPs that we described above. In order to incorporate it into the analysis, we introduce the notation hl ("hidden locus") to identify the lower threshold and fl ("fixed locus") to identify the upper threshold. Then, a locus i where $n_{[A1]i} = \sum_{j=1}^{J} n_{[A1],ij} < hl$ is called a hidden locus (almost no individuals have the $[A1]$ phenotype at locus i), and a locus i where $n_{[A2]i} = \sum_{j=1}^{J} n_{[A2],ij} < fl$ is called a fixed locus (almost all individuals have the $[A1]$ phenotype at locus i). Note that the intrinsic bias of AFLPs described at the beginning of this section sets a minimum lower bound of $hl = 1$.

The first consequence of these biases is that the observed phenotype frequencies are not actually drawn from a binomial distribution as assumed by equation 4.7. Faced with this, previous studies on single nucleotide polymorphisms (SNPs) have modified the likelihood by conditioning on only observing frequencies between fixed bounds (Nicholson *et al.* 2002, Nielsen *et al.* 2004). Using the same approach, this time for phenotype frequencies rather than allele frequencies, we can rewrite the likelihood as :

$$
\begin{aligned}
L\left(\widetilde{\boldsymbol{p}}, \boldsymbol{F_{IS}}\right) &= \prod_{i=1}^{I} P\left(n_{[A1],i1} \ldots n_{[A1],iJ} | g_{[A1],i1} \cdots g_{[A1],iJ}, n_{[A1]i} \geq hl, n_{[A2]i} \geq fl\right) \\
&= \prod_{i=1}^{I} \frac{P\left(n_{[A1],i1} \ldots n_{[A1],iJ} | g_{[A1],i1} \cdots g_{[A1],iJ}\right)}{P\left(n_{[A1]i} \geq hl, n_{[A2]i} \geq fl | g_{[A1],i1} \cdots g_{[A1],iJ}\right)} \\
&= \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} P\left(n_{[A1],ij} | g_{[A1],ij}\right)}{1 - P\left(n_{[A1]i} < hl | g_{[A1],i1} \cdots g_{[A1],iJ}\right) - P\left(n_{[A2]i} < fl | g_{[A1],i1} \cdots g_{[A1],iJ}\right)}
\end{aligned}
$$

The numerator is then the same product of binomial distributions as in the original likelihood function and the denominator can be calculated by considering all the possible cases, for example for the hidden loci (we have similar equations for fixed loci) :

$$
P\left(n_{[A1]i} < hl | g_{[A1],i1} \cdots g_{[A1],iJ}\right) = \sum_{k=0}^{hl-1} P\left(n_{[A1]i} = k\right)
$$

with :

$$
P\left(n_{[A1]i} = k\right) = \sum_{\substack{k_1 \ldots k_J \geq 0 \\ k_1 + \cdots + k_J = k}} \prod_{j=1}^{J} P\left(n_{[A1],ij} = k_j | g_{[A1],ij}\right)
$$

And then $P\left(n_{[A1],ij} = k_j | g_{[A1],ij}\right)$ is just a binomial density. However, as is demonstrated below, this modification of the likelihood does not correctly model the ascertainment process. This is most conveniently explained by considering the following algorithm for generating a sample that conforms to the model above.

Algorithm 3.

1. Simulate a from $logNormal(0,1)$.
2. For each population j in $1..J$:
   (a) Simulate F_{IS}^{j} from $\mathcal{U}[0,1]$.
   (b) Simulate F_{ST}^{j} from $\mathcal{U}[0,1]$ and calculate $\theta_j = 1/F_{ST}^{j} - 1$.
3. For each locus i in $1..I$:
   (a) Simulate allele frequency p_i in ancestral population from $Beta(a,a)$.
   (b) For each population j in $1..J$:
      i. Simulate allele frequency $\widetilde{p_{ij}}$ from $Beta(\theta p_i, \theta(1-p_i))$.
      ii. Calculate phenotype frequency $g_{[A1],ij} = \widetilde{p_{ij}}^2(1-F_{IS}^{j}) + F_{IS}^{j}\widetilde{p_{ij}} + \left(1-F_{IS}^{j}\right)2\widetilde{p_{ij}}\left(1-\widetilde{p_{ij}}\right)$.
   (c) For each population j in $1..J$:
      i. Simulate phenotype counts $n_{[A1],ij}$ from $Binomial(n_{[A1],ij}, g_{[A1],ij})$.
   (d) If $\sum_{j=1}^{J} n_{[A1],ij} < hl$ or if $\sum_{j=1}^{J} n_{[A2],ij} < fl$, go back to 3c.

It can be seen that the likelihood correction suggested by Nicholson *et al.* (2002) and Nielsen *et al.* (2004) implies a rather peculiar model for discovering loci : if a locus does not conform to the discovery criteria it will be discarded, and the next locus will have *exactly the same* parametric population frequencies as the one discarded. This process will continue until a locus is accepted. Intuitively this is not a reasonable process. Rather, once a locus is discarded the next locus should be drawn with completely independent frequencies in all populations. Thus, at step 3d the algorithm should move to step 3a. Unfortunately, we have not been able to provide an analytical expression for this biologically more realistic ascertainment model, but demonstrate that this is possible using likelihood-free inference (Beaumont *et al.* 2002, Marjoram *et al.* 2003).

Two main problems arise from the use of the correction of Nicholson *et al.* (2002) and Nielsen *et al.* (2004). Firstly there is a violation of the assumption of statistical independence among the allele frequency distributions of the different populations implicit in equation 4.3, and secondly, the ascertainment process modifies the distribution of ancestral allele frequencies.

As an illustration of the first effect, for a given ancestral allele frequency, we can simulate a large number of replicates of allele frequencies and num-

ber of bands in actual populations form the exact model, and then estimate the correlation coefficient between these sets of frequencies before and after having applied the ascertainment process described above with $hl = 1$ and $fl = 1$. We do this for 2 populations with 30 individuals and $F_{ST} = 0.2$ in each population for $F_{IS} = 0.1$ and $F_{IS} = 0.9$. The correlation coefficients are plotted on figure 3.4 against the value of the ancestral allele frequency. For unbiased data sets, the correlation is around zero because allele frequencies are independent in the two populations. But as expected, low and high ancestral allele frequencies produce a high correlation for biased data sets. For example when the ancestral frequency is low (respectively high), if the allele frequency is close to zero (resp. one) in the first population, it is likely to be far from zero (resp. one) in the second one because the locus was not hidden (resp. fixed).

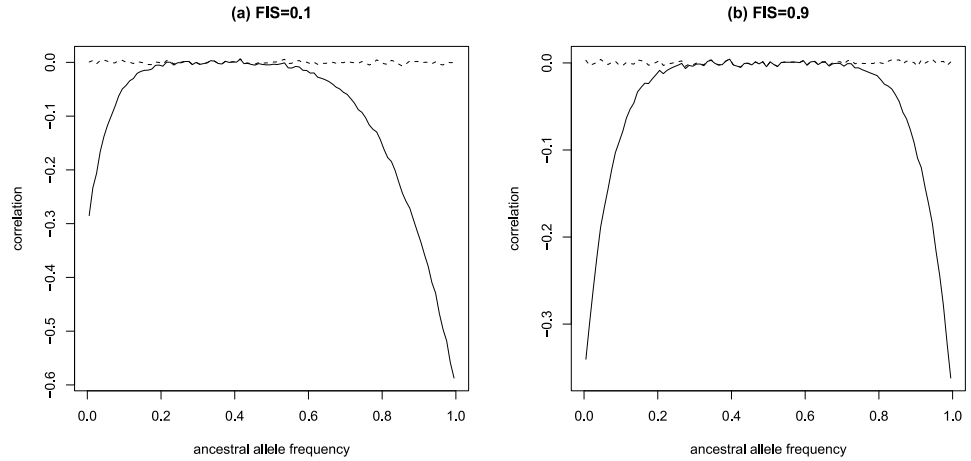

FIG. 3.4 – *Plot of the correlation coefficient between allele frequencies of 2 populations for unbiased (dashed lines) and biased (with* $hl = 1$ *and* $fl = 1$*, solid lines) data sets against the ancestral allele frequency. We simulated 30 individuals and* $F_{ST} = 0.2$ *in each population for* $F_{IS} = 0.1$ *(a) and* $F_{IS} = 0.9$ *(b).*

The second effect, the modification of the distribution of ancestral allele frequencies, arises because when we remove hidden and fixed loci, we remove at the same time the ancestral allele frequencies that produced them. For example, as hl increases (resp. fl), the probability of observing low (resp. high) ancestral allele frequencies decreases because they are more likely to produce hidden (resp. fixed) loci. As an illustration, with simulated data sets, we can draw the distribution of ancestral allele frequencies after having applied the ascertainment process because we know their true values. For this, we first simulate an unbiased large number of loci with ancestral allele frequencies drawn from a Beta distribution with $a = 0.5$. Then for each locus, we simulate the allele frequencies in 5 populations with $F_{ST} = 0.2$ from the beta distribution of equation 3.2. Finally for each locus in each population we

draw the corresponding number of bands observed for 30 individuals and $F_{IS} = 0.2$ from the Binomial distribution of equation 4.7. By this way, we know for each locus the true value of allele frequencies in each population and in the ancestral population. After that, we remove all hidden and fixed loci from this data set using the ascertainment process described above with $hl = 3$ and $fl = 3$ to obtain a biased data set. This allows us to plot the distribution of ancestral allele frequencies in the biased data set because we know the true values of each ancestral allele frequency in this simulated data set (we do not need to estimate them). We plot these distributions for both unbiased and biased data sets on figure 3.5. We can see that as expected, the loci with low and high frequencies are less likely to appear in the biased data set than in the original one.

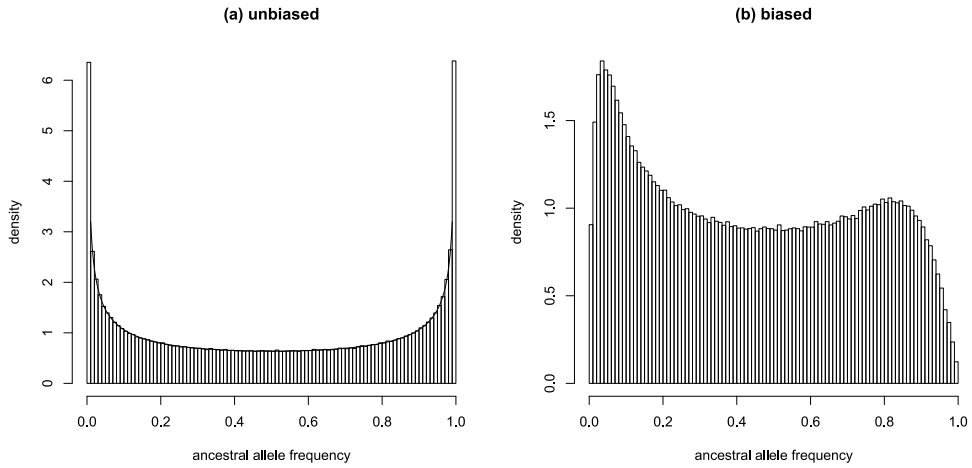

FIG. 3.5 – *Simulated distributions of ancestral allele frequencies for an unbiased data set (a) and a biased data set (b). The unbiased data was generated from the exact model with* $a = 0.5$, *5 popluations, 30 individuals,* $F_{IS} = 0.2$ *and* $F_{ST} = 0.2$ *in each population. Then we applied the ascertainment process to this data set using* $hl = 3$ *and* $fl = 3$ *to obtain the biased data set.*

Ignoring these effects, and continuing to use the modified likelihood function, following the approach of NICHOLSON *et al.* (2002) and NIELSEN *et al.* (2004), leads to strong biases in estimation. We illustrate the influence of the bias using a typical example that may be problematic : we consider 5 populations, a sample size of 30 individuals per population, and a high differentiation coefficient $F_{ST} = 0.25$ in each one. Ancestral allele frequencies are simulated from a "U-shaped" beta distribution with parameter $a = 0.7$. We simulate two series of data sets : the first one with $F_{IS} = 0.8$ and the second one with $F_{IS} = 0.2$ in each population. In order to introduce the ascertainment bias, we imposed the constraint that at each locus there should be at least hl and at most fl individuals with the band, and we generated datasets

with 100 loci. In each of the two series we simulated 50 replicates of different data sets with hl and fl varying independently from 0 to 3 (3 corresponds to 2% of the total number of 150 individuals). The results obtained for each of the 5 populations are very similar so we present the results for only one of them in figure 3.6. When no bias is introduced by the exclusion of loci from the analysis (i.e. $hl = 0$ and $fl = 0$), the boxplots are centered around the true values of 0.8 and 0.2. On the other hand, when hl is positive, F_{IS} is underestimated and when fl is positive, F_{IS} is overestimated. As expected, the bias is maximal for $F_{IS} = 0.2$ and $hl > 0$ because there is a very large number of fixed loci (see figure 3.1). The bias is strong even for $hl = 1$ or $fl = 1$, however, increasing these values further has a little effect on the estimates.

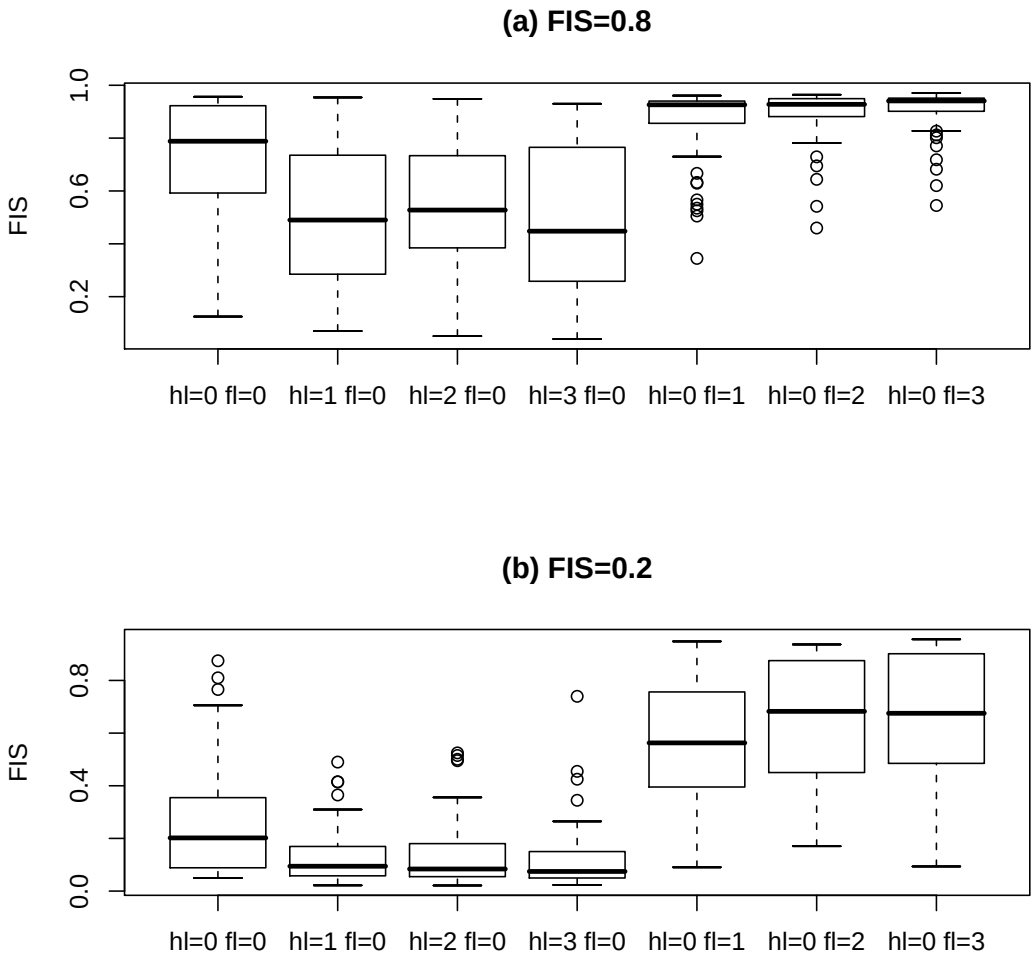

FIG. 3.6 – *Comparison of the estimates of F_{IS} based on 50 replicates of different data sets with the ascertainment bias hl and fl varying independently from 0 to 3. Estimates are made under the assumption that there is no ascertainment bias (supposing that $hl = 0$ and $fl = 0$). In (a) F_{IS} is fixed to 0.8 and in (b) F_{IS} is fixed to 0.2. Simulated data sets consist of 5 populations and 100 loci, $F_{ST} = 0.25$ and sample size of 30 individuals per population.*

A solution : an approximate Bayesian Computation (ABC) approach

The solution we propose to overcome the two biases described above is to use the ABC algorithm of Beaumont *et al.* (2002) instead of the classic MCMC scheme. The most valuable advantage of the ABC for our problem is that it does not require a closed form for the likelihood and internal priors (this will allow to include the ascertainment bias in the model). In addition the ABC algorithm has the advantage of being highly parallelizable. This is an important consideration because of the emergence of multicore processors and the availability of calculation clusters. The ease with which the model is simulated and the fact that the ABC algorithm is highly parallelizable makes the method proposed here very fast compared to the MCMC version and allows a very detailed sensitivity study.

The ABC algorithm is a rejection sampler : it generates data sets from given parameter values and accepts them when the simulated data set is "close" enough to the real data set. In the ABC framework, large data sets are reduced to a vector of summary statistics and "close" means that the distance between these statistics is below a given threshold value. If we have a data set D generated from a model M determined by parameters ϕ and the prior density is noted $\pi(\phi)$, the algorithm is given by :

Algorithm 4.

1. Choose a set of summary statistics S that will represent the data, and calculate s for D.
2. Generate ϕ from $\pi(\cdot)$.
3. Simulate D' from M with parameters ϕ and calculate s' for D'.
4. Accept ϕ if $\|s - s'\| < \delta$ and return to 2.
5. Stop when sufficient data sets have been accepted.

Our approach also implements the local linear regression method proposed by Beaumont *et al.* (2002) that greatly improves the accuracy of the estimates for higher acceptance rates. In our case, simulating from the model M is very easy, including simulating the biases discussed above. Note that the values of hl and fl need to be known. The complete algorithm for steps

2 and 3 is the one presented above, with step 3d modified to move back to step 3a instead of step 3c in order include the realistic ascertainment model we described.

In the ABC algorithm, the summary statistics are of primary importance and their choice determine directly the accuracy of the final results. Here we use different statistics to estimate F_{IS}^{j} and F_{ST}^{j} coefficients as proposed by Hamilton *et al.* (2005). Since the information about F_{IS}^{j} is contained in the shape of the phenotype frequency distribution of population j (see introduction), we use the n-quantiles of these distributions as summary statistics for each population. They are representative values of the shape of a distribution : points are taken at regular intervals from the cumulative distribution function. For example the five 6-quantiles of the distributions presented in figure 3.1 are, from left to right : $(0.005, 0.06, 0.25, 0.57, 0.87)$; $(0.04, 0.16, 0.38, 0.66, 0.90)$ and $(0.07, 0.25, 0.50, 0.75, 0.93)$. These vectors reflect well that the two first distributions are skewed to the left (but the effect is less pronounced in the second case), and that the third distribution is symmetrical. In the context of the fission model (resp. island model), the F_{ST}^{j} measure how divergent each local population is from the ancestral population (resp. from the metapopulation as a whole). For this reason we calculate the global phenotype frequency at each locus i as $g_{[A1],i} = \sum_{j=1}^{J} n_{[A1],ij} / \sum_{j=1}^{J} n_{ij}$ and we define the observed phenotype differentiation for population j at locus i as :

$$D_{ij} = \frac{g_{[A1],ij} - g_{[A1],i}}{g_{[A1],i}}$$

Then for each population j the summary statistics used are the n-quantiles of the distribution of D_{ij} among loci. If $g_{[A1],i} = 0$ we set $D_{ij} = 0$ because all populations will also have $g_{[A1],ij} = 0$.

