Méthodes bayésiennes en génétique des populations : relations entre structure génétique des populations et environnement

Flora Jay

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 7 août 2006

Présentée par

Flora JAY

Thèse dirigée par **Olivier FRANÇOIS**
et codirigée par **Michael BLUM**

préparée au sein **du laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG)**
et de **l'école doctorale « Ingénierie de la Santé, de la Cognition et Environnement » (EDISCE)**

Méthodes bayésiennes pour la génétique des populations :

relations entre structure génétique des populations et environnement

Thèse soutenue publiquement le **14 novembre 2011**,
devant le jury composé de :

Florence FORBES
DR INRIA, INRIA Grenoble, Présidente
Étienne KLEIN
DR INRA, INRA Avignon, Rapporteur
Renaud VITALIS
CR CNRS, INRA Montpellier, Rapporteur
Oscar GAGGIOTTI
Professeur, UJF Grenoble, Examinateur
Stéphanie MANEL
Professeur, université Aix-Marseille-I, Examinatrice
Bertrand SERVIN
CR INRA, INRA Toulouse, Examinateur
Olivier FRANÇOIS
Professeur, INP Grenoble, Directeur de thèse
Michael BLUM
CR CNRS, UJF Grenoble, Codirecteur de thèse

Titre : Méthodes bayésiennes en génétique des populations : relations entre structure génétique des populations et environnement.

Résumé : Nous présentons une nouvelle méthode pour étudier les relations entre la structure génétique des populations et l'environnement. Cette méthode repose sur des modèles hiérarchiques bayésiens qui utilisent conjointement des données génétiques multi-locus et des données spatiales, environnementales et/ou culturelles. Elle permet d'estimer la structure génétique des populations, d'évaluer ses liens avec des covariables non génétiques, et de projeter la structure génétique des populations en fonction de ces covariables. Dans un premier temps, nous avons appliqué notre approche à des données de génétique humaine pour évaluer le rôle de la géographie et des langages dans la structure génétique de populations amérindiennes. Dans un deuxième temps, nous avons étudié la structure génétique des populations pour 20 espèces de plantes alpines, et nous avons projeté les modifications intraspécifiques qui pourront être causées par le réchauffement climatique.

Mots-clés : structure génétique des populations, covariables environnementales, modèles bayésiens hiérarchiques, modèles à classes latentes, MCMC, modèles bioclimatiques.

Title: Bayesian methods for population genetics: relationships between population genetic structure and environment.

Abstract: We introduce a new method to study the relationships between population genetic structure and environment. This method is based on Bayesian hierarchical models which use both multi-loci genetic data, and spatial, environmental, and/or cultural data. Our method provides the inference of population genetic structure, the evaluation of the relationships between the structure and non-genetic covariates, and the prediction of population genetic structure based on these covariates. We present two applications of our Bayesian method. First, we used human genetic data to evaluate the role of geography and languages in shaping Native American population structure. Second, we studied the population genetic structure of 20 Alpine plant species and we forecasted intra-specific changes in response to global warming.

Keywords: population genetic structure, environmental covariates, Bayesian hierarchical models, latent class models, MCMC, bioclimatic models.

Sommaire

Résumé

Au cours de cette thèse, nous nous sommes intéressés aux relations entre structure génétique des populations et environnement, et plus particulièrement aux questions suivantes :

Comment déterminer ces relations ? Peut-on exploiter des variables spatiales, environnementales ou culturelles pour inférer ou prédire la structure génétique des populations ? En particulier, peut-on prédire les modifications de la structure en cas de changements des conditions environnementales ?

Pour y répondre, nous avons développé une nouvelle méthode statistique, implémentée dans le logiciel `POPS` (*Prediction Of Population Structure*). `POPS` utilise un algorithme bayésien de classification génétique reposant sur un modèle hiérarchique. La méthode intègre les informations supplémentaires apportées par des variables spatiales, environnementales et/ou culturelles dans une couche cachée qui est ajoutée au modèle hiérarchique. Cette « couche environnementale » consiste en un modèle de régression entre des variables latentes liées à la structure génétique et les variables non génétiques. De cette manière, les variables environnementales ou culturelles peuvent aider à l'inférence de la structure génétique. De plus, `POPS` peut prédire les probabilités d'appartenance ou les coefficients de métissage uniquement à partir de variables environnementales ou culturelles.

Dans un premier temps, nous avons utilisé cette méthode pour étudier les relations entre la structure génétique de populations amérindiennes, la géographie et les langues. À l'aide de `POPS` et d'une nouvelle procédure de sélection de modèle, nous avons montré que l'ajout de l'information linguistique à l'information spatiale améliorait la prédiction de la structure génétique de ces populations amérindiennes. Nous avons de plus comparé l'apport relatif de différentes classifications linguistiques existantes.

Dans un deuxième temps, nous avons étudié le rôle de variables climatiques, topographiques et spatiales dans la structure génétique de populations de plantes alpines. Appliquant `POPS` à 20 espèces de plantes, nous avons inféré la structure des populations de chaque espèce et, après avoir discuté des hypothèses simplificatrices réalisées, nous avons prédit les changements de cette structure en réponse à différents scénarios de réchauffement climatique. Chez la majorité des espèces, nous avons trouvé une zone de contact intraspécifique entre des populations du sud des Alpes, potentiellement adaptées à des environnements chauds, et des populations plus au nord. D'après les prédictions, ces zones de contact se déplacent vers le nord-est en cas de réchauffement climatique. Nous avons évalué l'amplitude de ces déplacements ainsi que le taux de modification de la structure pour chaque espèce.

Préambule

Dans le chapitre 1, nous présenterons tout d'abord la notion de structure génétique des populations et les domaines dans lesquels elle est utilisée. Nous passerons ensuite en revue les méthodes habituellement employées pour détecter cette structure, et enfin les méthodes consacrées à l'étude des relations entre structure génétique et environnement.

Le chapitre 2 sera consacré aux problématiques centrales de la thèse, à savoir : l'apport des informations environnementales pour l'estimation de la structure génétique ; la détection des facteurs environnementaux liés à une structure génétique ; l'utilisation de ces facteurs pour « pronostiquer » la structure et enfin la projection de celle-ci en cas de changements environnementaux. Nous y présenterons les modèles bayésiens avec lesquels nous avons choisi de travailler et les méthodes développées pour répondre aux différentes questions soulevées.

Dans le chapitre 3, nous nous intéressons aux interactions entre gènes, géographie et traits culturels. Nous y étudions les liens respectifs de la géographie et des langages avec la structure génétique de populations amérindiennes.

Le chapitre 4 est consacré à l'impact de changements environnementaux sur les espèces. Nous y expliquons comment notre méthode offre un angle d'approche nouveau pour aborder cette problématique, et nous étudions les réponses potentielles de 20 espèces de plantes alpines à un changement climatique.

Enfin, le dernier chapitre nous permettra de présenter le logiciel `POPS` que nous avons développé.

Chapitre 1

Introduction

1.1 Contexte

Les variations génétiques entre individus sont le résultat de différentes forces évolutives, plus ou moins dépendantes les unes des autres. Ces forces peuvent être induites par des processus neutres, comme la dérive génétique et les processus démographiques de migration et d'expansion, ou par des processus d'adaptation biologique, comme la sélection d'une mutation bénéfique au sein d'un environnement donné. Qu'ils soient neutres ou non, ces processus sont souvent influencés par l'environnement ; ainsi une barrière géographique sera une barrière aux migrations, et une variation climatique peut être à l'origine d'adaptations locales.

La plupart des espèces sont structurées, c'est-à-dire que l'ensemble de leurs individus ne forme pas une unité génétiquement homogène. Une espèce peut être constituée de plusieurs groupes, totalement ou partiellement isolés, pour diverses raisons : éloignement géographique, présence de barrières environnementales, reproduction préférentielle... Cet isolement, associé au phénomène aléatoire de dérive génétique et parfois à des phénomènes d'adaptation locale, amène les groupes à se différencier génétiquement. En effet, le hasard conduit les fréquences alléliques à évoluer différemment dans deux groupes distincts qui ne sont plus en contact, ce qui cause une différenciation génétique entre les populations. De même, si les pressions sélectives sont géographiquement différenciées, du fait de conditions environnementales différentes par exemple, l'espèce peut s'adapter localement, et des groupes génétiquement différenciés émergent. Des groupes isolés un certain temps peuvent être amenés à se rencontrer de nouveau, ce qui crée des zones de contact secondaire, dans lesquelles on trouvera généralement des individus métissés.

C'est donc l'histoire démographique d'une espèce, associée à diverses forces évolutives éventuellement influencées par l'environnement, qui produit la structure génétique des populations.

Nous présentons ci-dessous les principaux domaines d'application de la génétique des populations, et en particulier l'intérêt d'étudier la structure génétique des populations dans chacun de ces domaines.

Histoire évolutive des populations. L'étude de la structure génétique des populations fait partie, avec la paléontologie, l'archéologie et la linguistique, des disciplines de prédilection pour reconstituer l'histoire de l'Homme depuis ses origines. Dès son apparition, le potentiel de la génétique des populations pour décrypter l'histoire évolutive a été souligné (Wright 1943, 1949). Le célèbre généticien des populations Luigi Luca Cavalli-Sforza est considéré comme l'un des initiateurs des travaux reliant génétique, culture et histoire évolutive humaine (Cavalli-Sforza *et al.* 1964; Menozzi *et al.* 1978; Cavalli-Sforza *et al.* 1988, 1994). Il est à l'origine du projet *Human Genome Diversity Panel* (HGDP), lequel a permis de récolter des informations génétiques sur des populations réparties dans le monde entier. De nombreuses études ont par la suite été réalisées, dont nous mentionnerons ici quelques-unes. Certaines s'intéressent à l'origine africaine du peuplement humain et aux chemins de migration (Prugnolle *et al.* 2005; Ramachandran *et al.* 2005; Blum and Jakobsson 2011). D'autres s'intéressent aux relations entre populations (métissage, migration) à l'échelle de la planète (Rosenberg *et al.* 2002), à l'échelle plus fine du continent américain (Wang *et al.* 2007) ou africain (Tishkoff *et al.* 2009). L'étude de l'histoire évolutive ne s'est pas restreinte à l'Homme : des travaux ont retracé l'histoire démographique d'espèces modèles, comme la plante *Arabidopsis thaliana* (François *et al.* 2008) ou la drosophile (Das *et al.* 2004).

Génétique épidémiologique. La génétique épidémiologique — discipline en nette expansion depuis le développement spectaculaire des méthodes de séquençage et de génotypage — est, elle aussi, liée à la structure des populations. Les études d'association cherchent à détecter les marqueurs génétiques corrélés avec un trait phénotypique. Lorsqu'elles sont appliquées à l'Homme, le trait ciblé est très souvent une maladie (diabète, maladie de Crohn). Ces études comparent les données génétiques d'un groupe d'individus malades à celles d'un groupe d'individus témoins. Un marqueur génétique montrant des différences significatives entre malades et témoins sera supposé associé à la maladie. Néanmoins, si les individus sont structurés en populations et que cette structure n'est pas identique dans le groupe malade et dans le groupe témoin, une méthode aussi simple risque de détecter des faux positifs, à savoir des marqueurs génétiques liés à la structure de populations et non pas à la maladie (Cardon and Bell 2001). Des méthodes pour réaliser ces tests d'association tout en tenant compte de la structure ont donc été développées (Pritchard *et al.* 2000b; Hoggart *et al.* 2003; Marchini *et al.* 2004; Price *et al.* 2006).

Un autre objectif important dans le domaine de la génétique épidémiologique est de

faire bénéficier les individus d'une médecine hautement personnalisée. Il est connu que des médicaments peuvent avoir une efficacité variable d'un sujet à l'autre. Cette variabilité dépend de plusieurs facteurs, comme l'âge, le sexe, l'environnement et le matériel génétique (Wilson *et al.* 2001). Des études ont montré que certains allèles impliqués dans l'efficacité de traitements varient nettement en fréquence selon la population considérée (voir Bernal *et al.* 1999, pour un exemple de variation entre populations européennes). L'idée d'une médecine personnalisée tenant compte de l'influence de la structure génétique des populations est donc apparue. Ce concept a d'ailleurs sérieusement ravivé les débats sur l'intérêt en médecine des groupes génétiques, et encore plus sur l'emploi de variables non génétiques pour « pronostiquer » l'appartenance à ces groupes (Schwartz 2001; Wilson *et al.* 2001; Risch *et al.* 2002; Bamshad *et al.* 2003).

Biologie de la conservation. La problématique de la biologie de la conservation est de préserver la diversité biologique. Un de ses principaux travaux est donc d'identifier les espèces potentiellement en danger. Une espèce n'est souvent pas une unité homogène, il est donc important, pour conserver la biodiversité, de comprendre sa structuration en populations et de protéger les différents groupes qui la composent (Avise 1992; Moritz 1994; Crandall *et al.* 2000). Un autre objectif de ce domaine est de veiller à ce que les modifications de l'environnement (gestion des espaces, constructions...) n'affectent pas outre mesure l'écosystème ; pour cela, l'étude approfondie des relations entre environnement et structure génétique est primordiale.

Génétique du paysage. La génétique du paysage[1] cherche à comprendre comment l'hétérogénéité de l'environnement peut influencer la structure et la diversité génétiques (Manel *et al.* 2003; Balkenhol *et al.* 2009). Ses travaux se concentrent sur des problèmes aux échelles spatiale et temporelle plutôt fines, et s'intéressent donc plus à la diversité au sein des espèces qu'à la diversité entre espèces. À cette échelle, la génétique du paysage tente de mettre au jour les interactions entre l'environnement et différents processus évolutifs : flux de gènes, dérive génétique, sélection... Elle cherche à détecter plus précisément quels sont les facteurs environnementaux qui jouent un rôle dans ces interactions. La génétique du paysage a des applications évidentes en biologie de la conservation (Segelbacher *et al.* 2010). D'une part, elle permet de délimiter les unités biologiques d'intérêt pour les campagnes de conservation. D'autre part, elle met en évidence les facteurs environnementaux qui ont une influence sur la structure génétique et qui méritent d'être suivis attentivement. Ainsi l'étude de Epps *et al.* (2005) montre que les réseaux routiers, en bloquant les flux de gènes entre les populations de mouflons canadiens, entraînent un fort déclin de leur diversité génétique.

1. *Landscape genetics.*

1.2 Détection de la structure génétique

Lorsque l'on connaît les données génétiques d'un ensemble d'individus pour un ensemble de marqueurs (un certain nombre d'allèles, de microsatellites, de SNPs [2] ou encore des séquences d'ADN), un objectif est de détecter si les individus sont structurés en populations, et, si c'est le cas, d'identifier le nombre de groupes (*clusters*) isolés ou partiellement isolés, les individus composant chacun de ces groupes et éventuellement les individus métissés — qui ont des origines dans plusieurs groupes. Plusieurs méthodes permettent d'effectuer ces recherches. Nous les présentons ici, regroupées en trois grands types d'approche.

1.2.1 Arbre phylogénétique

Les méthodes historiques pour étudier la proximité génétique d'individus sont basées sur la construction d'arbres phylogénétiques. Dans les années 60, les premiers travaux sur la structure des populations, essentiellement publiés par Cavalli-Sforza et Edwards, reposent sur la reconstruction des phylogénies à partir de fréquences alléliques. La méthode est décrite dans Cavalli-Sforza and Edwards (1967). Il est intéressant de noter que Cavalli-Sforza et Edwards avaient aussi proposé une méthode de partitionnement de données multidimensionnelles, mais que celle-ci était difficilement applicable à l'époque d'un point de vue computationnel (Edwards and Cavalli-Sforza 1965; Scott and Symons 1971). Actuellement, l'algorithme fréquemment appliqué pour la construction d'arbres phylogénétiques est l'algorithme de *neighbor-joining*, proposé par Saitou and Nei (1987), implémenté dans divers logiciels (Felsenstein 1989; Kumar *et al.* 2004), et utilisé dans de nombreuses études (e.g. Bowcock *et al.* 1994; Wang *et al.* 2007) (voir Figure 1.1). La construction de l'arbre est faite à partir d'une matrice contenant les distances génétiques pour chaque paire d'individus. Ces distances génétiques sont classiquement calculées à l'aide des distances proposées par Nei *et al.* (1983) ou par Reynolds *et al.* (1983). Une fois l'arbre construit, on peut décider de le « couper » à un certain niveau pour définir les différents groupes génétiques. On peut aussi l'utiliser comme outil descriptif, en particulier pour vérifier si les groupes génétiques correspondent à des groupes prédéfinis, par exemple des groupes linguistiques (voir Figure 1.1).

1.2.2 Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode classique d'analyse de données multidimensionnelles. Elle permet de projeter des données sur un nombre réduit d'axes orthogonaux, tout en maximisant la variance des données projetées sur chacun des axes. Son application à des données génétiques remonte à l'étude de la distribution géographique des fréquences alléliques de 10 locus chez des individus européens (Menozzi

2. Single Nucleotide Polymorphisms.

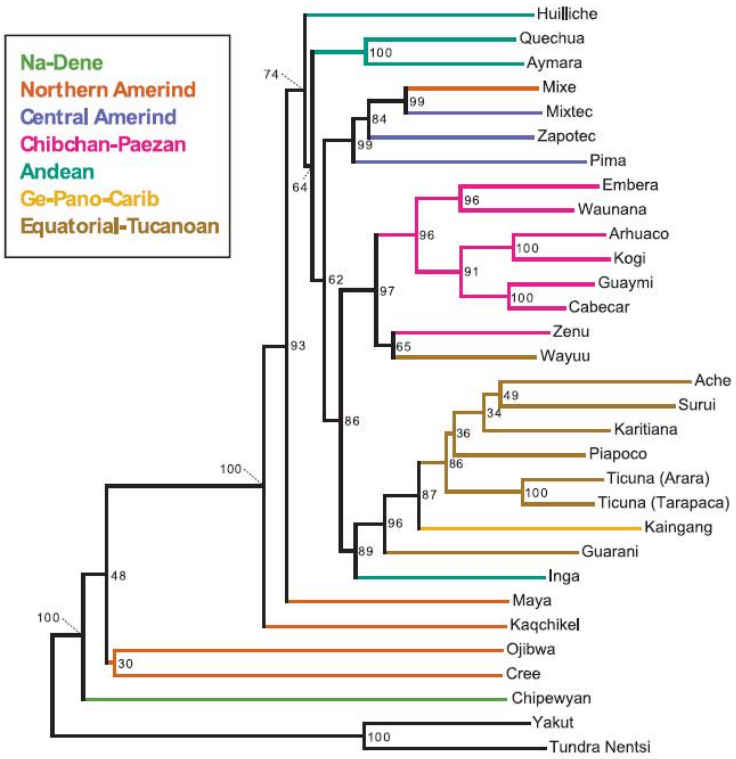


FIGURE 1.1 – Arbre phylogénétique de 29 populations amérindiennes (et une population de Sibérie) construit par *neighbor-joining* (WANG *et al.* 2007). Les distances génétiques ont été calculées à partir de 678 marqueurs microsatellites chez 530 individus. Les couleurs correspondent aux familles linguistiques des langues parlées par les différentes populations. Ces familles sont décrites dans la classification de Greenberg (GREENBERG 1987).

et al. 1978). L'utilisation de l'ACP en génétique des populations est redevenue populaire grâce à un article de PATTERSON *et al.* (2006) et à leur logiciel `EIGENSTRAT`. Les auteurs y proposent un test formel de la présence ou non de structure de populations et démontrent que l'ACP peut détecter cette structure à partir d'un niveau minimum de différenciation génétique qui dépend du nombre de marqueurs et de la taille de l'échantillon. De plus, ils établissent un lien entre le nombre de clusters détectables par l'algorithme `STRUCTURE`[3] et le nombre d'axes significatifs de l'ACP. Plus récemment, ENGELHARDT and STEPHENS (2010) établissent eux aussi des correspondances entre l'ACP et les méthodes basées sur un modèle explicite comme `STRUCTURE`. Comme applications récentes de l'ACP en génétique des populations, on peut citer notamment trois études, portant sur des populations européenne, qui ont montré que la structure génétique était fortement corrélée avec la géographie (LAO *et al.* 2008; NOVEMBRE *et al.* 2008; HEATH *et al.* 2008) (voir Figure 1.2).

3. Cet algorithme proposé par PRITCHARD *et al.* (2000a) est présenté dans la section suivante.

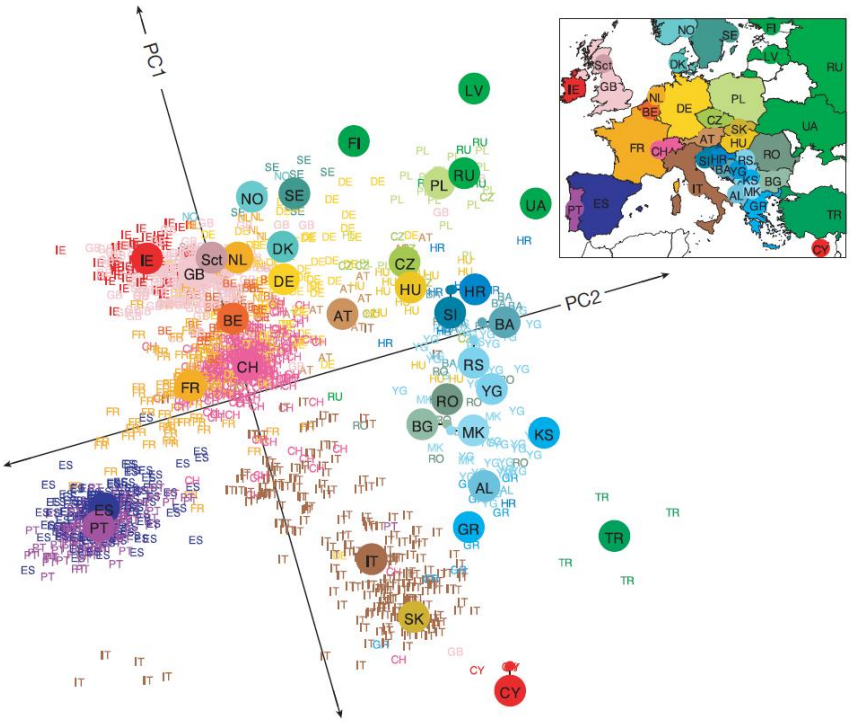


FIGURE 1.2 – Visualisation des deux premiers axes d'une ACP réalisée sur 197 146 SNPs chez 1 387 individus européens (NOVEMBRE *et al.* 2008). Les individus sont projetés dans le repère (PC1, PC2) et coloriés en fonction de leur pays d'origine. Les ronds colorés correspondent aux médianes des projections sur PC1 et PC2 de chaque pays. NOVEMBRE *et al.* (2008) mettent ici en avant la correspondance entre la projection sur l'ACP et la géographie européenne.

ACP et métissage. La possibilité de détecter le métissage à partir de l'ACP est aussi étudiée par PATTERSON *et al.* (2006). Les auteurs réalisent une ACP des données génétiques de deux populations parentales (sources du métissage) et d'une population métissée. Les positions relatives (par rapport aux populations ancestrales) sur le premier axe de l'ACP (PC1) des individus métissés correspondent aux coefficients de métissage. Par contre, le nombre d'axes significatifs ne semble plus relié au nombre de clusters (PATTERSON *et al.* 2006). Plus récemment, BRYC *et al.* (2010) ont utilisé cette méthode pour détecter le métissage le long du génome et l'ont appliquée à des individus afro-américains. Pour cela, ils partitionnent les données génétiques de populations européennes, africaines et afro-américaines en fenêtres d'une quinzaine de SNPs, et l'ACP est réalisée pour chaque fenêtre. La fenêtre d'un individu métissé est assignée à la population (européenne ou africaine) ayant le plus proche score moyen sur le premier axe principal. HENN *et al.* (2011) ont appliqué une méthode similaire pour détecter le métissage chez des chasseurs-cueilleurs sud-africains. Dans les applications récentes, comme celle de HENN *et al.* (2011), l'ACP est souvent réalisée sur les populations parentales uniquement, puis les individus métissés y sont projetés. Les résultats doivent être examinés avec prudence, car LEE *et al.* (2010)

ont démontré que la projection de nouveaux échantillons sur des axes de l'ACP peut être biaisée vers 0 lorsque les données sont constituées d'un grand nombre de marqueurs par rapport au nombre d'individus (ce qui est typiquement le cas des données génétiques actuelles).

ACP et étude d'association. Comme expliqué précédemment, les études d'association entre traits phénotypiques et gènes risquent de détecter de fausses associations dues à la structure cachée de populations. Pour corriger ce biais plusieurs méthodes ont été proposées, dont certaines reposent sur l'ACP (Zhu *et al.* 2002; Price *et al.* 2006; Heath *et al.* 2008). Le principe en est de corriger les tests en utilisant les projections des individus sur les axes principaux de l'ACP des données génétiques.

1.2.3 Méthodes avec modèle explicite

Une autre approche consiste à supposer que les données génétiques peuvent être expliquées par un modèle probabiliste dont les paramètres sont inconnus.

Une des premières méthodes basées sur un modèle explicite permettant de détecter la structure génétique des populations a été proposée par Pritchard *et al.* (2000a) ; elle est implémentée dans la première version du logiciel `STRUCTURE`. Il s'agit d'un modèle hiérarchique bayésien qui constitue la base de nombreuses méthodes actuelles. Le modèle de `STRUCTURE` suppose l'existence de K populations (non prédéfinies), et chacune d'entre elles est caractérisée à chaque locus par un ensemble de fréquences alléliques. Dans le modèle sans métissage, chaque individu échantillonné est assigné à une unique population. Les probabilités qu'un individu appartienne aux différentes populations sont appelées « coefficients d'appartenance » (ces probabilités reflètent les incertitudes de la classification). Dans le modèle avec métissage, chacun des locus d'un individu est assigné à une population. Un individu peut donc être assigné conjointement à plusieurs populations. Les pourcentages de locus d'un individu assignés aux différentes populations sont appelés « coefficients de métissage ». L'objectif de `STRUCTURE` est d'estimer conjointement les fréquences alléliques au sein des différentes populations et les coefficients d'appartenance ou de métissage des individus. Pour cela l'algorithme estime la distribution jointe de ces variables (et des variables cachées) à l'aide d'une méthode de Monte Carlo par chaîne de Markov (*Markov chain Monte Carlo*, MCMC). Nous ne détaillerons pas l'algorithme ici, mais l'idée est que chaque population devrait se trouver à l'équilibre de Hardy-Weinberg. `STRUCTURE` cherche donc à créer des populations afin de minimiser l'écart à l'équilibre de Hardy-Weinberg au sein de chacune d'elles. La Figure 1.3 montre un célèbre exemple de l'application de `STRUCTURE` à des individus issus de 52 populations humaines du HGDP (Rosenberg *et al.* 2002).

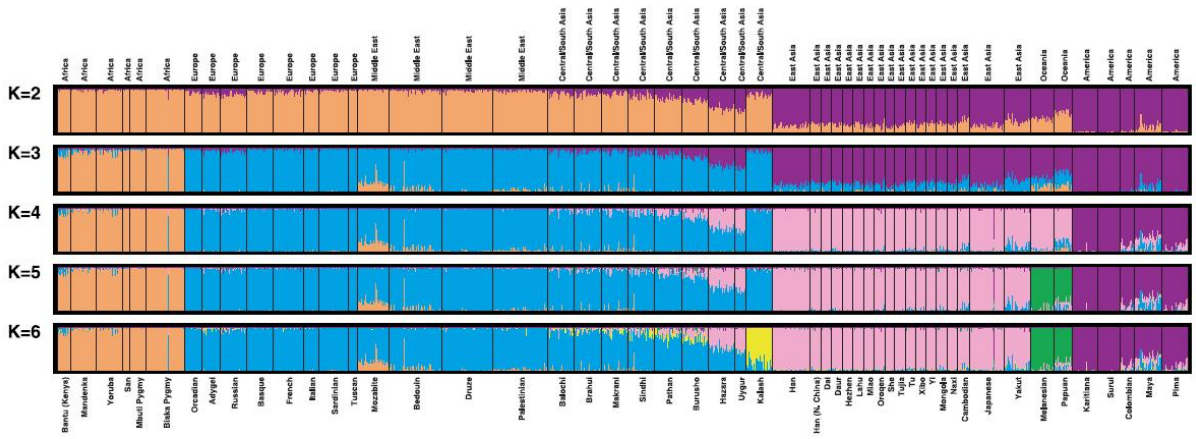


FIGURE 1.3 – ROSENBERG *et al.* (2002) ont appliqué le logiciel `STRUCTURE` à 1 056 individus issus de 52 populations humaines pour lesquels 377 marqueurs microsatellites ont été génotypés (données HGDP). Chaque individu est représenté par une fine barre verticale partitionnée en K segments correspondant aux coefficients d'appartenance estimés pour K populations. Avec un modèle à $K = 5$ populations, les auteurs retrouvent 5 régions géographiques majeures : Afrique, Eurasie, Asie de l'Est, Australie, Amérique.

Correction pour les études d'association gène-maladie. Pritchard et ses collaborateurs ont proposé d'utiliser `STRUCTURE` pour corriger les tests d'association gène-maladie. Leur approche STRAT (*STRuctured population Association Test*) consiste simplement à remplacer l'hypothèse nulle : « pas d'association entre fréquences alléliques et phénotype », classiquement testée dans les études d'association par « pas d'association au sein de chaque population identifiée par `STRUCTURE` » (PRITCHARD *et al.* 2000b).

Extensions du modèle. Des extensions de la première version de `STRUCTURE` ont été proposées dans le but de rendre le modèle plus réaliste en y intégrant certains processus biologiques. Entre autres : FALUSH *et al.* (2003) ont proposé une extension tenant compte des corrélations existantes entre les locus physiquement liés sur le génome ; FRANÇOIS *et al.* (2006) et GAO *et al.* (2007) ont modélisé la consanguinité ; FALUSH *et al.* (2007) ont pris en compte les marqueurs dominants ; ZHANG (2008) les migrations, et SHRINGARPURE and XING (2009) les mutations alléliques. D'autre part, des modifications visant à améliorer les performances ont été proposées. La majorité d'entre elles reposent sur la maximisation de la vraisemblance à partir d'algorithmes EM (*Expectation-Maximisation*) : TANG *et al.* (2005) ; CHEN *et al.* (2006) ; WU *et al.* (2006). Le logiciel `ADMIXTURE` utilise un algorithme plus complexe de relaxation par bloc (ALEXANDER *et al.* 2009 ; ALEXANDER and LANGE 2011). Le logiciel `ADMIXTURE` est bien plus rapide que `STRUCTURE` et ses extensions EM, mais surtout aussi rapide que le logiciel `EIGENSTRAT` réalisant l'ACP pour la détection de structure.

Ce dernier point est important, puisque la popularité des méthodes de détection de

structure basées sur l'ACP dépend principalement de leur rapidité. Celle-ci est un atout majeur compte tenu des améliorations des techniques de génotypage, lesquelles permettent aujourd'hui d'obtenir des données génétiques de très grandes dimensions. Quand elles sont en grande quantité (de l'ordre de millions de SNPs), les données sont difficilement analysables par `STRUCTURE` ou par les programmes dérivés (mis à part `ADMIXTURE`). En revanche, les algorithmes construits à partir de modèles explicites, et c'est leur avantage, produisent des résultats plus aisément interprétables. Entre autres, ils détectent explicitement les groupes génétiques, estiment les probabilités d'appartenance à ces groupes ou les coefficients de métissage et les fréquences alléliques dans les différents groupes. Il faut aussi rappeler que les dimensions des données génétiques restent pour l'instant réduites pour les espèces non modèles, le nombre de marqueurs disponibles étant en général inférieur à la centaine.

Extensions spatialement explicites. Pour améliorer la détection de la structure plusieurs extensions de `STRUCTURE` prennent en compte des données géographiques. L'idée sous-jacente est que, structure génétique et géographie étant souvent liées, des individus spatialement proches ont une probabilité *a priori* plus élevée d'appartenir à la même population. Les logiciels `GENELAND` (Guillot *et al.* 2005), `TESS` (Chen *et al.* 2007; Durand *et al.* 2009b) et `BAPS` (Corander *et al.* 2008) implémentent différentes méthodes spatialement explicites. Ils ont leurs spécificités : ainsi `GENELAND` s'intéresse particulièrement à la localisation des discontinuités génétiques au sein d'un échantillon et `TESS` à la détection de *clines*. Un cline correspond à une variation continue des fréquences alléliques le long d'un gradient spatial, qui apparaît lors d'un contact secondaire entre populations précédemment isolées ou lors d'une colonisation de zones déjà peuplées (Barton and Hewitt 1985). La plupart des algorithmes ont tendance à surestimer le nombre de populations en cas de cline (Serre and Pääbo 2004) ; `TESS` a été développé dans le but d'apporter une solution à ce problème. François and Durand (2010a) présentent les principales différences entre les logiciels `STRUCTURE`, `GENELAND`, `BAPS` et `TESS` et les testent dans différentes situations (absence ou présence de métissage dans les données, différenciation génétique plus ou moins grande, présence d'un cline).

Dans un état d'esprit un peu différent, Hubisz *et al.* (2009) proposent d'utiliser le groupe d'échantillonnage des individus comme information supplémentaire. Leur méthode fait partie des extensions géographiques puisque les groupes d'échantillonnage correspondent à des emplacements géographiques distincts et peuvent donc être considérés comme des observations d'une variable géographique qualitative. Hubisz *et al.* (2009) illustrent l'efficacité de leur méthode sur un sous-ensemble de 5 groupes d'échantillonnage (Surui, Han, Basque, Mélanésien et Mandenka) des données du HGDP. Ils montrent que la méthode aide à détecter la structure lorsque les données génétiques sont peu informatives et que les groupes d'échantillonnage sont corrélés avec la structure. De plus, les résul-

tats ne sont pas détériorés si cette corrélation n'existe pas. En revanche, la modélisation qualitative de la géographie est moins adaptée aux études de génétique du paysage, où le schéma d'échantillonnage est généralement différent (les individus sont échantillonnés plus régulièrement ; il y a donc de nombreux groupes constitués de peu d'individus, voire pas de groupe du tout).

1.3 Structure génétique et environnement

Nous décrivons dans cette section des méthodes qui s'intéressent aux relations entre structure génétique et environnement. Cette liste n'est pas exhaustive ; les méthodes ont été choisies en raison soit de leur fréquente utilisation, soit de leur intérêt méthodologique, soit de la proximité de leurs applications avec les applications présentées dans cette thèse.

1.3.1 Test de Mantel

Le test de Mantel est sûrement la méthode la plus couramment utilisée pour étudier les relations entre variables environnementales et structure génétique. Il permet de calculer la corrélation entre deux matrices et d'évaluer si cette corrélation est significative en la comparant à la distribution de valeurs obtenues à la suite de permutations au sein des matrices (Mantel 1967). Cette méthode, tout à fait générale, a rapidement été utilisée pour tester la corrélation entre une matrice contenant les distances génétiques pour chaque paire d'individus et une matrice contenant les distances géographiques entre ces mêmes individus (voir Figure 1.4). Développé plus récemment, le test de Mantel partiel permet d'évaluer l'effet d'une variable sur une autre, tout en contrôlant l'effet d'une troisième (Smouse *et al.* 1986). Depuis, il a souvent été appliqué en génétique des populations, notamment à des données de populations humaines pour tester la corrélation entre distances génétiques et distances linguistiques, en tenant compte des distances géographiques (Belle and Barbujani 2007; Wang *et al.* 2007).

Plusieurs difficultés sont néanmoins présentes lors de l'utilisation des tests de Mantel. Tout d'abord, le choix de la distance génétique ou environnementale peut influencer le résultat. Belle and Barbujani (2007) ont montré que les conclusions différaient selon qu'ils appliquaient leurs tests à des distances génétiques mesurées par l'indice de différenciation F_{ST} (mesure la plus couramment utilisée) ou par l'indice R_{ST} (Slatkin 1995). Toujours dans la même étude, le choix de la distance linguistique s'avérait primordial. Enfin, cette méthode nécessite de choisir quelle technique de permutation permettra d'obtenir la distribution « neutre » des corrélations (Legendre 2000).

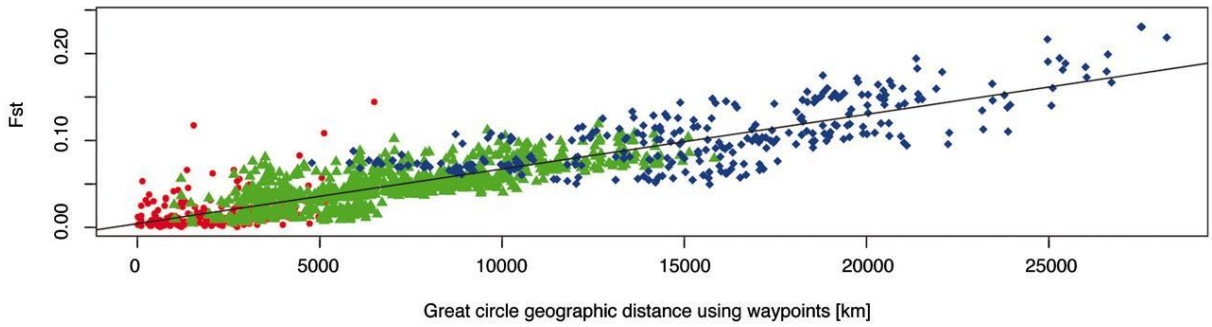


FIGURE 1.4 – RAMACHANDRAN *et al.* (2005) appliquent le test de Mantel pour évaluer les corrélations entre distances génétiques et distances géographiques calculées pour chaque paire de populations issues du HGDP. Sur ce graphe, les différenciations génétiques (F_{ST}) sont tracées en fonction de distances géographiques tenant compte des points de passage supposés des migrations humaines. La corrélation entre F_{ST} et distance géographique est de $0,8851$ (p-valeur du test de Mantel $< 10^{-4}$). Les ronds rouges correspondent aux comparaisons au sein d'une même région, les triangles verts aux comparaisons entre populations d'Afrique et d'Eurasie, et les losanges bleus aux comparaisons avec l'Amérique et l'Océanie.

1.3.2 GESTE (*GEnetic STructure inference based on genetic and Environmental data*)

FOLL and GAGGIOTTI (2006) ont développé une méthode bayésienne pour étudier l'effet de variables environnementales sur la différenciation génétique. Pour cela les auteurs ont intégré des données environnementales dans un modèle hiérarchique bayésien estimant les coefficients de différenciation génétique, F_{ST}, au sein de populations prédéterminées dans un modèle de structure en *continent-îles*. Le modèle permet d'estimer conjointement l'influence de différentes variables environnementales et les coefficients F_{ST} de chaque population. Cet algorithme est implémenté dans le logiciel `GESTE`. Contrairement aux tests de Mantel, `GESTE` ne nécessite pas le calcul de distances génétiques ou environnementales. Il nécessite en revanche de connaître à l'avance les populations pour lesquelles seront estimés les coefficients F_{ST}.

Pour illustrer l'intérêt de leur logiciel, FOLL and GAGGIOTTI (2006) l'appliquent à des données concernant l'arganier du Maroc et montrent que l'altitude n'affecte pas la différenciation génétique de cette espèce. Dans le même article, les auteurs trouvent que la latitude et la longitude ont une influence sur les F_{ST} des données de génétique humaine issues du HGDP. `GESTE` a été utilisé par la suite dans diverses études, dont celle de LECLERC *et al.* (2008), qui montre une potentielle influence de la température sur la structure génétique de la perche jaune (la probabilité *a posteriori* de cette hypothèse est égale à 0,29).

1.3.3 ACP et analyse canonique des corrélations

Pour tester l'association entre la structure génétique du chêne blanc de Californie et des variables climatiques, Sork *et al.* (2010) proposent une méthode qui repose sur l'ACP et l'analyse canonique des corrélations. Tout d'abord, les auteurs réalisent une ACP sur les marqueurs génétiques. Comme expliqué précédemment, les composantes principales (PCs) de l'ACP sont liées à la structure génétique, s'il y en a une. Ensuite, ils effectuent une analyse canonique des corrélations (ACC) entre les PCs et les variables climatiques. L'ACC est une méthode de statistique multidimensionnelle permettant d'étudier les relations entre deux ensembles de variables (ici l'ensemble des composantes principales et l'ensemble des variables climatiques). L'ACC détecte les paires de combinaisons linéaires (une combinaison de variables dans chaque ensemble) qui ont une corrélation maximale. Les auteurs s'intéressent ensuite à la contribution de chaque variable climatique aux premières paires canoniques (les paires présentant les corrélations les plus élevées entre composantes principales et variables climatiques). Suivant une procédure suggérée par Borcard *et al.* (1992), ils distinguent aussi les contributions respectives du climat et de la géographie.

1.3.4 Distances génétiques et classifications linguistiques

Nous présentons ici une méthode qui a été spécifiquement développée pour étudier les relations entre génétique et linguistique. À partir d'une méthode proposée par Cavalli-Sforza and Piazza (1975), Hunley *et al.* (2007) ont développé un test formel de la coévolution entre génétique et linguistique. Il s'agit d'établir si une classification linguistique, présentée sous la forme d'un arbre linguistique, est compatible avec une matrice de distance génétique. Hunley *et al.* (2007) ont proposé de calculer une matrice de distances génétiques attendues sous l'hypothèse que les divergences entre populations sont données par l'arbre de classification linguistique. Un test d'adéquation entre distances génétiques attendues et distances génétiques observées permet alors de voir si cette classification génétique est compatible avec les données génétiques. Néanmoins, Hunley *et al.* (2007) estiment que ce test n'est pas assez conservatif et l'utilisent donc pour comparer les performances de différentes classifications entre elles, plus que pour rejeter l'hypothèse d'une coévolution. Les auteurs suggèrent également une inspection visuelle du graphique des distances génétiques observées en fonction des distances attendues : cela permet d'une part d'analyser les résultats population par population (une population précise peut être à l'origine du manque de correspondance), d'autre part de voir si une classification linguistique a tendance à sur-estimer ou sous-estimer la différenciation génétique entre populations (voir Figure 1.5). Cette méthode a été appliquée pour les populations amérindiennes et pour les populations mélanésiennes (Hunley and Long 2005; Hunley *et al.* 2007, 2008).

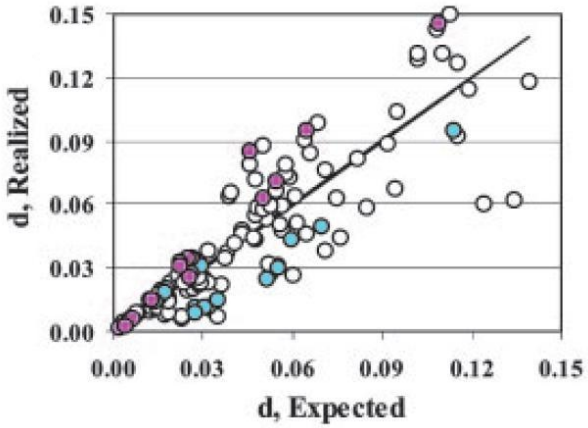


FIGURE 1.5 – HUNLEY and LONG (2005) ont calculé les distances génétiques pour des paires d'individus issus de 17 populations d'Amérique du Nord à partir de données d'ADN mitochondrial. Ces distances observées sont tracées en fonction des distances génétiques attendues, calculées à partir de l'arbre de classification linguistique de Greenberg (GREENBERG 1987), selon la méthode de CAVALLI-SFORZA and PIAZZA (1975). Ce graphe montre que la classification linguistique choisie surestime les distances pour les populations Navajo (points bleus) et les sous-estime pour les Inuits du Canada (points roses).

D'autres méthodes statistiques utilisées en génétique du paysage sont détaillées dans la revue de BALKENHOL *et al.* (2009).

Chapitre 2

Modèles bayésiens pour l'étude des relations entre structure génétique des populations et environnement

2.1 Problématique

Les questions sur les relations entre la structure génétique des populations et l'environnement sont multiples. Comment détecter les facteurs environnementaux et/ou culturels ayant un lien fort avec la structure génétique ? Peut-on établir qu'une variable environnementale en particulier a un rôle dans la structure génétique, tout en tenant compte de corrélations cachées avec d'autres variables ? Peut-on améliorer l'estimation de la structure génétique en utilisant des informations environnementales ? Comment prédire la structure génétique de nouveaux individus, ou encore comment prédire le changement de structure en fonction de changements de conditions environnementales ?

Nous nous sommes dès lors efforcés de développer une méthode qui permette de répondre à cet ensemble de questions. Nous l'avons appliquée à des problématiques précises : relations entre structure génétique, géographie et langages chez l'Homme ; rôle des variables climatiques dans la structure génétique de populations de plantes alpines, et modifications possibles de cette structure à la suite d'un changement climatique. Ces applications sont détaillées dans les chapitres 3 et 4.

Pourquoi développer une nouvelle méthode ? Des méthodes étudiant les relations entre structure génétique et environnement existent et ont été présentées dans le chapitre 1 (section 1.3). Elles diffèrent entre elles par leur manière de poser le problème, en particulier de caractériser la structure génétique. Certaines s'intéressent uniquement aux distances

génétiques ou à la différenciation génétique entre populations. C'est le cas des tests de Mantel — qui étudient la corrélation entre des matrices de distances génétiques et des matrices de distances géographiques ou environnementales — ou de la méthode proposée par Hunley and Long (2005) — qui teste si un arbre linguistique est compatible avec une matrice de distances génétiques. Dans ces approches, la structure génétique est en quelque sorte résumée par une unique mesure. D'autres méthodes permettent d'avoir une meilleure compréhension de la complexité de cette structure. La méthode `GESTE` de Foll and Gaggiotti (2006) repose, elle aussi, sur des mesures de différenciation génétique, mais celles-ci sont estimées au sein de populations prédéterminées (et non entre paires de populations). Dans la méthode « ACP-ACC » appliquée par Sork *et al.* (2010), une ACP des données génétiques est réalisée ; les projections sur les premiers axes, qui sont liées à la structure génétique [1], sont utilisées pour la suite de l'étude. On comprend bien que dans ce cas la structure génétique n'est pas réduite à une unique mesure. Remarquons que l'« ACP-ACC » est la seule méthode non supervisée, puisque toutes les autres approches nécessitent de définir des populations au préalable.

Nous avons choisi de modéliser différemment la structure génétique des populations. Notre objectif était de proposer une méthode non supervisée, qui permette à la fois d'estimer la structure génétique et les interactions avec l'environnement. Pour plusieurs raisons, il était naturel de s'orienter vers un modèle bayésien de détection de la structure proche du modèle du logiciel `STRUCTURE` [2] (Pritchard *et al.* 2000a).

- La méthode de `STRUCTURE` permet d'inférer une structure génétique des populations complexe. Elle identifie les populations (approche de classification non supervisée), estime les coefficients de métissage ou les probabilités d'appartenance à ces populations pour chaque individu, ainsi que les fréquences alléliques au sein des populations, et reflète l'incertitude qui existe autour de ces différents paramètres. Ce modèle nous offre donc une vision complète de la structure des populations, ainsi qu'une interprétation explicite des résultats (ce qui est encore un point sensible pour l'ACP).
- Le modèle et son cadre bayésien nous permettent d'intégrer aisément des variables environnementales en utilisant l'approche hiérarchique (Gelman 2004).
- Le modèle est individu-centré, il n'est donc pas nécessaire de définir des distances génétiques ou environnementales, entre populations ou individus.
- Ce modèle a été testé à de nombreuses reprises ; il est utilisé par une grande partie des écologues et des généticiens des populations.

Bien que de nombreuses extensions de `STRUCTURE` aient été développées, peu s'intéressent aux informations supplémentaires disponibles. Comme expliqué dans le chapitre 1, il existe des extensions spatialement explicites qui utilisent l'information géographique

1. Voir section 1.2.2.
2. Voir la section 1.2.3 pour la présentation du logiciel.

pour améliorer l'inférence de la structure, mais aucune méthode ne propose d'incorporer d'autres variables environnementales pour étudier plus précisément les relations entre la structure génétique et l'environnement.

Dans ce chapitre, nous présentons les modèles bayésiens que nous avons développés et implémentés dans le logiciel `POPS` (pour la présentation du logiciel et de son interface graphique, voir Jay 2011; Jay *et al.* 2011, section 5). Nous détaillons ensuite les apports de cette méthode pour les problématiques liant structure génétique des populations et environnement, à savoir :

- l'amélioration de l'estimation de la structure génétique à l'aide de covariables environnementales ;
- la détection des variables environnementales liées à la structure ;
- la prédiction de la structure à partir de données environnementales.

2.2 Modélisation statistique

Nous présentons les deux catégories de modèles considérés dans `POPS`. Dans les modèles sans métissage, un individu est originaire d'une unique population ; dans ces modèles, on cherche à estimer sa probabilité d'appartenir à chaque population. Dans les modèles avec métissage, différentes fractions du génome d'un individu peuvent provenir de différentes populations ; dans ces modèles, on cherche à estimer la part du génome de l'individu appartenant à chaque population.

2.2.1 Contexte : les modèles à classes latentes

Les modèles sur lesquels repose `POPS` sont liés aux familles des modèles de mélange et plus particulièrement à celle des modèles à classes latentes. Contrairement aux modèles traditionnels, qui décrivent uniquement les relations entre variables observées, les modèles de mélange introduisent une variable discrète, non observée, qui modélise l'existence de classes ou de groupes cachés. Ces classes pourront souvent être interprétées comme des clusters ayant une signification particulière (groupe génétique en génétique des populations, catégorie comportementale de clients en marketing...). L'idée est que la densité des observations a une forme spécifique à chaque classe. Cette densité peut, par exemple, être normale (Hasselblad 1966), exponentielle (Thomas 1966), de Bernoulli (Lazarsfeld 1954) ou multinomiale. Lorsque la loi de densité est discrète, on parle de modèle à classes latentes. L'apparition des modèles de mélange remonte à la fin du XIX^{e} siècle (Pearson 1894), mais le modèle à classes latentes pour des variables binomiales a été introduit en 1950 par Lazarsfeld [3] (Lazarsfeld 1950). Il a ensuite été adapté pour analyser des variables multinomiales.

3. Il a d'abord été présenté sous l'appellation de *latent structure analysis* et non de *latent class model*.

Notons $X = (X_1, \ldots, X_L)$ les L variables multinomiales observées et Z la variable latente discrète. L'idée du modèle classique à K classes latentes est d'exprimer la probabilité des observations de la manière suivante :

$$\Pr(X = x) = \sum_{k=1}^{K} \Pr(Z = k)\Pr(X = x|Z = k). \tag{2.1}$$

Sous l'hypothèse d'*indépendance locale* des variables (on parle aussi d'*indépendance conditionnelle*), les variables $X_1, \ldots, X_L$ au sein d'une classe sont indépendantes, on peut écrire :

$$\Pr(X = (x_1, \ldots, x_L)) = \sum_{k=1}^{K} \Pr(Z = k) \prod_{l=1}^{L} \Pr(X_l = x_l|Z = k). \tag{2.2}$$

Ajout de covariables

Admettons maintenant que l'on veuille utiliser des variables supplémentaires que l'on appellera prédicteurs. Notons $\tilde{X}$ cet ensemble de prédicteurs, alors la loi $\Pr(X, \tilde{X})$ est de la forme :

$$\Pr(X = x, \tilde{X} = \tilde{x}) = \Pr(\tilde{X} = \tilde{x}) \sum_{k=1}^{K} \Pr(Z = k|\tilde{X} = \tilde{x})\Pr(X = x|Z = k, \tilde{X} = \tilde{x}). \tag{2.3}$$

Le modèle particulier où la probabilité *a priori* d'appartenir à une classe ne dépend pas des prédicteurs, *i.e.* $\Pr(Z = k|\tilde{X} = \tilde{x}) = \Pr(Z = k)$, est appelé *modèle à classes latentes avec régression*[4] ; c'est un cas particulier des *modèles de mélange avec régression*[5] (DeSarbo and Cron 1988; Wedel and DeSarbo 1994). Ces modèles ont été implémentés par Leisch (2004) dans le package `R flexmix`.

Le modèle où les prédicteurs n'influencent pas la densité des observations au sein d'une classe donnée, *i.e.* $\Pr(X = x|Z = k, \tilde{X} = \tilde{x}) = \Pr(X = x|Z = k)$, est désigné sous le terme *concomitant-variable latent class* (Dayton and Macready 1988) ou sous le terme *latent class feed-forward* (Vermunt and Magidson 2003), mais parfois sous le terme *latent class regression* comme le modèle précédent (Linzer and Lewis 2011). Il a très récemment été implémenté par Linzer and Lewis (2011) dans le package `R poLCA`. C'est cette catégorie de modèles que nous allons utiliser dans le chapitre suivant.

2.2.2 Modèle sans métissage

Le modèle de `POPS` sans métissage et sans covariables est un modèle à classes latentes. Supposons que l'on observe les données génétiques de N individus haploïdes[6]. La géné-

4. *Latent class regression.*
5. *Finite mixture regression.*
6. Dont les cellules ne possèdent qu'un exemplaire de chaque chromosome.

ralisation à une ploïdie quelconque est aisée, et présentée dans le chapitre 5. On note $x^{(i)} = (x_1^{(i)}, \ldots, x_L^{(i)})$ les allèles observés pour l'individu i aux locus $1, \ldots, L$. À chaque locus, le marqueur génétique correspond à une variable multinomiale dont les catégories sont les allèles. Pour les SNPs les allèles sont 0 ou 1, tandis que pour les microsatellites les allèles correspondent aux différents nombres de répétition du motif (on ne prend pas en compte l'aspect quantitatif de ce nombre). Pour inférer les coefficients d'appartenance des individus à K populations, notre méthode considère un modèle à K classes latentes. On note Z_i la variable latente de l'individu i, $i = 1, \ldots, N$, et on note $P = (p_{k\ell j})$, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$, $j = 1, \ldots, J_\ell$, la matrice des fréquences alléliques, où $p_{k\ell j}$ donne la fréquence de l'allèle j au locus ℓ dans la classe k.

De manière similaire à l'équation (2.1), la probabilité, connaissant les fréquences alléliques P, d'observer les allèles $x^{(i)}$ de l'individu i est donnée par

$$\begin{aligned}\Pr(x^{(i)}|P) &= \sum_{k=1}^{K} \Pr(Z_i = k|P)\Pr(x^{(i)}|P, Z_i = k) \\ &= \sum_{k=1}^{K} \Pr(Z_i = k)\Pr(x^{(i)}|P, Z_i = k).\end{aligned}$$

Au sein de chaque population k, les marqueurs sont supposés indépendants[7], et chaque marqueur ℓ suit une loi multinomiale dont les paramètres sont les fréquences alléliques $p_{k\ell\cdot}$, d'où

$$\Pr(x^{(i)} = (x_1^{(i)}, \ldots, x_L^{(i)})|P) = \sum_{k=1}^{K} \Pr(Z_i = k) \prod_{\ell=1}^{L} \Pr(x_\ell^{(i)}|P, Z_i = k) \tag{2.4}$$

$$= \sum_{k=1}^{K} \Pr(Z_i = k) \prod_{\ell=1}^{L} p_{k\ell x_\ell^{(i)}}. \tag{2.5}$$

On note X la matrice de taille $N \times L$ contenant les allèles observés $x^{(i)}$ pour chaque individu. L'objectif de `POPS` est d'estimer la loi *a posteriori* des paramètres Z et P, à savoir, s'il n'y a pas de covariable,

$$\Pr(Z, P|X) \propto \Pr(X|Z, P)\Pr(Z)\Pr(P), \tag{2.6}$$

où la vraisemblance est donnée par

$$\Pr(X|Z = (z^{(1)}, \ldots, z^{(N)}), P) = \prod_{i=1}^{N} \prod_{\ell=1}^{L} p_{z^{(i)}\ell x_\ell^{(i)}} \ , \tag{2.7}$$

7. Cette hypothèse correspond à l'hypothèse d'indépendance locale des modèles à classes latentes.

la loi *a priori* de Z est uniforme :

$$\Pr(Z_i = k) = \frac{1}{K} \quad k = 1, \ldots, K \ ,$$

et la loi *a priori* des fréquences d'allèles est une loi de Dirichlet, comme proposé par Balding and Nichols (1995) :

$$p_{k\ell\cdot} \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_\ell}). \tag{2.8}$$

Ajout de covariables

Soit $\tilde{X}$ la matrice de dimension $N \times (p+1)$ contenant, pour chaque individu, la valeur 1 dans la première colonne et les valeurs des p covariables dans les p colonnes suivantes. Comme expliqué dans la section précédente, il y a plusieurs manières d'introduire des prédicteurs/covariables dans un modèle à classes latentes. Dans les modèles *latent class regression*, les prédicteurs ont une influence uniquement sur la densité des observations, ce qui correspondrait dans notre modèle à remplacer $\Pr(X|P, Z)$ par $\Pr(X|P, Z, \tilde{X})$ dans l'équation (2.6). Il faudrait donc remplacer la loi multinomiale $\Pr(x_\ell^{(i)}|P, Z_i = k)$ par une loi dépendant aussi de la valeur des covariables pour l'individu i, *i.e.* $\Pr(x_\ell^{(i)}|P, Z_i = k, \tilde{X}_i)$. Dans les modèles de type *concomitant-variable latent class*[8], les prédicteurs influencent uniquement la probabilité *a priori* d'être dans une classe, c'est la méthode que nous avons choisie et elle correspond à écrire la loi *a posteriori* de la manière suivante :

$$\Pr(Z, P, \beta|X, \tilde{X}) \propto \Pr(X|Z, P)\Pr(Z|\tilde{X}, \beta)\Pr(\beta)\Pr(P), \tag{2.9}$$

où β correspond aux coefficients de régression qui seront définis dans les équations (2.10) et (2.11). La vraisemblance et la loi *a priori* de P restent inchangées (équations (2.7) et (2.8)). Il reste à définir la probabilité d'appartenance à une classe sachant les covariables. C'est-à-dire qu'il faut choisir un modèle qui explique une variable multinomiale, Z, par p variables discrètes et/ou continues, $\tilde{X}^{(2)}, \ldots, \tilde{X}^{(p+1)}$. Dans le cas particulier où il n'y a que deux classes (pour tout i, $Z_i \in \{1, 2\}$), une modélisation naturelle est de considérer que chaque Z_i suit une loi de Bernoulli de probabilité de succès π_i. Les π_i sont reliés aux covariables par le modèle de régression :

$$\pi_i = \mathcal{H}(\tilde{X}_i\beta), \quad i = 1, \ldots, N, \tag{2.10}$$

où $\tilde{X}_i$ est la i-ième ligne de la matrice $\tilde{X}$, β désigne le vecteur colonne des coefficients de régression et $\mathcal{H}$ est une fonction dite *de lien*. Ce modèle de régression est un modèle *probit* si $\mathcal{H}$ est la fonction de répartition de la loi normale ; c'est un modèle *logit* si $\mathcal{H}$ est la fonction de répartition de la loi logistique.