Sensitivity study

The method has been implemented in a software written in C++. We provide a command line version for both Linux and Microsoft Windows operating system and a Graphical User Interface for the Windows version. The ABC algorithm is well adapted for parallel computing : with an acceptance rate of 0.005 and a sample size of 5000, the algorithm will simulate 1 000 000 independent data sets in steps 2 and 3. They can be generated independently on different computers or processors. We implemented the ABC algorithm on a computer cluster composed of 72 Itanium processors at 1.6 Ghz. As an example, it takes less than 15s for 48 processors to simulate 1 000 000 samples

for a data set composed of 5 populations and 100 loci. Multicore processors are now available on desktop computers and for example on a 2.66 Ghz quad core processor the same simulation would take less than 2 minutes. For each data set we present results based on 50 replicates of the same scenario.

We estimate parameters using 5000 independent samples from the ABC algorithm. We use the mode as a point estimate for the posterior distributions and estimate it using a Gaussian density kernel. Multi-parameter least square fitting for the local linear regression is performed using the GNU Scientific Library Galassi *et al.* (2006). The value of δ used in step 4 of the algorithm presented above is pilot tuned in a shorter run using a target acceptance ratio chosen by the user. For example, an acceptance rate of 0.01 means that the 1% of simulated s' that are closest to s are accepted in step 4.

ABC algorithm parameters : The algorithm we introduce requires the user to set the acceptance rate for the rejection algorithm. In general, smaller acceptance rates give more accurate results but also increases computation time because the user is forced to generate a larger number of data sets in order to obtain enough of them to estimate the parameters. Thus, it is important to investigate the influence of this parameter on the estimates. We simulated 50 synthetic data sets in which all local populations had the same values of $F_{IS} = 0.5$ and $F_{ST} = 0.15$ and estimated these parameters for each one of them using acceptance rates varying between 0.001 and 0.1. The results are illustrated in figures 3.7a and 3.7b, which show the relative mean square error (RMSE) of the estimates of F_{IS} and F_{ST} for one of the populations against the acceptance rate. We calculated the RMSE using $1/50 \sum_{n=1}^{50} (\widetilde{\phi_n} - \phi)^2 / \phi^2$, where ϕ is either $F_{IS} = 0.5$ or $F_{ST} = 0.15$. As expected, a lower acceptance rate give lower RMSEs but the effect of this parameter is not very strong : multiplying the acceptance rate by 100 (0.001 to 0.1, which makes the calculation 100 time faster) only increases the RMSE by 16% for F_{IS} (from 0.061 to 0.071) and doubles the one of F_{ST} (0.03 to 0.06). We also investigated the effect of the number of quantiles used for the summary statistics. Results are presented in figures 3.7c and 3.7d. The influence of the number of quantiles on RMSE is also small. Interestingly, for F_{IS} there is an optimal number of 15 quantiles while for F_{ST} RMSE decreases first very rapidly and then very slowly ; not much is gained by using values larger than 25. Thus, we used these values for the number of quantiles and an acceptance ratio of 0.005 for all the results we present below.

Ascertainment bias : In order to show that the ABC algorithm can efficiently solve the problem posed by ascertainment bias, we use the same data sets used to plot figure 3.6. The bias is corrected fairly well in all the scenarios

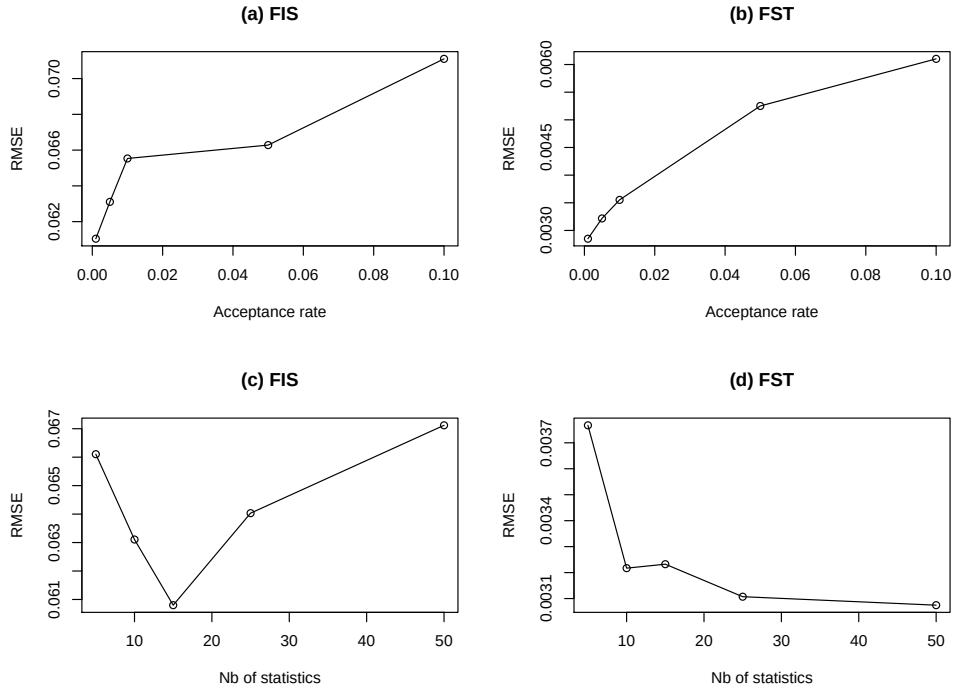

FIG. 3.7 – *Plot of RMSE for estimates of F_{IS} and F_{ST} against the two parameters of the ABC algorithm : the acceptance rate and the number of quantiles used.*

we explored (figure 3.8). Positive values of fl are the most problematic, especially when $F_{IS} = 0.2$. In this case it is clear that removing fixed loci leads to a loss of information for F_{IS} estimation as expected from figure 3.1. Note that the only intrinsic limitation of AFLP markers is $hl \geq 1$, thus, this bias can be avoided simply by including all the fixed loci in the analyzes.

Size of the data set used : We simulated many different data sets in order to investigate which kind of data can give the best results with AFLP. The starting point is a scenario with 100 loci, 5 populations and 50 individuals per population. Then we modified each one of these parameter at a time and calculated the RMSE based on 50 replicates of each data set. We fixed $F_{ST} = 0.25$ and $F_{IS} = 0.5$ in each population, and choose $a = 0.7$. We also included ascertainment bias with $hl = 1$ and $fl = 0$. We present the RMSE for F_{IS} and F_{ST} for data sets containing 50 to 500 loci in table 3.1. As expected increasing the number of loci greatly reduces the RMSE. More precisely, the RMSE is reduced by a factor of 4.2 for F_{IS} and by a factor of 11.6 for F_{ST}.

Number of loci :	50	100	200	500
F_{IS}	0.137	0.102	0.052	0.033
F_{ST}	0.0116	0.0044	0.0035	0.0010

TAB. 3.1 – *RMSE of the estimates of F_{IS} and F_{ST} using data sets with different number of loci.*

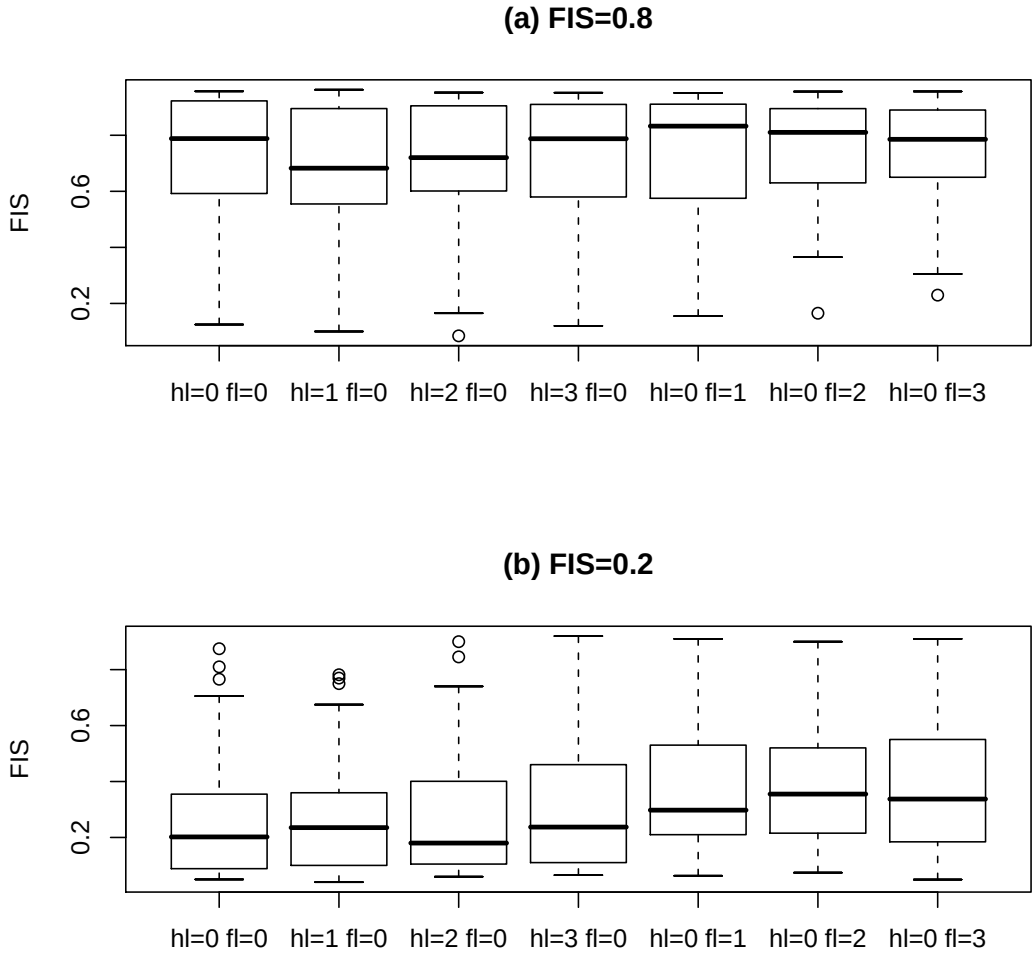

FIG. 3.8 – *Comparison of the estimates of F_{IS} based on the same simulated data sets presented in figure 3.6. There are 50 replicates of different data sets with the ascertainment bias hl and fl varying independently from 0 to 3. Estimates are made taking into account the ascertainment bias in the ABC algorithm presented in the text. In (a) F_{IS} is fixed to 0.8 and in (b) F_{IS} is fixed to 0.2. Data sets are based on 5 populations and 100 loci, $F_{ST} = 0.25$ and 30 individuals per population.*

Increasing the number of individuals per population is much less helpful than increasing the number of loci. There is no significant improvement for F_{IS} with 30, 50 or 100 individuals per population, and for F_{ST}, the RMSE is only reduced by 30% (0.0073 to 0.0051 ; data not shown). In terms of number of populations considered, results for F_{IS} are similar, the RMSE does not change much when the number of populations changes. However, the RMSE of F_{ST} is divided by 3 (0.0049 to 0.0016) when the number of populations is increased from 5 to 50 (data not shown). This can be easily explained by the fact that F_{ST} estimates are based on the estimation of the ancestral population

(resp. metapopulation) allele frequencies, which are better estimated with a large number of populations.

Influence of biased F_{IS} estimations on F_{ST} coefficients : The biases we described above do not have a direct effect on F_{ST} estimates but they can influence them indirectly if they lead to highly biased estimates of F_{IS} simply because then allele frequency distributions will also be biased. In order to show this effect, we simulated 50 replicates of two data sets : one with $F_{IS} = 0.2$ and the other with $F_{IS} = 0.8$ in each population. For both scenarios the simulated data sets considered 200 loci, 10 populations and 50 individuals per population, with $a = 0.7$ and F_{ST}=0.15 in each population. For both scenarios we used the new ABC algorithm presented here were F_{IS} is estimated, and the MCMC algorithm were instead of estimating F_{IS} we used a fixed value. We first fixed the value of F_{IS} to the true value (0.2 or 0.8) and then to the worst case possible false one (1 for the case of $F_{IS} = 0.2$ and 0 for the case of $F_{IS} = 0.8$). Results are presented in figure 3.9. It is clear that both the ABC algorithm and the MCMC algorithm with the correct value of F_{IS} give correct estimates of F_{ST} centered around the true value (0.15). However, the estimates of F_{ST} are clearly biased when using biased estimates of F_{IS}. F_{ST} is over-estimated when F_{IS} is over-estimated and F_{ST} is under-estimated when F_{IS} is under-estimated. Finally it is important to note that the ABC algorithm gives wider posterior density intervals for F_{ST} than the MCMC algorithm, but the latter can only be used if one knows the true value of F_{IS}.

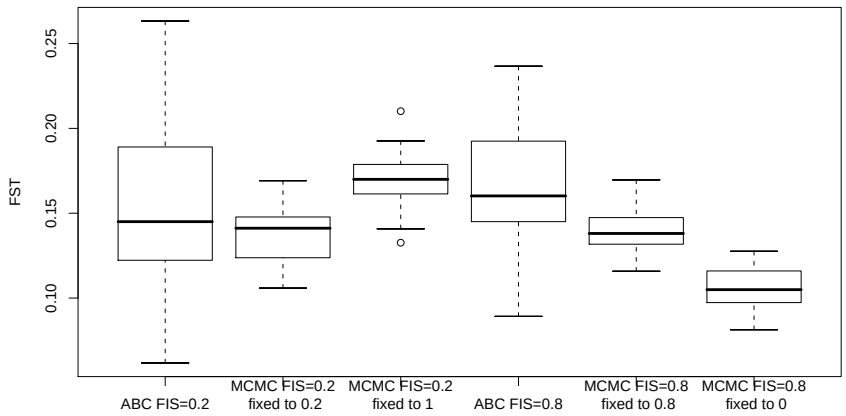

FIG. 3.9 – *Comparison of the estimates of F_{ST} based on 50 replicates of two data sets : one with $F_{IS} = 0.2$ and the other with $F_{IS} = 0.8$ in each population. We estimated F_{ST} with the ABC algorithm which estimates F_{IS} and with the MCMC algorithm with a fixed value of F_{IS} : either the true value or a false value.*

Discussion

In this article we identify two sources of bias that affect the estimation of F-statistics when using dominant markers and particularly AFLPs. More specifically, we show that when estimating inbreeding coefficients using dominant markers, (i) the use of MCMC techniques can not take into account the ascertainment process, (ii) flat priors for allele frequency distributions cannot be considered as "non-informative" , and (iii) monomorphic loci should be included in the analyzes whenever this is possible. In order to avoid these biases, we presented a new statistical method based on the ABC algorithm. Additionally, our method estimates population specific F_{IS} and F_{ST} coefficients and incorporates parameters to model ascertainment bias.

Our approach takes into account the fact that loci for which a band is not observed constitute hidden loci; it suffices to set $hl = 1$ (at least one individual have the band). Also, our study allows us to provide guidance to researchers developing AFLP markers. The common practice of excluding loci for which the frequency of the band is very low or very high should be avoided whenever possible. Sometimes it is necessary to impose stringent threshold values for technical reasons. For example a loci will be included only if at least 5 individuals have the band ($hl = 5$ and at least 5 don't have it ($fl = 5$). In these cases, our method will still be capable of giving unbiased estimates (c.f. figure 3.8) but only if one knows the values of hl and fl that were used by the people developing AFLPs. Thus, it is very important to choose the conditions under which a given loci will be included in the analyzes and then apply them in a consistent manner. These requirements may represent a problem when analyzing old datasets for which these values are not known. Thus, one desirable extension of our method would be to incorporate the uncertainty in hl and fl within the simulation step of the ABC algorithm. However, we note that doing this would lead to an increase in the RMSE of the estimations.

This article highlights the usefulness of the ABC algorithm for modeling ascertainment bias. We have demonstrated that earlier approaches to modeling demographic history in structured populations (Nicholson *et al.* 2002, Nielsen *et al.* 2004) have used formulae for taking into account ascertainment that are demonstrably problematic and do not conform to the underlying biological processes. Obtaining closed form solutions for the likelihood function when loci are not chosen at random can be very difficult. Using the ABC approach overcomes this problem and we demonstrate that this approach is

particularly well adapted to incorporate the complex ascertainment biases observed for markers such as AFLPs and SNPs.

An important challenge posed by the use of ABC methods is finding sufficient summary statistics for the estimation of model parameters. In fact, no such statistics are usually available for population genetics applications but, encouragingly, near-sufficient statistics provide a reasonable approximation (Tavaré *et al.* 1997). In general, studies that use the ABC approach consider summary statistics such as the mean or mode of a given parameter (e.g. F_{ST}, mean linkage disequilibrium between pair of loci, average number of differences between pairs of DNA sequences, etc.). Here, we propose to use the quantiles of the summary statistic distributions across loci, because they provide much more information than the mean or the mode.

Many methods have been proposed to estimate the frequency of the null allele and the genetic diversity when using AFLP data (see Bonin *et al.* (in press) for a review) but all of them except that of Holsinger *et al.* (2002) assume Hardy-Weinberg equilibrium. It has been argued that doing this for AFLP markers when no information on F_{IS} is available is not a problem when comparing F_{ST} values across species or populations (Bonin *et al.* (in press). However, this is only the case if the level of inbreeding is the same for all species/populations. Otherwise, the magnitude of the bias will be different among them.

It should be noted that the model of Holsinger *et al.* (2002), as in the present paper, implicitly assumes that the probability of observing a heterozygote at one locus in an individual, given F_{IS} and the allele frequencies, is independent of observing a heterozygote at a another locus in the same individual. This will only be true if departures from Hardy-Weinberg are due to cryptic population structure (often termed 'Wahlund Effect') because under a model of inbreeding loci within an individual are not independent in their probability of heterozygosity. Indeed, it is possible to use this information to distinguish between the two potential causes of departure from Hardy-Weinberg (Overall et Nichols 2001).

Among all the existing methods dedicated to dominant markers, the model of Holsinger *et al.* (2002), is the only one that simultaneously estimate all parameters but, as we have shown, it is affected by the two sources of bias we have uncovered in this study. Our method, therefore, represents an important improvement over all existing ones and we expect that it will be of great help for the many researchers that are interested in using AFLP markers to study population structure.

Acknowledgments

This work was supported by the Fond National de la Science (grant ACI-IMPBio- 2004-42-PGDA). M.F. holds a Ph.D. studentship from the Ministère de la Recherche. Most the computations presented in this paper were performed on the cluster HealthPhy (CIMENT, Grenoble, France). The software implementing the method is available at http://www-leca.ujf-grenoble.fr/logiciels.htm both for unix and windows platforms.

4 Détection de loci sous sélection : différents marqueurs et différents scénarios démographiques

Identifier des loci influencés par la sélection naturelle présente un enjeu très important dans différents domaines. Ces loci ont une plus grande probabilité d'avoir un rôle dans une fonction biologique importante, et sont également de bons candidats pour être impliqués dans des maladies ou des dysfonctionnements génétiques. La quantité de variation génétique entre populations est déterminée à la fois par des forces neutres, comme la dérive, qui agissent sur tout le génome de façon uniforme, et des forces adaptatives, comme la sélection, qui agissent uniquement sur des loci spécifiques. Dans le chapitre 2, nous avons présenté une méthode qui permet d'identifier les facteurs environnementaux responsables de la différenciation génétique résultante de ces différentes forces. Dans ce chapitre, nous allons séparer les effets de ces deux forces, ce qui peut permettre en particulier d'identifier les loci soumis à la sélection naturelle. Plus précisément, nous généralisons un modèle bayésien existant en utilisant les résultats du chapitre 3 pour l'étendre aux marqueurs dominants, et nous implémentons une méthode de Monte Carlo par Chaine de Markov à sauts réversibles pour estimer la probabilité que chaque locus soit soumis à la sélection.