8. Parfois désignés eux-aussi sous le terme *latent class regression*

Aitchison and Bennett (1970) ont introduit le modèle probit dans le cas de réponses multinomiales (voir aussi Hausman and Wise 1978; Daganzo 1979). Albert and Chib (1993) et McCulloch and Rossi (1994) ont proposé une méthode pour l'estimation des coefficients de régression de ce modèle dans un cadre bayésien. Leur méthode permet de simuler des réplicats selon la loi *a posteriori* de β à l'aide d'un algorithme d'échantillonnage de Gibbs. Elle a par exemple été utilisée par Zhou *et al.* (2006) pour la classification de cancers à partir de données d'expression de gènes. Comme les paramètres du modèle sans métissage de `POPS` sont tous estimés via un algorithme d'échantillonnage de Gibbs (voir section 2.2.4), il est facile d'intégrer l'algorithme proposé par Albert and Chib (1993) à `POPS`. L'intégration d'un modèle logit multinomial dans un modèle à classes latentes a été proposée par Chung *et al.* (2006), mais dans ce cas les coefficients de régression β ne peuvent être estimés via un algorithme d'échantillonnage de Gibbs, du moins pas de manière évidente. Nous avons donc choisi d'intégrer à `POPS` le modèle probit multinomial plutôt que logit multinomial, et nous le décrivons dans le paragraphe suivant.

Modèle probit multinomial. Pour régresser les réponses multinomiales à K catégories, $Z = (Z_1, \ldots, Z_N)$, à partir des covariables, $\tilde{X}$, le modèle probit multinomial définit un vecteur de variables latentes continues $C_i = (C_{i,1}, \ldots, C_{i,K-1})$ pour chaque individu i. Ces variables latentes sont considérées comme les réponses des $K-1$ régressions suivantes :

$$C_{i,k} = \tilde{X}_i \beta_k + \epsilon_{i,k}, \tag{2.11}$$

$$\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,K-1}) \sim \mathcal{N}(0, \mathrm{Id}),$$

où β_k est le vecteur colonne des coefficients de la k-ième régression et Id est la matrice identité. La classe latente Z_i est donnée par la fonction suivante :

$$Z_i = \begin{cases} K & \text{if } \max_h C_{i,h} < 0 \\ k & \text{if } \max_h C_{i,h} > 0 \text{ et } \max_h C_{i,h} = C_{i,k}\,. \end{cases} \tag{2.12}$$

Ce modèle n'est pas symétrique puisqu'il n'y a pas de régression associée à la K-ième classe. Les valeurs des coefficients de régression sont définies par rapport à cette K-ième classe, appelée « classe de référence » (ou « cluster de référence »).

Variables spatiales et environnementales. Pour mieux comprendre les enjeux de l'ajout de covariables, nous détaillons les régressions des variables latentes C_i du modèle probit multinomial. Pour mettre en avant les différences entre covariables spatiales et environnementales, on désigne par $\tilde{X}^{\mathrm{S}}$ le sous-ensemble des covariables spatiales et par $\tilde{X}^{\mathrm{E}}$ le sous-ensemble des covariables environnementales. Cette distinction est courante dans le domaine de l'écologie du paysage (Lichstein *et al.* 2002) et permet de séparer les effets spatiaux et environnementaux. Elle est particulièrement adaptée aux études où l'on

désire comprendre le rôle de l'environnement, tout en tenant compte de la géographie. L'équation (2.11) est ainsi redéfinie comme

$$C_{i,k} = g(\tilde{X}_i^{\mathrm{E}})\beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}} + \epsilon_{i,k}, \quad i = 1, \ldots, N, \ k = 1, \ldots, K-1, \tag{2.13}$$

où $\epsilon_{i,k}$ est défini comme précédemment, β_k^{S} et β_k^{E} sont les 2 vecteurs colonnes des coefficients de la k-ième régression, et f est une fonction permettant de modéliser l'influence spatiale (fonction linéaire, polynôme de degré 2...).

2.2.3 Modèle avec métissage

Dans le modèle avec métissage, le génome d'un individu peut provenir de plusieurs populations. Un certain nombre de notations sont communes avec le modèle sans métissage : X est la matrice contenant les données génétiques, K le nombre maximal de populations, P la matrice des fréquences alléliques de chaque locus dans chaque population. Toutefois, le vecteur Z ne contient plus la population d'appartenance de chaque individu, mais la population d'appartenance de chaque locus ℓ pour chaque individu i, notée $z_\ell^{(i)}$ [9]. On définit de plus une matrice Q de dimension $N \times K$ dont chaque coefficient q_{ik} est le coefficient de métissage de l'individu i dans la population k. S'il n'y a pas de covariable, la distribution *a posteriori* à estimer est

$$\Pr(Q, Z, P, \alpha | X) \propto \Pr(X|Z, P)\Pr(P)\Pr(Z|Q)\Pr(Q|\alpha)\Pr(\alpha), \tag{2.14}$$

où α est un hyper-paramètre du modèle. La vraisemblance est donnée par

$$\Pr(X|Z, P) = \prod_{i=1}^{N}\prod_{\ell=1}^{L} p_{z_l^{(i)}\ell x_\ell^{(i)}}. \tag{2.15}$$

La loi *a priori* des fréquences alléliques, P, est la même que dans le modèle sans métissage (équation (2.8)). La distribution des populations de chaque locus conditionnellement aux coefficients de métissage, Q, est donnée par

$$\Pr(Z_\ell^{(i)} = k|Q) = q_{ik}, \quad i = 1, \ldots, N, \ k = 1, \ldots, K.$$

Ces coefficients de métissage suivent la loi de Dirichlet :

$$q_{i.}|\alpha \sim \mathcal{D}(\alpha_{i1}, \ldots, \alpha_{iK}), \tag{2.16}$$

où α est la matrice composée des hyper-paramètres α_{ik} pour chaque individu i et chaque population k. Pritchard *et al.* (2000a) posent $\alpha_{11} = \ldots = \alpha_{1K} = \ldots = \alpha_{NK} = \eta$ et

9. Dans le cas général d'individus de ploïdie A, une population $z_\ell^{(i,a)}$ est définie pour chaque copie $a = 1, \ldots, A$ du locus ℓ.

imposent à η une distribution uniforme entre 0 et 10. Si η tend vers 0, le modèle est proche du modèle sans admixture, c'est-à-dire que les locus d'un individu proviendront essentiellement d'une seule population. Au contraire, si η est grand, un individu aura *a priori* des locus assignés dans les K populations.

Remarque. Le modèle avec métissage ne correspond pas à un unique modèle à classes latentes, il implémente en fait un modèle à classes latentes pour chaque locus. Toutefois, ces modèles ne sont pas traités indépendamment, *i.e.* les locus ne sont pas traités indépendamment, puisque le cadre hiérarchique bayésien permet de modéliser les dépendances entre locus. Les coefficients de métissage sont des paramètres communs à tous les locus. Un modèle similaire au modèle de métissage sans covariables que l'on vient de présenter a été désigné sous le terme *latent Dirichlet allocation* et est appliqué à la classification thématique de textes (Blei *et al.* 2003).

Ajout de covariables spatiales et environnementales

Nous supposons maintenant que des covariables sont disponibles. Rappelons que l'on note $\tilde{X}$ la matrice contenant les données spatiales (sous-ensemble $\tilde{X}^{\mathrm{S}}$) et environnementales (sous-ensemble $\tilde{X}^{\mathrm{E}}$).

Dans le modèle sans métissage, ces informations sont utilisées pour modifier la probabilité *a priori* qu'un individu appartienne à une population (voir section 2.2.2). De manière similaire, dans le modèle avec métissage elles sont utilisées pour influencer la loi *a priori* des coefficients de métissage, Q. Toutefois, ce n'est pas la distribution $P(Q|\alpha)$ qui est modifiée, mais plutôt la loi *a priori* des paramètres α_{ik}. Ceci évite une influence trop forte des covariables sur Q (cette astuce a été proposée par Gaggiotti *et al.* 2004, dans leur modèle hiérarchique bayésien). Étendant le modèle spatial proposé par Durand *et al.* (2009b) à tous types de covariables, nous modélisons les variables $\alpha_{\cdot k}$ par le modèle log-normal suivant :

$$\log(\alpha_{ik}) = \tilde{X}_i^{\mathrm{E}}\beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}} + y_{ik}, \quad i = 1, \ldots, N, \ k = 1, \ldots, K, \tag{2.17}$$

où β_k^{S} et β_k^{E} sont les 2 vecteurs colonnes des coefficients de la k-ième régression, f est une fonction polynomiale, et y_{ik}, est le résidu de la régression. Le modèle logarithmique permet d'avoir des valeurs α_{ik} toujours positives, ce qui est nécessaire pour que la distribution de Dirichlet des coefficients de métissage soit définie (équation (2.16)).

Dans l'équation (2.17), l'effet spatial global est modélisé par le second terme, $f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}}$, qui correspond à la « tendance » spatiale, tandis que l'effet spatial local est modélisé par le troisième terme, y_{ik}, qui est un résidu de moyenne nulle et autocorrélé spatialement. Ce type de modélisation de l'influence spatiale a été désigné sous le terme de *krigeage universel* (Ripley 1981; Cressie 1993). Le premier terme de l'équation, $\tilde{X}_i^{\mathrm{E}}\beta_k^{\mathrm{E}}$, mesure

l'effet des covariables environnementales après cette correction pour l'influence spatiale.

Durand *et al.* (2009b) suggèrent de modéliser $y_{.k}$ par un modèle d'autocorrélation conditionnelle gaussienne (*conditional autoregressive model*, CAR ; Besag 1975; Vounatsou *et al.* 2000). Dans un tel modèle la distribution de y_{ik} dépend des valeurs des individus voisins. La distribution conditionnelle est normale, de moyenne

$$E[y_{ik}|y_{jk}, j \neq i] = \rho_k \sum_{j \sim i} w_{ij} y_{jk}, \tag{2.18}$$

et de variance

$$Var(y_{ik}|y_{jk}, j \neq i) = \sigma_k^2, \tag{2.19}$$

où ρ_k est la magnitude de l'influence du voisinage dans la population k, w_{ij} sont les poids déterminant l'influence de j sur i, et σ_k^2 est le paramètre de variance dans la population k. Les relations de voisinage sont obtenues à l'aide d'un diagramme de Voronoï [10] construit à partir des coordonnées spatiales des individus (Voronoï 1908). Les poids entre deux individus voisins i et j dépendent de la distance d_{ij} les séparant :

$$w_{ij} = \exp(\frac{d_{ij}}{\theta}), \tag{2.20}$$

où θ est la moyenne des distances entre individus.

2.2.4 Estimation des paramètres

Les paramètres des modèles présentés peuvent être estimés avec différents algorithmes. La première version de STRUCTURE repose sur une méthode de Monte Carlo par chaîne de Markov (MCMC). Mais comme nous l'avons vu dans le chapitre 1, de nombreuses extensions de STRUCTURE ont été proposées. Certaines cherchent à maximiser la vraisemblance, à l'aide par exemple d'algorithmes EM (*Expectation-Maximisation*) (Tang *et al.* 2005; Chen *et al.* 2006) ou d'un algorithme de relaxation par bloc (Alexander *et al.* 2009). Pour notre part, nous avons choisi de nous placer dans un cadre bayésien. On s'intéresse donc aux lois jointes *a posteriori* des paramètres, à savoir $\Pr(Z, P, \beta|X, \tilde{X})$ dans le modèle sans métissage et $\Pr(Z, P, Q, \alpha, \beta|X, \tilde{X})$ dans le modèle avec métissage. Ces distributions ne sont pas calculables explicitement, mais peuvent être simulées avec un algorithme MCMC. Il s'agit de construire une chaîne de Markov dont la loi stationnaire est la loi *a posteriori* jointe des paramètres. Pour cela, chaque paramètre est mis à jour conditionnellement aux autres paramètres et aux données, à l'aide de l'algorithme de Metropolis-Hastings (MH) ou de l'échantillonneur de Gibbs. L'échantillonnage de Gibbs constitue un cas particulier de l'algorithme MH, il est applicable lorsque la loi conditionnelle est connue. Dans les modèles de `POPS`, seul le paramètre α (modèle avec métissage)

10. Parfois désigné par *diagramme de Dirichlet*.

ne peut être mis à jour par échantillonnage de Gibbs. Le détail des mises à jour des paramètres est donné par Jay *et al.* (2011) (voir section 5.3).

L'estimation des paramètres a été testée sur des jeux de données simulées. Jay *et al.* (2011a) présentent des exemples qui montrent que l'appartenance aux populations, Z, et les coefficients de régression, β, sont correctement estimés (voir Figures 3.3 et 3.4).

2.3 Améliorer l'estimation de la structure génétique à l'aide d'informations environnementales

Des extensions spatiales de `STRUCTURE`, comme `BAPS` et `TESS`, ont été développées dans le but d'améliorer l'estimation de la structure génétique des populations[11]. De manière analogue, les informations environnementales, si elles sont pertinentes, devraient pouvoir améliorer l'estimation. Nous avons testé `POPS` sur des jeux de données simulées et montré que les informations environnementales permettent d'améliorer l'estimation de la structure génétique. Le bénéfice est plus élevé lorsque la quantité de marqueurs génétiques est faible (Jay *et al.* 2011a).

Nous avons tout d'abord simulé des données selon le modèle de `POPS`. C'est-à-dire que nous avons choisi des variables (latitude, longitude, langage), et fixé un nombre de populations et des coefficients de régression dans chaque cluster. À partir des modèles de régression de `POPS`, il est possible de générer les populations d'appartenance pour un certain nombre d'individus, en fonction de leurs covariables. Nous avons ensuite simulé des données génétiques pour ces individus, selon des lois multinomiales dont les paramètres dépendent de leur cluster et des fréquences alléliques (choisies à l'avance) dans cette population. De cette manière nous avons créé deux séries de jeux de données avec un nombre de locus variant de 20 à 100. Dans la première série, la latitude et la longitude sont disponibles, mais seule la latitude a été utilisée pour générer les données. Dans la deuxième série, la latitude, la longitude et 3 classifications linguistiques sont disponibles. Seules la latitude et la classification définissant 5 familles linguistiques distinctes ont servi à la génération des données. Enfin, une troisième série a été simulée pour différentes valeurs de F_{ST} à partir d'un modèle en îles dans lequel la latitude et la longitude sont prises en compte.

Nous avons appliqué `POPS` aux trois séries en incluant différentes combinaisons de covariables (exemple pour la première série : aucune covariable, longitude uniquement, latitude uniquement, latitude et longitude ; voir Figure 2.1A), puis calculé le taux d'erreurs de classification dans chacun des cas. Les modèles sans covariables correspondent au modèle de `STRUCTURE`. On constate que, pour les trois séries, l'erreur est généralement

11. Voir chapitre 1

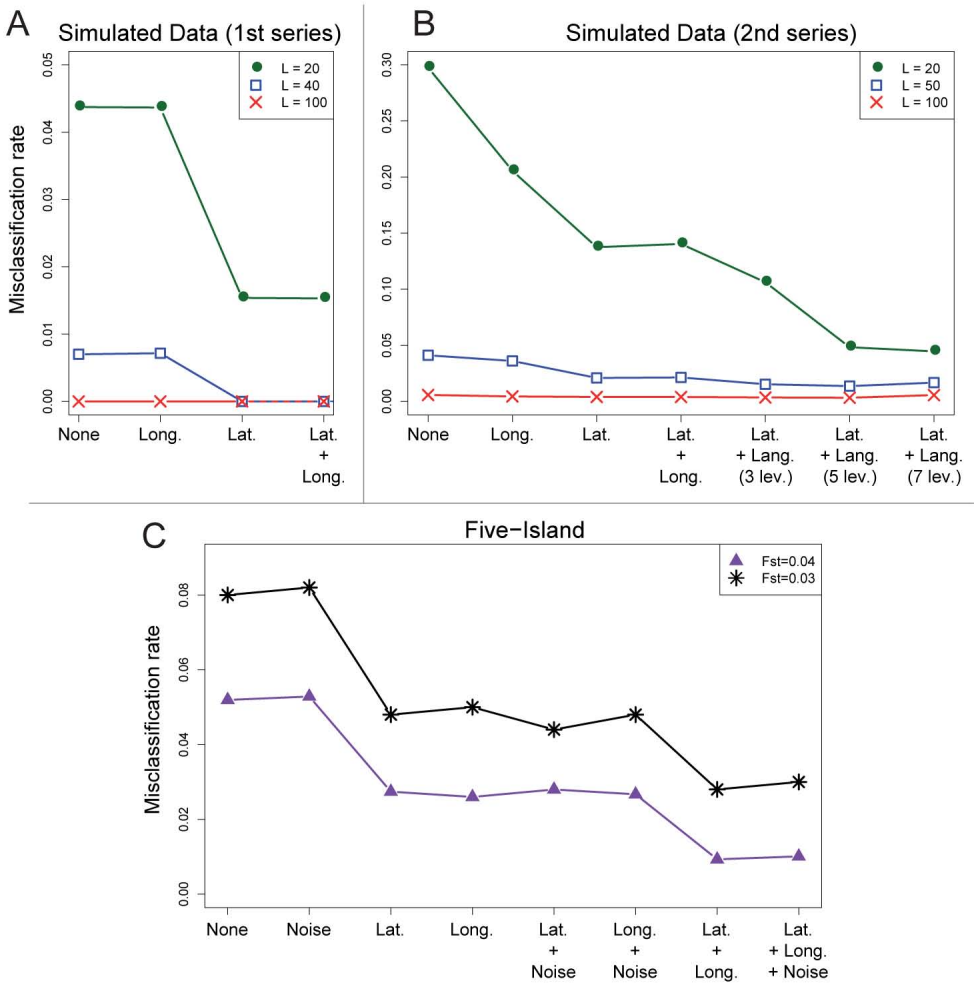


FIGURE 2.1 – **Erreurs de classification (taux) pour des jeux de données simulées en fonction des covariables utilisées dans POPS. A. Les données ont été simulées en utilisant la latitude et non la longitude. B. Les données ont été simulées à partir de la latitude et de langues issues d'une classification linguistique à 5 familles. C. Les données ont été simulées à l'aide d'un modèle à 5 îles (donc influencées par la latitude et la longitude) de manière à créer un jeu de $F_{ST} = 0,03$ et un jeu de $0,04$.**

plus faible dans les modèles avec covariables que dans les modèles sans covariable. Lorsque l'on ajoute des covariables n'ayant pas d'influence sur la structure (comme la longitude dans la première série), le taux d'erreur est généralement similaire au taux du modèle de STRUCTURE (Figure 2.1). Donc (i) l'ajout de covariables pertinentes peut améliorer les performances de détection de structure, (ii) l'ajout de covariables non pertinentes ne diminue pas ces performances. Toutefois, on constate sur la Figure 2.1 (panels A et B) que lorsqu'il n'y a que 20 locus la diminution de l'erreur est très nette, alors que pour 50 ou 100 locus cette diminution est faible, voire inexistante. Ceci est dû au fait que dans ces deux cas, les données génétiques sont suffisamment informatives pour estimer correctement la structure.

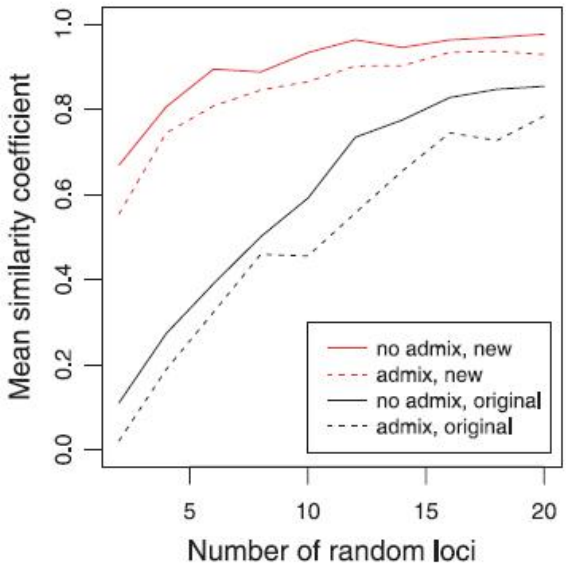


FIGURE 2.2 – HUBISZ *et al.* (2009) ont développé une extension de `STRUCTURE` intégrant comme information le lieu d'échantillonnage. Le graphique présente le score de similarité entre la structure réelle et la structure estimée en fonction du nombre de locus disponibles. Les auteurs concluent que leur méthode (courbes rouges) améliore les performances de `STRUCTURE` (courbes noires) lorsque les données génétiques sont en faible quantité.

Il est intéressant de faire le rapprochement avec l'étude de HUBISZ *et al.* (2009), dans laquelle le même phénomène a été mis en avant. Les auteurs ont développé une extension de `STRUCTURE` pour tenir compte des lieux d'échantillonnage. Par exemple, les « étiquettes » d'échantillonnage dans les données de génétique humaine du HGDP sont les populations Surui, Mandenka, etc. HUBISZ *et al.* (2009) ont appliqué leur nouvelle méthode aux données du HGDP en utilisant ces étiquettes et un nombre variable de marqueurs génétiques (de 2 à 20 locus). Ils ont constaté que les performances de `STRUCTURE` peuvent être principalement améliorées lorsque le nombre de marqueurs utilisés est faible (Figure 2.2).

2.4 Détecter les variables environnementales liées à la structure génétique

Nous détaillons ici différentes méthodes pour identifier les variables qui influencent la structure.

2.4.1 Analyser les estimations des modèles de régressions cachées

Rappelons que `POPS` modélise les relations avec l'environnement à l'aide de modèles de régressions cachées permettant d'expliquer des variables latentes liées à la structure par

des variables spatiales et environnementales. Dans ces modèles de régression (équation (2.13) pour le modèle sans métissage, équation (2.17) pour le modèle avec métissage), la valeur du coefficient associé à une covariable renseigne sur l'intensité et la direction de l'influence de cette covariable sur les variables latentes, et donc indirectement sur la structure génétique. On peut s'intéresser aux estimations ponctuelles des coefficients, mais surtout le cadre bayésien nous donne accès à leurs distributions *a posteriori* (simulées à l'aide de l'algorithme MCMC). À partir de ces distributions nous pouvons calculer les intervalles de crédibilité de chaque coefficient. Il peut être intéressant de déterminer la significativité de chaque coefficient de régression en regardant si son intervalle de crédibilité, à 95 % ou 99 % par exemple, contient la valeur 0. Des applications sont présentées dans Durand *et al.* (2009b); Jay *et al.* (2011). Toutefois l'interprétation de ces coefficients peut s'avérer difficile en pratique. Lorsque l'on veut intégrer des variables qualitatives, celles-ci sont codées par autant de variables que de catégories, moins une. Cette procédure, ajoutée au fait qu'il y ait une régression par cluster, rend le nombre de coefficients important et le sens des multiples coefficients associés à la variable qualitative est peu clair. Dans notre étude sur les relations entre langage et structure génétique des populations (Jay *et al.* 2011a), nous avons donc eu recours à d'autres critères, que nous détaillons plus loin.

2.4.2 Sélection de modèle

Une possibilité pour détecter les covariables expliquant de manière significative la structure génétique des populations est de définir plusieurs modèles qui incluent différentes combinaisons de covariables spatiales/environnementales et d'appliquer une procédure de sélection de modèle. On peut, par exemple, vouloir comparer un modèle avec covariables à un modèle sans aucune covariable, ou bien évaluer si l'ajout d'une covariable spécifique (e.g. la famille linguistique) améliore un modèle donné (e.g. n'incluant que des variables géographiques). Dans cet exemple, on voit que l'on peut évaluer l'influence des langages sur la structure en ayant corrigé pour l'effet de la géographie.

La comparaison de modèles est un problème statistique difficile. Dans le cadre des méthodes bayésiennes de classification génétique, il a particulièrement été étudié dans le but d'estimer le nombre de populations structurant les données (e.g. Smyth 2000; Pritchard *et al.* 2000a; Evanno *et al.* 2005; Guillot *et al.* 2005; Durand *et al.* 2009b; Gao *et al.* 2011). Dans ce cas, il s'agit de comparer des modèles avec différentes valeurs du nombre de classes, K. Nous allons présenter les principales catégories de méthodes utilisées pour la sélection de modèles hiérarchiques bayésiens. Bien qu'elles soient utilisées pour déterminer le nombre de populations, nous montrons que nous pouvons les adapter à notre problématique particulière du choix des covariables.

Critères basés sur la vraisemblance

Pour comparer deux modèles différents, on ne peut pas se contenter de comparer la vraisemblance des paramètres estimés dans les deux modèles. En effet, la vraisemblance augmente systématiquement avec la complexité du modèle. Elle est plus grande pour un modèle à $K+1$ classes que pour un modèle à K classes, et plus grande pour un modèle à $p+1$ covariables que pour un modèle à p covariables (si les p covariables sont communes aux deux modèles). Un modèle ayant plus de paramètres a plus de degrés de liberté pour expliquer les données, ce qui correspond à une augmentation de la vraisemblance. On comprend, par exemple, que plus un modèle est paramétré, plus les données atypiques pourront être expliquées. Si l'on pousse à l'extrême, dans un modèle à classes latentes on pourrait aller jusqu'à choisir un nombre de classes K égal au nombre d'individu N et obtenir une vraisemblance maximale. Néanmoins, ce modèle surparamétré n'a aucun intérêt pour ce qui est de la classification !

Dans `STRUCTURE` et dans la majorité de ses extensions ainsi que dans `POPS`, le paramètre K doit être fixé par l'utilisateur. Dès 2000, Pritchard *et al.* (2000a) discutent le choix de K et présentent un critère basé sur la vraisemblance marginale, $\Pr(X|K)$. En 2005, Evanno *et al.* (2005) montrent à l'aide de simulations que cette méthode n'est particulièrement pas performante pour les modèles d'îles complexes[12]. Ils proposent un nouveau critère, le critère d'Evanno, qui consiste à tracer le maximum de vraisemblance en fonction de K et à choisir la valeur de K pour laquelle le changement de pente est le plus important. Cette méthode est implémentée dans les versions récentes de `STRUCTURE` ; elle est donc couramment employée par les utilisateurs.

D'autres critères basés sur la vraisemblance ont été développés. Ils tiennent compte de l'adéquation du modèle aux données, mais pénalisent les modèles complexes. Parmi eux, l'*Aikaike Information Criterion* (AIC, Akaike 1974), le *Bayesian Information Criterion* (BIC, Schwarz 1978) et le *Deviance Information Criterion* (DIC, Spiegelhalter *et al.* 2002). Ces trois critères sont très proches. Nous détaillons ici le DIC, que nous avons implémenté dans `POPS`.

Deviance Information Criterion. Pour calculer le DIC, il faut ajouter à la moyenne *a posteriori* de la déviance, qui est une mesure directement liée à la vraisemblance, un terme de pénalité, p_D, qui correspond au nombre de paramètres « effectifs » du modèle (Spiegelhalter *et al.* 2002). Le DIC est le critère à minimiser suivant :

$$DIC = E^{\theta|X}[D(\theta)] + p_D,$$

12. C'est-à-dire pour les modèles où les îles sont elles-mêmes structurées.

où X correspond aux données, θ aux paramètres et $D(\theta)$ est la déviance, définie comme

$$D(\theta) = -2\log(\Pr(\mathrm{X}|\theta)) + \text{constante},$$

où $\Pr(\mathrm{X}|\theta)$ est la fonction de vraisemblance. Le terme de pénalité est défini comme la différence entre la moyenne *a posteriori* de la déviance et la déviance évaluée pour un estimateur *a posteriori* des paramètres, $\tilde{\theta}$,

$$p_D = E^{\theta|X}[D(\theta)] - D(\tilde{\theta}).$$

La valeur p_D pénalise le modèle par le nombre « effectif » de paramètres, qui est inférieur ou égal au nombre réel de paramètres ; plus les paramètres sont indépendants et peu contraints par les lois a priori, plus le nombre effectif des paramètres est proche du nombre réel. Pour sélectionner parmi plusieurs modèles de mélange (choix de K et des fonctions de densité du modèle), dont ils estiment les paramètres à l'aide d'un algorithme EM, Fraley and Raftery (1998) se basent sur le BIC. Le BIC nécessite de calculer le maximum de la fonction de vraisemblance, qui n'est pas forcément accessible dans notre modèle. On remarque que le DIC repose, lui, sur la moyenne *a posteriori* de la déviance, qui est très facile à calculer dans un algorithme MCMC. Cette moyenne peut en effet être estimée par la moyenne de la déviance calculée en M instants de l'algorithme (avec M suffisamment grand) :

$$E^{\theta|X}[D(\theta)] = \frac{\sum_{m=1}^{M} D(\theta^{(m)})}{M},$$

où $\theta^{(m)}$ désigne les paramètres simulés à l'instant m du MCMC. Il est intéressant de savoir que dans un modèle donné la déviance et donc le DIC sont reliés à la divergence de Kullback-Leibler [13] entre la loi des données dans le modèle considéré $\Pr(\mathrm{X}|\theta)$ et la vraie loi des données $\Pr(\mathrm{X})$ (qui est inconnue). Il existe de nombreuses variantes du DIC. En particulier dans les modèles hiérarchiques, les définitions de la déviance et de la pénalité ne sont pas évidentes. Elles dépendent de l'intérêt porté aux différents paramètres, intérêt qui guide le choix du *focus* du modèle (Spiegelhalter *et al.* 2002). La déviance peut être définie à partir de la vraisemblance complète, mais si l'on désire placer le focus sur un paramètre en particulier, on utilisera la vraisemblance marginale (qui permet d'intégrer sur les autres paramètres, considérés comme des paramètres de nuisance). Pas moins de 8 variantes sont détaillées par Celeux *et al.* (2006). Une variante encore différente, que l'on peut trouver dans Gelman (2004), consiste à utiliser comme pénalité la moitié de variance de la déviance [14]. Cette version du DIC s'avère très proche du critère de la vraisemblance marginale $\Pr(\mathrm{X}|\mathrm{K})$ proposé par Pritchard *et al.* (2000a), car l'approxi-

13. La divergence de Kullback-Leibler est une mesure de l'écart entre deux distributions introduite en 1951 par Kullback and Leibler (1951). Ce n'est pas une distance, car elle n'est pas symétrique.

14. Cette version du DIC est contestée, comme le montrent les échanges entre Andrew Gelman, Brad Carlin et Angelika Van der Linde (voir http://andrewgelman.com/2006/07/number_of_param/).

mation suggérée par les auteurs revient en fait à pénaliser la déviance par un quart de la variance de la déviance (donc la pénalité est plus forte pour le DIC que pour le critère de `STRUCTURE`). Néanmoins, le DIC est basé sur des approximations parfois critiquées, et, d'après Ando (2007), il a tendance à favoriser le surapprentissage car les données sont utilisées deux fois : pour l'estimation des paramètres et pour l'évaluation du critère.

Durand *et al.* (2009b) proposent l'utilisation du critère de DIC dans le logiciel `TESS`. Il permet d'une part de sélectionner K, d'autre part de sélectionner, à K fixé, le modèle spatial. Pour sélectionner le modèle spatial, Durand *et al.* (2009b) comparent 3 modèles pour chaque valeur de K : un modèle nul sans information spatiale, un modèle incluant la tendance spatiale globale — terme $f(\tilde{X}_i^{(S)})\beta_k^{\mathrm{S}})$ dans l'équation (2.17) — et un modèle incluant cette tendance ainsi que le terme d'autocorrélation spatiale — terme $y_{i,k}$ dans l'équation (2.17). Une autre étude a démontré récemment l'intérêt du DIC pour le choix de K (Gao *et al.* 2011). Les auteurs ont implémenté dans leur logiciel `INSTRUCT` (Gao *et al.* 2007) une version du DIC légèrement différente. Suivant une suggestion de Celeux *et al.* (2006), ils ont remplacé la pénalité p_D,

$$p_D = E^{\theta|X}[D(\theta)] + 2\sum_{i=1}^{N} \log \mathrm{Pr}(\mathrm{X_i}|\tilde{\theta})$$

par une pénalité jugée plus stable :

$$p_D = E^{\theta|X}[D(\theta)] + 2\sum_{i=1}^{N} \log \overline{\mathrm{Pr}(\mathrm{X_i}|\theta)},$$

où

$$\overline{\mathrm{Pr}(\mathrm{X_i}|\theta)} = \frac{\sum_{m=1}^{M} \mathrm{Pr}(\mathrm{X_i}|\theta^{(\mathrm{m})})}{M}$$

est la vraisemblance pour l'individu i moyennée en M pas du MCMC.

Remarque. Dans les modèles incluant des informations spatiales et/ou environnementales, comme `TESS` ou `POPS`, un phénomène de régularisation du nombre effectif de populations peut avoir lieu. C'est-à-dire que la méthode peut estimer un nombre de populations inférieur au K défini par l'utilisateur. Durand *et al.* (2009b) proposent donc de tenir compte du phénomène de régularisation lors du choix de modèle, *i.e.* de choisir le modèle en utilisant le DIC, mais de corriger ce choix dans les cas où K est plus grand que le nombre effectif de populations détectées.

Approches entièrement bayésiennes

Certaines approches considèrent le modèle, non pas comme la conséquence d'un choix *a priori* de l'utilisateur, mais comme une variable d'un « super-modèle » qui contient tous les modèles. Le nombre K de populations peut être modélisé comme une variable aléatoire

dont la loi *a posteriori* est estimée au cours de l'algorithme. Par exemple, le logiciel `STRUCTURAMA` de Huelsenbeck and Andolfatto (2007) implémente une extension de `STRUCTURE` dans laquelle K est modélisé à l'aide d'un processus de Dirichlet (Pella and Masuda 2006). Dans le cas de modèles utilisant des covariables, ce sont les covariables incluses que l'on veut faire varier au cours de l'algorithme, ce qui peut être réalisé à l'aide d'une méthode de Monte Carlo par chaîne de Markov à sauts réversibles (*reversible jump MCMC*, RJMCMC)

Le principe de la méthode RJMCMC est de visiter l'espace des paramètres non pas d'un modèle mais de plusieurs modèles. L'idée est de combiner les algorithmes MCMC de chaque modèle avec un pas (ou saut) permettant de passer d'un modèle à un autre. On parle de RJMCMC parce qu'il peut y avoir un changement de dimension d'un pas à un autre. Cet algorithme a été introduit par Green (1995). Guillot *et al.* (2005) utilisent un algorithme RJMCMC pour parcourir des modèles avec différentes valeurs de K. Dans le modèle hiérarchique bayésien `GESTE`, les coefficients de différenciation génétique sont régressés par des variables environnementales. Pour parcourir les modèles avec différentes combinaisons de covariables, Foll and Gaggiotti (2006) ont implémenté un RJMCMC. À partir d'un tel algorithme, ils obtiennent les distributions *a posteriori* des paramètres dans chaque modèle, ainsi que la probabilité de chaque modèle sachant les données. Cette probabilité est déduite du nombre de pas que la chaîne a passés dans un modèle donné. Les covariables incluses dans le modèle de plus grande probabilité sont alors considérées comme ayant une influence sur la structure. L'algorithme RJMCMC est donc intéressant, il permet, dans un même cadre, l'estimation de paramètres et la sélection de modèle. Toutefois, cela nécessite un changement assez radical dans l'implémentation de la méthode, contrairement aux approches reposant sur les critères AIC, BIC, DIC, etc.. De plus, l'exploration de l'espace des paramètres peut s'avérer difficile lorsqu'un grand nombre de modèles est envisageable, ce qui serait le cas si l'on décidait d'implémenter un RJMCMC pour sélectionner à la fois K et les covariables.

Pouvoir prédictif du modèle

Nous nous sommes particulièrement intéressés à l'utilisation des covariables environnementales pour prédire la structure génétique. Pour cela, il est utile de déterminer quelles sont les variables qui constituent le meilleur *proxy*, c'est-à-dire qui permettent le mieux de pronostiquer l'appartenance d'un individu à un groupe génétique, ou de prédire ses coefficients de métissage. Ayant cette approche prédictive en tête, nous avons naturellement cherché des procédures qui permettent de sélectionner les modèles ayant une capacité prédictive élevée.

L'avantage de `POPS` est que l'on peut utiliser les modèles de régression pour prédire *a posteriori* les coefficients d'appartenance ou de métissage d'un nouvel individu, pour lequel on ne dispose pas de données génétiques. À partir de nouvelles données spatiales, $\tilde{X}^{(S)}_{new}$, et

environnementales, $\tilde{X}^{(E)}_{new}$, et de la distribution *a posteriori* des coefficients de régression $\beta^{(S)}$ et $\beta^{(E)}$, on peut utiliser les équations (2.12) et (2.13) du modèle sans métissage pour prédire les coefficients d'appartenance du nouvel individu aux populations. De manière analogue, en utilisant les équations (2.16) et (2.17) du modèle avec métissage, on peut prédire les coefficients de métissage de ce nouvel individu.

Coefficient de corrélation. Pour un modèle donné on peut calculer la corrélation entre les coefficients d'appartenance (ou de métissage) estimés pour les individus et les coefficients prédits à partir des covariables de ces mêmes individus. Il s'agit donc de calculer la corrélation entre deux vecteurs de taille $N \times K$. Plus le modèle est capable de reproduire la structure génétique des individus à partir des covariables, plus le coefficient de corrélation est proche de 1. De plus, il est possible de calculer un intervalle de confiance pour ce coefficient. En effet, sous certaines conditions, la loi de la transformation de Fisher du coefficient de corrélation entre deux vecteurs de taille M de variables aléatoires est une gaussienne de moyenne $\text{arctanh}(r)$ et de variance $1/\sqrt{M-3}$ (Fisher 1915), où $\text{arctanh}(r)$ est la transformation de Fisher, définie comme

$$\text{arctanh(r)} = \frac{1}{2}\ln\frac{1+\text{r}}{1-\text{r}}.$$

Faire cette hypothèse nous permet de calculer les intervalles de confiance pour le coefficient de corrélation en appliquant la transformation inverse. Calculer le coefficient de corrélation est une méthode simple, mais les données sont utilisées deux fois : pour l'estimation des paramètres et pour la prédiction. Comme avec le DIC, il y a un risque de surévaluer les capacités prédictives d'un modèle. Il ne faut donc pas se contenter de chercher le modèle pour lequel la corrélation est maximale, mais plutôt tracer la corrélation en fonction de la complexité du modèle (fonction croissante, donc) et observer à partir de quelle complexité le score demeure stable. Ceci est illustré par une étude de simulations à la fin de la section 2.4.2 (voir Figure 2.3).

Validation croisée. Une méthode statistique couramment utilisée pour évaluer les performances prédictives d'un modèle est la validation croisée (Ripley 1996; Hastie *et al.* 2009). Il s'agit de séparer le jeu de données en deux parties (pas forcément de tailles égales) : une partie constitue l'ensemble d'apprentissage et sert au calibrage du modèle ; l'autre constitue l'ensemble de validation, ou ensemble test, et permet de valider le modèle. Dans notre cas, il s'agit d'utiliser une partie des données pour estimer les distributions *a posteriori* des paramètres Z, P, β et éventuellement Q et α (dans le modèle avec métissage), puis de voir si le modèle ainsi paramétré se généralise à de nouvelles données, en calculant la vraisemblance des paramètres pour ces nouvelles données. La log-vraisemblance de ces données est aussi appelée *log predictive score* (Good 1952). Pour que la validation

soit *croisée*, ce score doit être moyenné pour différents couples (ensemble d'apprentissage, ensemble de validation). Pour la validation croisée *2-fold*, par exemple, il suffit d'inverser les rôles de l'ensemble d'apprentissage et de l'ensemble de validation.

L'approche la plus classique consiste à utiliser un certain nombre d'individus pour l'apprentissage, le reste des individus pour la validation. C'est cette approche que Smyth (2000) décrit dans le cadre des modèles de mélange, pour sélectionner le nombre de clusters représentés dans les données. Toutefois, dans notre situation les données contiennent des informations génétiques et des covariables. Si certaines covariables sont qualitatives, le découpage « selon les individus » peut être problématique. En effet, les paramètres risquent d'être mal estimés si une ou plusieurs des catégories de la variable sont mal représentées dans l'ensemble d'apprentissage. Pour étudier les relations entre les familles linguistiques et la structure génétique (Jay *et al.* 2011a), ce type de découpage n'était donc pas adéquat. Nous avons choisi de partitionner « selon les locus », c'est-à-dire que tous les individus sont représentés dans l'ensemble d'apprentissage, noté X^{training}, mais le nombre de locus est limité. Les données pour les mêmes individus mais aux locus restants constituent l'ensemble de validation, noté $X^{\text{validation}}$. Le score de validation est alors défini comme l'espérance *a posteriori* de la log-vraisemblance, sachant les données de validation, des paramètres estimés à partir de X^{training}, c'est-à-dire :

$$E^{\theta|X^{\text{training}}}\left[\log(\Pr(X^{\text{validation}}|\theta))\right],$$

où θ correspond à l'ensemble des paramètres du modèle. Dans le modèle sans métissage, on peut écrire que

$$\log(\Pr(X^{\text{validation}}|\theta)) = \log(\Pr(X^{\text{validation}}|Z)) \tag{2.21}$$

$$= \sum_{k=1}^{K} \sum_{\ell \in X^{\text{validation}}} \log \Pr(x_\ell^{[k]}), \tag{2.22}$$

où $x_\ell^{[k]}$ correspond aux allèles observés au locus ℓ dans la population k. Pour calculer la vraisemblance, la difficulté ici est que les fréquences des allèles n'ont pas été estimées. On peut donc considérer ces fréquences alléliques comme des paramètres de nuisance, et intégrer sur ces fréquences. On note $n_{k\ell} = (n_{k\ell 1}, \ldots, n_{k\ell J_\ell})$ le vecteur des nombres de chaque allèle observés au locus ℓ dans la population k. La variable $n_{k\ell}$ est de loi multinomiale

$$n_{k\ell} \sim \text{Multinomial}(m_{k\ell}, p_{k\ell 1}, \ldots, p_{k\ell J_\ell}),$$

où $m_{k\ell} = \sum_{j=1}^{J_\ell} n_{k\ell j}$ est le nombre d'allèles différents au locus ℓ dans la population k. En intégrant sur les fréquences alléliques, on obtient la distribution conjuguée Dirichlet

multinomiale [15] :

$$\begin{aligned}\Pr(x_\ell^{[k]}) &= \frac{\prod_j (n_{k\ell j}!)}{m_{k\ell}!}\Pr(n_{k\ell}|\lambda) \\ &= \frac{\Gamma(\lambda J_\ell)}{\Gamma(m_{k\ell}+\lambda J_\ell)} \prod_j \frac{\Gamma(n_{k\ell j}+\lambda)}{\Gamma(\lambda)}.\end{aligned}$$

Cette modélisation des données à l'aide d'un modèle multinomial Dirichlet a été proposée par Rannala and Mountain (1997) et utilisée par Balding (2003), Corander *et al.* (2003) et Foll and Gaggiotti (2006) dans leurs méthodes.

Si l'on voulait définir le score de validation pour le modèle avec métissage, on ne pourrait pas procéder exactement de la même façon. En effet, l'équation (2.21) n'est plus valable puisqu'on ne connait pas les clusters d'appartenance des locus de l'ensemble de validation, ie les $Z_\ell^{(i)}, i = 1, \ldots, N,\ \ell = 1, \ldots, L$. En plus d'intégrer sur les fréquences alléliques, il faudrait donc intégrer sur les clusters d'appartenance des locus. Le score de validation dans le modèle avec métissage s'écrit

$$\begin{aligned}\log(\Pr(X^{\text{validation}}|\theta)) &= \log(\Pr(X^{\text{validation}}|Q)) \\ &= \sum_{\ell \in X^{\text{validation}}} \sum_{i=1}^{N} \log \Pr(x_\ell^{(i)}|q_{i\cdot}) \\ &= \sum_{\ell \in X^{\text{validation}}} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \Pr(x_\ell^{(i)}|Z_\ell^{(i)} = k)\ q_{ik} \right)\end{aligned}$$

On ne peut retrouver directement la distribution conjuguée Dirichlet multinomiale comme dans le modèle sans métissage. Une solution approchée repose sur une méthode de Monte Carlo, et consiste à simuler des clusters d'appartenance pour tous les individus et les nouveaux locus selon la distribution

$$\Pr(Z_\ell^{(i)} = k|Q) = q_{ik}, \quad i = 1, \ldots, N,\ k = 1, \ldots, K$$

Il sera alors possible d'écrire le score sous la forme de l'équation 2.22 et de le calculer à l'aide de la distribution conjuguée Dirichlet multinomiale. Il est en revanche nécessaire de renouveler les étapes plusieurs fois, et de moyenner les scores. Nous ne l'avons pas implémenté dans `POPS`, car nous avons eu recours à la validation croisée dans l'étude des relations entre structure génétique et langues, étude dans laquelle le modèle avec métissage n'a pas été pas utilisé (Jay *et al.* 2011a).

Ayant défini le score de validation, on peut choisir de réaliser une validation croisée *d-fold* : les données sont partitionnées en d ensembles qui serviront à tour de rôle d'ensemble de validation, tandis que les $(d-1)$ ensembles restants serviront d'ensemble

15. *Multivariate Pólya distribution.*

d'apprentissage. Le score de validation croisée est donné par la moyenne des scores définis dans l'équation (2.21) pour chacun des d ensembles de validation. La procédure *leave-one-out* correspond au cas particulier où d est égal à la taille de l'échantillon (ici au nombre de locus, L).

La validation croisée est reconnue comme une excellente méthode pour sélectionner un modèle en évitant le sur-apprentissage (Ripley 1996; Hastie *et al.* 2009). Un modèle trop complexe explique bien les données d'apprentissage, mais se généralise mal et est donc pénalisé par le score de validation croisée. Toutefois, elle est assez gourmande en temps de calcul, ce qui est un point négatif quand on l'applique à des algorithmes reposant sur des méthodes MCMC. Dans le cas d'une validation croisée *d-fold*, il faut en effet relancer d fois l'algorithme de base ; et si cet algorithme est stochastique, il est en général exécuté à plusieurs reprises, et ce pour chacun des modèles que l'on souhaite étudier. Les critères DIC ou de corrélation que nous avons présentés plus haut sont certainement moins sophistiqués que la procédure de validation croisée ; ils ont en revanche l'avantage d'être extrêmement faciles à implémenter et sont très peu coûteux en temps de calcul. De plus, Smyth (2000) met en avant le fait que le score de validation croisée est relié à la divergence de Kullback-Leibler, ce qui a aussi été démontré pour le DIC. L'espérance du score de validation est égale, à une constante près, à la divergence de Kullback-Leibler de la fonction de densité du modèle par rapport à la vraie densité (inconnue). L'espérance de la déviance est, elle aussi, proportionnelle à cette divergence de Kullbak-Leibler (Gelman 2004). La validation croisée et le DIC cherchent donc à approcher le même critère.

Remarque. La méthode de validation croisée a très récemment été implémentée dans un logiciel de classification génétique, `ADMIXTURE`, pour la sélection du nombre de populations (Alexander and Lange 2011). Leur algorithme, plus rapide que les algorithmes reposant sur un MCMC comme `POPS`, permet de rendre possible l'utilisation systématique de la procédure de validation croisée.

Étude de simulations. Jay *et al.* (2011a) testent l'utilisation des critères de corrélation et de validation croisée *2-fold*, pour la sélection de modèle. Pour cela, nous avons simulé des données selon le modèle de `POPS`, de sorte que la structure génétique soit influencée par un certain nombre de covariables spatiales ou environnementales. Ces simulations sont décrites dans la section 2.3. Pour chaque jeu de données nous faisons varier les combinaisons de variables incluses dans le modèle avant d'appliquer `POPS` et de calculer le coefficient de corrélation entre structure génétique estimée et structure génétique prédite, et le score de validation croisée. Sur la Figure 2.3, les trois panels correspondent à trois séries de simulations différentes, et dans chaque panel le vrai modèle est surligné en rouge. Quel que soit le nombre de locus dans les données, les scores de corrélation et de validation croisée atteignent un plateau lorsque la bonne combinaison de covariables

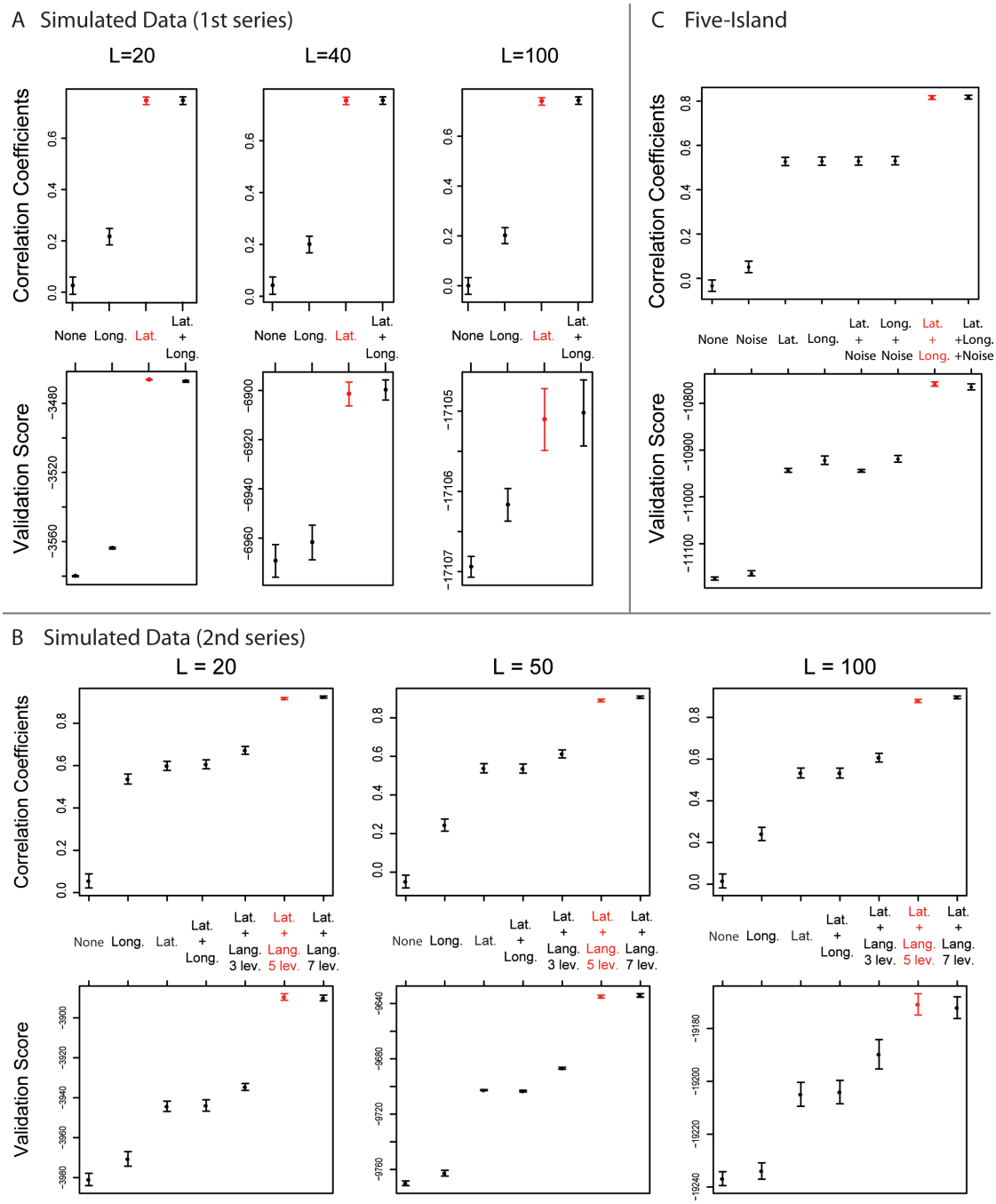


FIGURE 2.3 – **Sélection de variables pour des données simulées.** Les coefficients de corrélation r mesurent les corrélations entre les coefficients d'appartenance estimés et prédits. Leurs intervalles de confiance sont calculés à partir de la transformation de Fisher. Les scores de validation sont estimés à partir d'une validation croisée *2-fold* et leurs mesures de déviation standard à partir d'une méthode de *bootstrap*. Pour chacune des trois séries de données (panels A, B, ou C), la combinaison de covariables ayant servi à la simulation est surlignée en rouge.

est incluse ; c'est-à-dire que tant que l'on ajoute une covariable liée à la structure, les scores augmentent, mais lors de l'ajout d'une covariable supplémentaire n'influençant pas la structure, les scores n'augmentent plus, ou très peu. Ces critères permettent donc de sélectionner le bon modèle, à savoir celui qui contient la latitude dans la première série (Figure 2.3A), la latitude et la classification linguistique à 5 familles dans la deuxième série (Figure 2.3B), la latitude et la longitude dans la troisième série (Figure 2.3C).

2.5 Projeter la structure génétique des populations

2.5.1 Objectifs

Médecine personnalisée. En génétique humaine, prédire la structure génétique de populations d'individus à partir de leurs données environnementales uniquement pourrait être utile pour la médecine personnalisée. Comme expliqué dans le chapitre 1, il a été montré que certaines variations dans la réponse aux traitements sont liées à la structure génétique. Si la médecine personnalisée se développe, il peut donc être intéressant d'avoir de bons prédicteurs de la structure génétique. À cette fin, on pourra utiliser des variables environnementales, culturelles et spatiales comme « substituts » de la structure (Wilson *et al.* 2001; Risch *et al.* 2002; Bamshad *et al.* 2003). Toutefois, si l'information génétique des nouveaux individus est accessible, il est préférable de l'utiliser (Schwartz 2001; Wilson *et al.* 2001). N'ayant pas directement appliqué `POPS` à ce type de problématique, nous ne détaillerons pas d'avantage.

Changements environnementaux. En écologie, de nombreuses études s'intéressent à la prédiction de la distribution d'une espèce en cas de modification de l'environnement. Il peut s'agir de la distribution d'une espèce invasive qui vient d'apparaître dans une nouvelle région, où les conditions environnementales sont différentes de celles de sa région d'origine (Peterson 2003a), ou encore de la distribution d'espèces en cas de changement climatique important (Peterson 2003b; Thuiller *et al.* 2008, voir Figure 2.4). En général, ces études reposent sur des modèles de distribution d'espèce, appelés aussi « modèles bioclimatiques » (Guisan and Thuiller 2005; Jeschke and Strayer 2008). À part quelques exceptions, elles ne tiennent pas compte des variations génétiques intraspécifiques et considèrent que tous les individus d'une espèce sont liés à l'environnement de la même manière. Nous avons pourtant vu que structure génétique et variables environnementales peuvent être reliées ; et d'ailleurs, plusieurs articles soulignent qu'il serait utile d'inclure l'aspect génétique dans les modèles bioclimatiques habituellement utilisés pour prédire la distribution d'espèce (Davis and Shaw 2001; Davis *et al.* 2005; Aitken *et al.* 2008).

Le modèle de `POPS`, et en particulier le modèle avec métissage, nous permet de proposer

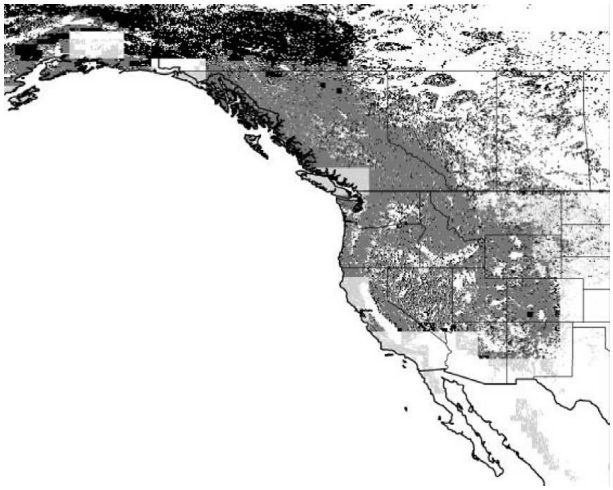


FIGURE 2.4 – À l'aide d'un modèle bioclimatique, PETERSON (2003b) projette sur une carte la future distribution du cincle d'Amérique (une espèce de passereaux) en cas de changement climatique. Gris clair : zones habitables actuellement, prédites comme n'étant plus habitables dans le futur ; gris foncé : zones habitables actuellement, prédites comme restant habitables dans le futur, noir : zones non habitables, prédites comme devenant habitables.

une nouvelle approche pour étudier les conséquences de modifications environnementales. Dans cette approche, nous tenons compte des variations génétiques au sein d'une espèce, grâce aux coefficients de métissage des individus. Si l'environnement et les coefficients de métissage sont liés, un changement environnemental à venir (par exemple une augmentation de température) aura probablement des répercussions sur la structure génétique des populations, répercussions qui peuvent être prédites par `POPS`.

Nous présentons brièvement dans cette section les aspects méthodologiques. Les enjeux et les hypothèses biologiques des modèles bioclimatiques, d'une part, et de l'approche que nous proposons, d'autre part, sont détaillés dans le chapitre 4. Une étude des effets du changement climatique sur la structure génétique des populations de 20 plantes alpines (JAY *et al.* 2011b) est aussi présentée dans le chapitre 4.

2.5.2 Modèle de projection de la structure génétique des populations

Contexte : les modèles de distribution d'espèce. Dans les modèles de distribution d'espèce (*species distribution models*, SDMs) ce sont les données d'absence/présence, parfois de présence uniquement, qui sont reliées à l'environnement. De nombreuses méthodes sont actuellement employées pour estimer cette relation (GUISAN and ZIMMERMANN 2000). La méthode la plus évidente est certainement la régression logistique. Les données d'absence/présence peuvent en effet être codées comme des variables binaires (absence = 0, présence = 1) que l'on souhaite expliquer à l'aide de combinaisons de me-

sures environnementales. Pour cela, une solution simple est de calibrer un modèle linéaire généralisé (GLM (*generalized linear model*), Nelder and Wedderburn 1972) avec comme fonction de lien la fonction logit. Un modèle de régression plus souple, le modèle additif généralisé (GAM (*generalized additive model*), Hastie and Tibshirani 1990), est aussi utilisé. Dans ce modèle la réponse binaire est expliquée par l'addition de fonctions, à ajuster, des variables environnementales (Yee and Mitchell 1991). D'autres types d'approche ne reposent pas sur les modèles de régression. Par exemple, l'algorithme de type CART (*classification and regression tree*) est appliqué pour construire un arbre de décision binaire expliquant la présence ou l'absence (Thuiller 2003). Les règles de décision (à chaque nœud de l'arbre) sont construites à partir des variables environnementales. Des méthodes basées sur l'estimation de l'enveloppe minimale de la distribution d'une espèce dans l'espace multidimensionnel des variables climatiques ont aussi été développées (Busby 1991; Walker and Cocks 1991; Carpenter *et al.* 1993). Enfin, d'autres approches s'appuyant sur des méthodes de réseaux de neurones, d'analyse canonique des correspondances ou encore d'analyse discriminante sont présentées par Guisan and Zimmermann (2000) et Jeschke and Strayer (2008).