« Si des variations utiles à un être organisé quelconque apparaissent, les individus affectés doivent assurément avoir une meilleure chance de l'emporter dans la lutte pour l'existence, de survivre et, en vertu de l'hérédité, de produire des descendants semblablement caractérisés. C'est ce principe de conservation, de survivance du mieux adapté, que j'appelle sélection naturelle. »

Charles Darwin

4.1 Problématique et démarche scientifique

De nombreuses méthodes ont été développées pour identifier des régions du génomes soumises à la sélection naturelle (voir Nielsen 2005, pour une revue). On peut distinguer celles utilisant des données comparatives (provenant d'espèces différentes) qui permettent de détecter une signature de sélection ancienne, de celles utilisant des données de génomique des populations qui permettent de détecter une signature de sélection récente. La première catégorie utilise des données de séquences nucléotidiques et l'outil principal utilisé est la comparaison des différents taux de substitutions nucléotidiques.

Nous nous intéressons ici plus particulièrement à la deuxième catégorie de méthode utilisant des données de criblage génomique (utilisation d'un grand nombre de loci). En effet, elles peuvent s'appliquer aux données typiquement disponibles dans le domaine de l'écologie et issues d'espèces non modèles. L'idée générale est d'identifier les loci qui présentent un coefficient F_{ST} avec une valeur différente de ce que l'on attend dans un modèle neutre. Plus précisément, un locus sous l'influence de sélection directionnelle présentera une différenciation génétique plus élevée que les loci neutres, et au contraire, un locus sous l'influence d'une sélection balancée présentera une différenciation génétique plus faible.

Le problème principal de ces méthodes est de choisir le modèle neutre qui permet de distinguer les loci qui s'en écartent. En effet la distribution des coefficients F_{ST} entre les loci dépend de l'histoire démographique des populations. Cet effet de confusion entre histoire démographique et sélection naturelle complique fortement l'estimation de la sélection. Pour cette raison, nous utilisons la méthode proposée par Beaumont et Balding (2004) et basée sur le modèle Dirichlet-multinomial présenté en introduction. Nous avons montré au chapitre 2 que ce modèle était robuste à différents scénarios démographiques. Cette méthode ne s'applique qu'aux marqueurs codominants, et nous utilisons les résultats du chapitre 3 pour l'étendre aux marqueurs dominants. De plus, nous intégrons une méthode de Monte Carlo par Chaine de Markov à sauts réversibles (Green 1995) pour estimer la probabilité que chaque locus soit soumis à la sélection.

4.2 Contribution scientifique : article C

Foll, M., et O. Gaggiotti, 2007 A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. Submitted to Genetics

4.2.1 Matériels et méthodes

Modèle

On a montré que le modèle Dirichlet-multinomial pour la différenciation génétique était robuste à différents scénarios dans le chapitre 2. Pour cette raison c'est un bon candidat pour éviter l'effet de confusion entre histoire démographique et sélection. Ce modèle estime un coefficient de différenciation génétique (F_{ST}^{j}) dans chaque population locale j. Dans le but d'identifier les loci soumis à la sélection, nous avons besoin d'estimer aussi un coefficient pour chaque locus i. Ainsi, la méthode devra estimer des coefficients F_{ST}^{ij} spécifiques à chaque couple locus/population. Cependant, la quantité de données disponibles pour estimer de tels coefficients est alors très réduite et cela se révèle impossible en pratique, en particulier pour les marqueurs présentant peu d'allèles différents et les marqueurs dominants. Pour résoudre ce problème, Balding *et al.* (1996) ont proposé de décomposer ces coefficients en une partie spécifique à chaque population mais commune à tous les loci, et une partie spécifique à chaque locus mais commune à toutes les populations. Ici nous utilisons le modèle de régression proposé par Beaumont et Balding (2004) et basé sur cette idée :

$$\ln\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right) = \alpha_i + \beta_j \tag{4.1}$$

Dans le cas de J populations et I loci, cette formulation permet de remplacer l'estimation de $I \cdot J$ coefficients F_{ST}^{ij} par I coefficients α_i et J coefficients β_j. Dans le cas d'un modèle neutre, tous les coefficients α_i sont nuls et le modèle est équivalent à celui présenté dans les chapitres 2 et 3. Un coefficient α positif augmente la différenciation génétique et représente donc un cas de sélection directionnelle. A l'inverse, un coefficient négatif est attendu dans un cas de sélection directionnelle.

Dans la méthode originale, Beaumont et Balding (2004) utilisaient la distribution à posériori des α_i pour identifier les loci soumis à la sélection en utilisant une méthode approchée. Nous proposons de généraliser le modèle en considérant pour chaque locus i l'hypothèse de neutralité ($\alpha_i = 0$) et d'en estimer la probabilité. La méthode fournit alors une probabilité à postériori que chaque locus soit soumis à la sélection en plus de la distribution à postériori des α_i. Les probabilités des différents modèles sont calculées à l'aide d'un algorithme de Monte Carlo par Chaine de Markov à sauts réversibles (Green 1995).

Pour des marqueurs codominants, les fréquences alléliques nécessaires à l'estimation des coefficients F_{ST}^{ij} (ici remplacés par α_i et β_j) sont modélisées à l'aide d'une distribution multinomiale comme dans le chapitre 2. Pour les marqueurs dominants, l'estimation de ces fréquences nécessite d'inclure dans le modèle le coefficient de consanguinité F_{IS} de chaque population. On a montré dans le chapitre 3 que la méthode proposée par Holsinger *et al.* (2002) souffrait du biais de recrutement affectant les marqueurs AFLP. La solution que nous avons proposée en utilisant un algorithme du type « Approximate Bayesian Compuation » (ABC) n'est pas satisfaisante ici car elle conduit à une incertitude trop grande pour pouvoir estimer correctement les coefficients α_i. En effet, l'information disponible pour estimer ces coefficients spécifiques à chaque locus est beaucoup plus réduite que dans le cas du chapitre 3. De plus, l'algorithme ABC ne permet pas d'estimer les probabilités à postériori des différents modèles. Ici nous proposons une alternative qui permet à la fois d'inclure l'incertitude sur la consanguinité, et d'utiliser un algorithme de type MCMC pour estimer les α_i et les probabilités à postériori $P(\alpha_i \neq 0)$.

Méthode de validation

Nous complétons l'étude de sensibilité de la méthode publiée par Beaumont et Balding (2004) tout en y intégrant les nouveautés présentées ici. Pour cela, nous utilisons deux approches différentes pour générer des données simulées.

La première approche utilise le même modèle statistique que la méthode proposée et permet d'étudier l'effet de trois paramètres clés sur la précision des résultats obtenus : la taille de l'échantillon, le nombre de populations et le niveau de la différenciation génétique. Nous utilisons aussi cette approche pour comparer l'efficacité des trois types de marqueurs moléculaires les plus fréquents : les AFLP, les SNP et les microsatellites. Les AFLP sont des marqueurs dominants, alors que les deux autres sont des marqueurs codominants. Les SNP sont bi-alléliques et au contraire, les microsatellites sont très polymorphes. Toutes ces simulations sont utilisées par la suite pour mesurer le gain apporté par la méthode de Monte Carlo par Chaine de Markov à sauts réversibles (Green 1995) permettant d'estimer les probabilités $P(\alpha_i \neq 0)$.

La deuxième approche utilisée vise, comme dans le chapitre 2, à illustrer l'effet d'une forte violation du modèle en ile supposé ici. Pour cela nous avons utilisé SPLATCHE (Currat *et al.* 2004) pour générer des données neutres selon un scénario d'expansion spatiale des populations. Nous utilisons alors ces données neutres pour étudier dans quelle mesure l'effet de confusion entre histoire démographique et sélection influence notre méthode.

Jeux de données analysés

Nous avons appliqué la méthode à deux jeux de données différents qui illustrent deux types d'applications possibles. Tout d'abord nous avons analysé les données de génétique humaine du HGDP-CEPH (Cann *et al.* 2002) présentées au chapitre 2. Cet exemple permet d'illustrer le cas d'une étude portant sur des données codominantes pour lesquelles le génome est connu et où les marqueurs sont localisés. Ensuite nous avons re-analysé les données de littorine (*Littorina saxatilis*) publiées par Wilding *et al.* (2001). Cette espèce de bigorneaux a été la première espèce non modèle pour laquelle un criblage génomique a été réalisé à l'aide des marqueurs AFLP.

- Humain : Nous avons utilisé ici les 835 marqueurs microsatellites publiés dans la dernière version disponible. Le jeu de données contient 1056 individus, répartis dans 51 populations (voir la figure 2.2 du chapitre 2). Nous n'avons choisi comme étant sous sélection que les marqueurs pour lesquels l'estimation à postériori de α_i était située à l'une des extrémités de la distribution des α_i entre les loci. Nous avons montré avec les simulations du modèle d'expansion spatiale, que cette approche permettait de limiter très fortement le nombre de faux positifs attendus avec ces données.
- Littorine (*Littorina saxatilis*) : Les données consistent en 290 marqueurs AFLP polymorphes étudiés dans quatre zones côtières rocheuses de Grande-Bretagne. Dans cette région il existe deux morphotypes distincts notés H et M de *L. saxatilis* qui semblent être en isolement reproductif partiel. Un échantillon de chacun des deux morphotypes a été prélevé dans chaque zone, sauf une, où deux échantillons de M ont été prélevés. Les huit populations ainsi obtenues sont composées de 43 à 51 individus. Wilding *et al.* (2001) ont identifié 15 loci soumis à la sélection, et cet exemple illustre aussi l'effet de confusion entre sélection et histoire démographique. En effet, l'arbre « Neighbour-Joining » construit en utilisant tous les loci conduit à une histoire totalement différente de celui construit en excluant les 15 loci sélectionnés (figure 4.1).

4.2.2 Résultats

Les simulations selon le modèle exact ont montré que les résultats avec les marqueurs AFLP (dominants) n'étaient que très légèrement inférieurs à ceux obtenus avec des marqueurs SNP (bi-alléliques codominants). De plus la méthode proposée est insensible au niveau de consanguinité rencontré dans les

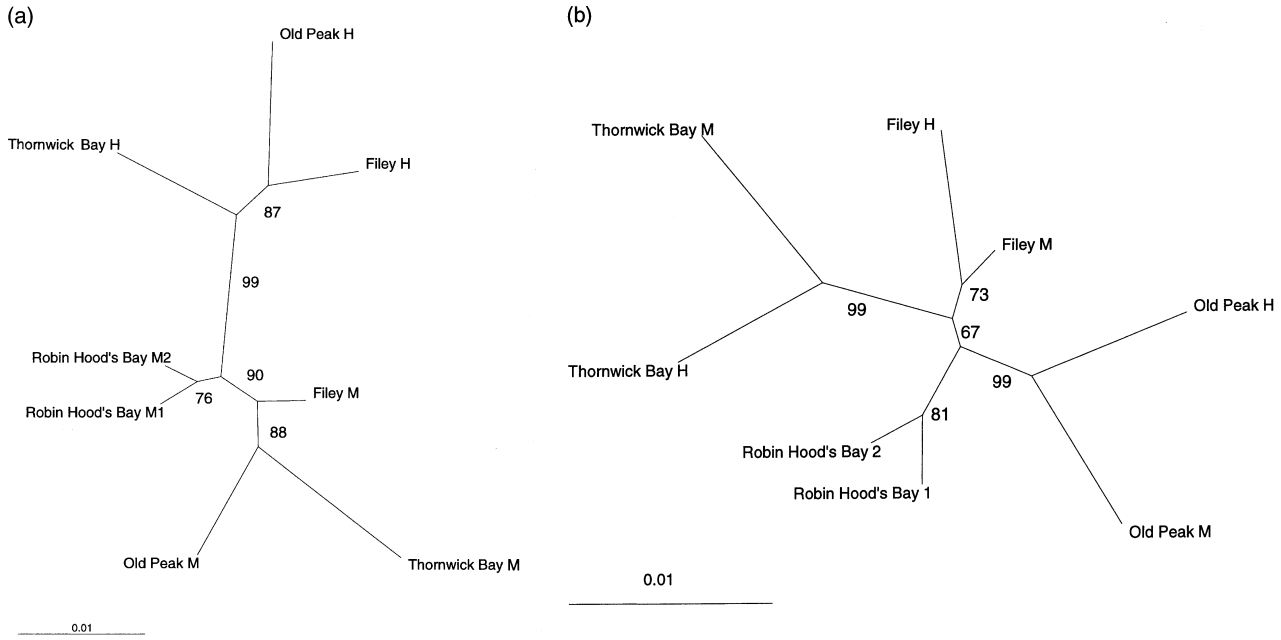

FIG. 4.1 – *Arbres « Neighbour-Joining » estimés entre les populations de Littorina saxatilis de deux morphotypes (H et M) (a) en utilisant tous les marqueurs disponibles ; (b) après avoir retiré 15 marqueurs identifié comme étant sous sélection, d'après* WILDING et al. *(2001).*

populations. Pour les marqueurs microsatellites, simulés avec sept allèles différents, les résultats sont beaucoup plus précis, ce qui permet de détecter des niveaux plus faibles de sélection. Ils peuvent aussi détecter de la sélection avec des jeux de données de moins bonne qualité. En particulier, le nombre de populations, la taille de l'échantillon, et le niveau de différenciation génétique sont des paramètres très importants pour une identification correcte, surtout pour la sélection balancée.

L'analyse des données simulées selon le modèle d'expansion spatiale a montré que la méthode était sensible à la violation de l'hypothèse du modèle en ile considéré. En effet, nous obtenons 10.6% de faux positifs avec une probabilité $P(\alpha_i \neq 0) > 0.99$, et la quasi totalité de ces faux positifs est identifiée comme étant sous sélection directionnelle (92%).

Parmi les 835 marqueurs microsatellites des données humaines, 337 marqueurs ont une probabilité $P(\alpha_i \neq 0) > 0.99$. Nous n'avons conservé que les loci ayant une valeur extrême de α_i en comparaison de celles observées pour les faux positifs de l'analyse de sensibilité du modèle d'expansion spatiale. Nous avons alors obtenu 104 marqueurs sous sélection directionnelle, et 160 marqueurs sous sélection balancée. Parmi les 264 marqueurs identifiés, 114 sont situés sur des gènes connus dont 69 sont référencés comme étant impliqués dans des maladies génétiques, et 39 sont associés à des QTL.

L'histoire démographique des populations de Littorine semble s'éloigner

du modèle en ile. Afin de limiter les faux positifs, Wilding *et al.* (2001) ont procédé à trois analyses distinctes, pour chacune des trois zones où l'on trouve les deux morphotypes H et M. En suivant la même approche, nous avons identifié 13 loci soumis à la sélection. Ces 13 loci font partie de la liste des 15 identifiés par Wilding *et al.* (2001) à l'aide de la méthode de Beaumont et Nichols (1996) adaptée pour les marqueurs dominants. En simulant des données selon l'histoire démographique supposée des populations (figure 4.1b), nous avons montré que le fait de conserver les marqueurs identifiés dans deux analyses seulement au lieu des trois (Wilding *et al.* 2001) semblait être une bonne approche, ce qui nous a permis d'identifier six loci supplémentaires.

La méthode a été implémentée dans un logiciel écrit en langage C++ avec une interface graphique conviviale, le rendant simple d'utilisation. L'interface permet aussi de visualiser les résultats sous forme graphique. Le logiciel est diffusé librement sur le site internet du Laboratoire d'Ecologie Alpine à l'adresse `http://www-leca.ujf-grenoble.fr/logiciels.htm`.

4.2.3 Discussion

Nous avons montré au chapitre 3 que les estimations de F_{IS} à partir des marqueurs dominants sont fortement influencées par le biais de recrutement qui touche les marqueurs AFLP. Dans notre cas, nous ne sommes intéressés que par le biais potentiel sur les estimations des effets locus-spécifiques et les simulations que nous avons effectuées montrent qu'il suffit d'incorporer l'incertitude sur le coefficient de consanguinité pour éviter un tel biais. Ainsi, au lieu d'estimer le coefficient F_{IS}, nous le laissons explorer librement l'intervalle $[0, 1]$. Il serait possible de tenir compte de connaissances à priori sur le niveau de consanguinité en utilisant un intervalle plus restreint, mais le fait que les résultats fournis par les AFLP soient très similaires à ceux des SNP suggère que l'on ne peut pas espérer une amélioration très significative de cette façon.

Nous avons étudié le biais qui résulte de données issues d'un modèle démographique très différent de celui supposé dans notre modèle en utilisant SPLATCHE (Currat *et al.* 2004) avec le cas de l'expansion humaine comme exemple. Les résultats produits indiquent un taux de faux positifs de l'ordre de 10%, pour la plupart sous sélection directionnelle. Cette étude a permis d'appliquer la méthode aux données humaines microsatellites du HGDP-CEPH (Cann *et al.* 2002) en limitant les risques de faux positifs. Pour cela nous n'avons retenu que les loci ayant une distribution à postériori pour α plus extreme que celles obtenues avec les données simulées de SPLATCHE.

Bien que le nombre de gènes utilisés soit faible, ces résultats sont en accord avec des études précedentes. Par exemple, nous avons obtenu le même résultat que Clark *et al.* (2003) qui a montré, en classant 7645 genes avec la base de données « Panther » (`www.pantherdb.org`), que la catégorie contenant le plus grand nombre de gènes sélectionnés est celle de la transduction de signaux.

Dans le cas des données de *Littorina saxatilis* publiées par Wilding *et al.* (2001), nous avons obtenu des résultats différents en effectuant des analyses par paire ou en utilisant toutes les populations dans une seule analyse. Afin de trancher sur la meilleure stratégie à adopter, nous avons simulé des données avec une histoire démographique similaire à celle des données réelles. Cela a montré que la première approche conduisait à des faux négatifs et la deuxième à des faux positifs. Ainsi on a proposé une troisième stratégie qui, sur les données simulées, identifie correctement tous les marqueurs. En l'utilisant sur les données réelles, on a obtenu six loci de plus que Wilding *et al.* (2001) soumis à la sélection.

Ces deux exemples d'application sont une bonne illustration de l'effet de confusion entre sélection et histoire démographique. Dans les deux cas, des simulations préliminaires ont permis d'éviter d'obtenir des faux positifs. Cependant, l'approche utilisée pour les données humaines est très restrictive et conduit certainement à de nombreux faux négatifs. Au contraire, l'histoire démographique simple des populations de *Littorina saxatilis* conduit à penser que l'on a correctement identifié la majeure partie des marqueurs. Ainsi, il semble important d'effectuer ce type de simulations avant toute analyse de ce type. De cette façon, il est possible de choisir la valeur seuil à utiliser pour identifier les marqueurs et de définir la meilleure stratégie pour les analyses.

Une extension naturelle de notre méthode serait d'inclure l'histoire démographique au sein de l'analyse au lieu d'imposer un modèle démographique simple. Dans les cas où cette histoire est connue, il serait possible de l'incorporer dans les estimations, sinon, il serait nécessaire de l'estimer avec les données génétiques.