Une fois le modèle choisi calibré, la présence (ou parfois la probabilité de présence) d'une espèce dans un emplacement peut être prédite à partir des mesures environnementales (actuelles, passées, futures) de cet emplacement. De cette manière, on peut, par exemple, obtenir la distribution d'une espèce en cas de changement climatique et la comparer à la distribution actuelle. La comparaison mettra en évidence les zones dans lesquelles l'espèce risque de disparaître et l'étendue des zones qui vont rester ou devenir habitables dans le futur (e.g. Figure 2.4).

Utiliser `POPS` pour projeter la structure génétique de populations. Le principe de notre approche est d'étudier non pas la distribution d'une espèce dans sa globalité, mais la distribution intraspécifique. Pour cela nous nous intéressons aux données génétiques multilocus, plutôt qu'aux données d'absence/présence. La méthode de `POPS`, que nous avons présentée dans ce chapitre, consiste à estimer conjointement la structure génétique de population d'une espèce et son lien éventuel à des covariables non génétiques. Si ce lien est significatif, la structure peut être prédite à partir des covariables, en particulier à partir de covariables dont les valeurs ont été modifiées (pour cause d'un changement climatique par exemple). Le modèle de régression cachée de `POPS` permet en effet de simuler *a posteriori* des coefficients de métissage en utilisant les équations (2.16) et (2.17), la distribution *a posteriori* des coefficients de régression et les nouvelles valeurs de covariables. La projection sur une carte des coefficients de métissage donne alors un aperçu de la structure génétique des populations en cas de changement des conditions environnementales.

De manière analogue aux SDMs qui comparent distributions actuelle et future de l'espèce, il est possible de comparer projections actuelle et future de la structure génétique

des populations. Nous définissons ci-dessous deux critères que nous avons utilisés pour réaliser ces comparaisons.

Déplacement de la zone de contact entre deux clusters. Nous pouvons évaluer des phénomènes comme des mouvements de clusters dans l'espace ou encore la disparition d'un cluster spécifique. En particulier, si une zone de contact existe entre deux clusters voisins géographiquement (i.e. une zone où les individus sont métissés et où l'on peut observer une variation graduelle des coefficients de métissage), il est possible d'étudier son déplacement en cas de changement climatique. On localise la zone de contact à l'aide d'une courbe correspondant aux valeurs des coefficient de métissage égales à $0,5$. L'amplitude du déplacement en cas de changement climatique est définie comme la distance entre la courbe pour la zone de contact actuelle et la courbe pour la zone de contact prédite.

Renouvellement génétique de l'espèce (*intraspecific turnover*). Ce critère nous permet d'évaluer la modification globale de la structure génétique des populations. Il est mesuré à l'aide du coefficient de corrélation entre la matrice des coefficients de métissage pour les conditions actuelles et la matrice des coefficients de métissage prédits en cas de changement climatique. Plus la corrélation est faible, plus le renouvellement génétique est important, *i.e.* plus le changement climatique risque d'avoir un impact élevé sur la structure génétique des populations.

Chapitre 3

Relations entre structure génétique et langages dans des populations amérindiennes

3.1 Contexte

Pour mieux comprendre l'histoire du peuplement humain, les chercheurs combinent souvent les informations apportées par diverses disciplines : paléontologie, archéologie, génétique des populations mais aussi linguistique. L'histoire évolutive des langages constitue en elle-même un centre d'intérêt et peut être étudiée à la lueur de données génétiques ou de méthodes utilisées en génétique des populations (par exemple, la reconstruction d'un arbre de divergence des langues indo-européennes par Gray and Atkinson 2003). Déjà Darwin, dans son ouvrage *L'Origine des espèces*, déclarait que la généalogie des populations humaines, si elle était parfaitement connue, procurerait la meilleure classification possible des langues actuelles (Darwin 1859). C'est-à-dire que si les humains sont reliés entre eux selon un processus évolutif donné, les langues devraient avoir évolué selon ce même processus. Pour expliquer cela, on a supposé que les populations sont reliées entre elles essentiellement par une suite d'événements d'expansion-fission ayant eu lieu au cours de la colonisation de la planète. Durant ces événements, gènes et langues auraient colonisé simultanément de vastes zones et y auraient coévolué (Cavalli-Sforza *et al.* 1988, 1992). De là viendrait la forte corrélation actuelle entre génétique, linguistique et géographie. Cavalli-Sforza et ses collaborateurs ont, entre autres, trouvé d'importantes correspondances entre l'arbre évolutif construit à partir de marqueurs génétiques de 38 populations du HGDP réparties dans le monde et la classification linguistique établie par Ruhlen (1987) (voir Figure 3.1 ; Cavalli-Sforza *et al.* 1992). Si les avis ne sont pas toujours aussi tranchés, il n'en demeure pas moins que les relations entre gènes et langages sont au centre de nombreuses études (Sokal *et al.* 1988; Excoffier *et al.* 1991; Belle

and Barbujani 2007; Hunley *et al.* 2008).

Dans notre étude, nous nous intéressons particulièrement aux relations entre langues, géographie et structure génétique de populations amérindiennes. La classification des langues amérindiennes a fait l'objet de vifs débats. Dans leur étude, basée sur des données génétiques, linguistiques et dentaires, Greenberg *et al.* (1986) suggèrent que le peuplement américain a eu lieu en trois vagues de migration venues d'Asie, à l'origine de trois groupes linguistiques différents (Amerinde, Na-Dene et Esquimo-Aléoute). Cette théorie, et en particulier l'existence d'une super-famille linguistique Amerinde regroupant un grand nombre de langages, est fortement controversée (Bolnick *et al.* 2004). À ce jour, il n'existe toujours pas de consensus quant à la classification linguistique des langues amérindiennes. Nous avons donc tenu compte de deux classifications : d'une part, celle proposée par Greenberg et Ruhlen (Greenberg 1987; Ruhlen 1991) ; d'autre part, celle de Gordon (2005), disponible sur le site *The Ethnologue*[1].

Les précédentes études des relations entre génétique et langages sont principalement basées sur deux types de méthode détaillés dans la section 1.3. La méthode de Mantel teste la significativité de la corrélation entre distances génétiques et distances linguistiques. Elle peut être accompagnée du test de Mantel partiel pour tenir compte des distances géographiques. Ces tests ont été appliqués aux populations du HGDP par Belle and Barbujani (2007) et aux populations amérindiennes par Wang *et al.* (2007). Belle and Barbujani (2007) trouvent que les données linguistiques expliquent une faible partie de la variance génétique du HGDP lorsque la géographie est prise en compte, mais surtout que la significativité de la corrélation dépend des définitions choisies pour les distances. Wang *et al.* (2007) mettent en avant le fait que la corrélation chez les populations amérindiennes est significative seulement si les populations, comparées deux à deux, parlent des langues du même *stock* linguistique. Un deuxième type d'approche repose sur des arbres de classification linguistique. Cavalli-Sforza *et al.* (1992) trouvent une forte association entre arbres génétiques et arbres linguistiques de 42 populations, en utilisant une mesure appelée *constitency index*. Hunley and Long (2005) et Hunley *et al.* (2007) ont, eux, développé un test pour déterminer si un arbre linguistique est compatible avec des distances génétiques. Lorsqu'ils l'appliquent à différentes classifications linguistiques chez les populations amérindiennes, ils ne trouvent aucune classification compatible avec les distances génétiques ; toutefois, certaines classifications fournissent une bien meilleure prédiction des distances génétiques que d'autres.

Pour notre part, nous proposons d'étudier les relations entre structure génétique et langages dans un cadre méthodologique nouveau, et de répondre aux questions suivantes. Dans quelle mesure la géographie et les langues peuvent-elles expliquer la structure génétique des populations amérindiennes ? Ajoutées aux informations géographiques, les

1. http://www.ethnologue.com/

informations linguistiques permettent-elles d'améliorer la prédiction de la structure génétique des populations ? Y a-t-il une classification des langages amérindiens qui constitue un meilleur prédicteur de cette structure ?

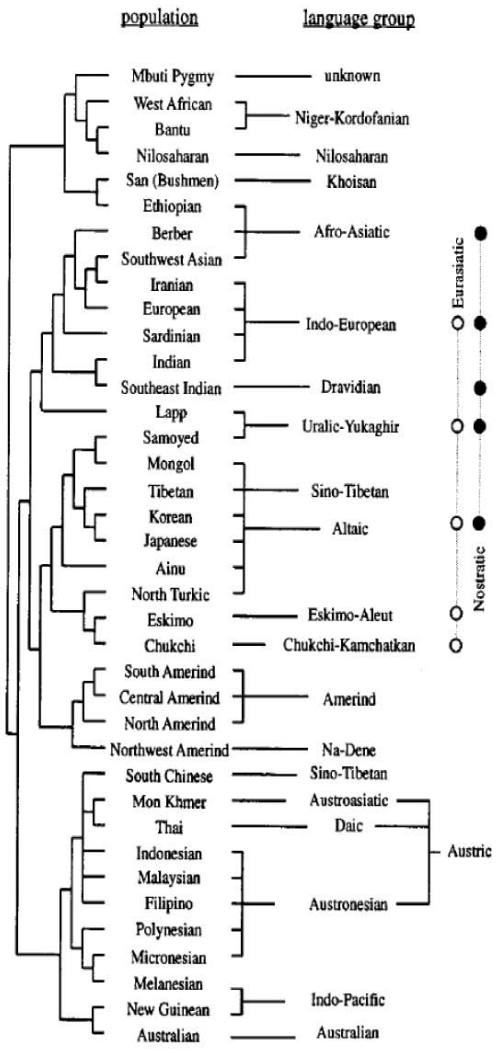


FIGURE 3.1 – À gauche, un arbre montrant l'évolution génétique des populations construit par CAVALLI-SFORZA *et al.* (1988). Seule la topologie de l'arbre est présentée, *i.e.* les longueurs de branches ne correspondent pas aux temps de divergence. À droite, les familles linguistiques de la classification de GREENBERG (1987). CAVALLI-SFORZA *et al.* (1992) présentent cette figure dans le cadre de leur étude de la coévolution entre gènes et langages.

3.2 Résultats et discussion

Après avoir testé l'efficacité du modèle sans métissage de `POPS` sur des données simulées (voir Figures 3.3, 3.4 et 3.5), nous l'avons appliqué à 512 individus issus de 28 populations amérindiennes, pour lesquelles 678 marqueurs microsatellites ont été génotypés

(données issues de Wang *et al.* 2007). Le principe de `POPS` est d'estimer conjointement la structure génétique des populations et l'influence de covariables non génétiques. Les covariables sont utilisées comme prédicteurs dans une régression cachée des coefficients d'appartenance aux clusters. Nous avons considéré 4 modèles qui n'utilisent pas les mêmes covariables. Le modèle A inclut les informations géographiques via des combinaisons quadratiques des coordonnées spatiales des individus. Les modèles B à D incluent à la fois ces informations géographiques et des informations linguistiques via une variable multinomiale. Dans le modèle B les catégories de la variable linguistique correspondent aux *stocks* de la classification de Greenberg (8 catégories) ; dans le modèle C elles correspondent aux *groupes* de cette classification, ce qui correspond à une échelle plus fine (14 catégories) ; et dans le modèle D elles correspondent aux *familles* de la classification de *The Ethnologue* (16 catégories). L'utilisation de l'information géographique dans tous les modèles permet d'étudier les relations entre structure génétique et langages tout en corrigeant pour la géographie.

Nous avons comparé les structures génétiques de populations estimées par les 4 modèles. La quantité d'information génétique étant suffisante pour détecter la structure, les covariables n'ont pas beaucoup influencé les estimations, et la Figure 3.9 montre que les coefficients d'appartenance varient peu d'un modèle à l'autre. Nous avons ensuite calculé, pour chaque modèle, une mesure de corrélation entre la structure génétique estimée et la structure génétique prédite par les covariables[2]. La corrélation pour le modèle A (géographie uniquement) est de $0,81$, ce qui montre que la géographie est un bon prédicteur de la structure génétique (Figure 3.7A). Toutefois, la mesure de corrélation pour les modèles incluant des informations linguistiques est plus élevée (0,94-0,98), et la structure prédite est nettement différente de celle prédite par le modèle A (Figures 3.6A et 3.9). L'information linguistique améliore donc la prédiction de la structure génétique. Enfin, nous avons utilisé un critère reposant sur une technique de validation croisée *2-fold*[3] pour comparer l'apport des différentes classifications linguistiques utilisées. La figure 3.6B montre que le score de validation croisée est plus élevé pour le modèle incluant la classification de *The Ethnologue*. Cette classification est donc plus adaptée que la classification de Greenberg pour prédire la structure génétique des populations amérindiennes. Néanmoins, certaines familles linguistiques de la classification de *The Ethnologue*, comme les familles Chibchan, Choco et Tupi, ne présentent pas une correspondance univoque avec les clusters génétiques (Figure 3.6).

Il est intéressant de remarquer que la classification *The Ethnologue* est celle qui ne contient pas de « super-groupes » linguistiques (i.e. qui n'impose pas de liens entre les familles linguistiques). La conclusion que cette classification est la plus adaptée va donc

2. Voir la section 2.4.2 pour plus de détails.
3. Voir la section 2.4.2 pour plus de détails.

dans le même sens que les résultats de Wang *et al.* (2007) et Hunley *et al.* (2007). En effet, Wang *et al.* (2007) ont trouvé que la corrélation partielle entre distances linguistiques et génétiques est seulement de $0,01$, mais qu'elle monte à $0,40$ si les comparaisons entre populations amérindiennes ne se font qu'au sein d'un même *stock*. Hunley *et al.* (2007), eux, ont montré que les arbres de classification linguistique expliquent mieux les distances génétiques observées s'ils n'ont pas de structure interne ancienne, c'est-à-dire si les classifications ne contiennent pas de super-familles ou de super-groupes linguistiques (voir Figure 3.2).

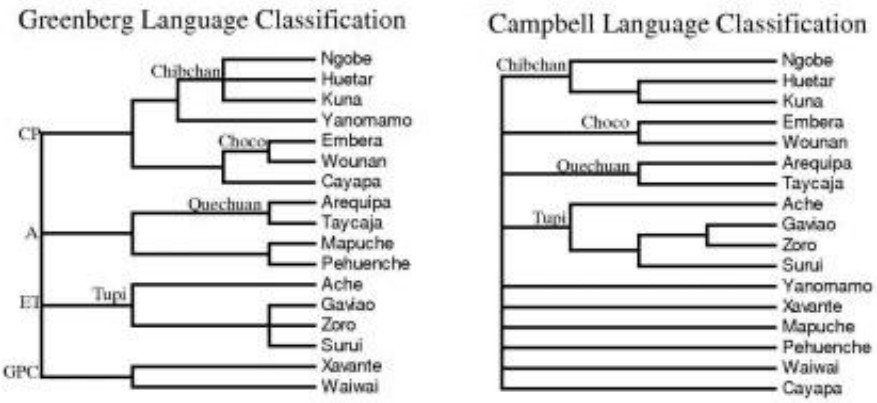


Figure 3.2 – Deux des classifications linguistiques considérées par Hunley *et al.* (2007) pour tester la coévolution entre gènes et langages en Amérique centrale et en Amérique du Sud. À gauche, la classification de Greenberg (Greenberg 1987; Ruhlen 1991). À droite, la classification de Campbell (Campbell 1987). Les auteurs montrent que la classification de Campbell, qui a moins de « structure interne » (i.e. moins de super-familles linguistiques), explique mieux les distances génétiques que les autres classifications.

Pour comprendre cet effet d'échelle, nous avons réalisé une étude similaire sur 77 populations du HGDP, mondialement réparties. En comparant le modèle avec géographie au modèle avec géographie et langages, nous avons trouvé que l'apport des langages pour améliorer la prédiction de la structure était très faible.

Les études de Hunley *et al.* (2007), Wang *et al.* (2007) et Belle and Barbujani (2007), ainsi que nos résultats tendent donc à montrer qu'un lien entre langages et structure génétique des populations amérindiennes ne pourrait être détectable qu'à fine échelle, en dessous d'un certain seuil de de différenciation linguistique, tandis qu'à plus grande échelle les différenciations génétiques refléteraient les événements démographiques anciens et non l'influence de traits culturels.

3.3 Article A

F. Jay, O. François, and M.G.B. Blum. Predictions of native American population structure using linguistic covariates in a hidden regression framework. PLoS ONE, 6, 2011.

Abstract

Background

The mainland of the Americas is home to a remarkable diversity of languages, and the relationships between genes and languages have attracted considerable attention in the past. Here we investigate to which extent geography and languages can predict the genetic structure of Native American populations.

Methodology/Principal Findings

Our approach is based on a Bayesian latent cluster regression model in which cluster membership is explained by geographic and linguistic covariates. After correcting for geographic effects, we find that the inclusion of linguistic information improves the prediction of individual membership to genetic clusters. We further compare the predictive power of Greenberg's and *The Ethnologue* classifications of Amerindian languages. We report that *The Ethnologue* classification provides a better genetic proxy than Greenberg's classification at the stock and at the group levels. Although high predictive values can be achieved from *The Ethnologue* classification, we nevertheless emphasize that Choco, Chibchan and Tupi linguistic families do not exhibit a univocal correspondence with genetic clusters.

Conclusions/Significance

The Bayesian latent class regression model described here is efficient at predicting population genetic structure using geographic and linguistic information in Native American populations.

Introduction

Comparing genetic and linguistic data provides information about various aspects of American prehistory, the process by which the Americas were originally colonized (Greenberg *et al.* 1986) or migration across linguistic barriers (Hunley and Long 2005). In addition to anthropological applications, evaluating the relationships between genes and languages has potential biomedical applications since language could be used as a *proxy* for genetic ancestry in various epidemiological contexts (Bamshad *et al.* 2003; Tishkoff and Kidd 2004).

Previous analyses comparing genetic to linguistic differentiation in the Americas yielded equivocal results. Cavalli-Sforza *et al.* (1994) reported that, prior to the publication of their book, three of seven studies supported congruence between genes and languages (Spuhler 1972; Spielman *et al.* 1974; Chakraborty *et al.* 1976; Murillo *et al.* 1977; Salzano *et al.* 1977; Spuhler 1979; Barrantes *et al.* 1990). At that time, Ward *et al.* (1993) found that rates of linguistic diversification are faster that rates of genetic differentiation in mtDNA, and concluded that there is little congruence between linguistic and genetic relationships in the Americas. In more recent studies also using mtDNA, the hypothesis that language classifications reflect the genetic structure of Native American populations was also rejected (Hunley and Long 2005; Hunley *et al.* 2007). Lastly, an analysis of autosomal microsatellite markers in 28 Native American populations from the Human Genome Diversity Panel (HGDP) provided a qualitative correspondence between linguistic and genetic groupings (Wang *et al.* 2007). However, tests of correlation were not significant for these data.

To investigate the relationships between genes and languages, the previous studies made use of tree-based or distance-based methods. Hunley and Long (2005) and Hunley *et al.* (2007) applied a test of treeness developed by Cavalli-Sforza and Piazza (Cavalli-Sforza and Piazza 1975) to decide if a matrix of genetic distances is compatible with a language tree. These authors dealt with various hierarchical classifications of American languages, and they found that none of them were consistent with the mitochondrial genetic distances. Adopting another approach, Cavalli-Sforza *et al.* (1992) found a high degree of association between linguistic and genetic trees using a consistency index. Alternatively, the association of genes and languages can be assessed by Mantel tests (Mantel 1967). Mantel tests are used to reject the absence of correlation between a matrix of genetic distances and a matrix of linguistic distances, and do not require reconstructing population trees. Since a spurious association between genetic and linguistic distances may be detected when geography is not accounted for, more elaborate procedures called partial Mantel tests can be applied in order to control for geography (Smouse *et al.* 1986). Partial Mantel tests were applied to the HGDP and did not provide strong evidence of association in Native American populations (Wang *et al.* 2007).

By definition, the results obtained from tree-based and distance-based methods are influenced by specific choices of tree reconstruction methods or particular genetic and linguistic distances. The validity of population trees depends on the reliability of their reconstruction method and on the hypothesis that genetic differentiation results from population fission. Whereas trees are well-suited for describing evolutionary relationships of non-recombining sequences like mtDNA, they may be sensitive to distortion due to gene flow between populations when nuclear data are analyzed (Ayub *et al.* 2003). In addition, we still lack an evolutionary tree for languages as linguists have not yet reached a clear consensus on their classification (Campbell 2006), and even questioned the validity of branching trees as an adequate representation of linguistic patterns of divergence (Heggarty *et al.* 2010). Finally, there are several pairwise measures of population differentiation or of linguistic divergence, and the choice of a specific measure can have a significant impact on Mantel tests (Belle and Barbujani 2007). Linguistic distances can, for instance, be based on a hierarchical linguistic classification (Excoffier *et al.* 1991), or they can be directly derived from structural linguistic features such as aspects of sound systems and grammar (Hunley *et al.* 2008; Colonna *et al.* 2010).

In this study, we introduce a novel method for investigating the relationships between genes and languages that avoids genetic and linguistic distances as well as tree reconstruction methods. We consider Bayesian *latent class regression* models (Bandeen-Roche *et al.* 1997) where we regress the unobserved genetic structure on linguistic and geographic variables. The principle of the method is to group individuals into genetic clusters at the same time as their latent cluster labels are regressed. To evaluate the predictive capacity of different sets of linguistic and geographic covariates, we also propose procedures of variable selection. Using this approach, the following questions are addressed. To what extent can geographic or linguistic origin explain individual membership to genetic clusters? Do languages contribute to a better prediction of cluster membership than geography alone? Among the classifications of Native American languages that have been proposed by linguists (Greenberg 1987; Gordon 2005), which one is the best predictor of population genetic structure? Although some of these questions have received considerable attention in the context of evolutionary trees or evolutionary distance comparisons (Cavalli-Sforza *et al.* 1988; Hunley and Long 2005; Belle and Barbujani 2007), examining their answers from a latent class individual-based model is new and potentially highly informative.

Methods

Several Bayesian model-based approaches have been proposed to assign individuals to genetic clusters (Pritchard *et al.* 2000a; Dawson and Belkhir 2001; Corander *et al.* 2003). To assess the effects of geographic and linguistic covariates on the assignment

of individuals to genetic clusters, we considered a Bayesian latent class regression model (DeSarbo and Cron 1988; Bandeen-Roche *et al.* 1997; Chung *et al.* 2006). This new model incorporates a hidden regression model within the framework proposed by Pritchard *et al.* (2000a) and implemented in the computer program `STRUCTURE`.

Bayesian model

Consider a genotypic data set, X, for a sample of n diploid individuals genotyped at L loci, and assume that there are K clusters, each of which is characterized by a set of allele frequencies at each locus. Let $Z = (Z_1, \ldots, Z_n)$ be the vector of cluster labels of each individual in the sample, and let P be the set of allele frequencies. In addition, assume that a set of covariates is measured for each individual, and stored in a design matrix, $\tilde{X}$. The covariates represent the geographic and linguistic information that is available to build predictors of the population genetic structure that is encoded in vector Z. Regarding geography, predictors can be defined as linear or quadratic trend surfaces as proposed by Durand *et al.* (2009b). Linear trend surfaces include two covariates, latitude and longitude, while quadratic surfaces also include squared and cross-product terms. Languages are coded as factors defined as binary dummy variables in the design matrix (Suits 1957). The factor levels will be dependent on the choice of the linguistic classifications considered further in this study. Remark that in regression models using factors, a linear constraint (or contrast) must be defined for identifiability reasons. In our study, we assumed that the sum of effects is null.

For algorithmic reasons, the latent regression model was implemented through a hidden *multinomial probit model* (Albert and Chib 1993). In the multinomial probit model, there are $K-1$ regression equations

$$W_{i,k} = \tilde{X}_i \beta_k + \epsilon_{i,k}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K-1, \tag{3.1}$$

$$\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,K-1}) \sim \mathcal{N}(0, \mathrm{Id}),$$

each corresponding to a genetic cluster. The $(W_{i,k})$ are "augmented" continuous variables defined for each individual and each cluster, β_k is a column vector of regression coefficients, and Id denotes the identity matrix. For each individual i, a cluster label Z_i can be obtained from the augmented variables as follows

$$Z_i = \begin{cases} K & \text{if } \max_\ell W_{i,\ell} < 0 \\ k & \text{if } \max_\ell W_{i,\ell} > 0 \text{ and } \max_\ell W_{i,\ell} = W_{i,k}\,. \end{cases} \tag{3.2}$$

In the multinomial probit model the role of the clusters is not symmetric. The estimates of the regression coefficients are defined with respect to the K^{th} cluster, called the *reference cluster*.

Given the above latent class model framework, we used a Markov Chain Monte Carlo (MCMC) algorithm based on Gibbs sampling to compute the joint posterior distribution on individual cluster labels, regression coefficients and allele frequencies

$$\Pr(Z, \beta, P|X) \propto \Pr(X|Z, P)\Pr(Z|\beta)\Pr(\beta)\Pr(P)\,.$$

In this equation, the likelihood $\Pr(X|Z, P)$ and the prior distribution on allele frequencies $\Pr(P)$ are computed in the same way as in the model without admixture of the program `STRUCTURE` (equations (2) and (4) in (Pritchard *et al.* 2000a)). The distribution $\Pr(\beta)$ is a noninformative prior distribution (see Appendix S1), and $\Pr(Z|\beta)$ corresponds to the distribution of cluster labels obtained from the multinomial probit model. The algorithm was implemented in the software `POPS`, and is described in more details in Appendix S1.

For each subset of covariates, we additionally computed a matrix of posterior predictive membership probabilities using a Monte Carlo method. To perform the computations, we simulated cluster labels from the generative model described in equation (3.1) and (3.2) where the regression coefficients are sampled from their posterior distribution. To display predicted and inferred membership probabilities graphically, we used barplot representations. In these graphics, each individual is represented by K aligned colored segments, and the segment lengths are proportional to their estimated or predicted membership probabilities.

Variable selection

To investigate whether a particular subset of covariates is a suitable proxy for genetic assignment, we used two distinct measures. Both measures are based on the posterior of regression coefficients and cluster labels. The first measure is a Pearson correlation coefficient, ρ. For a given subset of covariates, the ability of the model to predict genetic structure was evaluated by computing the correlation between the matrix of predicted membership coefficients and the matrix of estimated membership coefficients. The second measure is based on cross-validation, a technique used in the field of machine learning (Ripley 1996; Hastie *et al.* 2009) and for latent class models (Smyth 2000). In our analyses, a 2-fold cross-validation was implemented. More specifically, we divided the genotypic data set, X, into two non-overlapping data sets containing complementary subsets of loci. We considered one of these data sets as the training set, X^{training}, and the other one as the validation set, $X^{\text{validation}}$. The rationale of the cross-validation approach is that the demographic processes that shaped population genetic structure have affected all loci across the genome. Thus the training and validation sets are exchangeable, as they provide the same amount of information about population structure. We performed 500 runs of the Gibbs sampling algorithm using the training set, and retained the 50 runs having reached the highest likelihood values. For each of the retained runs, a predictive

score was computed by averaging the log-probability of the validation set over the posterior distribution given the training set

$$\text{Predictive Score} = E\left[\log(\Pr(X^{\text{validation}}|Z)) \mid X^{\text{training}}\right].$$

The computation of predictive scores is detailed in Appendix S2. Another series of 50 scores was computed after exchanging the role of the validation and training sets, and a cross-validation score was obtained by averaging the resulting $2 \times 50 = 100$ predictive scores.

Simulated data

We ran a first series of simulations using the generating model of the program `POPS`. Assuming three clusters, cluster labels of 300 individuals were simulated using the following regression equations

$$W_{i,1} = 1 + 3\tilde{X}_i^{\text{Lat}} + \epsilon_{i,1} \quad (3.3)$$
$$W_{i,2} = -4 + 12\tilde{X}_i^{\text{Lat}} + \epsilon_{i,2} \quad (3.4)$$

where $\epsilon_{i,k}$ is a standard Gaussian noise. The interpretation of the above linear trend model is that latitude is the only variable that influences individual cluster labels. Biallelic genotypes were simulated at $L = 20$, 40, 100 loci. Allele frequencies were dependent on the population of origin, and were equal to 30% and 70%, $70\% - 30\%$ and $50\% - 50\%$ in each population respectively. We implemented four hidden regression models: one model without covariates, one with latitude, one with longitude and one with both covariates.