Conclusion

La méthode présentée ici permet de détecter des loci soumis à la sélection à partir de marqueurs dominants et codominants. De plus, elle intègre une estimation rigoureuse de la probabilité à postériori que chaque locus soit soumis à la sélection. Beaumont et Balding (2004) avaient montré que cette méthode n'était pas appropriée pour détecter de la sélection balancée en utilisant des marqueurs bialléliques. Cependant, notre analyse prouve que cela est possible en utilisant des marqueurs multialléliques comme les microsatellites. De plus, nous avons établi que les marqueurs dominants (comme les AFLP) sont quasiment aussi performants que les marqueurs codominants bialléliques (comme les SNP) pour détecter des loci sélectionnés. Ainsi, le niveau de polymorphisme est un des facteurs les plus importants pour la puissance de la méthode. Les deux exemples d'application illustrent qu'il est possible d'éviter les faux positifs dûs à l'histoire démographique des populations en effectuant des simulations préliminaires qui imitent les données réelles.

Chapitre 4 - Article C

Foll, M., et O. Gaggiotti, 2007 A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. Submitted to Genetics

Abstract

Identifying loci under natural selection from genomic surveys is of great interest in different research areas. Commonly used methods to separate neutral effects from adaptive effects are based on locus-specific population differentiation coefficients to identify outliers. Here we extend such an approach to estimate directly the probability that each locus is subject to selection using a Bayesian method. We also extend it to allow the use of dominant markers like AFLPs. It has been shown that this model is robust to complex demographic scenarios for neutral genetic differentiation. However, we show here that strong violations of the demographic model assumed leads to a high rate of false positive. Nevertheless, we demonstrate that it is possible to avoid them by carrying out a preliminary simulation study. We re-analyze two previously published data sets : a human data set of codominant markers and a *Littorina saxatilis* data set of dominant markers. We also perform a detailed sensitivity study to compare the power of the method using AFLP, SNP and microsatellite markers. The method has been implemented in a new software available at our website (http ://www-leca.ujf-grenoble.fr/logiciels.htm).

Key words : Natural selection, AFLP, demographic history, Bayesian statistics, Reversible Jump MCMC.

Running title : Detecting selection.

Introduction

One of the main challenges of modern biology is to dissect and understand the molecular basis for naturally occurring genetic variation. Recent advances in the fields of computational biology and molecular biology techniques have led to the emerging field of 'population genomics' whose main objective is to characterize the parts of the genome subject to natural selection. This new discipline has important applications in many domains such as medical genetics and the improvement of agricultural crops and breeds. Additionally, ignoring the effect of natural selection in population genetics studies can lead to wrong estimates of the demographic history of species. Therefore, separating the effect of neutral drift and adaptive genetic differentiation is a necessary preliminary step in most analyzes of genome wide data sets, and this distinction can also help to understand speciation processes.

A wide variety of methods have been developed to identify regions of the genome that have been subject to natural selection (see Nielsen *et al.* 2005, for a review). Among them, we can distinguish those based on comparative data (taken from different species) that can detect old signatures of selection and those using population genomics data that allow the detection of more recent ones. This latter family of methods has became very popular in the last decade and has been applied to a large number of non-model species (see Wilding *et al.* 2001, for example).

Many of the existing methods for detecting recent selection from population genomics data are based on an idea first introduced by Lewontin et Krakauer (1973) (see for example Bowcock *et al.* 1991, Beaumont et Nichols 1996, Vitalis *et al.* 2001, Beaumont et Balding 2004). The basic rationale is that loci influenced by directional (also called adaptive or positive) selection will show a larger genetic differentiation than neutral loci, and loci that have been subject to balancing (also called negative or purifying) selection will show a lower genetic differentiation. Then the methods generally consist of identifying loci that present F_{ST} coefficients which are 'significantly' different from those expected under the neutral theory (they are called outlier loci).

Lewontin et Krakauer (1973)'s method has raised many criticisms (see Beaumont 2005, for more details about them) and finally fell out of use. More recently, Bowcock *et al.* (1991) and Beaumont et Balding (2004)showed that problems of a purely statistical nature can be easily solved. In particular, the problem related to the correlation of allele frequencies among demes can be overcome by adopting a Bayesian approach that implements the Dirichlet-

multinomial model of genetic differentiation Beaumont et Balding (2004). This distribution describes an island model (Wright 1931) in which subpopulation allele frequencies are correlated and allows to estimate population specific F_{ST} coefficients. These properties make it more robust to complex demographic scenarios (Foll et Gaggiotti 2006).

Here we extend Beaumont et Balding (2004) approach to estimate the posterior probability of a given locus being under the effect of selection by defining two alternative models, one that includes the effect of selection and another that excludes it ; we then estimate their respective posterior probabilities using a Reversible Jump MCMC approach. We also implement the possibility of using dominant markers like AFLPs and perform a detailed sensitivity study to compare the power of the method using AFLP, SNP and microsatellite markers. Finally, we present examples that illustrate how the method can be applied to study particular cases where the demographic scenario differs from the one assumed in the model.

Methods

Bayesian model for locus-population specific F_{ST} coefficients

The model for genetic differentiation used is based on ideas first introduced by Balding et Nichols (1995) and Beaumont et Balding (2004) first used it to detect loci under natural selection. For the sake of simplicity we describe the details of our approach using the terminology of a fission model but it should be kept in mind that, even if the two models have the same first two moments of allele frequencies, it only exactly applies to an island model (Wright 1931). We consider a collection of J subpopulations that evolved in isolation after splitting from an ancestral population. The derived subpopulations may have been subject to different amounts of genetic drift and, therefore, their allele frequencies will show different degrees of differentiation from the ancestral allele frequency. We consider a set of I loci and let K_i be the number of alleles at the i^{th} locus. The extent of differentiation between subpopulation j and the ancestral population at locus i is measured by F_{ST}^{ij} and is the result of its demographic history. Let $\boldsymbol{p_i} = \{p_{ik}\}$ denote the allele frequencies of the ancestral population at locus i, where p_{ik} is the frequency of the allele k at locus i ($\sum_k p_{ik} = 1$). We use $\boldsymbol{p} = \{\boldsymbol{p_i}\}$ to denote the entire set of allele frequencies of the ancestral population and $\widetilde{\boldsymbol{p_{ij}}} = \{\widetilde{p_{ijk}}\}$ to denote the current alleles frequencies at locus i for subpopulation j. Under

these assumptions, the allele frequencies at locus i in subpopulation j follows a Dirichlet distribution with parameters $\theta_{ij}\boldsymbol{p_i}$,

$$\widetilde{\boldsymbol{p_{ij}}} \sim \mathcal{D}ir(\theta_{ij}p_{i1}, \ldots, \theta_{ij}p_{iK_i}) \tag{4.2}$$

where $\theta_{ij} = 1/F_{ST}^{ij} - 1$. The parameters F_{ST}^{ij}s are very closely related to Wright's F_{ST} (Wright 1951) parameter and are interpreted as measures of the shared ancestry within each of the subpopulations (see Balding 2003, for a more detailed explanation). The full prior distribution can be obtained by multiplying across loci and populations :

$$\pi\left(\widetilde{\mathbf{p}}|\mathbf{p}, \grave{}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} \pi\left(\widetilde{\boldsymbol{p_{ij}}}|\boldsymbol{p_i}, \theta_{ij}\right) \tag{4.3}$$

Bayesian regression model for selection

The amount of data available to estimate all locus-population specific F_{ST} coefficients is reduced and this leads to inaccurate estimates, especially for loci with a small number of different alleles. As an alternative, Balding *et al.* (1996) proposed to decompose locus-population specific F_{ST} coefficients into a population specific component shared by all loci and a locus specific component shared by all populations. We use the regression model proposed by Beaumont et Balding (2004) and based on the following equation :

$$log\left(\frac{F_{ST}^{ij}}{1 - F_{ST}^{ij}}\right) = log\left(\frac{1}{\theta_{ij}}\right) = \alpha_i + \beta_j \tag{4.4}$$

The advantage of this formulation is that instead of estimating $I \cdot J$ F_{ST}^{ij} coefficients, we only have to estimate the I parameters α_i and the J parameters β_j. In case of absence of natural selection, all α_i coefficients are null, and the above model is equivalent to the one used by Foll et Gaggiotti (2006), where the term β_j is replaced by a generalized linear model. Note that with this formulation, F_{ST}^{ij}, or equivalently θ_{ij}, are no longer model parameters that need to be estimated because they are replaced by α_i's and β_j's parameters. In what follows we use θ_{ij} for the sake of simplicity but note that it can be replaced directly by $\theta_{ij} = exp(-(\alpha_i + \beta_j))$.

Beaumont et Balding (2004) originally proposed to include a locus-population parameter γ_{ij} in their formulation. However, they noticed that the posterior probability for this parameter was very similar to the prior used. This indicates that there is not enough information to estimate this parameter and, therefore, we choose here to exclude the γ_{ij}s from the model.

Reversible jump

In order to infer which loci are subjected to selection we focus on the posterior distribution of α_i : a positive value suggests that locus i is subject to directional selection, whereas a negative value suggests balancing selection. However, before deciding on the type of selection we need to decide wether or not there is selection at all. In their original formulation, Beaumont et Balding (2004) focused on the posterior distribution of α_i and from this they identified locus subjected to selection using an approximate method (see below). Here we present a rigorous way of estimating the posterior probability of a given locus being under the effect of selection. Equation 4.4 can lead to two alternative models, one that includes both effects and another one that does not include the effect of selection. Thus, we use a Reversible Jump MCMC algorithm (Green 1995) to estimate the posterior probability of each one of these models. At each iteration of the MCMC algorithm, for each locus i we propose to remove α_i from the model if it is currently present, or to add it if it is not included. For example if we propose to add α_i to the vector $\boldsymbol{\alpha}$ of locus effect, we draw a proposed value from a distribution q. Then we accept to add this locus in the model with probability $min(1, A)$, where :

$$A = \frac{\pi(\widetilde{\mathbf{p}}|\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta})\pi(\alpha_i)}{\pi(\widetilde{\mathbf{p}}|\mathbf{p}, \boldsymbol{\alpha} \text{ with } \alpha_i = 0, \boldsymbol{\beta})q(\alpha_i)}$$

Because we only have two models and we choose them uniformly, the ratio of prior model probability simplifies to one. The Jacobian is one because of the canonical jump function used. To consider the reverse move, we simply accept the move deleting α_i with probability $min(1, 1/A)$. The proposal distribution q is a normal distribution with its mean and variance pilot tuned to improve convergence (see below).

With this method, we have posterior estimates of the probability that a locus is subject to selection : $P(\alpha_i \; included)$ corresponds to $P(\alpha_i \neq 0)$. The reversible jump algorithm gives us $P(\alpha_i \neq 0)$ for each α_i simply by counting the number of times α_i is not included in the model from the MCMC output.

Estimating allele frequencies

Beaumont et Balding (2004) original formulation considered codominant markers ; here we extend it to dominant ones. Note that we have to estimate the allele frequencies of each subpopulation and that of the ancestral population because they are unknowns. Thus, we present two different formulations

depending of the type of markers used : codominant (like microsatellites or SNPs) or dominant (like RFPLs or AFLPs).

Codominant markers

The data consist of allele counts obtained from samples of size n_{ij}. We use a_{ijk} to denote the number of allele k observed at locus i in the sample from subpopulation j. Thus, $n_{ij} = \sum_k a_{ijk}$. The full data set can be presented as a matrix $\boldsymbol{N} = \{\boldsymbol{a_{ij}}\}$, where $\boldsymbol{a_{ij}} = \{a_{ij1}, a_{ij2}, \ldots, a_{ijK_i}\}$ is the allele count at locus i for subpopulation j. The observed allele frequencies, $\boldsymbol{a_{ij}}$, can be considered as sampled from the true alleles frequencies $\widetilde{\boldsymbol{p_{ij}}}$ and, therefore, can be described by the multinomial distribution (Holsinger 1999) :

$$\boldsymbol{a_{ij}} \sim \mathcal{M}ultinomial\{n_{ij}; \widetilde{p_{ij1}}, \widetilde{p_{ij2}}, \ldots, \widetilde{p_{ijK_i}}\} \tag{4.5}$$

In principle, we could use as likelihood the multinomial distribution (equation 4.5) and consider equation 4.2 as a Bayesian prior. However, in our case, we can calculate exactly the marginal distribution of $\boldsymbol{a_{ij}}$ because the Dirichlet distribution is the conjugate prior of the multinomial. This allows us to eliminate the nuisance parameters $\widetilde{\boldsymbol{p_{ij}}}$ which are not of immediate interest but are needed by the model. Thus, we obtain the multinomial-Dirichlet distribution :

$$P(\boldsymbol{a_{ij}}|\boldsymbol{p_i}, \alpha_i, \beta_j) = \frac{n_{ij}!\Gamma(\theta_{ij})}{\Gamma(n_{ij}+\theta_{ij})} \prod_{k=1}^{K_i} \frac{\Gamma(a_{ijk}+\theta_{ij}p_{ik})}{a_{ijk}!\,\Gamma(\theta_{ij}p_{ik})}$$

The likelihood is obtained by multiplying across all loci and populations :

$$L(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} P(\boldsymbol{a_{ij}}|\boldsymbol{p_i}, \alpha_i, \beta_j) \tag{4.6}$$

Since the allele frequencies in the ancestral population are unknown, we have to estimate them by introducing a non-informative Dirichlet prior, $\boldsymbol{p_i} \sim \mathcal{D}ir(1, \ldots, 1)$, into our Bayesian model.

Dominant markers

Estimating allele frequencies from dominant markers is more difficult because of the inability to distinguish heterozygous individuals from those that are homozygous for the dominant allele. Nevertheless, they have became very popular in the last decade, mostly due to the development of the Amplified

Fragment Length Polymorphism (AFLP), an inexpensive and easy way of obtaining large number of genetic markers from a wide variety of organisms (Bensch et Akesson 2005, Meudt et Clarke 2007). For each individual the information is "band-presence" or "band-absence", which can be viewed as a phenotype. One possible solution is to suppose Hardy-Weinberg equilibrium to estimate allele frequencies but this imposes the strong hypothesis of no inbreeding. Holsinger *et al.* (2002) first proposed a general method that includes the estimation of the inbreeding coefficient F_{IS}.

In the context of dominant markers, the data $\mathbf{N}$ consists of the sample counts of observed phenotypes instead of alleles counts. Let $n_{[A1],ij}$ and $n_{[A2],ij}$ be the observed number of phenotypes $[A1]$ and $[A2]$ at locus i for population j. The full data set is presented as a matrix $\mathbf{N} = \left\{n_{[A1],ij}, n_{[A2],ij}\right\}$ and the sample size at locus i for population j is $n_{ij} = n_{[A1],ij} + n_{[A2],ij}$. We can consider that the number of phenotypes $n_{[A1],ij}$ follows a binomial distribution with parameters $g_{[A1],ij}$ and n_{ij}, where $g_{[A1],ij}$ is the unknown $[A1]$ phenotype frequency at locus i in population j :

$$n_{[A1],ij} \sim \mathcal{B}inomial\left(g_{[A1],ij}, n_{ij}\right) \tag{4.7}$$

Note that the binomial distribution is a particular case of the multinomial distribution with only two alleles, and the Dirichlet distribution of equation 4.2 reduces to a Beta distribution. The Beta distribution is the conjugate prior of the Binomial distribution, but contrary to the case of codominant markers with the multinomial-Dirichlet distribution, we can not calculate exactly the marginal distribution. This is due to the parameters of the binomial distribution that are the phenotype frequencies in the case of dominant markers, instead of having directly the allele frequencies. If we assume independence we can multiply across loci and populations to obtain the likelihood function :

$$L\left(\widetilde{\mathbf{p}}, \mathbf{F_{IS}}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} P\left(n_{[A1],ij} | g_{[A1],ij}\right)$$

The phenotype frequency $g_{[A1],ij}$ can be linked to the corresponding frequency p_{ij} of allele $A1$ and the inbreeding coefficient F_{IS}^{j} of population j using the following equations :

$$\begin{aligned} g_{[A1],ij} &= \widetilde{p_{ij}}^2(1 - F_{IS}^{j}) + F_{IS}^{j}\widetilde{p_{ij}} + \left(1 - F_{IS}^{j}\right) 2\widetilde{p_{ij}}\left(1 - \widetilde{p_{ij}}\right) && (4.8)\\ g_{[A2],ij} &= \left(1 - F_{IS}^{j}\right)\left(1 - \widetilde{p_{ij}}\right)^2 + F_{IS}^{j}\left(1 - \widetilde{p_{ij}}\right) && (4.9)\\ &= 1 - g_{[A1],ij} && (4.10) \end{aligned}$$

However, Foll *et al.* (2007) showed that estimates obtained from this model are strongly influenced by the ascertainment bias of AFLPs. They proposed an alternative Approximate Bayesian Computation (ABC) approach that gives unbiased estimates of population specific F_{ST} and F_{IS} coefficients. This solution leads to more uncertainty on posterior distributions, which precludes the estimation of locus specific α_is. Additionally, the ABC algorithm can not be used to estimate the posterior probability of each hypothesis of the form $\alpha_i = 0$. Because here values of F_{IS} are not of immediate interest, we propose an intermediate solution : we do not estimate F_{IS} coefficients but we incorporate the full uncertainty on F_{IS} by letting it move freely between 0 and 1 during the MCMC process. This approach has also been proposed for the software Hickory implementing the method of Holsinger *et al.* (2002) and is discribed in the online manual (Holsinger et Lewis 2002). Of course if some other source of information suggests that inbreeding can be bounded within a narrower interval it is possible to restrict it to reduce uncertainty on parameter estimates. We use the prior on ancestral allele frequencies proposed by Foll *et al.* (2007) (in review) : $p_i \sim \mathcal{B}eta(a, a)$. The parameter a describes the shape of allele frequencies in the ancestral population Wright (1931) and is estimated using a log-normal positive prior : $a \sim log\mathcal{N}ormal(0, 1)$.

Implementation

For codominant markers, the full Bayesian model represented by the Directed Acyclic Graph (DAG) in Figure 4.2a is given by :

$$\pi(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{N}) \propto L(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta})\pi(\boldsymbol{p})\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta}) \tag{4.11}$$

For dominant markers, the full Bayesian model represented by the DAG in Figure 4.2b is given by :

$$\pi\left(\boldsymbol{p}, \boldsymbol{F_{IS}}, \widetilde{\boldsymbol{p}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a|\boldsymbol{N}\right) \propto L\left(\widetilde{\boldsymbol{p}}, \boldsymbol{F_{IS}}\right) \pi\left(\widetilde{\boldsymbol{p}}|\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) \pi\left(\boldsymbol{F_{IS}}\right) \pi\left(\boldsymbol{p}|a\right) \pi\left(\boldsymbol{\alpha}\right) \pi\left(\boldsymbol{\beta}\right) \pi\left(a\right) \tag{4.12}$$

Following Beaumont et Balding (2004), for the population effects β_j, we used a Gaussian prior with mean -2 and standard deviation respectively 1.8 ; for the locus effects ,α_i, we use a Gaussian prior with a zero mean and a standard deviation of 1. As explained above, F_{IS}^j are not estimated during the MCMC algorithm but are used to incorporate the uncertainty on inbreeding in the model with dominant markers.

The estimation of model parameters is carried out using a combination of

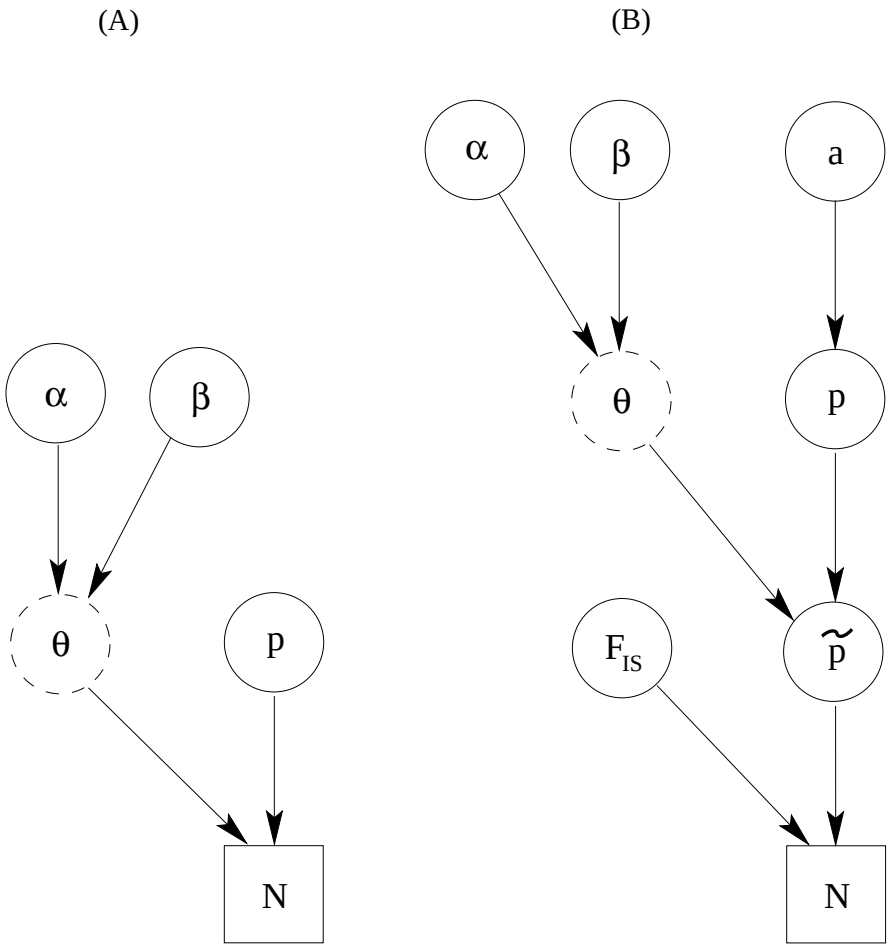

FIG. 4.2 – *DAG of the models given in equation 4.11 (A) and equation 4.12 (B). Square node denotes know quantity (i.e. data) and circles represents parameters to be estimated. Lines between nodes represent direct stochastic relationships within the model. The variables within each node correspond to the different model parameters discussed in the text. N is the genetic data, that is allele frequency counts for codominant markers or phenotype frequency counts for dominant data.* $\tilde{\boldsymbol{p}}$ *and* $\boldsymbol{p}$ *are respectively the allele frequencies in each local population and in the ancestral population.* $\boldsymbol{\theta}$ *is the vector of the genetic differentiation coefficient for each local population.* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *are respectively the vectors of locus and population specific effects of the genetic differentiation. The vector* $\boldsymbol{\theta}$ *is represented within a dotted circle because it is not actually a parameter of the model : it can be calculated directly from equation 4.4, but we represent it for a better understanding of the diagram.* $\boldsymbol{F_{IS}}$ *is the vector of inbreeding coefficients and a is the hyper-prior determining the shape of the ancestral allele frequencies.*

MCMC and RJMCMC GREEN (1995) techniques that are described in the Supporting Information. We evaluated the convergence of the method using the diagnostic tests implemented in the R BOA package (SMITH 2005). The tests indicated that a burn-in of 10000 iterations was enough to attain convergence and it has been implemented as part of the pilot-tuning process (see below). We used a sample size of 10000 and a thinning interval of 50 as suggested by an autocorrelation analysis. With these parameter values, the total length of the chain was 500 000 iterations. The method has been implemented in a

software written in C++. We provide a command line version for Linux and a version with a user friendly Graphical User Interface for Microsoft Windows.

Proposal distribution have to be adjusted in order to have acceptance rates between 0.25 and 0.45. If we propose values in a very wide interval, most moves will be rejected because they will correspond to areas of low posterior probability. On the other hand, if we propose values very close to the current one, the move will be almost always accepted but the chain will take a long time to explore all the parameter space. These values are automatically tuned on the basis of short pilot runs : we run 2000 iterations, and for each parameter, the proposal is adjusted in order to reduce or increase the acceptance rate. We make 10 such pilot runs before starting the sampling, which also play the role of a burn-in period. At the same time, we can choose the proposal distribution q for the reversible jump. Brooks *et al.* (2003) showed that the best choice is to take $q(\boldsymbol{\alpha})$ to be the full conditional distribution of $\boldsymbol{\alpha}$ in the saturated model. Because we don't know this distribution, we use the pilot run to get rough estimates of the mean m_i and the variance v_i for all α_i under the saturated model (in which all parameters α_i are included). Then we propose new value for α_i from $\mathcal{N}(m_i, v_i)$ which is generally close to the full conditional distribution.

Simulation study

We investigated the performance of our method under different scenarios using a simulation study and also we compared its performance with that of Beaumont et Balding (2004) approach. Our first simulation approach uses the same statistical model assumed by our method (the inference model) and allows us to study the effect of 3 critical parameters of the model in the identification of selection : the sample sizes, the number of populations and the level of genetic differentiation. We also use this simulation scheme to compare the power of three different types of markers : AFLPs, SNPs and microsatellites. The first marker is a dominant marker while the two others are codominant.

We also used a second simulation approach to investigate the effect of departures from the demographic model assumed by our method. For this purpose we generated neutral marker data sets under a population expansion model that assumes a stepping-stone colonization process (SPLATCHE, Currat *et al.* 2004). This allows us to investigate if the confounding effect of

selection and demographic history can lead the method to identify as selected loci that are in fact neutral (false positive detection of selection).

Data

Our first simulation scheme assumes 1000 loci of which 100 are under directional selection and 100 are under balancing selection. We introduced selection using $\alpha = 2$ and $\alpha = -2$ for directional and balancing selection respectively. These values correspond to a strong level of selection compared to the selection coefficients used in Beaumont et Balding (2004) but they are similar to values estimated from real data sets (see below). In order to investigate the performance of the method under different scenarios, we considered a default set of values for parameters that were common to both codominant and dominant markers and then changed values of one parameter at a time. This procedure led to 10 different data sets that are described in Table 4.1. The default values were 6 population, a sample size of 30 individuals per population, and a F_{ST} coefficient of 0.10.

Name	Populations	F_{ST}	Sample size
pop-2	2	0.15	30
pop-6*	6	0.15	30
pop-10	10	0.15	30
pop-20	20	0.15	30
fst-0.05	6	0.05	30
fst-0.10	6	0.1	30
fst-0.15*	6	0.15	30
fst-0.25	6	0.25	30
size-15	6	0.15	15
size-30*	6	0.15	30
size-50	6	0.15	50
size-100	6	0.15	100

Tab. 4.1 – *Parameters used in data simulated under the inference model discussed in text. The four parameters marked with an asterisk (*) are in fact the same data set used as a reference. In all other data sets, we modified parameters one by one from this reference.*

In the particular case of AFLP markers we also need to consider the effect of inbreeding so we used a default value of 0.5 and two additional values (0 and 1 ; corresponding data sets are called Fis-0, Fis-0.5 and Fis-1 in Table 4.2) leading to a total of 12 different data sets. Additionally, we included in the simulation the ascertainment bias process observed for AFLP markers and described by Foll *et al.* (2007) (in review). The bias we imposed assure that at least 2% of the total number of individuals have a band and that at most

2% do not have a band. We used default parameters that we modified one by one to obtain the 12 data sets.

Allele frequencies in the ancestral population for both AFLPs and SNPs were generated from a U-shaped beta distribution with both parameters equal to 0.7. In the case of microsatellites, we could simply use the non-informative Dirichlet prior with all parameters equals to 1 assumed by our inference model for multiallelic markers. However, Pritchard et Feldman (1996) showed that the stepwise mutation model describes better the mutation process of microsatellites, and Graham *et al.* (2000) found that the Dirichlet distribution is not appropriate in that case. In particular, simulating ancestral allele frequencies from their prior distribution would lead to a higher variability than what is generally observed in real data sets, and this would artificially increase the power of our method. In order to take into account this findings and at the same time evaluate the influence of a violation of the underlying infinite alleles model assumption, we follow the approach of Lockwood *et al.* (2001) to simulate allele frequencies similar to those observed in real microsatellites. They considered a maximum of seven different alleles and fixed the vector of allele frequencies in the ancestral population at each locus to $(0.05, 0.1, 0.2, 0.3, 0.2, 0.1, 0.05)$. Although Moran (1975) showed that no equilibrium distribution can be obtained under a stepwise mutation model, this provides a practical way to simulate realistic microsatellites data sets. In order to allow variability in the ancestral population, we simulate the vector of allele frequencies at each locus from a Dirichlet distribution with parameters $(10, 20, 40, 60, 40, 20, 10)$.

For some of the scenarios considered for AFLPs and SNPs, we observed a true positive rate of 1 for microsatellite data sets. Thus, we decided to enlarge the range of parameter values considered for this marker and instead of presenting results for data sets where all the 200 loci under selection were correctly identified we simulated additional data sets corresponding to samples of lower quality. More specifically, we added simulations with : 4 populations, $F_{ST} = 0.01$, $F_{ST} = 0.03$, 10 individuals in each population, $|\alpha| = 0.5$, $|\alpha| = 1.0$ and $|\alpha| = 1.5$ (respectively called Pop-4, Fst-0.01, Fst-0.03, size-10, alpha-0.5, alpha-1.0 and alpha-1.5 in Table 4.4).

In order to determine whether or not a locus i is influenced by selection, we need to choose a cutoff value for the posterior probability of including the selection term in $P(\alpha_i \neq 0)$. All loci for which this posterior probability is larger than the cut of value are considered as outliers. However, the choice of a "good" cutoff value depends on the quality of the data set considered. For example, retaining only loci with a posterior probability above 0.99 using mi-

crosatellites with a high sample size and many populations will lead to both a very low false positive rate and a very high true positive rate. By contrast, using the same cutoff value with dominant markers with only two populations will not allow us to detect any marker under selection. For this reason, a pragmatic way to compare results between different kind of data sets is to choose cutoff values that give the same false positive rate and to compare the rate of true positive. With a fixed false positive rate, the true positive rate is directly related to the Positive Predictive Value (PPV, or precision rate) defined as the proportion of markers detected as being under selection that are correctly classified. Here we choose to present results using a cutoff value that lead to a global false positive rate of 5% in each of the three sets of simulations (it corresponds respectively to posterior probabilities of 0.79, 0.86 and 0.85 for AFLPs, SNPs and microsatellites). For AFLPs and SNPs we also show results using a false positive rate of 10% that increases the true positive rate (it corresponds respectively to posterior probabilities of 0.69 for AFLPs and 0.79 for SNPs). For microsatellites, because the true positive rate is already very high with 5% of false positives, we also show results with a false positive rate of 1% (it corresponds to a posterior probability of 0.98).

Results

Comparison among markers

The detailed results obtained for AFLPs, SNPs and microsatellites are presented in Tables 4.2, 4.3 and 4.4, respectively. Table 4.5 presents a summary for comparing the power to detect selection among markers. The first interesting observation is the very similar results obtained for AFLPs and SNPs, which indicates that they have similar power to detect selection. Moreover, the fact of being dominant does not seem to be a big handicap for AFLPs. The precision rate is slightly higher for both balancing and directional selection with SNPs. With microsatellites, the results are much better than with the two other diallelic markers. Note that in Table 4.5 the results for microsatellites are presented for poor quality data sets and are still much better than those of AFLPs or SNPs. Our study shows that the polymorphism of microsatellites is a very strong advantage for the detection of selection. For example, with only two populations, we did not identify any loci under balancing selection for SNPs, but we obtain a true positive rate of 75% with microsatellites. We were also able to detect weak effects of selection ($|\alpha| = 0.5$) with microsatellites whereas no loci were detected at this level with SNPs (results not

shown). Beaumont et Balding (2004) concluded from simulations of diallelic codominant markers that the method could not distinguish loci under balancing selection even when the selection coefficient is 20 times the migration rate. The results we obtained here show that microsatellites can be used to detect balancing selection, specially with data sets containing a large number of populations.

True :	Balancing selection			Neutral			Directional selection		
Classified :	bal.	neut.	direc.	bal.	neut.	direc.	bal.	neut.	direc.
pop-2	0 (0)	100 (100)	0 (0)	0 (0)	794 (788)	6 (12)	0 (0)	70 (59)	30 (41)
pop-6*	37 (60)	63 (40)	0 (0)	20 (61)	758 (702)	22 (37)	0 (0)	29 (28)	71 (72)
pop-10	72 (81)	28 (19)	0 (0)	21 (45)	755 (721)	24 (34)	0 (0)	21 (19)	79 (81)
pop-20	95 (97)	5 (3)	0 (0)	26 (50)	749 (709)	25 (41)	0 (0)	14 (10)	86 (90)
fst-0.05	1 (14)	99 (86)	0 (0)	0 (19)	784 (752)	16 (29)	0 (0)	33 (22)	67 (78)
fst-0.10	27 (49)	73 (51)	0 (0)	17 (46)	761 (712)	22 (42)	0 (0)	26 (22)	74 (78)
fst-0.15*	37 (60)	63 (40)	0 (0)	20 (61)	758 (702)	22 (37)	0 (0)	29 (28)	71 (72)
fst-0.25	51 (67)	49 (33)	0 (0)	23 (47)	763 (724)	14 (29)	0 (0)	30 (29)	70 (71)
size-15	15 (35)	85 (65)	0 (0)	8 (33)	776 (736)	16 (31)	0 (1)	38 (32)	62 (67)
size-30*	37 (60)	63 (40)	0 (0)	20 (61)	758 (702)	22 (37)	0 (0)	29 (28)	71 (72)
size-50	47 (65)	53 (35)	0 (0)	38 (65)	734 (684)	28 (51)	0 (0)	21 (17)	79 (83)
size-100	61 (73)	39 (27)	0 (0)	40 (66)	732 (688)	28 (46)	0 (0)	19 (18)	81 (82)
Fis-0	39 (59)	61 (41)	0 (0)	23 (50)	760 (722)	17 (28)	0 (0)	29 (24)	71 (76)
Fis-0.5*	37 (60)	63 (40)	0 (0)	20 (61)	758 (702)	22 (37)	0 (0)	29 (28)	71 (72)
Fis-1	37 (57)	63 (43)	0 (0)	25 (60)	755 (703)	20 (37)	0 (0)	25 (21)	75 (79)

Tab. 4.2 – *Numbers of AFLP loci simulated under ("true") balancing selection, neutrality, and directional selection that were classified in each category using a reversible jump cutoff of 0.79 (0.69) that give false positive rates of 5% (10%). Data sets marked with an asterisk (*) represent the same reference data set that is replicated in the table in order to make comparaisons between results.*

Influence of data set characteristics

The number of population is a key parameter for the identification of selection, especially for balancing selection. For directional selection, we observed that for all the data sets 6 populations are enough to have a good true positive rate. However, for balancing selection, we need 10 populations with AFLPs and SNPs to reach a comparable result. Microsatellites, on the other hand, perform fairly well even with only two populations.

The level of genetic differentiation also plays an important role for the detection of balancing selection. Weak genetic differentiation ($F_{ST} \leq 0.05$) makes it almost impossible to detect balancing selection with AFLP or SNP data. On the other hand, with microsatellites, even a small amount of genetic differentiation $F_{ST} = 0.01$ allows to detect balancing selection. Here we did

True :	Balancing selection			Neutral			Directional selection		
Classified :	bal.	neut.	direc.	bal.	neut.	direc.	bal.	neut.	direc.
pop-2	0 (0)	100 (98)	0 (2)	0 (0)	799 (743)	1 (57)	0 (0)	86 (62)	14 (38)
pop-6*	58 (73)	42 (26)	0 (1)	34 (44)	751 (714)	15 (42)	0 (0)	39 (29)	61 (71)
pop-10	82 (85)	18 (13)	0 (2)	31 (44)	747 (719)	22 (37)	0 (0)	18 (11)	82 (89)
pop-20	97 (98)	3 (2)	0 (0)	25 (40)	756 (728)	19 (32)	0 (0)	8 (7)	92 (93)
fst-0.05	10 (25)	90 (74)	0 (1)	10 (34)	772 (728)	18 (38)	0 (0)	27 (18)	73 (82)
fst-0.10	33 (49)	67 (50)	0 (1)	22 (39)	763 (720)	15 (41)	0 (0)	38 (28)	62 (72)
fst-0.15*	58 (73)	42 (26)	0 (1)	34 (44)	751 (714)	15 (42)	0 (0)	39 (29)	61 (71)
fst-0.25	62 (76)	38 (23)	0 (1)	34 (51)	756 (701)	10 (48)	0 (0)	32 (17)	68 (83)
size-15	32 (52)	68 (47)	0 (1)	14 (33)	772 (729)	14 (38)	0 (0)	32 (25)	68 (75)
size-30*	58 (73)	42 (26)	0 (1)	34 (44)	751 (714)	15 (42)	0 (0)	39 (29)	61 (71)
size-50	61 (78)	39 (22)	0 (0)	31 (51)	741 (710)	28 (39)	0 (0)	22 (17)	78 (83)
size-100	61 (76)	39 (24)	0 (0)	30 (52)	745 (711)	25 (37)	0 (0)	25 (22)	75 (78)

TAB. 4.3 – *Numbers of SNP loci simulated under ("true") balancing selection, neutrality, and directional selection that were classified in each category using a reversible jump cutoff of 0.86 (0.79) that give false positive rates of 5% (10%). Data sets marked with an asterisk (*) represent the same reference data set that is replicated in the table in order to make comparaisons between results.*

True :	Balancing selection			Neutral			Directional selection		
Classified :	bal.	neut.	direc.	bal.	neut.	direc.	bal.	neut.	direc.
pop-2	75 (47)	25 (53)	0 (0)	23 (4)	765 (792)	12 (4)	0 (0)	3 (12)	97 (88)
pop-4	100 (98)	0 (2)	0 (0)	20 (7)	764 (788)	16 (5)	0 (0)	0 (0)	100 (100)
fst-0.01	27 (1)	73 (99)	0 (0)	21 (1)	763 (793)	16 (6)	0 (0)	0 (1)	100 (99)
fst-0.03	90 (67)	10 (33)	0 (0)	35 (12)	744 (782)	21 (6)	0 (0)	0 (0)	100 (100)
fst-0.05	100 (100)	0 (0)	0 (0)	17 (9)	761 (786)	22 (5)	0 (0)	0 (0)	100 (100)
size-10	97 (88)	3 (12)	0 (0)	17 (2)	761 (792)	22 (6)	0 (0)	0 (0)	100 (100)
size-15	100 (100)	0 (0)	0 (0)	20 (6)	766 (791)	14 (3)	0 (0)	0 (0)	100 (100)
alpha-0.5	31 (13)	69 (87)	0 (0)	25 (4)	755 (792)	20 (4)	1 (0)	49 (69)	50 (31)
alpha-1.0	92 (79)	8 (21)	0 (0)	15 (3)	762 (791)	23 (6)	0 (0)	7 (14)	93 (86)
alpha-1.5	100 (97)	0 (3)	0 (0)	22 (4)	761 (792)	17 (4)	0 (0)	0 (2)	100 (98)

TAB. 4.4 – *Numbers of microsatellites loci simulated under ("true") balancing selection, neutrality, and directional selection that were classified in each category using a reversible jump cutoff of 0.85 (0.98). The 0.85 cutoff give the same false positive rate of 5% used for ALPF and SNP data sets. The 0.98 cutoff gives a false positive rate of only 1%.*

not notice a negative influence of high genetic differentiation on the detection of directional selection but we conducted further simulations (not presented here) with only two populations, and in that case, having a high genetic differentiation (0.25) leads to low power to detect directional selection.