In the second series of simulations, we extended the model by including a factor with five levels representing five languages. The hidden regression equations were defined as

$$W_{i,1} = 1 - 0.2\tilde{X}_i^{\text{Lat}} + 0.5L_i^1 + 1L_i^2 - 1.5L_i^3 - 2L_i^4 + 2L_i^5 + \epsilon_{i,1} \quad (3.5)$$
$$W_{i,2} = -3 + 9\tilde{X}_i^{\text{Lat}} + 6L_i^1 - 1.5L_i^2 + 3L_i^3 - 1.5L_i^4 - 6L_i^5 + \epsilon_{i,2} \quad (3.6)$$

where L_i^k is equal to 1 if individual i speaks the language k and is 0 otherwise. When running `POPS` to predict population genetic structure, we considered three linguistic classifications. The first classification contained five languages corresponding to the indicator variables used in the simulation. The second classification contained seven languages obtained after splitting the second and the third languages of the first classification into two sublanguages. The last classification contained three languages because we merged two pairs of unrelated languages from the first classification.

In the third series of experiments, we studied two previously published data sets simulated from a five-island model (Chen *et al.* 2007). The simulated data represented one

population structured into five subpopulations differentiated at F_{ST} levels equal to 0.03 and 0.04. Five hundred individuals (100 per subpopulation) were simulated using allele frequency distributions across 10 codominant unlinked loci. Spatial coordinates were simulated using Gaussian distributions. The subpopulations were adjacent to each other and arranged on a ring. We ran `POPS` using the spatial coordinates of each individual as covariates. In addition, we introduced a spurious noisy covariate independent on the subpopulation of origin. We considered the models defined by all the possible inclusions of those three covariates ($2^3 = 8$ models). These data enabled us to compare the performances of `POPS` to other programs using spatial covariates (Chen *et al.* 2007; Corander *et al.* 2008; François and Durand 2010a).

Native American data

We applied `POPS` to 512 Native American individuals from the Human Genome Diversity Panel (HGDP) data set (Wang *et al.* 2007). Individuals from 28 populations were genotyped at 678 microsatellite loci. Fourteen Siberian individuals from the Tundra Nentsi population were also included in the study. In the regression models we considered three linguistic classifications. The first and second linguistic classifications corresponded to Greenberg's classification at the stock level and at the group level (Greenberg 1987; Ruhlen 1991). The third linguistic classification was given by the website *The Ethnologue* (`www.ethnologue.com`) (Campbell 1997; Gordon 2005). The three linguistic classifications were encoded with factors having 8, 14 and 16 levels respectively (see Supporting Information Table 3.1). To account for geography, all models included quadratic trend surfaces. The combinations of geographic and linguistic variables resulted in the following four latent cluster regression models. Model A included geographic information only. Models B-D included geographic and linguistic information: Model B used Greenberg's classification at the stock level (8 levels), Model C used Greenberg's classification at the group level (14 levels), and Model D used *The Ethnologue* classification at the family level (16 levels).

MCMC parameters

For the simulated data, the runs of `POPS` used 2,000 sweeps with an initial burn-in period of 1,000 sweeps. For the human data, the runs used 5,000 sweeps with an initial burn-in period of 2,500 sweeps. These values ensured that the likelihoods stabilized around their stationary values. For the HGDP data and for each model, we ran a total of 500 MCMC runs. We retained the 50 runs with the largest likelihood values, and we averaged the resulting estimated and predicted membership coefficients using the computer program CLUMPP (Jakobsson and Rosenberg 2007).

The number of clusters was set to $K = 9$ (Wang *et al.* 2007). Among these nine

clusters, there were eight Native American clusters plus the reference cluster. For Native American population samples, we chose the Siberian population (Tundra-Nentsi) to represent the reference group. Individuals in the reference cluster were not allowed to switch to other clusters during the MCMC runs.

Results

Simulation results

Using simulated data sets, we investigated whether including geographic and linguistic covariates can improve the estimation of membership probabilities or not, and we evaluated which subsets of variables best predict the estimated population genetic structure.

For the simulations where latitude was influential (equations (3.3) and (3.4)), we found that the true values of the regression coefficients were close to the mode of the posterior distributions (Figure 3.3). The influence of each covariate was thus correctly ascertained by `POPS` when the data were generated under its underlying statistical model. To further evaluate if missing the true set of covariates modifies the inference and the prediction of membership coefficients, we evaluated the performances of `POPS` using various hidden regression models. For all models, the misclassification rates were less than 4%. The upper bound was obtained under a model without covariates and for the smallest number of loci ($L = 20$, Figure 3.4A). The misclassification rates never increased when we included a spurious longitude variable. With $L = 20$ loci, the misclassification rate decreased to 2% when the correct covariate (latitude) was used. With $L = 40$ loci, the misclassification rates were less than 1% for all hidden models. All individuals were perfectly assigned to their population of origin when latitude was included. For $L = 100$, the misclassification rate was equal to 0% for all models. In the second series of simulations, linguistic covariates were added to the generating model (equations (3.5) and (3.6)). The misclassification rates were less than 30%, a value obtained for $L = 20$ loci in a model without covariates (Figure 3.4B). With $L = 20$, the misclassification rate decreased to 5% when including latitude and a linguistic variable with five levels. With $L = 100$ loci, the misclassification rate of the model without covariates was around 1%. We conclude that when the data are generated from a hidden regression model, including covariates in `POPS` increases the performances of the program. This is particularly true when the number of loci is relatively small.

Finally, we studied the variable selection criteria for the data where latitude was influential (equations (3.3) and (3.4)) as well as linguistic covariates (equations (3.5) and (3.6)). Whatever the number of loci we considered, the increase of the correlation coefficient was larger when including latitude rather than longitude in the regression model. Figure 3.5 shows that the correlation coefficient and the cross-validation score reach a plateau when the true predictors are included in the hidden regression model.

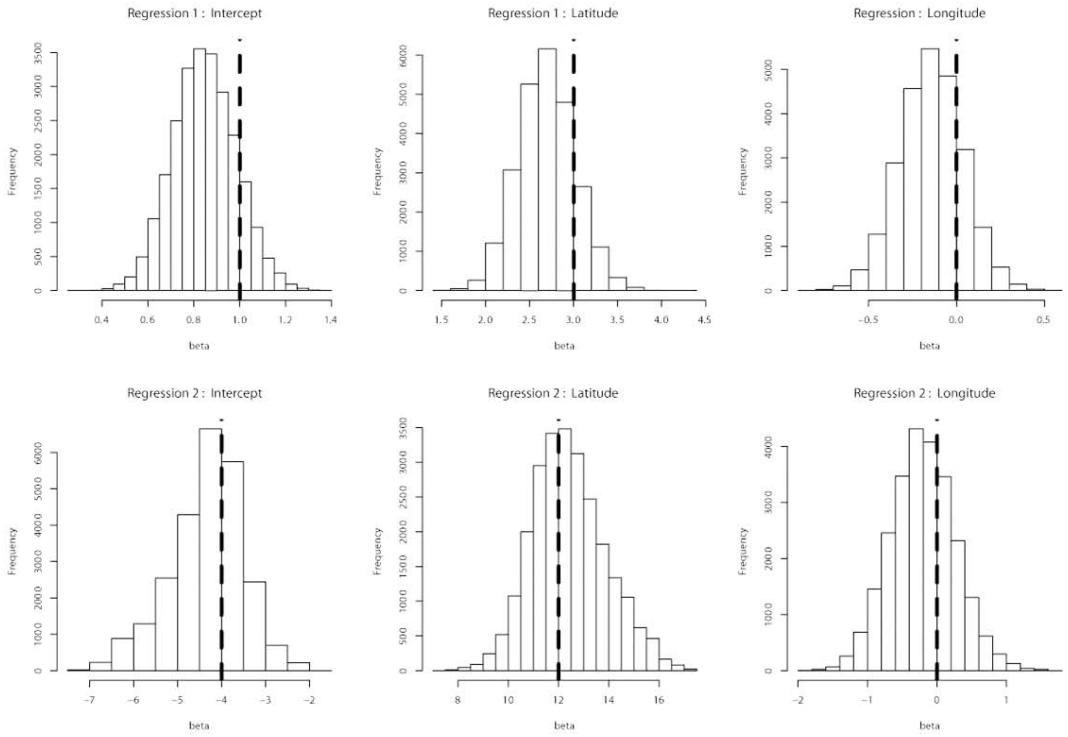


Figure 3.3: **Posterior distributions of the regression coefficients for a data set simulated with the hidden regression model ($K = 3$).** The dashed vertical lines correspond to the regression coefficients used for generating the data. Two spatial covariates (latitude and longitude) are included in the regression model but only the first one influences genetic structure.

This plateau was found when latitude was the sole determinant of genetic structure and when linguistic covariates had an additional contribution to genetic differentiation.

For the five-island data with a level of differentiation of $F_{ST} = 0.04$, the misclassification rates were less than 5% (Figure 3.4C). The worst performances were obtained for a model without covariates. When latitude (or longitude) was included in the hidden regression model, the misclassification rate decreased to 3%. When both latitude and longitude were included in the model, the misclassification rate decreased to 1%. The addition of a spurious noisy covariate did not impact the performance of the program. Regarding variable selection, Figure 3.5C shows that the correlation coefficients and the validation scores reach a plateau when longitude and latitude are included in the hidden regression model. For the five-island data with a level of differentiation of $F_{ST} = 0.03$, a model including latitude and longitude was also selected. In this case, the misclassification rate was equal to 2.8%. For these data, `POPS` compared favorably to the spatial versions of `BAPS` (misclassification rate = 3.9%) and `TESS` (misclassification rate = 4.4%) (Chen *et al.* 2007; Corander *et al.* 2008; François and Durand 2010a).

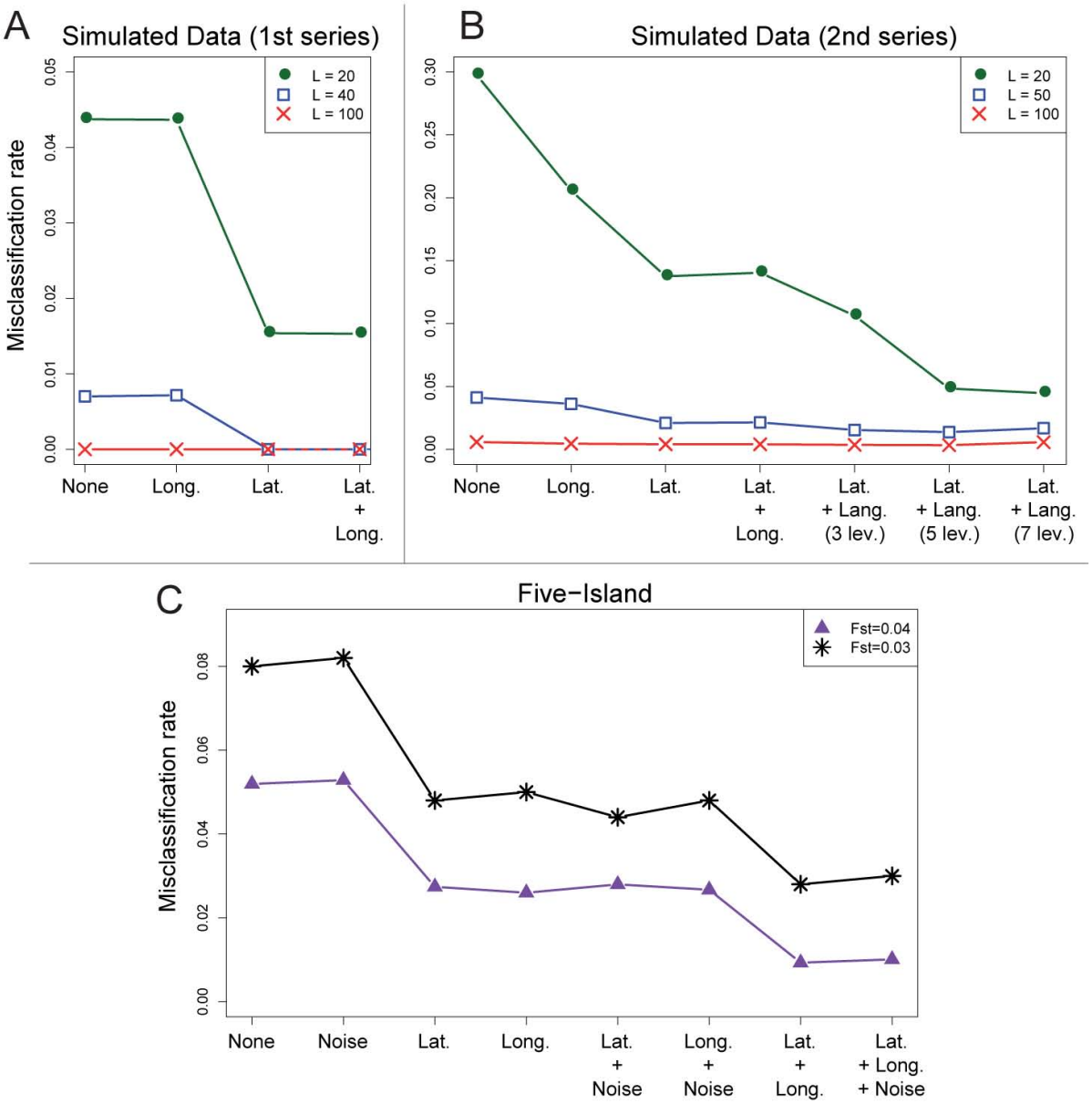


Figure 3.4: **Misclassification rates for simulated data as a function of the covariates included in the clustering algorithm.** A. The cluster memberships are influenced by latitude but not by longitude. B. The data are generated using latitude and a 5-level linguistic classification. C. The data are generated in a five-island model for which $F_{\mathrm{ST}} = 0.03$ or 0.04.

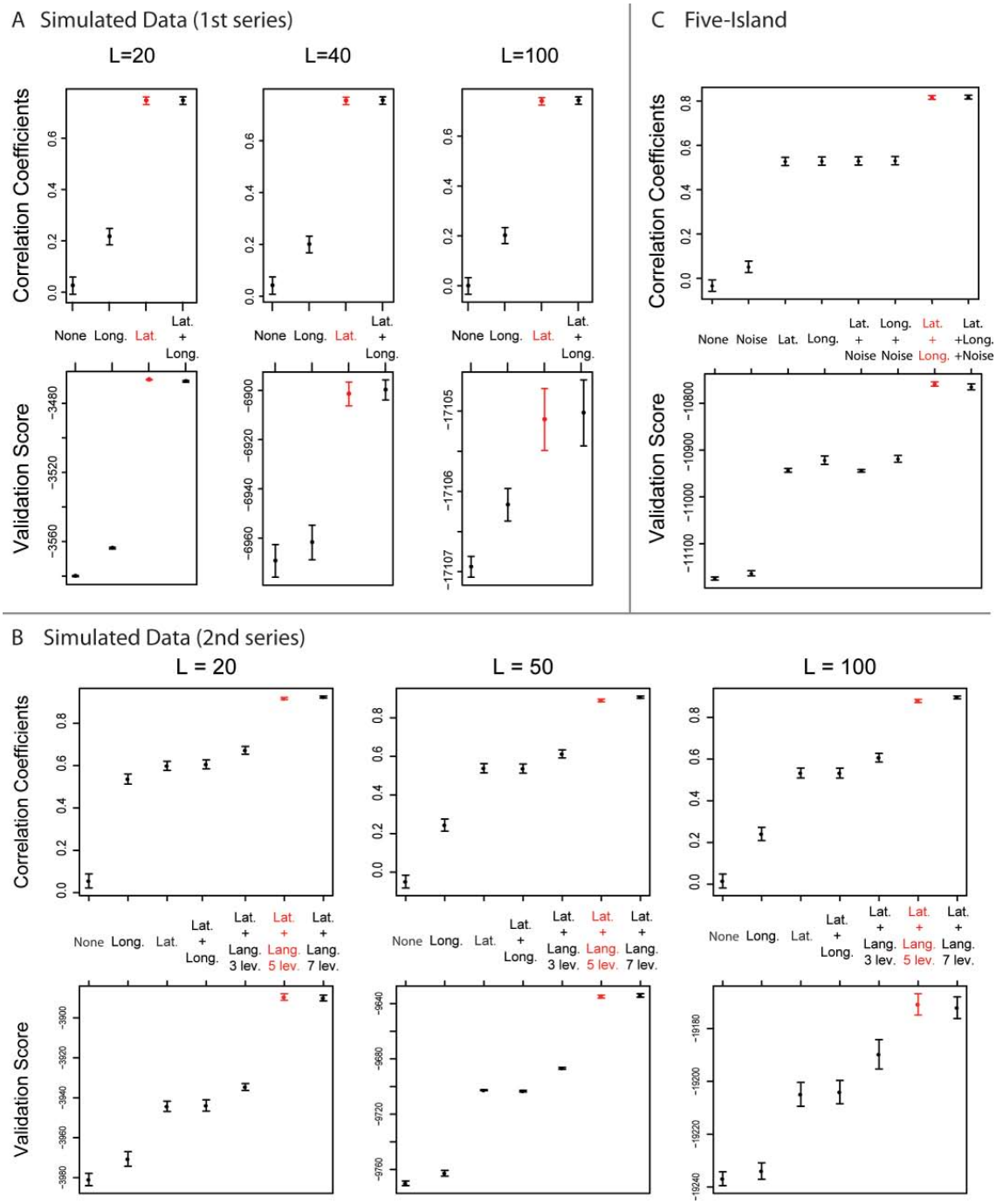


Figure 3.5: **Variable selection for simulated data.** The correlation coefficients ρ correspond to the correlations between the estimated and predicted membership probabilities. Confidence intervals of the correlation coefficients are estimated by assuming that the Fisher's transform $\operatorname{arctanh}(\rho)$ follows a Gaussian distribution (Fisher 1915). The validation scores are estimated with the 2-fold cross-validation method. Their standard deviations are estimated by using a non-parametric bootstrap method. A. The cluster memberships are influenced by latitude but not by longitude. B. The data are generated using latitude and a 5-level linguistic classification. C. The data are generated in a five-island model for which $F_{\mathrm{ST}} = 0.04$.

Native American HGDP data

To investigate the relationships between geography, languages and genes in Native American populations, we applied POPS to a multilocus genotype data set including 512 individuals from the HGDP. We compared the posterior membership coefficients predicted by four different models that use distinct linguistic classifications and we computed two variable selection criteria in order to discriminate among models (see Material and Methods).

The four clustering models resulted in highly similar patterns of estimated membership coefficients, and these patterns were also similar to the pattern found with STRUCTURE (Figure 3.6, Supporting Information Figure 3.9, Wang *et al.* (2007)). As we used a large number of microsatellite loci, these results are not surprising, and they warrant that the predictive power of the three linguistic classifications will be ascertained consistently.

Using a quadratic trend surface to correct for geographic effects, we compared the predictions of a model without languages (Model A) to the predictions of a model using Greenberg's classification at the stock level (Model B), a model using Greenberg's classification at the group level (Model C), and a model using *The Ethnologue* classification (Model D). Figure 3.6A compares the predictions of Model A and Model D. For many population samples, the membership probabilities predicted by Model A were close to the estimated coefficients ($\rho = 81\%$, Figure 3.7A). The predictions of Model A for every geographic location in the American mainland are displayed in Figure 3.8. The value of the correlation coefficient and the map of predicted membership coefficients confirmed that geography is a good predictor of genetic structure in Native American populations. When including linguistic covariates (Models B-D), the predictions of cluster membership were closer to the estimates of the MCMC algorithm than those obtained without languages (Model A) except for the Pima. The correlation coefficient increased from $\rho = 0.81$ to $\rho = 0.94 - 0.98$ (Figure 3.7A), and the predicted genetic structure changed substantially (Figure 3.6A and Supporting Information Figure 3.9). For several populations the predictions obtained from linguistic covariates (Models B-D) differed from the predictions obtained with the geographic covariates only: Model A predicted that the Kaqchikel and the Wayuu samples shared substantial ancestry with a group comprising Cabecar, Guaymi, Kogi, Arhuaco, Waunana and Embera populations; Model A also predicted that the Kaingang and Guarani samples clustered with the Ache population, and that the Inga and Piapoco samples were grouped with the Ticuna sample.

Figure 3.6 B-D displays the membership coefficients predicted by POPS using Greenberg's and *The Ethnologue* classifications (Models B-D), grouping populations with the same linguistic taxon. At the exception of the Andean and Ge-Pano-Carib stocks, Greenberg's linguistic stocks were associated with multiple clusters (Figure 3.6B). Refining Greenberg's classification at the group level improved the characterization of genetic clus-

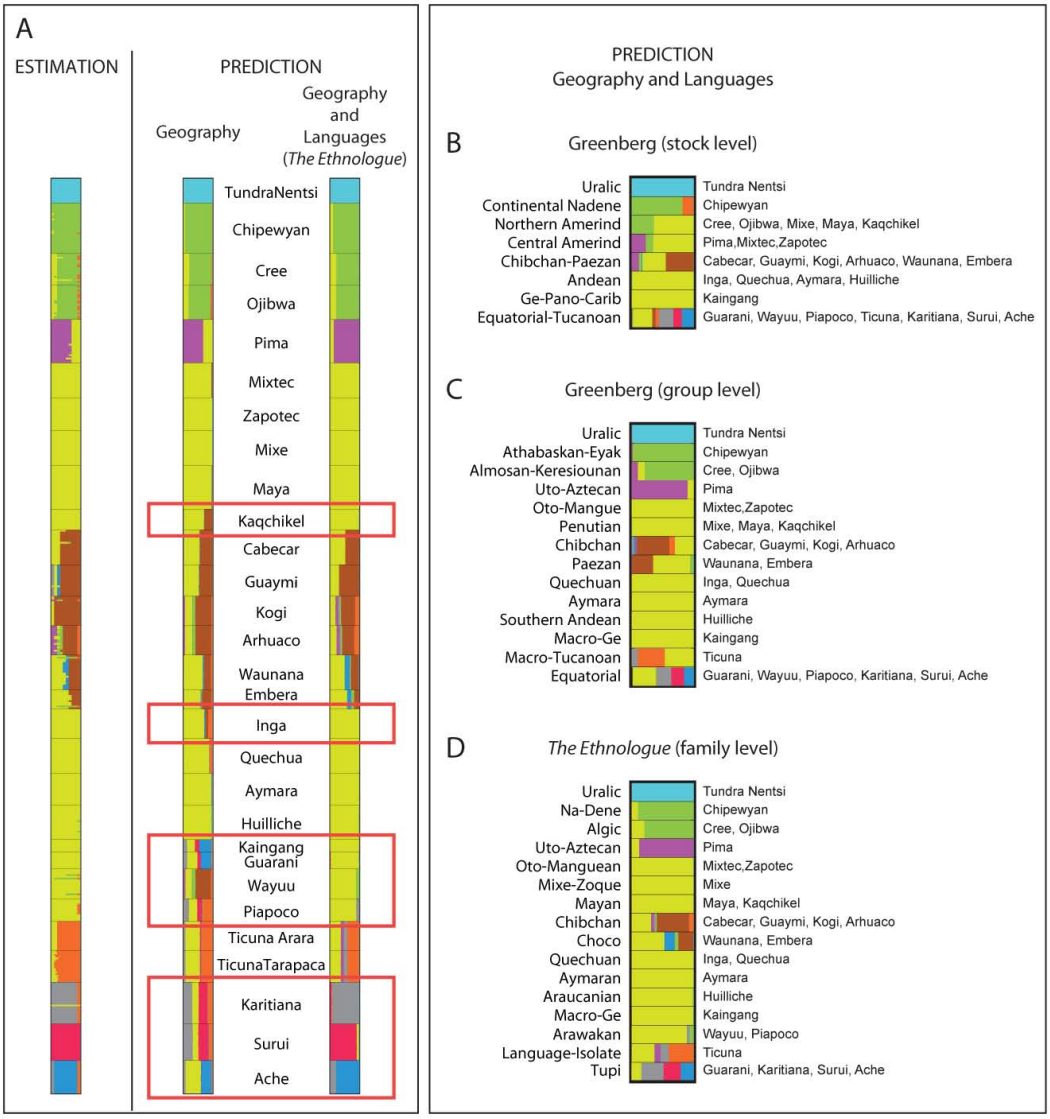


Figure 3.6: **Estimated and predicted population genetic structure for 28 Native American populations.** A. The membership coefficients are estimated in a model that includes spatial information (longitude, latitude). Inference of genetic structure is unchanged when we include additional linguistic covariates (Supporting Information Figure 3.9). The main differences between predictions obtained with or without linguistic information are framed in red. B-D. Membership coefficients predicted by Models B-D. The membership coefficients are averaged over individuals within the same linguistic unit.

ters by linguistic taxa (Model C, Figure 3.6C). At the group level, the Northern Amerind stock split into Almosan-Keresiouan and Penutian groups that correspond to genetically divergent clusters. Similarly, the Central Amerind stock split into Uto-Aztecan and Oto-Mangue groups which are also genetically divergent. However, the split of the Equatorial-Tucanoan stock into the Macro-Tucanoan and Equatorial groups, and the split of the Chibchan-Paezan stock into the Chibchan and Paezan groups, did not improve the prediction of genetic clusters. In *The Ethnologue* classification (Model D), the Equatorial

group split into the Arawakan and Tupi families. This separation improved the prediction of genetic clusters since the Arawakan family was associated with a unique genetic cluster. In contrast, the separation of the Penutian group into the Mixe-Zoque and Mayan families did not improve the characterization of genetic groups. Overall *The Ethnologue* classification provided better predictions of genetic groups than Greenberg's classification. Among the 16 families of *The Ethnologue* classification, only the Tupi, Choco and Chibchan families were not associated to a unique genetic cluster (Figure 3.6D). Supporting these comparisons, Figure 3.7B shows that the cross-validation score increases when using *The Ethnologue* (Model D). The values of the cross-validation scores are approximately equal to $-485,100$ for Models B and C, and around $-484,750$ for Model D. These scores provide quantitative evidence that the classification of *The Ethnologue* leads to better predictions of genetic structure than Greenberg's classification at the stock or group levels.

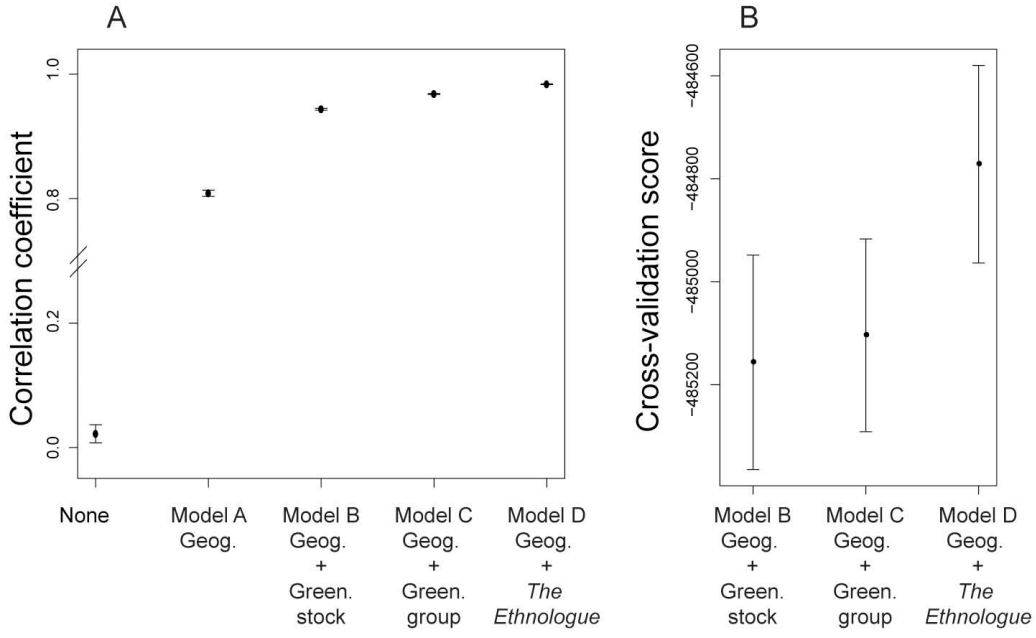


Figure 3.7: **Variable selection for the Native American HGDP data.** Geographic information includes longitude and latitude. Green. stands for Greenberg and Geog. stands for geography. The best model uses *The Ethnologue* linguistic classification.

Discussion

We proposed a Bayesian latent class regression model to investigate to which extent geographic and linguistic information can predict population genetic structure in Native American populations. The originality of this approach was to model individual responses, i.e., the unobserved genetic cluster labels for each individual, using spatial and linguistic

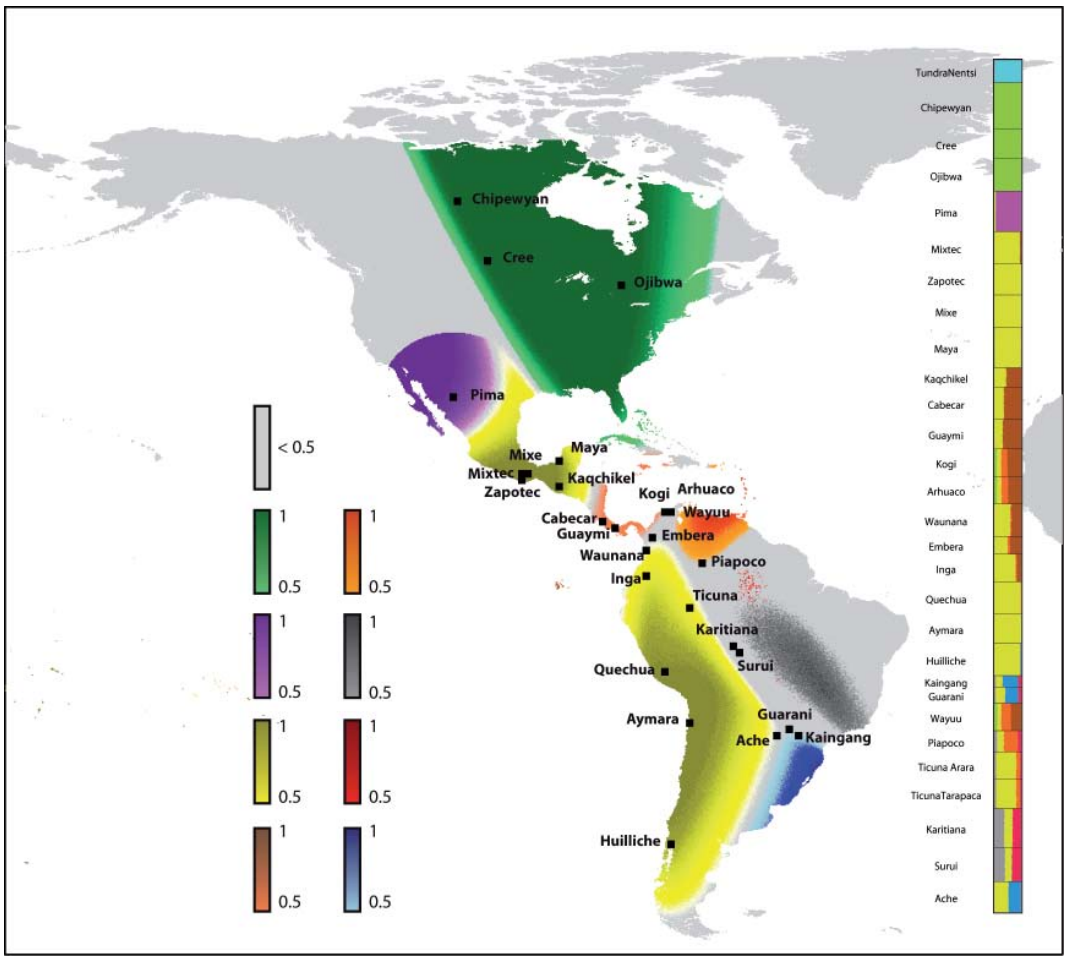


Figure 3.8: **Genetic structure of Native American populations as predicted by geographical covariates.** Geographical covariates include latitude, longitude, quadratic terms and an interaction term. Locations for which there is a cluster with a predicted membership coefficient larger than 0.5 are colored with the cluster color. Locations for which there is no cluster that reaches the 0.5 threshold or that are too distant from a sampled population are colored in grey. The barplot displays the membership probabilities as predicted by geographical covariates.

variables.

Our simulation study provided evidence that a hidden regression layer can improve the inference of genetic structure in addition to allowing their predictions from covariates. We also tested two criteria of variable selection based on correlation coefficients and cross-validation scores and found that these statistical indices reached a plateau when the true set of covariates was included in the `POPS` model. With small numbers of loci, the use of covariates decreased the misclassification rates of the clustering program significantly. For large numbers of loci, the estimation performances were hard to improve, especially when the likelihood dominated the prior distribution. However, using large numbers of loci made predictions and the use of the variable selection criteria reliable.

Using 678 microsatellite markers from the HGDP data set, we evaluated the suitability of geographic and linguistic predictors for Native American population genetic structure. Geography predicted genetic clusters rather accurately. However considering linguistic origin in addition to geographic origin improved the prediction of genetic structure. After correcting for geographic effects, we evaluated the predictive capabilities of three linguistic classifications: Greenberg's classification at two distinct levels and *The Ethnologue* classification. We did not consider Greenberg's tripartite classification (Amerind, Na-Dene, and Eskimo-Aleut) because, in addition to being controversial (Lewin 1988), all Native American HGDP populations, except the Chipewyan, belong to the Amerind family. We rather focused our analysis on taxonomically lower levels of Greenberg's classification: linguistic stocks and groups. Considering those refined levels, *The Ethnologue* provided better predictions of population genetic structure than Greenberg's classification.

Though *The Ethnologue* classification provided a better genetic proxy than Greenberg's classification, some linguistic families were not perfectly characterized in terms of genetic clustering. The Chibchan and Choco families were grouped in a Chibchan-Paezan stock by Greenberg (Greenberg 1987). These populations shared genetic ancestry with northern Mesoamerican populations (Mixtec, Zapotec, Mixe, Maya and Kaqchikel) and with southern Andean populations (Inga, Quechua, Aymara and Huilliche) (Figure 3.6A). Based on mtDNA data, Melton *et al.* (2007) also found genetic relationships between Chibchan speakers and a Mayan population from Mesoamerica. To explain these relationships, it has been argued that Chibchan and Mesoamerican languages were all interrelated at one time into a larger Proto-Mesoamerican linguistic group that subsequently splintered into different language families after the intensification of agriculture in Mesoamerica (Witkowski and Brown 1981; Bellwood 2005). The shared genetic relationships between Mesoamerican populations and Chibchan-Choco populations would result from their shared common history. Another family lacking genetic characterization was the Tupi. The Tupi family encompasses approximately 41 languages that spread throughout eastern South America several millennia ago (Noelli 1998, 2008). Since the Tupi expansion involved language replacement, it may have blurred the relationships between genes

and languages. Additionally, the Surui and Ache are populations with Tupi languages and small effective population sizes (Wang *et al.* 2007). The 'genetic patchwork' of the Tupi would then result from genetic drift essentially.

Despite the intrinsic difference between methods, our analysis confirmed previous findings that a sizeable correspondence between genetic and linguistic differentiation may exist only below a certain level of linguistic differentiation. The tests of treeness indicated that language classifications provide the best fit to mitochondrial data when they included external features of language classification trees and no deeper internal relationships between languages (Hunley *et al.* 2007). Using partial Mantel tests, Wang *et al.* (2007) found a low partial correlation ($r = 0.01$) between linguistic (Greenberg's stock level) and genetic dissimilarities, but the correlation increased to $r = 0.40$ when the authors considered pairs of populations within stocks. Our analysis revealed that the congruence between genetic and linguistic diversification is more evident when considering a finer grain of linguistic differentiation than the stock level.

To further investigate potential scale effects, we applied `POPS` to 77 world-wide population samples from the HGDP data set excluding two language isolates (Basque and Burushaski) and grouping the sub-Saharian samples in a reference cluster (Supporting Information Table 3.2). The genetic clusters detected by `POPS` agreed with those detected by `STRUCTURE` (Supporting Information Figure 3.10) (Rosenberg *et al.* 2002; Wang *et al.* 2007). The geographic predictions of a quadratic trend surface model were highly correlated to the estimated membership coefficients ($\rho = 0.97$). The high value of the correlation coefficient confirmed that geography is a good predictor of genetic structure at the world-wide scale (Dugoujon *et al.* 2004; Manica *et al.* 2005; Prugnolle *et al.* 2005; Ramachandran *et al.* 2005; Foll and Gaggiotti 2006; Handley *et al.* 2007; Novembre *et al.* 2008). Adding the linguistic covariates taken from *The Ethnologue* classification increased the correlation coefficient from $\rho = 0.97$ to $\rho = 0.98$. Thus it improved the prediction of genetic structure only marginally. These results provided evidence that the effects of language on the prediction of genetic structure are dependent on the scale considered. The results of `POPS` were also comparable to those obtained by Belle and Barbujani (Belle and Barbujani 2007) reporting that languages have a small effect on the pattern of molecular variation at the world-wide scale. At the global scale, the patterns of genetic population structure are likely to reflect ancient demographic events, such as population divergence associated with the colonization of major geographic regions of the world (Hunley *et al.* 2008). At the continental scale, cultural traits contribute to the mediation of gene flow between human groups (Premo and Hublin 2009). The predictive power provided by languages in the Americas could thus result from preferential mating within linguistic groups.

The examination of linguistic and genetic relationships in the Americas would obviously benefit from a more extensive sampling from the Na-Dene linguistic stock and from

the inclusion of the Eskimo-Aleut stock. In a regression framework, a large dispersion of the explanatory variables is preferable. Though the sampling design of the HGDP was not optimal in our framework, our approach provided evidence that linguistic proxies improved the prediction of Native American population genetic structure. As human genomic data expand in genetic and geographic coverage (Jakobsson *et al.* 2008; Li *et al.* 2008; Novembre *et al.* 2008), the use of latent class regression models could result in a more detailed picture of the role of geography and cultural factors in shaping human genetic variation.

Acknowledgments

The software `POPS` implementing the algorithm described in this article is available at `http://membres-timc.imag.fr/Olivier.Francois/tess.html`. We thank Eric Durand for his comments at various stages of this work. Simulations were run on the the UJF-CIMENT cluster of computers (`http://healthphy.grenoble.cnrs.fr/`).

Supporting information

Table 3.1: Coordinates and linguistic entities of 28 Native American populations from the Human Genome Diversity Panel.

Population	Latitude	Longitude	Green.[a] stock	Green.[a] group	Linguistic family (*The Ethnologue*)
Chipewyan	59.55	-107.3	Continental Na-Dene	Athabaskan-Eyak	Na-Dene
Cree	50.33	-102.5	Northern Amerind	Almosan-Keresiouan	Algic
Ojibwa	46.5	-81	Northern Amerind	Almosan-Keresiouan	Algic
Mixe	17	-96	Northern Amerind	Penutian	Mixe-Zoque
Maya	19	-91	Northern Amerind	Penutian	Mayan
Kaqchikel	15	-91	Northern Amerind	Penutian	Mayan
Pima	29	-108	Central Amerind	Uto-Aztecan	Uto-Aztecan
Mixtec	17	-97	Central Amerind	Oto-Mangue	Oto-Manguean
Zapotec	16	-97	Central Amerind	Oto-Mangue	Oto-Manguean
Cabecar	9.5	-84	Chibchan-Paezan	Chibchan	Chibchan
Guaymi	8.5	-82	Chibchan-Paezan	Chibchan	Chibchan
Kogi	11	-74	Chibchan-Paezan	Chibchan	Chibchan
Arhuaco	11	-73.8	Chibchan-Paezan	Chibchan	Chibchan
Waunana	5	-77	Chibchan-Paezan	Paezan	Choco
Embera	7	-76	Chibchan-Paezan	Paezan	Choco
Inga	1	-77	Andean	Quechua	Quechuan
Quechua	-14	-74	Andean	Quechua	Quechuan
Aymara	-22	-70	Andean	Aymara	Aymaran
Huilliche	-41	-73	Andean	Southern	Araucanian
Kaingang	-24	-52.5	Ge-Pano-Carib	Macro-Ge	Macro-Ge
Wayuu	11	-73	Equatorial-Tucanoan	Equatorial	Arawakan
Piapoco	3	-68	Equatorial-Tucanoan	Equatorial	Arawakan
Guarani	-23	-54	Equatorial-Tucanoan	Equatorial	Tupi
Karitiana	-10	-63	Equatorial-Tucanoan	Equatorial	Tupi
Surui	-11	-62	Equatorial-Tucanoan	Equatorial	Tupi
Ache	-24	-56	Equatorial-Tucanoan	Equatorial	Tupi
Ticuna Tarapaca	-4	-70	Equatorial-Tucanoan	Macro-Tucanoan	Language isolate
Ticuna Arara	-4	-70	Equatorial-Tucanoan	Macro-Tucanoan	Language isolate

[a] Green. stands for Greenberg

Table 3.2: Coordinates, distance to Addis-Abeba, and linguistic families of 77 worldwide populations from the Human Genome Diversity Panel.

Population	Latitude	Longitude	Distance to Addis-Abeba (km)	Linguistic family (*The Ethnologue*)
Bantu South East Africa	−28.40	27.6	4340	Niger-Congo
Bantu South West Africa	−21.00	18.7	4011	Niger-Congo
BantuKenya	−3.00	37.0	1354	Niger-Congo
Mandenka	12.00	−12.0	5585	Niger-Congo
Yoruba	8.00	5.0	3744	Niger-Congo
Biaka Pygmy	4.00	17.0	2495	Niger-Congo
Mbuti Pygmy	1.00	29.0	1422	Niger-Congo
San	−21.00	20.0	3934	Khoisan
Orcadian	59.00	−3.0	6189	Indo-European
Adygei	44.00	39.0	2990	North-Caucasian
Russian	61.00	40.0	4821	Indo-European
French	46.00	2.0	5524	Indo-European
Italian	46.00	10.0	4917	Indo-European
Sardinian	40.00	9.0	4896	Indo-European
Tuscan	43.00	11.0	4768	Indo-European
Mozabite	32.00	3.0	4507	Afro-Asiatic
Bedouin	31.00	35.0	2484	Afro-Asiatic
Druze	32.00	35.0	2594	Afro-Asiatic
Palestinian	32.00	35.0	2594	Afro-Asiatic
Balochi	30.50	66.5	3835	Indo-European
Brahui	30.50	66.5	3835	Dravidian
Makrani	26.00	64.0	3775	Indo-European
Sindhi	25.50	69.0	4251	Indo-European
Pathan	33.50	70.5	4152	Sino-Tibetan
Hazara	33.50	70.0	4105	Indo-European
Uygur	44.00	81.0	5105	Altaic
Kalash	36.00	71.5	4227	Indo-European
Melanesian	−6.00	155.0	14030	Austronesian
Papuan	−4.00	143.0	12792	Austronesian
Han	32.50	114.0	8112	Sino-Tibetan
Han-NChina	39.00	114.0	7875	Sino-Tibetan
Dai	21.00	100.0	7338	Austronesian
Daur	48.50	124.0	8338	Altaic
Hezhen	47.50	133.5	9040	Altaic
Lahu	22.00	100.0	7289	Sino-Tibetan
Miao	28.00	109.0	7852	Hmong-Mien
Oroqen	50.50	126.5	8444	Altaic
She	27.00	119.0	8799	Hmong-Mien
Tujia	29.00	109.0	7810	Sino-Tibetan
Tu	36.00	101.0	6856	Altaic
Xibo	43.50	81.5	5137	Altaic
Yi	28.00	103.0	7304	Sino-Tibetan
Mongola	48.50	119.0	7982	Altaic
Naxi	26.00	100.0	7110	Sino-Tibetan
Cambodian	12.00	105.0	8289	Austro-Asiatic
Japanese	38.00	138.0	9838	Japonic
Tundra Nentsi	66.08	76.5	5956	Uralic
Yakut	63.00	129.5	8258	Altaic
Chipewyan	59.55	−107.3	10251	Na-Dene
Cree	50.33	−102.5	11182	Algic
Ojibwa	46.50	−81.0	12490	Algic
Pima	29.00	−108.0	12898	Uto-Aztecan
Mixtec	7.00	−97.0	15606	Oto-Manguean
Zapotec	16.00	−97.0	14698	Oto-Manguean
Mixe	17.00	−96.0	14644	Mixe-Zoque
Maya	19.00	−91.0	14679	Mayan
Kaqchikel	5.00	−91.0	16085	Mayan
Cabecar	9.50	−84.0	15959	Chibchan
Guaymi	8.50	−82.0	16154	Chibchan
Kogi	11.00	−74.0	16277	Chibchan
Arhuaco	11.00	−73.8	16286	Chibchan
Waunana	5.00	−77.0	16742	Choco
Embera	7.00	−76.0	16587	Choco
Zenu	9.00	−75.0	16432	Choco
Inga	1.00	−77.0	17145	Quechuan
Quechua	−14.00	−74.0	18797	Quechuan
Aymara	−22.00	−70.0	19782	Aymaran
Huilliche	−41.00	−73.0	21482	Araucanian
Kaingang	−24.00	−52.5	20797	Macro-Ge
Guarani	−23.00	−54.0	20629	Tupi
Wayuu	11.00	−73.0	16323	Arawakan
Piapoco	3.00	−68.0	17363	Arawakan
Ticuna Tarapaca	−4.00	−70.0	17978	Language-Isolate
Ticuna Arara	−4.00	−70.0	17978	Language-Isolate
Karitiana	−10.00	−63.0	18910	Tupi
Surui	−11.00	−62.0	19057	Tupi
Ache	−24.00	−56.0	20636	Tupi

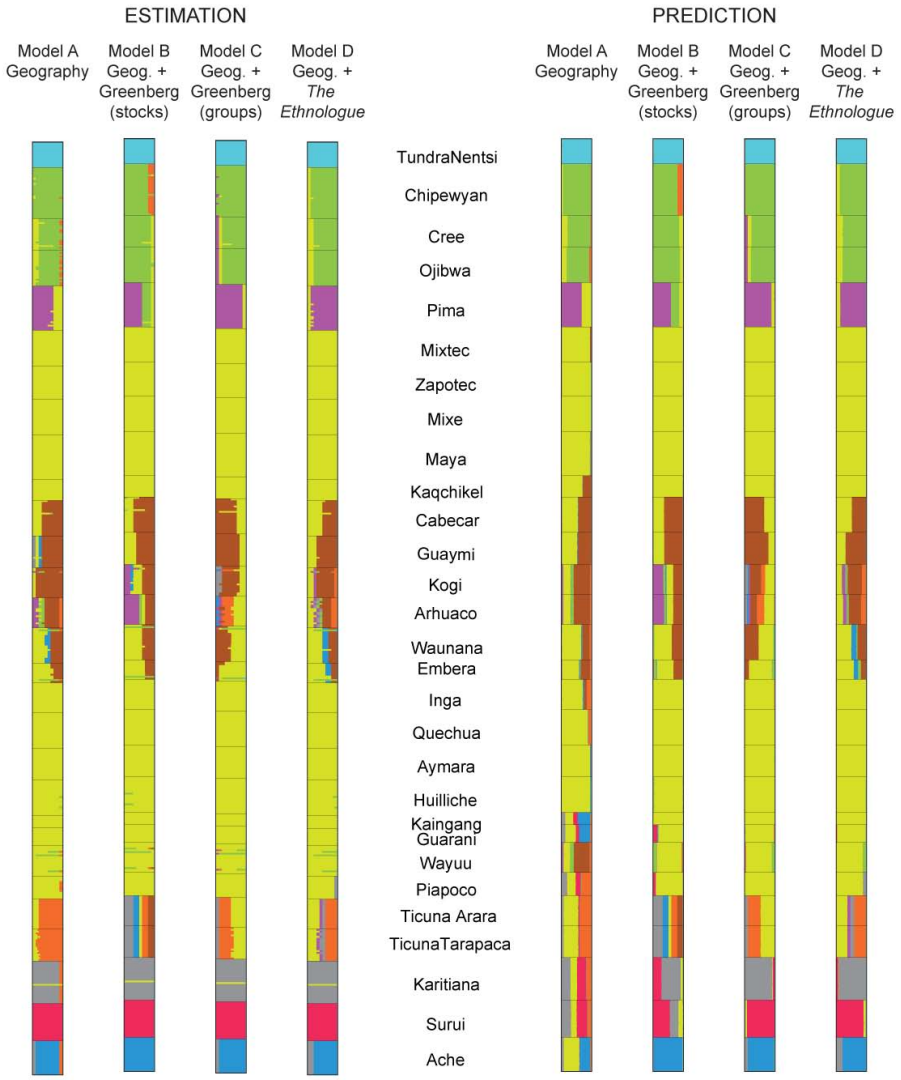


Figure 3.9: Estimated and predicted genetic structure of Native American populations, with $K = 9$ clusters, using different set of covariates in the probit model (Model A-D).

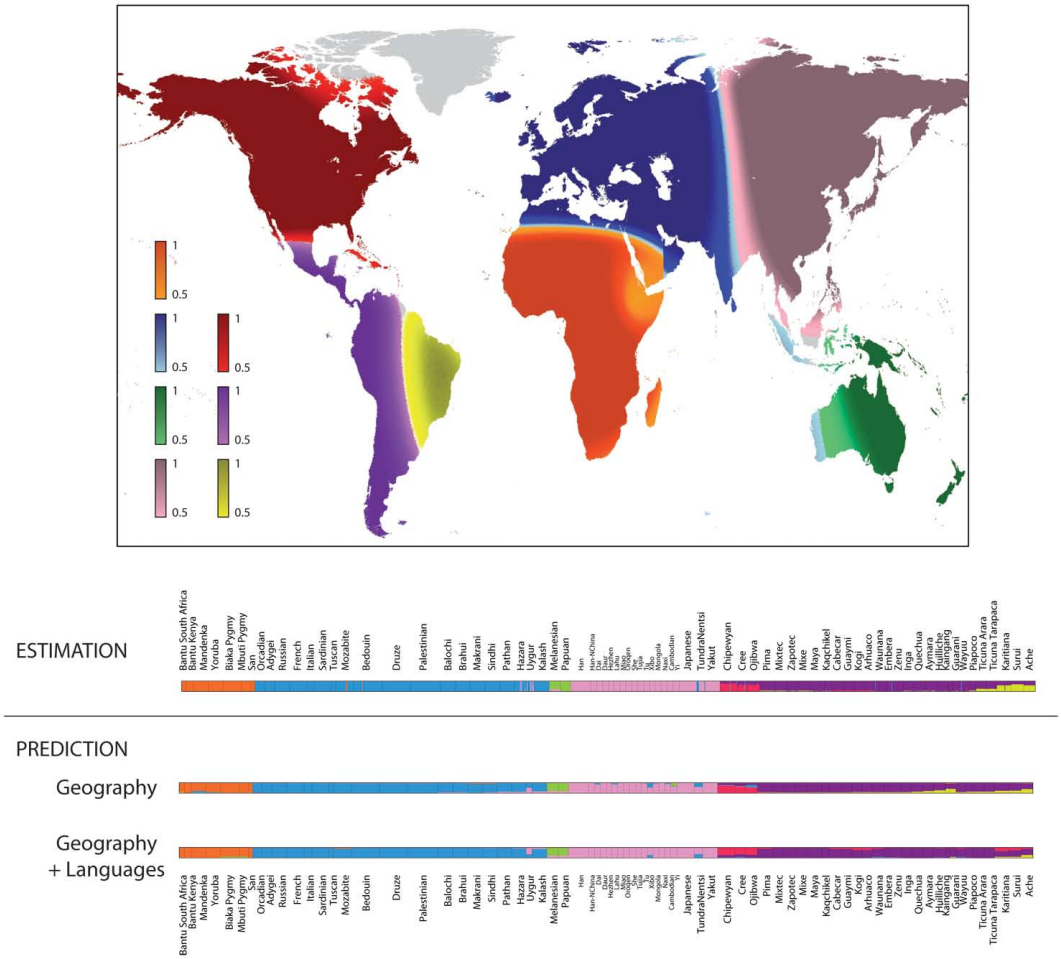


Figure 3.10: Genetic structure at a worldwide scale as predicted by geographical covariates when K = 7. Geographical covariates include latitude, longitude and distance to the Addis Abeba, which is computed by included five obligatory waypoints. The three barplots correspond to 1) the genetic structure as inferred with genetic data and both spatial and linguistic covariates, 2) the structure as predicted with spatial information and 3) the structure as predicted with spatial and linguistic information. The linguistic variable is a qualitative variable corresponding to *The Ethnologue* classification.

Appendix S1: Gibbs sampler

To sample from the posterior distribution of the cluster labels Z, the allelic frequencies P and the regression coefficients β, we implemented a Markov Chain Monte Carlo algorithm with Gibbs sampling steps.

UPDATING P. This step is the same as in the software `structure`. It is performed by simulating the set of frequencies as

$$p_{kl.}|X, Z \sim \mathcal{D}(\lambda + n_{kl1}, \ldots, \lambda + n_{klJ_l}), \tag{3.7}$$

where $p_{kl.}$ denotes the vector of allele frequencies in the cluster k at the locus l, and n_{klj} denotes the number of copies of the allele j in population k at the locus l, $k = 1, \ldots, K$, $l = 1, \ldots, L$, $j = 1, \ldots, J_l$. For our analysis, we considered $\lambda = 1$.

UPDATING (W, Z). Since Z can be obtained from W in a deterministic fashion, Z and W are updated simultaneously. Using the Bayes formula, the joint conditional distribution of (W, Z) can be written as

$$\Pr(W, Z|\beta, P, X) \propto \Pr(X|\beta, P, Z)\Pr(W|\beta)\Pr(Z|W)$$

To simulate the couples (W, Z), we use the following rejection algorithm.

- Step 1. For $i = 1, \ldots, n$, simulate the couple (W_i, Z_i) from the multinomial probit model by generating W_i from regression equation and determine $Z_i = k$ with its max-rule (see Methods, equations (1) and (2)).
- Step 2. Accept the couple (W_i, Z_i) with probability

$$\frac{\Pr(X_i = x_i|P, Z_i = k)}{\max_k \Pr(X_i = x_i|P, Z_i = k)},$$

  and return to step 1. The likelihood function $\Pr(X_i = x_i|P, Z_i = k)$ is given by equation (2) in PRITCHARD *et al.* (2000a).

UPDATING BETA We choose a noninformative prior distribution for β, $\beta \sim \mathcal{N}(0, A^{-1})$, with $A = 0$. The Gibbs sampler proceeds by updating values of β using its conditional distribution ALBERT and CHIB (1993)

$$\beta|W \sim \mathcal{N}(V\tilde{X}^T W,\ V), \text{ where } V = (\tilde{X}^T\tilde{X})^{-1}. \tag{3.8}$$

Appendix S2: Computation of the predictive score for cross-validation

Let X^{training} be a subset of the loci used for inferring the parameters of the clustering model. The log-probability of the complementary set of loci $X^{\text{validation}}$ is a function of the cluster labels Z given by

$$\log(\Pr(X^{\text{validation}}|Z)) = \sum_{k=1}^{K} \sum_{l \in X^{\text{validation}}} \log \Pr(x_l^{[k]}) \tag{3.9}$$

where $x_l^{[k]}$ denote the observed genotypes at locus l in cluster k. We denote by $n_{kl} = (n_{kl1}, \ldots, n_{klJ_l})$ the allele count at locus l in cluster k. The allele counts follow a multinomial distribution $n_{kl} \sim \text{Multinomial}(m_{kl}, p_{kl1}, \ldots, p_{klJ_l})$, where $m_{kl} = \sum_{j=1}^{J_l} n_{klj}$ is the number of different genotypes at locus l in cluster k . By integrating over the alleles frequencies, we find that

$$\begin{aligned} \Pr(x_l^{[k]}) &= \frac{\prod_j (n_{klj}!)}{m_{kl}!} \Pr(n_{kl}|\lambda) \\ &= \frac{\Gamma(\lambda J_l)}{\Gamma(m_{kl} + \lambda J_l)} \prod_j \frac{\Gamma(n_{klj} + \lambda)}{\Gamma(\lambda)}. \end{aligned}$$

We computed the predictive score after averaging the quantities in equation (3.9) over the posterior distribution of the cluster labels Z given by the training data set.

Chapitre 4

Projections de la structure génétique des populations en réponse aux changements climatiques

4.1 Contexte : la projection des distributions d'espèces

De forts changements climatiques ont été prévus pour les années à venir. Des travaux groupés dans les publications du GIEC (Intergovernmental Panel on Climate Change 2007) montrent que ces changements comprennent, entre autres, des risques de modification des températures, d'augmentation des précipitations et du CO_2 dans l'air. Selon les différents scénarios envisagés, la température risque d'augmenter de $1,8$ à $4\,°C$ degrés en 100 ans. Or, l'influence du climat sur les espèces, de plantes en particulier, a été établie à plusieurs reprises (Walther *et al.* 2002; Parmesan and Yohe 2003; Jump and Penuelas 2005, pour une synthèse). Une conséquence observée des changements climatiques passés est la modification des distributions des espèces de plantes, qui « migrent » vers de plus hautes altitudes, ou vers de nouvelles latitudes (Grabherr and Gottfried 1994; Sturm *et al.* 2001; Peñuelas and Boada 2003). Pour ces raisons, les études se sont multipliées pour tenter de prédire les futures distributions d'espèces. Ces prédictions reposent majoritairement sur les modèles bioclimatiques (parfois appelés *modèles de distribution d'espèces*, *modèles d'enveloppe*, ou encore *modèles de niche écologique*) qui ont été présentés dans la section 2.5. Ces modèles permettent de prédire les aires géographiques de distribution des espèces en fonction du climat et sont utilisés principalement pour prévoir les mouvements des espèces en cas de changement climatique, mais aussi pour prévoir les aires potentielles de distribution d'espèces invasives, extrapoler la distribution actuelle d'une espèce à partir d'observations en un nombre réduit de points, ou inférer le paléoclimat (e.g. Huntley 1995; Manel *et al.* 1999; Guisan

and Zimmermann 2000; Peterson and Vieglais 2001; Walther *et al.* 2004; Marra *et al.* 2004; Thuiller *et al.* 2008). Le principe est d'expliquer la présence/absence d'une espèce par des variables climatiques (parfois aussi topographiques) (voir, dans la section 2.5, les méthodes classiquement utilisées). Cela permet d'estimer l'*enveloppe climatique* occupée par l'espèce. Si les conditions environnementales changent, on peut alors prévoir le déplacement (géographique) de l'enveloppe climatique et donc la future distribution de l'espèce (voir Figure 4.1). Ceci repose sur la théorie de *niche* (ou enveloppe) écologique d'une espèce ; une niche correspond à un espace réunissant les conditions biotiques et abiotiques nécessaires à la survie d'une espèce (Hutchinson 1957). C'est cette niche (plus précisément, la *niche réalisée*) que les modèles bioclimatiques tentent d'estimer. Or, une hypothèse est que la niche d'une espèce est conservée au cours du temps, donc si les conditions climatiques changent, on suppose que l'espèce ne pourra persister que si elle se déplace là où sa niche sera réalisée (Peterson *et al.* 1999). Cette hypothèse de conservatisme de niche est généralement implicite dans les modèles bioclimatiques. En revanche, d'autres hypothèses-clés sont souvent mises en avant (Pearson and Dawson 2003; Hampe 2004; Guisan and Thuiller 2005; Jeschke and Strayer 2008) :

- Les interactions biotiques sont constantes dans l'espace et dans le temps. Pour prédire que dans le futur une espèce risque de se déplacer vers une nouvelle aire, il est supposé que dans la nouvelle aire les interactions biotiques seront les mêmes que dans l'aire actuelle (par exemple, les compétitions éventuelles avec de nouvelles espèces ne sont pas prises en compte).
- Le génotype et le phénotype des individus sont constants dans l'espace et le temps. Donc les individus au sein d'une espèce sont génétiquement identiques (constance dans l'espace), et l'évolution génétique rapide ou la plasticité phénotypique ne sont pas prises en compte dans les réponses possibles aux changements climatiques (constance dans le temps).
- Il n'y a pas de limitation de la dispersion. L'espèce peut atteindre n'importe quelle zone du moment que les conditions climatiques y sont favorables.