The sample size is also important for the detection of balancing selection. However, we observed that a larger sample size also leads to a higher rate of false positives with AFLPs. For directional selection, increasing the sample size is less valuable ; it is possible to obtain a correct true positive rate with

only 15 individuals per population for AFLPs or SNPs and with only 10 individuals per population for microsatellites. Note that this result is only valid because we used six populations and $F_{ST} = 0.15$, but for example if we had only two population and a higher genetic differentiation, it would be necessary to have larger sample sizes.

In terms of the effect of inbreeding on the power to detect selection using dominant markers such as AFLPs, the results are very similar for all the F_{IS} values considered (Table 4.2), which suggests that inbreeding is not an issue for the application of our method.

Comparison with Beaumont et Balding (2004)'s method

Instead of using a RJMCMC approach such as the one we propose here, Beaumont et Balding (2004) adopted a simple informal criterion for identifying values of α_i that are "significant". More precisely, they define α_i to be "significant at level P" if its equal-tailed $100(1-P)\%$ posterior interval excludes zero. For example, if $P = 5\%$ then α_i is significantly positive if its 2.5% quantile is positive, and is significantly negative if its 97.5% quantile is negative. Then the estimated "P-value" is an empirical estimation of $P(\alpha_i < 0)$ via counting the proportion of negative values among the MCMC outputs for α_i. In order to have a similar interpretation for $\alpha_i > 0$, they transform this "P-value" using $2\,|p - 0.5|$ which is expected to be an empirical estimate of the probability that α_i is subject to selection, i.e. $P(\alpha_i \neq 0)$.

We use the 3 series of data sets presented above to compare the two different ways of detecting selection. We applied the informal criterion on all these simulated data sets using the same false positive rate of 5% (it corresponds respectively to a cutoff value of the informal criterion of 0.95, 0.96 and 0.98 for AFLPs, SNPs and microsatellites). A summary of the results is presented in Table 4.5. The global PPV are sightly higher for the reversible jump than the informal criterion in all cases. The results are very similar between the two approaches for microsatellites and the new method seems to be particularly useful for AFLPs and SNPs.

Spatial population expansion model

To evaluate the performance of the method when the true evolutionary model differs radically from the inference model we used SPLATCHE Currat *et al.* (2004). More specifically, SPLATCHE simulates a population expansion from a single origin in a two-dimensional habitat (strict two-dimensional stepping-

	Marker :	AFLPs		SNPs		Microsatellites	
	Method used :	IC	RJ	IC	RJ	IC	RJ
Direc.	False positive	3.2%	2.5%	2.3%	2.1%	2.5%	2.3%
	True positive	70.8%	70.4%	65.6%	67.3%	94.1%	94.0%
	PPV	73.5%	78.0%	78.5%	80.1%	82.4%	83.7%
Bal.	False positive	1.9%	2.5%	2.8%	2.9%	2.5%	2.7%
	True positive	33.7%	40.2%	45.8%	49.6%	78.7%	81.2%
	PPV	69.4%	66.7%	67.1%	68.2%	79.8%	79.1%
Total.	False positive	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
	True positive	52.3%	55.3%	55.7%	58.5%	86.4%	87.6%
	PPV	72.2%	73.5%	73.3%	74.6%	81.2%	81.5%

TAB. 4.5 – *Summary of the results for AFLP, SNP and microsatellite data sets in Tables 4.2, 4.3 and 4.4. The PPV is defined as the proportion of markers detected as being under selection that are correctly classified. Results are presented using cutoff values that lead to a 5% total false positive rate for both the Reversible Jump (RJ) method introduced here and the Informal Criterion (IC) originally proposed by* BEAUMONT *et* BALDING *(2004).*

stone model) and generates genetic samples for geographic locations chosen by the user. CURRAT *et al.* (2006) showed that a spatial population expansion can lead to false positive detection of selection when using a simple comparison of haplotype frequencies. Here we use SPLATCHE to simulate a neutral scenario of spatial expansion to check if our method suffers from the same problem. Thus, we are interested in estimating the proportion of neutral loci that could be identified as being under the effect of selection by our method if the population underwent a recent spatial expansion.

We used the human example as a template and simulated the human population expansion with origin in East Africa. We used a growth rate of 0.10, a carrying capacity of 100 for all demes, and a migration rate of 0.20. With these settings, the whole world is colonized after around 4000 generations. Following FOLL et GAGGIOTTI (2006), we "collected" genetic data from 50 individuals for each one of 36 populations chosen uniformly on the map (c.f. Figure 3 in FOLL et GAGGIOTTI 2006). Because SPLATCHE simulates fully linked markers, we performed 1000 independent genetic simulations of a single microsatellite based on the same demographic scenario. Then we grouped this 1000 independent markers in one data set and analyzed it with our method.

The distribution of probabilities $P(\alpha_i \neq 0)$ in the 1000 loci, together with their associate posterior mean F_{ST} values, are illustrated in Figure **??**. In the figure, for improved visualization, we plot the logit of the probabilities where $logit(x) = log(x/(1-x))$. Because we have microsatellite markers, a high

number of populations and an important sample size, we used a cutoff probability value of 0.99 to detect selection (with the logit transformation, we have $logit(0.99) \approx 4.6$). Not surprisingly, spatial population expansion model is a strong violation to the demographic model assumed by the method and this leads to a high number of false positive, especially for directional selection. More precisely we detected 97 loci under directional selection and only 9 loci under balancing selection. This correspond to a global false positive rare of 10.6%. The mean F_{ST} for all loci is 0.09, the locus with the strongest directional selection detected has a F_{ST} coefficient of 0.16 and a posterior estimate of $\alpha = 0.76$, while the locus with the strongest balancing selection detected has a F_{ST} coefficient of 0.062 and a posterior estimate of $\alpha = -0.37$.

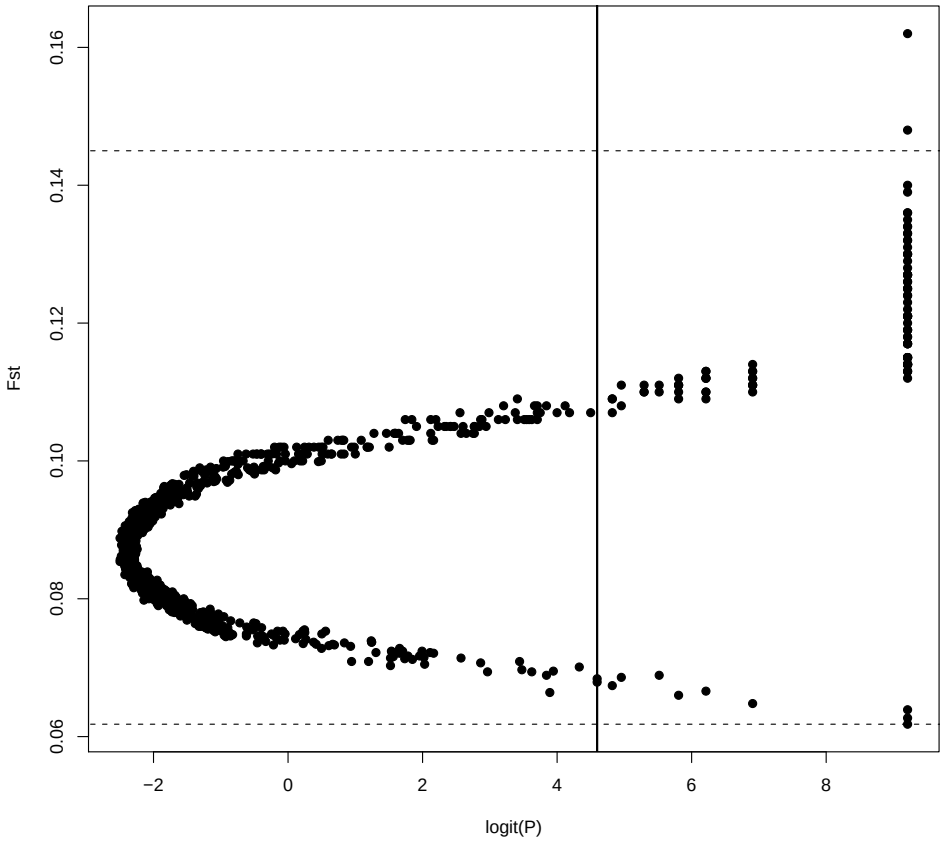

FIG. 4.3 – *Summary of the results of analysis of the simulated data set from the human spatial expansion model with the software Splatche. The data set is composed of 36 populations chosen uniformly on the world map at 1000 microsatellites. An estimate of F_{ST} is plotted against the logit of the probability $P(\alpha \neq 0)$. The vertical bar shows the critical probability of 0.99 used for identifying outlier loci, as described in the text. The two horizontal bars show the critical values of F_{ST} used to remove false positive in the analysis of the human data set.*

Application

Humans

We use the HGDP-CEPH Human Genome Diversity Cell Line Panel presented by Cann *et al.* (2002) in order to identify regions of the human genome that may be subject to selection. The last version of this data set consists of 1056 individuals from 51 subpopulations, which were scored for 835 microsatellites. We used the same cutoff value of 0.99 as in the simulated data set and the results are presented in Figure 4.4. We found 337 loci under selection : 181 were detected as under directional selection and 161 under balancing selection. This value represents 40% of the studied loci and is much higher than the proportion of false positives estimated from the simulation study that considered a similar demographic scenario. This suggests that there is a large number of selectd genes in the human genome. In order to remove as much of the false positives as possible, we only identified as "selected" the loci with a posterior estimate of α that was more extreme than the lowest and largest values observed in the simulation study. In the case of directional selection, we only included loci with $\alpha > 0.6$. After applying this correction, we were left with 104 loci out of the 181 originally detected as being under directional selection. For balancing selection, we kept only loci with a posterior estimate of $\alpha < -0.37$, the lowest value found in the simulated data set, and only one marker was excluded from the 161 loci originally detected. The two cutoff values of $\alpha = 0.6$ and $\alpha = -0.37$ we used correspond respectively to F_{ST} coefficients of 0.097 and 0.042, while the mean F_{ST} coefficient is 0.07. Thus, even after applying this correction, there are still 32% of loci that are influenced by selection. Moreover, these results suggest that a high number of loci have been subject not only to directional but also to balancing selection in the course of human evolution. 23 out of the 104 loci that we identified as being under directional selection are located on the X chromosome and 9 on the Y chromosome. On the other hand, all of the 160 markers identified as under the infulence of balancing selection are located on autosomes. Interestingly,we found that 9 out of the 10 loci with the strongest directional selection ($1.59 < \alpha < 2.64$ and $0.20 < F_{ST} < 0.38$), are located on the Y chromosome.

We identified the microsatellite loci that are located within a known gene using the NCBI UniSTS data base (`www.ncbi.nlm.nih.gov/sites/entrez?db=unists`). We found 31 distinct known genes under directional selection of which 5 were located on the X chromosome (Table 4.7) and 83 known

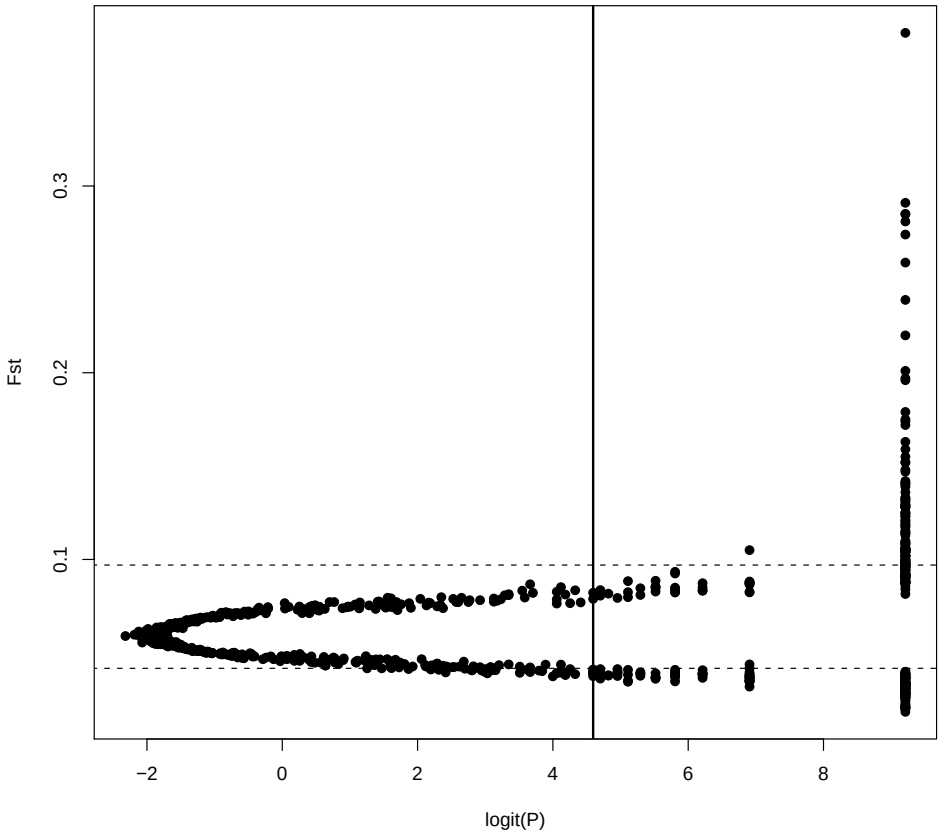

FIG. 4.4 – *Summary of the results of analysis the human data set composed of 1056 individuals from 51 populations at 835 microsatellites. An estimate of F_{ST} is plotted against the logit of the probability $P(\alpha \neq 0)$. The vertical bar shows the critical probability of 0.99 used for identifying outlier loci, as described in the text. The two horizontal bars show the critical values of F_{ST} corresponding to the critical α values of the simulated data set that give only 3 directional false positives.*

genes under balancing selection, all located on autosomes (Table 4.6). We then used the OMIM database (Online Mendelian Inheritance in Man, `ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap`) to establish the putative function of the 114 genes identified using NCBI and established that 69 genes (46 under balancing and 23 under directional selection) are referenced as implicated in a genetic disease. These results match with those of CLARK *et al.* (2003) who showed that the genes under selection are overrepresented in this data base. We also found 39 markers (30 under balancing and 9 under directional selection) associated with QTL's. 17 markers are associated to a QTL of rheumatoid arthritis or osteoarthritis, 10 markers to a QTL of blood pressure, 7 markers to a QTL of body weight, mass index or stature, and 4 markers to a QTL of body fluid distribution. Finally, we used the Panther classification of biological processes (`www.pantherdb.org`) to identify the function of these

114 known genes. Only 66 of them were in the database (Table 4.8) and they covered 17 different biological processes. The best represented were those involved in signal transduction (14 genes).

alpha	Symbol	Full Name
-1.23	TRA@	T cell receptor alpha locus
-1.18	ASPG3	Asperger syndrome, susceptibility to, 3
-1.18	PAND1	Panic disorder 1
-1.18	SCZD7	schizophrenia disorder 7
-1.16	ASPG3	Asperger syndrome, susceptibility to, 3
-1.16	GPD1L	glycerol-3-phosphate dehydrogenase 1-like
-1.15	ERC1	ELKS/RAB6-interacting/CAST family member 1
-1.09	SCZD2	schizophrenia disorder 2
-1.09	SLEH1	systemic lupus erythematosus with hemolytic anemia 1
-1.09	TRPC6	transient receptor potential cation channel, subfamily C, member 6
-1.03	TMEM16B	transmembrane protein 16B
-0.953	FLJ42117	FLJ42117 protein
-0.953	HSCRS2	Hirschsprung disease, short-segment, 2
-0.913	NIDDM3	Noninsulin-dependent diabetes mellitus 3
-0.866	OIT3	oncoprotein induced transcript 3
-0.848	ZPLD1	zona pellucida-like domain containing 1
-0.844	CALN1	calneuron 1
-0.83	BCS1L	BCS1-like (yeast)
-0.808	SCZD7	schizophrenia disorder 7
-0.804	PRKAG2	protein kinase, AMP-activated, gamma 2 non-catalytic subunit
-0.801	PAPA4	Polydactyly, postaxial, type A4
-0.773	BULN	Bulimia nervosa, susceptibility to
-0.773	CAMK1D	calcium/calmodulin-dependent protein kinase ID
-0.766	HDLC2	High density lipoprotein cholesterol level QTL on chromsome 8
-0.721	ARCC1	Age-related cortical cataract 1
-0.721	HLA-C	major histocompatibility complex, class I, C (psoriasis susceptibility 1)
-0.682	PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
-0.676	IBD6	inflammatory bowel disease 6
-0.66	FIMG1	myasthenia gravis, familial infantile, 1
-0.66	ITGAE	integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1 ; alpha polypeptide)

alpha	Symbol	Full Name
-0.651	SPAG16	sperm associated antigen 16
-0.646	MYOM1	myomesin 1, 185kDa
-0.643	NIDDM2	non-insulin-dependent diabetes mellitus (common, type 2) 2
-0.639	MCTP2	multiple C2 domains, transmembrane 2
-0.639	MDD2	Major depressive disorder 2
-0.632	NRG3	neuregulin 3
-0.614	ANIB2	aneurysm, intracranial berry 2
-0.61	NKD1	naked cuticle homolog 1 (Drosophila)
-0.559	LPO	lactoperoxidase
-0.556	KCTD8	potassium channel tetramerisation domain containing 8
-0.555	FOP	Fibrodysplasia ossificans progressiva
-0.552	R3HDM1	R3H domain containing 1
-0.546	OPCML	opioid binding protein/cell adhesion molecule-like
-0.538	DISC1	disrupted in schizophrenia 1
-0.538	SLEB1	systemic lupus erythematosus susceptibility 1
-0.527	ABCB5	ATP-binding cassette, sub-family B (MDR/TAP), member 5
-0.512	PLXNA4	plexin A4
-0.502	TAAR6	trace amine associated receptor 6 (schizophrenia disorder 5)
-0.494	GABRA4	gamma-aminobutyric acid (GABA) A receptor, alpha 4
-0.49	ANIB2	aneurysm, intracranial berry 2
-0.49	LRRC20	leucine rich repeat containing 20
-0.486	SLEN3	systemic lupus erythematosus with nephritis 3
-0.464	SLEN2	systemic lupus erythematosus with nephritis 2
-0.456	OR51A6P	olfactory receptor, family 51, subfamily A, member 6 pseudogene
-0.452	PDE9A	phosphodiesterase 9A
-0.415	ALSFTD	Amyotrophic lateral sclerosis with frontotemporal dementia

TAB. 4.6 – *Genes identified as under balancing selection with the corresponding posterior estimate of α. Lowest values of α suggest a stronger effect of selection.*

Littorina saxatilis

In order to present an application to AFLPs, we reanalyzed the *Littorina saxatilis* data set of WILDING *et al.* (2001), studied also by GRAHAME *et al.* (2006). The data consists of 290 polymorphic AFLP loci, surveyed in 4 different rocky shores in Britain : Thornwick Bay, Flamborough (TH), Filey Brigg (FY), Old Peak (OP) and Robin Hood's Bay (RB). In this region *L. saxatilis* is found as two morphological forms ("H" and "M") that show good evidence of partial reproductive isolation. One set of individuals of each morphological form was sampled in each shore, with the exception of the RB shore where two sets of "M" were sampled. Each of the 8 resulting samples is composed of 43 to 51 individuals.

In each shore two hypotheses can explain the observed divergence bet-

alpha	Symbol	Full Name
1.85	PHACTR1	phosphatase and actin regulator 1
1.56	CHDS3	Coronary heart disease, susceptibility to, 3
1.25	EBM	epidermolysis bullosa, macular type
1.15	RHOA	ras homolog gene family, member A
1.09	IDDMX	Diabetes mellitus, insulin-dependent, X-linked, susceptibility to
1.09	PQBP1	polyglutamine binding protein 1
1.01	E2F6	E2F transcription factor 6
0.934	F13A1	coagulation factor XIII, A1 polypeptide
0.881	ALS7	Amyotrophic lateral sclerosis 7
0.87	SLEB3	systemic lupus erythematosus susceptibility 3
0.87	STGD4	Stargardt disease 4 (autosomal dominant)
0.801	LYST	lysosomal trafficking regulator
0.801	RMD1	rippling muscle disease 1
0.801	SLEB1	systemic lupus erythematosus susceptibility 1
0.713	CHDS3	Coronary heart disease, susceptibility to, 3
0.696	SOS1	son of sevenless homolog 1 (Drosophila)
0.661	PSORS3	psoriasis susceptibility 3
0.661	RAB37	RAB37, member RAS oncogene family
0.636	RELN	reelin
0.626	GLIS3	GLIS family zinc finger 3
0.617	GARS	glycyl-tRNA synthetase
0.617	JAZF1	JAZF zinc finger 1

Tab. 4.7 – *Genes identified as under directional selection with the corresponding posterior estimate of α. Highest values of α suggest a stronger effect of selection.*

Biological Process	Number of Genes
Signal transduction	14
Nucleoside, nucleotide and nucleic acid metabolism	7
Protein metabolism and modification	5
Immunity and defense	5
Cell structure and motility	5
Developmental processes	5
Neuronal activities	4
Transport	4
Intracellular protein traffic	3
Cell adhesion	3
Cell cycle	2
Cell proliferation and differentiation	2
Sensory perception	2
Lipid, fatty acid and steroid metabolism	2
Muscle contraction	1
Electron transport	1
Protein targeting and localization	1

Tab. 4.8 – *Biological process categories of genes identified as under selection.*

ween the two morphological forms (Grahame *et al.* 2006) : an allopatric divergence followed by a secondary contact, or a primary parapatric divergence Wilding *et al.* (2001). In both cases populations are likely to be exchanging genes only in the region of contact, and using the 8 populations in a single analysis would lead to a violation of the demographic model assumed by our inference method. This is also supported by the Neighbor-Joining tree constructed by Wilding *et al.* (2001) from the loci they identified as neutral : populations were clustered by site (they also constructed a tree using all loci which lead to a grouping of populations by morphotype H and M).

Wilding *et al.* (2001) used a modified version of the Fdist model (Beaumont et Nichols 1996) to detect selection from dominant markers. They analyzed three data sets, corresponding to the three shores where both morphotypes were sampled, each one containing two populations. One potential problem of the Beaumont et Nichols (1996) method is the necessity to estimate Nm from the data set in order to perform simulations with this target value. However, the estimation of Nm assumes neutrality and is overestimated in presence of directional selection. In order to avoid this problem, they used an iterative procedure whereby the mean F_{ST} calculated from the full data set is used as input of a first Fdist run, and then it is iteratively modified as outlier loci are removed. After four such steps, Wilding *et al.* (2001) retained only loci that were lying above the 0.99 quantile in all three H-M comparisons, and identified 15 loci under selection.

We made the same three analysis of each two-populations data sets using our method. The Bayesian model we used takes explicitly into account the loci under selection in the estimation of F_{ST} coefficients in equation 4.4 and, therefore, does not suffer from the problem mentioned above. Beaumont et Balding (2004) compared the critical p-values between the Bayesian method and Fdist by matching the false positive rate of 6800 neutral loci. They showed that a level of 1% for Fdist is equivalent to a level of 10% for the Bayesian model. Here, the sensitivity study above indicates that a 10% level for the informal criterion used by Beaumont et Balding (2004) is equivalent to a cutoff value of 0.7 for the posterior probability estimated by our reversible jump version of the method. We identified 13 loci with a probability above 0.7 and they all belong to the list of 15 loci identified by Wilding *et al.* (2001). The two missing loci are named "A37" and "F11" by Wilding *et al.* (2001) and, according to our method, both are identified as outlier in only two of the data sets. More precisely, the "A37" locus has a posterior probability of only 0.53 in the Filey data set, and the "F11" locus has a posterior probability of 0.65 in the Old Peak data set. These loci are at the lower tail of the allele frequency

distribution estimated by Wilding *et al.* (2001) in two of the three data sets considered. If we were to use a cutoff value of 0.65 instead of 0.7 we would include the "F11" locus in the list of selected loci but also an additional marker not found by Wilding *et al.* (2001).

We also analyzed these three sets of two populations as a single data set of 6 populations to investigate the influence of the violation of the demographic model assumed by the method. Using a cut off value of 0.99, all the 13 loci found in the pairwise analyzes are identified as outliers, but we also find 4 additional loci. The results of the simulations of the spatial expansion model suggest that these loci could be false positives due to the violation of the demographic model assumed. As it was the case for the human data set, these 4 loci have a posterior estimate of α situated at the tail of the distribution of α values for loci with a posterior probability above 0.99. More precisely, the maximum estimated value of α for these 4 additional loci is 1.89, while most of the loci identified as outliers (7 out of 13) in the pairwise analyzes have a posterior estimate of α greater that this value.

In order to establish which of the two approaches is the most appropriate one, we modified the simulation scheme presented above to incorporate a different demographic scenario. More precisely, instead of simulating the six populations under an island model, we simulated first three populations from this model (for the three shores), and then, from each one of them generated allele frequencies for two populations (corresponding to the two different morphotypes). This demographic history mimics the Neighbor-Joining tree constructed by Wilding *et al.* (2001) from the loci they identified as neutral. We chose simulation parameters to obtain a data sets close to the real one. We simulated 290 such loci and 50 individuals in each population. We used $F_{ST} = 0.05$ between the ancestral population and the three intermediate populations and $F_{ST} = 0.03$ between the intermediate populations and the six populations sampled. The ancestral allele frequencies were simulated from a beta distribution with both parameters equal to 0.5 and we chose $F_{IS} = 0.5$. We added selection to 20 loci using $\alpha = 2.5$.

We performed the same analysis on this data set than with the real one : 17 loci over the 20 loci under selection had a posterior probability above 0.7 in all three pairwise analyzes. We did not detect any false positive and all the three false negative loci were identified as outliers in two of the three analyzes. We then carried out an analysis with the 6 populations as a single data set and identified all the 20 loci as selected with a posterior probability above 0.99. However, we also identified 4 additional false positive loci. The maximum estimated value of α for these 4 additional loci is 2.02, while only 11 of the 20

true outliers loci have a posterior estimate of α greater than this value. These results suggest that, under this particular demographic model, it is better to carry out pairwise analyzes instead of a single global one. Moreover, it seems that the best strategy is to identify as selected all loci that are outliers in two of the three pairwise analyzes. Indeed, if we use such an approach, then we retrieve all the 20 loci under selection without identifying any false positives. Note that we can obtain the same result even if we raise the cutoff probability to 0.78.

Applying this approach to the periwinkle dataset of Wilding *et al.* (2001) we identify as selected all the 15 loci originally found by them, and also 6 additional outliers. We obtain the same result even if we raise the cutoff probability up to 0.81. Thus, our analyzes suggest that there is a total of 21 loci that are influenced by selection in this species.

Discussion

We present an extension of Beaumont et Balding (2004) method to detect outlier loci that is applicable to both dominant and codominant markers. Additionally, it implements a rigorous way of estimating the posterior probability of a given locus being under the effect of selection. In their original formulation, Beaumont et Balding (2004) focused on the posterior distribution of locus-specific effects, α_i, and use an approximate method to determine if a given locus is significantly influenced by selection. On the other hand, the RJMCMC method we implemented is based on the idea that Equation 4.4 gives rise to two alternative models, one that includes both locus- and population-specific effects and another one that only includes the latter. Thus, it is possible to estimate the posterior probability of each alternative model and based on them decide which are the loci subject to selection.

Beaumont et Balding (2004) show that the method is not appropriate to detect balancing selection. However, our analyzes showed that we can indeed do so if we use microsatellites or some other multiallelic marker. Additionally, we showed that dominant markers perform almost as well as codominant biallelic ones. Thus, the degree of polymorphism is one of the most important factors determining the power to detect outlier loci. We also identified three other parameters that are particularly important determinants of power : the sample size, the number of populations and the level of genetic differentiation. In general, a sample size of 30 individuals seems enough when the study considers six or more populations. In terms of the effect of neutral

genetic differentiation, extremely low F_{ST} values decrease the power to detect balancing selection. On the other hand, very high values limit the detection of directional selection. Note, however, that these problems are avoided by using multiallelic markers.

Foll *et al.* (2007) showed that the estimation of F_{IS} from dominant markers is strongly biased by the ascertainment of markers when assuming the island model. However, in our case we are only concerned with the potential bias in the estimation of the locus specific effect and our simulation study shows that it suffices to incorporate the uncertainty about the inbreeding coefficient in order to avoid such a bias. In the present formulation we let F_{IS} to move freely between 0 and 1. It would be possible to incorporate prior knowledge about the degree of inbreeding by using a narrower interval for the prior distribution of F_{IS}. However, the fact that AFLP results are very similar to those of SNPs suggests that we cannot expect a large improvement by implementing this modification.

We investigated the potential biases that could be introduced if the demographic history of the species under study does not follow the model assumed by our method. For this purpose we generated synthetic data with SPLATCHE (Currat *et al.* 2004) using the human population expansion example as a template. The results indicate that violations to this assumption lead to a false positive rate of 10% and this problem seems to be more acute for directional than for balancing selection.

As an example of an application of the method with codominant markers, we analyzed the HGDP microsatellite database (Cann *et al.* 2002) and found that at least 32 % of the markers are influenced by selection. This is likely to be an underestimate because in order to minimize the false positive rate (see above) we only identified as "selected" the loci with a posterior estimate of α that was more extreme than the lowest and largest values observed for the synthetic data set generated with SPLATCHE. Many of the outlier loci (43%) are located within known genes and of these 60% are implicated in genetic diseases and 34% are QTLs. Although the number of genes considered in our study is small, our results are in agreement with previous studies. For example, Clark *et al.* (2003) classified 7645 genes using the Panther data base and identified the ones showing the strongest evidence for positive selection. He found that the category with the highest number of genes was "Signal transduction", which is also found as the most frequent in our study using the same classification. Nielsen *et al.* (2005) expanded Clark *et al.* (2003) analysis to consider 13 731 genes and obtained somewhat different results. In particular, the most well represented class of genes are involved in immunity

and defense. This difference is probably due to the fact that the inclusion of an outgroup (mouse) in the former study allowed the authors to make inferences regarding human specific processes. On the other hand, as explained by Nielsen *et al.* (2005), their analysis could not distinguish between selection that is particular to the human evolutionary lineage and positive selection that tends to occur in both chimps and humans. Given that we analyzed a human database, our study is more similar to that of Clark *et al.* (2003). Note, however, 5 genes associated to immunity and defense represented the third most common ones in our analysis.

We also present an example of an application with dominant markers. We analyzed the *Littorina saxatilis* dataset published by Wilding *et al.* (2001) and consisting on two samples from each of three shores in which two different morphotypes coexisted. Following Wilding *et al.* (2001) we first carried out separate analyzes for each shore, each including a sample from each morphotype and choose to identify as selected only the loci that were outliers in all three pairwise analyzes. Using this approach we detected only 13 out of the 15 loci found by Wilding *et al.* (2001). We also conducted an analysis with all six populations and identified a total of 17 outlier loci. In order to decide which of the two approaches was more appropriate we carried out a simulation study based on the demographic history of *L. saxatilis* and found that the use of the first approach can lead to false negatives while the use of the second one can lead to false positives. Thus, we propose an third strategy consisting in identifying as selected all loci that are outliers in at least two of the three pairwise analyzes. Using this approach on the simulated data recovers all the selected loci and does not lead to false positives. In the case of the *L. saxatilis* dataset we identified 21 selected loci, six more than Wilding *et al.* (2001).

The human and the *Littorina saxatilis* data sets are two good examples of how to deal with the violation of the demographic model assumed here. For the human data set we showed that keeping only loci with extreme posterior estimate of α eliminates most of the false positives. However, this is a very conservative approach and can lead to many false negative loci with moderate selection coefficients. For the *Littorina saxatilis* data set, taking into account the demographic history by carrying out three pairwise analyzes, one for each shore, allows to avoid all false positive and false negative loci at the same time. Thus, making a preliminary simulation study with a synthetic data set that mimics the real one is an efficient way of avoiding false positives. In this way, it is possible to choose the cutoff value as explained in the results section or to design the best strategy to analyze the samples.

A natural extension of our method would be to incorporate the full demographic history into the analysis instead of imposing a simple demographic model. If the demographic history is known, it may be possible to simply incorporate it into the estimation process, otherwise it would be necessary to estimate it from the genetic data.

Acknowledgments

We thank Mark Beaumont for the many helpful discussions we had on the subject of this article. This work was supported by the Fond National de la Science (grant ACI-IMPBio- 2004-42-PGDA). M.F. holds a Ph.D. studentship from the Ministère de la Recherche. Most the computations presented in this paper were performed on the cluster HealthPhy (CIMENT, Grenoble, France). The software implementing the method is available at http ://www-leca.ujf-grenoble.fr/logiciels.htm both for unix and windows platforms.

Discussion

« Il existe trois sortes de mensonges : les mensonges, les gros mensonges et les statistiques. »

Leonard Courtney

Plusieurs aspects des modèles proposés ici méritent d'être discutés. Nous avons regroupé ces différentes questions en trois parties : les problèmes d'ordre statistique, les problèmes liés à l'histoire démographique des populations et les problèmes spécifiques à l'utilisation de certains marqueurs génétiques. Nous présentons pour terminer quelques pistes sur les aspects qu'il reste à développer dans ce cadre.

Les outils statistiques

Le fait que les populations humaines partagent une histoire commune pour la dérive pose un problème statistique général dans ce type d'analyse. En effet, plusieurs populations étudiées sont issues du même goulot d'étranglement, comme par exemple les populations amérindiennes, et vont alors toutes porter le même signal génétique. Ainsi, les échantillons venant de la même zone géographique peuvent être en partie considérés comme dupliqués et peuvent affecter la significativité de la régression (Felsenstein 1985). Différentes simulations ont été effectuées à l'aide du logiciel SPLATCHE (Currat *et al.* 2004) pour étudier ce problème. Nous avons considéré des scénarios avec de multiples populations appartenant à la même région (12 population en Amérique du Sud), et au contraire des scénarios avec une unique population dans les continents nord et sud américains. Les résultats indiquent que la méthode est capable d'identifier correctement l'effet de la distance ou d'autres facteurs malgré ce problème, et ce, même quand le modèle démographique réel diffère radicalement du modèle en ile.

La méthode ABC utilisée dans le chapitre 3 est une bonne illustration d'un problème simple où l'on ne peut pas obtenir de fonction de vraisemblance. (Nicholson *et al.* 2002, Nielsen *et al.* 2004) ont proposé des formules qui tiennent compte du biais de recrutement des marqueurs, mais nous avons montré qu'elles ne correspondaient pas au processus biologique sous-jacent. Nous avons proposé une solution innovante qui peut s'appliquer dans beaucoup d'autres cas. En effet, si le calcul de la vraisemblance limite l'utilisation de modèles complexes, l'algorithme ABC ne nécessite que de pouvoir simuler ce modèle, ce qui se révèle réalisable dans la plupart des cas.

Cependant, les méthodes de type ABC doivent encore être améliorées pour permettre des estimations plus performantes. Par exemple, les modèles qui nécessitent l'estimation d'un grand nombre de paramètres comme celui du chapitre 4 semblent difficilement pouvoir être utilisés avec cet algorithme. En particulier, le choix des statistiques descriptives qui résument l'informa-

tion contenue dans les données est difficile. Un défi important à relever est de trouver des statistiques sommaires exhaustives pour chaque paramètre à estimer. En général les statistiques utilisées dans ces méthodes sont la moyenne ou le mode d'un paramètre donné (F_{ST}, la moyenne du déséquilibre de liaison entre paire de loci, le nombre moyen de différences entre deux séquences ADN etc.). Ici nous avons proposé d'utiliser des quantiles de la distribution d'une statistique sommaire et montré qu'ils fournissaient bien plus d'information qu'une unique estimation ponctuelle.

Pour terminer, on peut noter qu'il faut interpréter avec précaution les résultats de la méthode permettant d'inclure les facteurs environnementaux dans les estimations. Comme toute méthode basée sur la corrélation entre plusieurs variables, il est tentant de vouloir en déduire directement une causalité. On pourra avoir par exemple deux phénomènes corrélés à un même troisième, qui en est la cause, et qui n'est pas observé.

L'histoire démographique

Plusieurs exemples ont été donnés concernant l'influence de l'histoire démographique des populations considérées sur les estimations. Deux comportements différents ont été observés selon que l'on mesure la différenciation génétique globale ou que l'on cherche à séparer les effets neutres des effets adaptatifs. Dans le premier cas, il apparait qu'un modèle démographique très différent du modèle en ile ne vient pas particulièrement perturber l'identification de la structure spatiale ou des facteurs environnementaux qui en sont responsables. De ce fait, même si les coefficients F_{ST} estimés ne peuvent pas être mis directement en relation avec le taux de migration, comme dans le modèle en ile, ils restent une mesure informative de la structure génétique. Cependant, cette « robustesse » du modèle Dirichlet-multinomial ne se vérifie pas de la même façon dans le second cas. Par exemple, nous avons montré, en simulant un modèle démographique proche de celui de l'expansion humaine (SPLATCHE, Currat *et al.* 2004), que l'on obtenait un taux de faux positifs de l'ordre de 10%, pour la plupart sous sélection directionnelle.

Nous avons montré qu'il était possible d'éviter ces faux positifs en effectuant des simulations préliminaires. La méthode la plus simple a été utilisée pour les données humaines, où n'ont été conservés que les loci avec un niveau de sélection plus élevé que celui observé sur les faux positifs des données simulées. Néanmoins cette approche est très conservative et limite la détection de sélection modérée. Pour les données de Littorine, l'histoire démogra-

phique des populations considérées nous a servi de guide pour concevoir une stratégie d'analyse qui a permis vraisemblablement d'identifier correctement la quasi-totalité des marqueurs. Pourtant cette approche ne peut s'appliquer que dans des cas très spécifiques et comporte un sérieux point faible. La stratégie d'analyse s'est basée sur l'histoire démographique qui a été estimée par un arbre issu d'un algorithme de type « Neighbor-joining ». Cependant, l'estimation de cet arbre nécessite l'utilisation de marqueurs neutres et le problème devient circulaire. Evidemment cela ne s'applique pas dans les cas où l'on connait l'histoire démographique par d'autres moyens.

L'utilisation de facteurs liés au processus de colonisation dans le modèle du chapitre 2 a permis de retracer l'histoire démographique de l'expansion humaine. Toutefois, cette approche comporte le paradoxe de supposer un modèle démographique simple (en ile) dans le but d'en estimer un autre à partir des données. Les méthodes basées sur un algorithme de type Neighbor-joining sont plus adaptées à ce type de problème, mais résument toute l'information des données dans des statistiques sommaires (les coefficients F_{ST} par paire de populations). On peut retenir que la possibilité d'estimer l'histoire démographique à partir de notre méthode doit être réservée à des données avec une échelle spatiale importante.

Les marqueurs utilisés

Notre étude a montré que l'utilisation de marqueurs dominants ne peut se faire qu'en tenant compte de leur spécificité. En particulier, les estimations dans le modèle utilisé ici sont extrêmement sensibles au choix non aléatoire des marqueurs. Par exemple, la pratique courante de ne pas retenir les marqueurs monomorphes, ou de ne retenir que les marqueurs dont le nombre d'individus possédant une bande se situe entre deux bornes, a un effet important. Si ces contraintes sont nécessaires pour des raisons techniques, nous avons montré qu'il est possible de les intégrer dans le modèle à travers la méthode ABC.

Cette dernière méthode permet d'obtenir des estimations non biaisées des coefficients de consanguinité à partir de marqueurs AFLP. Cependant, l'algorithme ABC qui a permis cela, ne peut pas facilement être utilisé pour l'identification de la sélection. Pour résoudre ce problème, nous avons montré qu'il suffisait d'incorporer l'incertitude sur le coefficient de consanguinité pour obtenir des estimations non biaisées des effets locus-spécifiques de la différenciation génétique, liés à la sélection. La méthode ainsi proposée offre

une puissance quasiment équivalente à celle des marqueurs SNP. Malheureusement, cette remarque ne concerne que la composante locus-spécifique, et pour cette raison la méthode du chapitre 2 n'a pas pu être étendue aux marqueurs AFLP. Des analyses supplémentaires seront nécessaires pour connaitre l'effet précis qu'aurait ce biais sur les estimations.

Les techniques de biologie moléculaire semblent aujourd'hui évoluer vers la possibilité d'obtenir un séquençage systématique de nombreux individus. Les méthodes présentées ici peuvent s'appliquer à ces données en utilisant par exemple des SNP éloignés dans le génome car nos modèles supposent l'absence de déséquilibre de liaison. Toutefois cette approche a le défaut de n'utiliser qu'une petite partie de l'information disponible dans ces nouveaux types de données.

Les questions non résolues

La méthode du chapitre 2 permet d'identifier les facteurs environnementaux responsables de la différenciation génétique. Pour cela, on considère un unique coefficient F_{ST} dans chaque population. Si les marqueurs utilisés sont neutres, ce coefficient est constant pour tous les loci, mais si un certain nombre sont soumis à la sélection, le coefficient F_{ST} mesurera la différenciation génétique résultante à la fois des forces neutres et des forces adaptatives. Si on cherche par exemple à n'estimer que les processus d'ordre démographique, il est souhaitable de commencer par retirer du jeu de données les marqueurs soumis à la sélection. La méthode du chapitre 4 peut être utilisée dans ce but, mais cette approche en deux temps ne permet pas de prendre en compte l'incertitude sur le caractère neutre de chaque locus.

Pour cette raison, il serait intéressant d'inclure les variables environnementales dans le modèle du chapitre 4 sur la partie population-spécifique de la différenciation génétique. Cette approche permettrait de prendre en compte l'incertitude sur la neutralité de chaque marqueur dans les estimations. Une autre extension serait d'ajouter l'histoire démographique au sein de l'analyse au lieu d'imposer un modèle démographique simple. Dans les cas où cette histoire est connue, il serait possible de l'incorporer dans les estimations, sinon, il serait nécessaire de l'estimer avec les données génétiques.

Pour finir, les méthodes introduites ne permettent pas d'identifier les facteurs environnementaux qui sont responsables spécifiquement de la différenciation adaptative. On peut noter qu'il est possible d'obtenir un tel résultat en effectuant une analyse comprenant tous les marqueurs et une analyse avec

uniquement les marqueurs neutres, mais cette approche se limite à des cas très particuliers.

Conclusion générale

Les différentes méthodes introduites ici représentent de nouveaux outils performants dans le domaine de la génomique des populations. Elles fournissent de nombreuses informations qui peuvent aider à mieux comprendre l'histoire démographique et évolutive des espèces étudiées. Pour cela, elles permettent d'identifier les facteurs environnementaux responsables de la différenciation génétique et de détecter des régions du génome soumises à la sélection naturelle. Ces méthodes ont été implémentées dans trois logiciels distribués librement et vont pouvoir être appliquées dans un grand nombre de disciplines, allant de la génétique humaine, à la biologie de la conservation et à l'agronomie.

Nous avons aussi introduit un cadre général pour obtenir des estimations non biaisées en prenant en compte le biais de recrutement des marqueurs AFLP. La méthode qui en découle représente une amélioration importante par rapport à celles qui existaient jusqu'alors et sera très utile pour l'étude de la structure génétique à partir de ces marqueurs. De plus la méthode que nous avons développée rend possible l'identification des régions du génome sélectionnées à partir de ces marqueurs, avec une puissance quasiment équivalente aux marqueurs SNP. Ces deux points offrent de nouveaux outils qui n'étaient jusqu'alors pas disponibles pour les espèces non-modèles.

Beaucoup de choses restent à faire pour améliorer ces méthodes, et en particulier, concernant l'effet de confusion entre histoire démographique et sélection naturelle, et l'identification des déterminants environnementaux de la sélection. La génomique des populations lance de grands défis mathématiques et, en raison des connaissances fondamentales qu'elle peut apporter, motive à elle seule l'élaboration de nouveaux outils statistiques qui lui sont dédiés.

Bibliographie

Amos, W., et A. Manica, 2006 Global genetic positioning : Evidence for early human population centers in coastal habitats. Proc. National Acad. Sciences United States Am. **103** : 820–824. (Cité page 67.)

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biol. **63** : 221–230. (Cité pages 21, 33, 43, 44, 45, 46, 55, 64, 87 et 125.)

Balding, D. J., M. Greenhalgh, et R. A. Nichols, 1996 Population genetics of STR loci in caucasians. Int. J. Legal Medicine **108** : 300–305. (Cité pages 114 et 125.)

Balding, D. J., et R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica **96** : 3–12. (Cité pages 43, 44, 46, 64, 87, 90 et 124.)

Balloux, F., 2001 Easypop (version 1.7) : A computer program for population genetics simulations. J. Heredity **92** : 301–302. (Cité pages 35 et 50.)

Beaumont, M. A., 2005 Adaptation and speciation : what can F-st tell us ? Trends In Ecology & Evolution **20** : 435–440. (Cité pages 29 et 123.)

Beaumont, M. A., et D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecology **13** : 969–980. (Cité pages 29, 87, 113, 114, 115, 120, 123, 124, 125, 126, 129, 131, 132, 135, 137, 138, 145 et 147.)

Beaumont, M. A., et R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. Royal Soc. London Series B-biological Sciences **263** : 1619–1626. (Cité pages 118, 123 et 145.)

Beaumont, M. A., W. Y. Zhang, et D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162** : 2025–2035. (Cité pages 11, 92, 97 et 101.)

Bensch, S., et M. Akesson, 2005 Ten years of AFLP in ecology and evolution : why so few animals ? Mol. Ecology **14** : 2899–2914. (Cité pages 79, 87 et 128.)

Bonin, A., D. Ehrich, et S. Manel, (in press) Statistical analysis of aflp data : a toolbox for molecular ecologists and evolutionists. Molecular Ecology . (Cité page 108.)

Bowcock, A., J. Kidd, J. Mountain, J. Herbert, L. Carotenuto, *et al.*, 1991 Drift, Admixture, and Selection in Human Evolution : A Study with DNA Polymorphisms. Proceedings of the National Academy of Sciences **88** : 839–843. (Cité page 123.)

Brooks, S. P., P. Giudici, et G. O. Roberts, 2003 Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. J. Royal Statistical Soc. Series B-statistical Methodology **65** : 3–39. (Cité pages 75 et 131.)

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel, *et al.*, 2002 A human genome diversity cell line panel. Science **296** : 261–262. (Cité pages 6, 35, 37, 60, 116, 118, 140 et 148.)

Cavalli-Sforza, L. L., et M. W. Feldman, 2003 The application of molecular genetic approaches to the study of human evolution. Nat. Genet. **33** : 266–275. (Cité page 47.)

Cheverud, J. M., G. P. Wagner, et M. M. Dow, 1989 Methods for the comparative-analysis of variation patterns. Systematic Zoology **38** : 201–213. (Cité page 63.)

Ciofi, C., M. A. Beaumont, I. R. Swingland, et M. W. Bruford, 1999 Genetic divergence and units for conservation in the Komodo dragon Varanus komodoensis. Proc. Royal Soc. London Series B-biological Sciences **266** : 2269–2274. (Cité page 87.)

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, *et al.*, 2003 Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302** : 1960–1963. (Cité pages 119, 141, 148 et 149.)

Crow, J., et M. Kimura, 1970 *An introduction to population genetics theory*. Harper and Row, New York. (Cité page 17.)

Currat, M., L. Excoffier, W. Maddison, S. P. Otto, N. Ray, *et al.*, 2006 Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in

homo sapiens" and "microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". Science **313** : 172. (Cité page 138.)

Currat, M., N. Ray, et L. Excoffier, 2004 Splatche : a program to simulate genetic diversity taking into account environmental heterogeneity. Mol. Ecology Notes **4** : 139–142. (Cité pages 35, 38, 39, 50, 56, 65, 115, 118, 131, 137, 148, 152 et 153.)

Excoffier, L., 2001 *Handbook of Statistical Genetics*, chapter Analysis of population subdivision. John Wiley & Sons, New York, 271–308. (Cité page 43.)

Excoffier, L., 2002 Human demographic history : refining the recent African origin model. Current Opinion In Genetics & Development **12** : 675–682. (Cité page 60.)

Excoffier, L., et G. Heckel, 2006 Computer programs for population genetics data analysis : a survival guide. Nature Rev. Genetics **7** : 745–758. (Cité pages 79 et 87.)

Falush, D., M. Stephens, et J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data : Linked loci and correlated allele frequencies. Genetics **164** : 1567–1587. (Cité pages 18, 44, 46, 64, 87 et 90.)

Felsenstein, J., 1985 Phylogenies and the comparative method. Am. Naturalist **125** : 1–15. (Cité pages 67 et 152.)

Foll, M., M. Beaumont, et O. Gaggiotti, 2007 An approximate bayesian computation approach to overcome biases that arise when using aflp markers to study population structure. Submitted to Genetics . (Cité pages 129, 132 et 148.)

Foll, M., et O. Gaggiotti, 2006 Identifying the environmental factors that determine the genetic structure of populations. Genetics **174** : 875–891. (Cité pages 87, 90, 124, 125 et 138.)

Foll, M., et O. Gaggiotti, 2007 A genome scan method to identify selected loci appropriate for both dominant and codominant markers : A bayesian perspective. Submitted to Genetics .

Gaggiotti, O. E., S. P. Brooks, W. Amos, et J. Harwood, 2004 Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. Mol. Ecology **13** : 811–825. (Cité pages 26, 34 et 47.)

Gaggiotti, O. E., F. Jones, W. M. Lee, W. Amos, J. Harwood, *et al.*, 2002 Patterns of colonization in a metapopulation of grey seals. Nature **416** : 424–427. (Cité page 33.)

Galassi, M. J., J. Davies, et J. T. et al., 2006 *GNU Scientific Library Reference Manual (2nd Ed.), ISBN 0954161734*. (Cité page 103.)

Geffen, E., M. J. Anderson, et R. K. Wayne, 2004 Climate and habitat barriers to dispersal in the highly mobile grey wolf. Mol. Ecology **13** : 2481–2490. (Cité page 47.)

Graham, J., J. Curran, et B. S. Weir, 2000 Conditional genotypic probabilities for microsatellite loci. Genetics **155** : 1973–1980. (Cité pages 24 et 133.)

Grahame, J. W., C. S. Wilding, et R. K. Butlin, 2006 Adaptation to a steep environmental gradient and an associated barrier to gene exchange in Littorina saxatilis. Evolution **60** : 268–278. (Cité pages 143 et 145.)

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82** : 711–732. (Cité pages 12, 27, 35, 49, 73, 113, 114, 115, 126 et 130.)

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont, *et al.*, 2005 Bayesian estimation of recent migration rates after a spatial expansion. Genetics **170** : 409–417. (Cité page 102.)

Hastings, W., 1970 Monte carlo sampling methods using markov chains and their applications. Biometrika **57** : 97–109. (Cité pages 9 et 34.)

Hill, W. G., et B. S. Weir, 2004 Moment estimation of population diversity and genetic distance from data on recessive markers. Mol. Ecol. **13** : 895–908 Erratum in : Mol. Ecol. 2004 13 : 3617. (Cité pages 22 et 88.)

Holsinger, K. E., 1999 Analysis of genetic diversity in geographically structured populations : A Bayesian perspective. Hereditas **130** : 245–255. (Cité pages 20, 45 et 127.)

Holsinger, K. E., et P. O. Lewis, 2002 *Hickory : A Package for Analysis of Population Genetic Data v1.1*. (Cité pages 27, 79 et 129.)

Holsinger, K. E., P. O. Lewis, et D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol. Ecology **11** : 1157–1164. (Cité pages 22, 27, 77, 79, 80, 88, 89, 90, 91, 93, 94, 108, 115, 128 et 129.)

Hunley, K., et J. C. Long, 2005 Gene flow across linguistic boundaries in Native North American populations. Proc. National Acad. Sciences United States Am. **102** : 1312–1317. (Cité page 47.)

Legendre, P., 2000 Comparison of permutation methods for the partial correlation and partial Mantel tests. J. Statistical Computation Simulation **67** : 37–73. (Cité pages 33, 63 et 64.)

Lewontin, R., et J. Krakauer, 1973 Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics **74** : 175–195. (Cité pages 29 et 123.)

Lockwood, J. R., K. Roeder, et B. Devlin, 2001 A Bayesian hierarchical model far allele frequencies. Genetic Epidemiology **20** : 17–33. (Cité page 133.)

Lynch, M., et B. G. Milligan, 1994 Analysis of population genetic-structure with rapd markers. Mol. Ecology **3** : 91–99. (Cité pages 22 et 88.)

Marchini, L. R., et A. Cardon, 2002 Discussion on the meeting on "statistical modelling and analysis of genetic data.". J. Royal Statistical Soc. Series B-statistical Methodology **64** : 737–775. (Cité pages 18 et 43.)

Marjoram, P., J. Molitor, V. Plagnol, et S. Tavare, 2003 Markov chain Monte Carlo without likelihoods. Proc. National Acad. Sciences United States Am. **100** : 15324–15328. (Cité page 97.)

Meudt, H. M., et A. C. Clarke, 2007 Almost forgotten or latest practice ? AFLP applications, analyses and advances. Trends Plant Sci. **12** : 106–117. (Cité pages 79, 87, 95 et 128.)

Moilanen, A., et M. Nieminen, 2002 Simple connectivity measures in spatial ecology. Ecology **83** : 1131–1145. (Cité pages 36, 46 et 55.)

Moran, P. A. P., 1975 Wandering distributions and electrophoretic profile. Theoretical Population Biol. **8** : 318–330. (Cité pages 24 et 133.)

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson, *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. Royal Statistical Soc. Series B-statistical Methodology **64** : 695–715. (Cité pages 18, 27, 28, 43, 44, 46, 64, 81, 82, 95, 97, 99, 107 et 152.)

Nielsen, R., 2005 Molecular signatures of natural selection. Ann. Rev. Genetics **39** : 197–218. (Cité page 113.)

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. **3** : 976–985. (Cité pages 123, 148 et 149.)

Nielsen, R., M. J. Hubisz, et A. G. Clark, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168** : 2373–2382. (Cité pages 27, 79, 82, 95, 97, 99, 107 et 152.)

Overall, A. D. J., et R. A. Nichols, 2001 A method for distinguishing consanguinity and population substructure using multilocus genotype data. Mol. Biol. Evolution **18** : 2048–2056. (Cité page 108.)

Petit, R. J., A. E. Mousadik, et O. Pons, 1998 Identifying populations for conservation on the basis of genetic markers. Conservation Biol. **12** : 844–855. (Cité pages 6, 35, 36, 58 et 59.)

Piazza, A., P. Menozzi, et L. L. Cavallisforza, 1981 Synthetic gene-frequency maps of man and selective effects of climate. Proc. National Acad. Sciences United States America-biological Sciences **78** : 2638–2642. (Cité pages 38, 39, 60 et 68.)

Pritchard, J. K., et M. W. Feldman, 1996 Statistics for microsatellite variation based on coalescence. Theoretical Population Biol. **50** : 325–344. (Cité page 133.)

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, et M. W. Feldman, 1999 Population growth of human Y chromosomes : A study of Y chromosome microsatellites. Mol. Biol. Evolution **16** : 1791–1798. (Cité page 11.)

Prugnolle, F., A. Manica, et F. Balloux, 2005 Geography predicts neutral genetic diversity of human populations. Current Biol. **15** : R159–R160. (Cité pages 36, 37, 60, 65, 66 et 67.)

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. Proc. National Acad. Sciences United States Am. **102** : 15942–15947. (Cité pages 65 et 67.)

Raufaste, N., et F. Rousset, 2001 Are partial mantel tests adequate ? Evolution **55** : 1703–1705. (Cité pages 33 et 63.)

Ray, N., M. Currat, P. Berthier, et L. Excoffier, 2005 Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res. **15** : 1161–1167. (Cité pages 39, 65 et 66.)

Rousset, F., 2001 *Handbook of Statistical Genetics*, chapter Inference from spatial population genetics. John Wiley & Sons, New York, 239–270. (Cité page 43.)

Rousset, F., 2002 Partial Mantel tests : Reply to Castellano and balletto. Evolution **56** : 1874–1875. (Cité pages 33 et 63.)

Smith, B. J., 2005 Bayesian Output Analysis program (BOA), version 1.1.5. http ://www.public-health.uiowa.edu/boa. (Cité pages 49 et 130.)

Smouse, P. E., J. C. Long, et R. R. Sokal, 1986 Multiple-regression and correlation extensions of the mantel test of matrix correspondence. Systematic Zoology **35** : 627–632. (Cité pages 25, 33 et 43.)

Sokal, R. R., N. L. Oden, et B. A. Thomson, 1992 Origins of the indo-europeans - genetic-evidence. Proc. National Acad. Sciences United States Am. **89** : 7669–7673. (Cité page 47.)

Tavaré, S., D. J. Balding, R. C. Griffiths, et P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145** : 505–518. (Cité pages 82 et 108.)

Vitalis, R., K. Dawson, et P. Boursot, 2001 Interpretation of Variation Across Marker Loci as Evidence of Selection. Genetics **158** : 1811–1823. (Cité page 123.)

Wilding, C. S., R. K. Butlin, et J. Grahame, 2001 Differential gene exchange between parapatric morphs of Littorina saxatilis detected using AFLP markers. J. Evolutionary Biol. **14** : 611–619. (Cité pages 6, 79, 116, 117, 118, 119, 123, 143, 145, 146, 147 et 149.)

Wright, S., 1931 Evolution in mendelian populations. Genetics **16** : 97–159. (Cité pages 16, 17, 18, 25, 80, 90, 124 et 129.)

Wright, S., 1951 The genetic structure of populations. Annals of Eugenic **15** : 323–354. (Cité pages 45 et 125.)

Wright, S., 1969 *Evolution and genetics of populations. The theory of gene frequencies*, volume Vol. 2. Univ. of Chicago Press,Chicago. (Cité page 17.)

Yu, N., Y. X. Fu, N. Sambuughin, M. Ramsay, T. Jenkins, *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evolution **18** : 214–222. (Cité page 60.)

Zhao, Z. M., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins, *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. National Acad. Sciences United States Am. **97** : 11354–11358. (Cité page 60.)

Zhivotovsky, L. A., 1999 Estimating population structure in diploids with multilocus dominant DNA markers. Mol. Ecology **8** : 907–913. (Cité pages 22 et 88.)