L'hypothèse stipulant que les individus sont génétiquement identiques nous intéresse particulièrement. Elle implique que les modèles bioclimatiques ignorent la diversité génétique et le fait qu'au sein d'une espèce les individus peuvent répondre différemment aux changements climatiques. Plusieurs articles ont attiré l'attention sur le fait que la dimension génétique n'est que trop rarement envisagée dans les modèles bioclimatiques, alors que les associations entre variation génétique et environnement sont de plus en plus étudiées (Davis and Shaw 2001; Davis *et al.* 2005; Aitken *et al.* 2008; Sork *et al.* 2010; Atkins and Travis 2010). Sork *et al.* (2010) présentent une des rares études utilisant des modèles bioclimatiques mais tenant aussi compte de la structure génétique. Ils montrent qu'il existe un lien entre la variation génétique du chêne blanc de Californie et un certain nombre de variables climatiques, ce qui révèle le potentiel d'une réponse aux

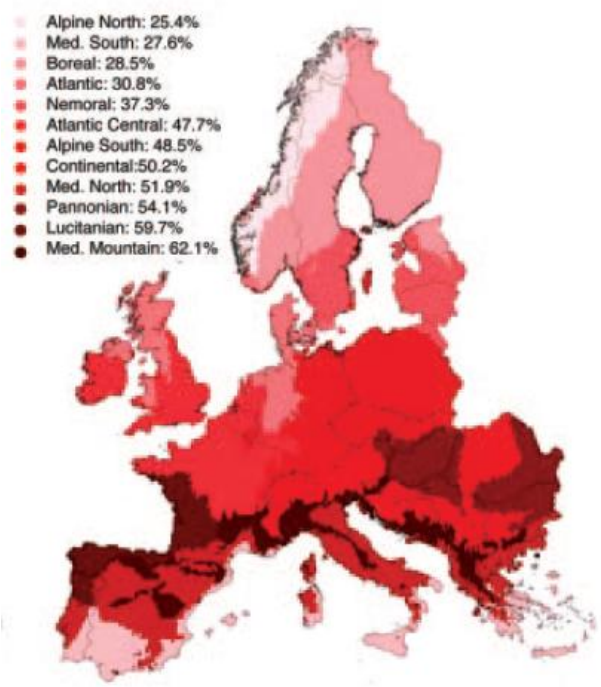


FIGURE 4.1 – THUILLER *et al.* (2005) utilisent des modèles bioclimatiques pour prédire les distributions d'espèces de plantes en Europe en cas de changement climatique. Cette carte présente le pourcentage d'espèces risquant de disparaître par zone pour le scénario de changement climatique le plus extrême.

changements climatiques diversifiée au sein de l'espèce. Pour prévoir la future distribution de l'espèce, ils proposent alors de calibrer, non pas un modèle bioclimatique, mais quatre modèles. Ces modèles correspondent à quatre régions géographiques (Nord, Est, Ouest, Sud) qui sont génétiquement différenciées et qui répondront donc peut-être différemment aux changements climatiques. Les projections de la distribution de l'espèce pour chaque région sont montrées sur la Figure 4.2. L'idée de SORK *et al.* (2010) est donc d'affiner la méthode pour améliorer la prévision de la future distribution d'espèce.

4.2 Projection de la structure génétique des populations

L'approche que nous avons développée pour intégrer la variation génétique est différente de celle de SORK *et al.* (2010), car l'aspect génétique n'est pas étudié simplement en amont de l'application d'un modèle bioclimatique. En effet, nous proposons de projeter, non pas la distribution des espèces, mais plutôt la structure génétique des populations. Comme expliqué dans la section 2.5, le modèle de `POPS` permet d'estimer conjointement les coefficients de métissage et leur lien avec l'environnement à partir des données génétiques et environnementales actuelles. Les coefficients de métissage peuvent ensuite être prédits

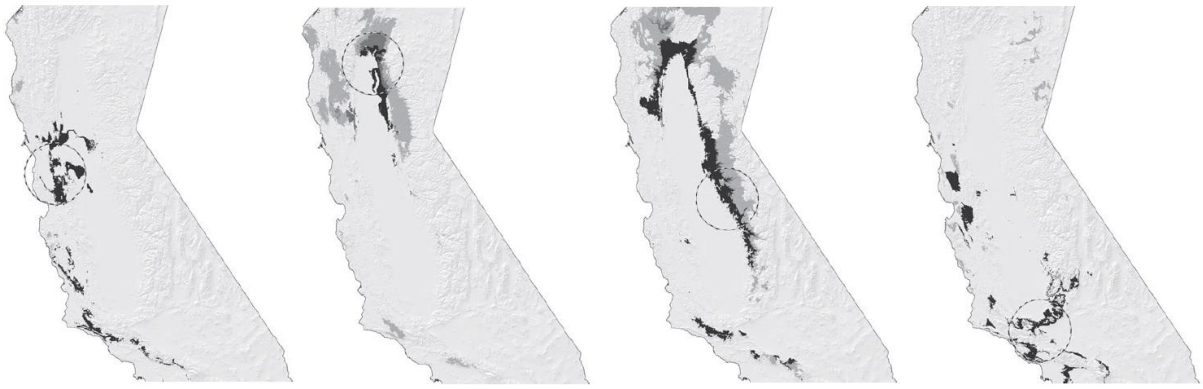


FIGURE 4.2 – Superposition de la distribution actuelle estimée de l'espèce du chêne blanc de Californie et de la distribution prévue pour un scénario fixé de changement climatique (SORK *et al.* 2010). Les prévisions ont été établies indépendamment pour 4 régions géographiques (Ouest, Nord, Est, Sud), abritant des individus génétiquement différenciés (d'une région à l'autre). Gris foncé : zones ayant une probabilité supérieure à 0,5 ($P > 0,5$) d'être occupées actuellement et une probabilité $P > 0,5$ d'être occupées après le changement climatique (zones restant habitables). Noir : zones prédites comme occupées actuellement (P>0,5), mais désertées dans le futur. Gris clair : zones prédites comme non occupées actuellement mais ayant une probabilité $P > 0,5$ d'être occupées après le changement climatique.

en cas de changement climatique. Comme ils correspondent à la part du génome assignée à chaque population estimée, les coefficients de métissage, et surtout leurs évolutions, sont reliés à la migration des gènes des différentes populations. Cette démarche nous permet donc d'évaluer indirectement la migration potentielle des gènes.

Un certain nombre d'hypothèses formulées pour prédire la structure génétique des populations en cas de changement climatique sont communes avec les modèles bioclimatiques, à savoir :

1. les interactions biotiques sont supposées constantes ;
2. il n'y a pas d'évolution génétique rapide, ni de plasticité phénotypique ;
3. les limites de dispersion des espèces ne sont pas modélisées explicitement.

La troisième hypothèse est moins forte que celle des modèles bioclimatiques qui, eux, n'imposent aucune limite à la dispersion. Contrairement aux modèles bioclimatiques où aucune variable spatiale n'intervient, il est possible d'utiliser les coordonnées spatiales et donc de tenir compte de l'éloignement géographique entre aire actuelle et aire future lors de la prédiction. Enfin, au lieu de supposer l'existence d'une niche écologique commune à une espèce, nous faisons l'hypothèse qu'il existe des niches intraspécifiques : les populations au sein d'une même espèce ont chacune leur propre niche, c'est-à-dire qu'elles sont adaptées à leur environnement local. On parlera d'« hypothèse d'adaptation locale ou de préadaptation », qu'il ne faut pas confondre avec le processus d'adaptation rapide, que l'on a écarté de notre modèle (cf. hypothèse 2).

4.3 Application : projection de la structure génétique d'espèces de plantes alpines en réponse au changement climatique

Nous avons appliqué notre méthode à 20 espèces de plantes alpines, issues des données du projet INTRABIODIV (Gugerli *et al.* 2008). Pour chaque espèce, des individus ont été échantillonnés sur une grille définie par des sites régulièrement espacés dans les Alpes européennes, et génotypés en une centaine de locus en moyenne. De plus, des variables climatiques et topographiques, telles que la température, les précipitations, l'ensoleillement, l'altitude, la pente, ont été extraites aux points d'échantillonnage à l'aide d'un système d'information géographique (`DAYMET`, Thornton *et al.* 1997).

Les structures génétiques estimées par le modèle avec métissage de `POPS` sont constituées, pour la majorité des espèces, de populations (ou clusters) correspondant à des zones géographiques séparées (Figure 4.3). Pour 90 % des espèces, un cluster significativement chaud est détecté, c'est-à-dire un cluster pour lequel les températures aux points d'échantillonnage des individus assignés au cluster sont significativement plus élevées que dans les autres clusters. Pour 15 espèces, les populations chaudes sont localisées au sud-ouest des Alpes, à une latitude inférieure à 46 °N et une longitude inférieure à 8 °E (Figure 4.3A, Table 4.2). De plus, les coefficients de métissage estimés pour les individus au sud-ouest présentent une variation graduelle le long de la direction sud-ouest/nord-est, ce qui indique l'existence d'une zone de contact entre les clusters chauds et les clusters voisins (voir la première ligne de la Figure 4.4).

Pour chaque espèce, nous avons utilisé `POPS` pour prédire la structure génétique après une augmentation de température variant de $0,25$ à $4\,^{\circ}\mathrm{C}$. Des coefficients de métissage sont prédits à partir des paramètres estimés par `POPS`, et des valeurs actuelles des covariables, à l'exception des températures moyennes qui sont augmentées graduellement (cf. section 2.5). Les prédictions des coefficients de métissage avec le cluster chaud sont projetées sur une carte des Alpes à l'aide d'une technique de krigeage. Pour la majorité des espèces, on observe des modifications de la structure génétique des populations et, en particulier, un déplacement de la zone de contact entre le cluster chaud et son voisin plus froid (illustration pour 3 espèces sur les cartes de la Figure 4.4).

Pour quantifier ces modifications de la structure génétique des populations, nous avons mesuré l'étendue du déplacement de la zone de contact le long d'un axe sud-ouest/nord-est, pour les espèces ayant un cluster chaud au sud-ouest ou à l'ouest (Figure 4.5A). La valeur moyenne du déplacement dépend de l'espèce, mais le sens est toujours le même. Le déplacement moyen prédit est de 92 km (minimum : 5 km, maximum : 212 km) vers le

nord-est pour une augmentation de température de 2 °C, et de 188 km (minimum : 11 km, maximum : 393 km) vers le nord-ouest pour une augmentation de 4 °C. Pour l'espèce *Hypochaeris uniflora* la direction du déplacement est inversée puisque son cluster chaud se trouve à l'est. Comme les hypothèses de notre modèle impliquent que ni l'adaptation rapide ni la plasticité phénotypique ne sont possibles, la seule réponse envisageable pour les populations est la migration. Il faut donc garder à l'esprit que l'ampleur des déplacements prédits constitue probablement une borne supérieure pour les migrations envisagées.

D'autre part, pour évaluer la modification globale des structures génétiques des populations, nous avons mesuré pour chaque espèce le *renouvellement intraspécifique*, mesuré à partir de la corrélation entre les coefficients de métissage actuels et les coefficients prédits pour chaque augmentation de température (Figure 4.5B). Pour une augmentation de 2 °C, la corrélation reste au dessus de 70 % chez 18 des 20 espèces, ce qui suggère un faible renouvellement génétique et un impact modeste du changement climatique. En revanche, pour une augmentation de 4 °C, chez 10 des 20 espèces la corrélation passe au-dessous de 60 %, ce qui indique un renouvellement intraspécifique plus important.

En projetant la structure intraspécifique plutôt que la distribution globale de l'espèce, l'idée est de tenir compte de variations intraspécifiques de la niche. Comme nous l'avons vu, l'existence de niches intraspécifiques correspond en fait à l'hypothèse stipulant que les populations sont localement préadaptées à l'environnement. Pour examiner cette hypothèse, nous avons étudié les variations des fréquences alléliques en fonction de la latitude. Même s'ils sont neutres, les marqueurs pour lesquels une forte corrélation est trouvée peuvent être liés à des gènes sélectionnés, du fait de l'*auto-stop génétique* par exemple, mais aussi à des barrières aux flux géniques créées par l'adaptation locale. De tels marqueurs sont donc la signature d'une éventuelle adaptation locale (Joost *et al.* 2007; Manel *et al.* 2010b; Coop *et al.* 2010). Pour la majorité des espèces, les variations les plus extrêmes correspondent à des clines localisés à une latitude d'environ 45-46 °N (voir les clines présentés pour 3 espèces, Figure 4.6). Cette latitude concorde avec la zone de contact entre le cluster chaud et son voisin, ce qui indique que l'hypothèse de préadaptation locale est vraisemblable. Toutefois, il est théoriquement possible que seule l'histoire démographique, et non l'adaptation locale, soit à l'origine de ces clines et de la structure de populations que nous observons. Pour tester cette idée, nous avons simulé des données génétiques à partir d'un scénario d'expansion démographique dans les Alpes et comparé, entre ces données simulées et les données observées chez les plantes, des mesures portant sur la forme des clines et sur la structure génétique des populations. Les résultats des tests statistiques montrent que le rôle de l'histoire démographique dans la formation de ces clines ne peut être totalement écarté, mais qu'elle ne peut expliquer à elle seule les corrélations gènes-environnement observées.

Dans notre étude nous avons mis en évidence des réponses au changement climatique pour un ensemble d'espèces. Bien qu'une tendance commune soit trouvée, comme le déplacement de la zone de contact vers le nord-est et le renouvellement intraspécifique, les mesures quantitatives associées varient entre les espèces. Il faut garder à l'esprit que notre modèle repose sur un certain nombre d'hypothèses et ne donne donc pas une image exacte des réponses aux changements climatiques. Ses prévisions, axées sur des données à la fois environnementales et génétiques, permettent toutefois une grande avancée dans le domaine des prévisions bioclimatiques.

4.4 Article B

F. Jay, S. Manel, N. Alvarez, E.Y. Durand, W. Thuiller , R. Holderegger, P. Taberlet, O. François. Forecasting changes in population genetic structure of Alpine plants in response to global warming. *Submitted*

Abstract

Species range shifts in response to climate and land use change are commonly forecasted with species distribution models based on species occurrence or abundance data. These models relate distribution data to climatic and habitat factors to forecast where species requirements would occur under different environmental change scenarios. Although appealing, these models ignore the genetic structure of species, and the fact that different populations might respond in different ways due to adaptation to their environment. Here, we introduced new models for forecasting intra-specific changes based on genetic ancestry and population structure instead of species distribution data. Using multi-locus genotypes and extensive geographic coverage of distribution data across the European Alps, we applied this approach to 20 alpine plant species considering a global increase in temperature from 0.25 °C to 4 °C. We measured the magnitudes of displacement of contact zones between plant populations potentially adapted to warmer environments and other populations. While a global trend of movement in a northeast direction was observed, the magnitude of displacement was species-specific. For a temperature increase of 2 °C, contact zones were predicted to move by 92 km on average (minimum of 5 km, maximum of 212 km), and by 188 km for an increase of 4 °C (minimum of 11 km, maximum of 393 km). Intra-specific turnover — measuring the extent of change in global population genetic structure — was generally found to be moderate for 2 °C of temperature warming. For 4 °C of warming, however, the models indicated substantial intra-specific turnover for ten species. These results illustrate that, in spite of unavoidable simplifications, ancestry distribution models open new perspectives to forecast population genetic changes within species in combination with more traditional distribution-based models.

Key-words: Alpine plants, Climate change, Intra-specific variation, Landscape genetics, Ancestry distribution models.

Introduction

The impact of both climate and land use changes on biodiversity and more specifically on species distribution is widely acknowledged (Parmesan and Yohe 2003; Jump and Penuelas 2005; Thuiller *et al.* 2008). For alpine plants, available evidence includes range shifts that trigger the movement of plants to higher elevations or latitudes (Grabherr and Gottfried 1994; Walther *et al.* 2002; Walther 2003). Modification of range limits, however, may occur not only among species (Thuiller *et al.* 2005), but also among genetically differentiated clusters of populations within species (Sork *et al.* 2010; Aitken *et al.* 2008).

Predictions of the effects of climate change commonly rely on species distribution models (Guisan and Zimmermann 2000; Guisan and Thuiller 2005). Species distribution models are correlative models relating field observations to environmental predictors. Field observations are usually occurrence or abundance observations that are used to infer the realized niche of a species. Geographic locations that satisfy the habitat requirements of a given species can then be identified by projecting the niche on a map with modified environmental data according to specified climate change scenarios. The impact of climate change can thus be evaluated by comparing the locations where niche requirements are presently satisfied to those where they could be satisfied in the future. Despite their obvious appeal, there are many shortcomings to species distribution models for predicting future distributions (Aitken *et al.* 2008). In particular, these models do not account for genetic variation within the species range and for adaptation of populations to past and contemporary environments.

Association of genetic variation with environmental variables has been frequently reported in the recent literature (Hedrick *et al.* 1976; Aitken *et al.* 2008; Balkenhol *et al.* 2009). For example, Duminil *et al.* (2007) tested the influence of a set of life-history traits on population genetic structure and gene flow in seed plants. Richardson *et al.* (2009) analyzed climate-related genetic patterns in the western white pine suggesting that divergent climatic selection has influenced phenotypic traits associated with tree growth. Several methods can be used to identify environmental factors determining population structure, for example through estimates of population divergence (Foll and Gaggiotti 2006). Alternatively, other approaches identify loci with clear correlations between allele frequencies and ecological variables, and interpret these loci as being potentially involved in local adaptation (Joost *et al.* 2007; Poncet *et al.* 2010; Manel *et al.* 2010b,a; Coop *et al.* 2010). Among the climatic variables that contribute to genetic differentiation in plant species, the influence of temperature has recurrently been demonstrated since the pioneering work of Turesson (1925). An example of the influence of climatic conditions is the timing of phenological events, such as flowering, which depend on temperature, and which are obvious targets for natural selection Stanton and Galen (1997); Stinson

(2004); Jump *et al.* (2009a,b); Wang *et al.* (2009); Doi *et al.* (2010); Scherrer and Körner (2011). From this perspective, appropriate standing variation and the ability of long-distance gene dispersal are key factors for plant species to keep track with rapidly changing environments (Davis and Shaw 2001; Salamin *et al.* 2010).

The objective of this study was to introduce a model-based approach to identify environmental variables that influence individual ancestry and to apply this approach to predict intra-specific genetic variation in alpine plants in response to climate change. Forecasts of change in population genetic structure require defining ancestry distribution models based on ecological characteristics and multi-locus genotypes instead of occurrence or abundance data of plant species. To illustrate this novel approach, we considered a set of 20 widespread, mostly perennial alpine plant species genotyped using amplified fragment length polymorphism (AFLPs; Gugerli *et al.* 2008; Alvarez *et al.* 2009; Thiel-Egenter *et al.* 2011). More specifically, we developed Bayesian models that incorporate hidden regression models of genetic admixture (ancestry) on climatic and topographic variables. The motivation behind ancestry distribution models is that individual ancestry can be correlated to environmental variables and that estimations of change in individual ancestry can provide indirect estimates of future rates of gene migration. The need for those models was underlined in two recent reviews (Storfer *et al.* 2007; Thomassen *et al.* 2010), and a related approach was applied to California valley oaks where regional distribution models were fitted to genetically differentiated groups with distinct responses to climate change (Sork *et al.* 2010). Implemented in the computer program `POPS` (Jay *et al.* 2011), ancestry distribution models simultaneously estimate genetic admixture based on neutral genetic markers, and the effects of varying environmental covariates on population genetic structure.

Plant populations can avoid extinction by tolerating climate change through phenotypic plasticity, adapting to new conditions through selection on genetic standing variation or favorable mutations, or migrating to locations with favorable conditions (Davis *et al.* 2005; Aitken *et al.* 2008). Regarding these three alternatives, ancestry distribution models assume that gene migration represents the main component in the response of plant populations to environmental change. One mechanism for migrating populations to track moving environments is through the dispersal of alleles adapted to the local environments prior to climate change (Davis and Shaw 2001). Therefore, a careful interpretation of ancestry distribution model predictions asks for an evaluation of whether the genetic clusters inferred from the models also correlate with genetic variation adapted to changing environmental pressures.

Focusing on 20 alpine plant species with these caveats in mind, we addressed the following questions. (1) How large are the changes in population genetic structure of alpine plants if temperatures rise from 2 °C up to 4 °C (Intergovernmental Panel on Cli-

MATE CHANGE 2007)? (2) What is the rate of forecasted gene migration for clusters of populations adapted to warmer environments? To answer these questions, we measured the extent of intra-specific turnover, defined by the correlation between current and predicted ancestry coefficients, and the magnitude of displacement of contact zones, where individuals with mixed ancestry were found, along a central transect of the European Alps. Regarding future environments, the level of turnover indicates the importance of change in population genetic structure, and contact zone shift were used as an indirect way to quantify gene migration through projections of ancestry coefficients on geographic maps. Finally, we attempted to disentangle the relative contributions of past adaptations versus demography in ancestry distribution model estimates.

Materials and Methods

Plant material and data.

Plant genotypes and environmental covariates were extracted from the INTRABIODIV data base (GUGERLI *et al.* 2008). We considered a subset of 20 mostly perennial plant species studied by (ALVAREZ *et al.* 2009) on the basis of their extensive geographic coverage of the Alps, including their south-western range: *Androsace obtusifolia* All., *Arabis alpina* L., *Campanula barbata* L., *Cerastium uniflorum* Clairv., *Dryas octopetala* L., *Gentiana nivalis* L., *Geum montanum* L., *Geum reptans* L., *Gypsophila repens* L., *Hedysarum hedysaroides* (L.) Schinz & Thell. s.l., *Hypochaeris uniflora* Vill., *Juncus trifidus* L., *Ligusticum mutellinoides* (Cr.) Vill., *Loiseleuria procumbens* (L.) Desv., *Luzula alpinopilosa* (Chaix) Breistr., *Phyteuma hemisphaericum* L., *Rhododendron ferrugineum* L., *Saxifraga stellaris* L., *Sesleria caerulea* (L.) Ard., *Trifolium alpinum* L. Individual-based sampling was performed within a rectangular grid system with cell surfaces of 22.3 km x 25 km (12' latitude and 20' longitude). Only cells with areas of elevation higher than 1500 m above sea level were considered. Three individuals per species were sampled within cells, respecting at least 10 m distance between successive individuals. All plant samples were genotyped using amplified fragment length polymorphisms as detailed in (GUGERLI *et al.* 2008). In the INTRABIODIV database, habitats are modeled by using DAYMET, a computer program that analyzes daily records of climate variables in a spatial context (THORNTON *et al.* 1997; GUGERLI *et al.* 2008). Geographic and environmental covariates extracted from the INTRABIODIV database included latitude and longitude, annual average of minimum and maximum daily temperatures (years 1980-1989), spring and summer seasonal precipitations (years 1980-1989), and an additional set of topographic variables (slope, orientation) measured in each cell. We selected those variables based on their relative importance in generating gene-environment associations in the 20 species studied here. A short description of the data set used in this study,

including the number of sampled cells and individuals, and the number of loci genotyped, is provided in Table 4.1.

Ancestry distribution models.

To identify environmental variables that influence individual ancestry and use these variables to forecast responses to climate change, we developed POPS, a Bayesian algorithm that incorporates genetic and geographic data and infers admixture coefficients based on correlation with environmental variables (Durand *et al.* 2009b; Jay *et al.* 2011,a). More specifically, for each individual, the admixture coefficients represent the fraction of ancestry shared between K putative source populations or clusters. The `POPS` algorithm includes a hidden regression framework where admixture coefficients are regressed on geographic and ecological data. Using standard notations, admixture coefficients are stored in a matrix, Q, with elements, q_{ik}, corresponding to the genome proportion of individual i that originates from the source population k ($k = \ldots, K$). In addition to admixture coefficients, `POPS` estimates a multidimensional vector of regression coefficients, β, measuring the effects of each covariate on individual ancestry. For the parameters of interest, Q and β, the posterior distribution is given by

$$\Pr(Q, \beta | X, \tilde{X}) = c\Pr(X|Q)\Pr(Q|\tilde{X}, \beta)\Pr(\beta),$$

where X is the matrix of genotypes, $\tilde{X}$ is the matrix of geographic and environmental covariates, and c is a constant of proportionality. In this model, the likelihood function, the prior distribution of allele frequencies and the conditional distribution of cluster labels given Q are described by the same statistical models as implemented in the software `STRUCTURE` (Pritchard *et al.* 2000a). To define the conditional probability distribution, $\Pr(Q|\tilde{X}, \beta)$, we used a hierarchical regression model. For each individual (i) and each cluster (k), we introduced a parameter (α_{ik}) proportional to the expected value of the ancestry coefficient (q_{ik}) under a Dirichlet distribution (Durand *et al.* 2009b). Separating $\tilde{X}$ into a set of geographic variables, $\tilde{X}_g$, and a set of habitat variables, $\tilde{X}_h$, the hierarchical model was defined by K hidden regression equations

$$\log(\alpha_{ik}) = \tilde{X}_{ih}\beta_{hk} + g(\tilde{X}_{ig})\beta_{gk} + \epsilon_{ik}, i = 1, \ldots, N, k = 1, \ldots, K.$$

where β_{hk}, β_{gk} are vectors of regression coefficients modeling environmental and geographic effects, $g(\tilde{X}_{ig})$ represents a non-linear spatial trend surface, ϵ_{ik} is a zero-mean spatially auto-correlated residual (Besag 1975), N is the sample size. In the right-hand side of the regression equations, the last two terms account for broad-scale and local spatial patterns in unobserved ancestries. The first term models the effects of habitat variables once the spatial effects have been removed. Similar approaches — separating habitat

	Number of cells	Number of individuals	Number of markers	Temperature (°C)	Altitude (m a.s.l.)
Androsace obtusifolia	45	131	138	0.48	2280
Arabis alpina	129	385	140	2.99	1874
Campanula barbata	104	307	114	2.75	1857
Cerastium uniflorum	44	130	93	-0.42	2445
Dryas octopetala	124	370	100	2.62	1934
Gentiana nivalis	74	218	158	1.45	2155
Geum montanum	122	363	86	2.43	1982
Geum reptans	51	153	57	0.12	2404
Gypsophila repens	107	319	94	3.58	1759
Hedysarum hedysaroides	66	144	123	2.13	1984
Hypochaeris uniflora	59	177	74	2.43	1951
Juncus trifidus	91	269	86	1.74	2109
Loiseleuria procumbens	80	239	199	1.31	2191
Luzula alpinopilosa	42	92	93	0.85	2213
Ligusticum mutellinoides	90	270	116	1.50	2088
Phyteuma hemisphaericum	76	225	234	1.54	2111
Rhododendron ferrugineum	126	377	111	2.89	1891
Saxifraga stellaris	113	265	70	3.13	1816
Sesleria caerulea	100	283	187	2.15	1984
Trifolium alpinum	39	76	97	2.13	2060

Table 4.1: Sampling design, number of AFLPs, average minimal daily temperature, mean elevation for 20 Alpine species from INTRABIODIV.

and spatial covariates — have been adopted in landscape ecology where the responses consist of abundance or occurrence data (Lichstein *et al.* 2002). Using a landscape genetic framework (Holderegger and Wagner 2008), individual ancestry coefficients are used instead of species distribution data in ancestry distribution models. A Markov chain Monte Carlo algorithm allowed us to jointly estimate the ancestry coefficients and the effects of the environmental variables. Like species distribution models, `POPS` also provides routines to project ancestry coefficients on geographic maps under scenarios of environmental change. These projections were used to estimate changes in population genetic structure and the magnitudes of shift for areas of mixed ancestry separating pairs of genetic clusters (*i.e.*, contact zones). Forecasts from regression models were obtained by varying the climatic variables under various scenarios, considering temperatures increasing from 0 °C to 4 °C by increments of 0.25 °C (Intergovernmental Panel on Climate Change 2007). The projections of the future distributions of ancestry coefficients were displayed on geographic maps using a kriging method in R (R Development Core Team 2011).

Population genetic structure estimates obtained under current environmental conditions were controlled by applying the Bayesian program `TESS 2.3` (Chen *et al.* 2007) and the Neighbor-Joining algorithm using shared allele distance (Saitou and Nei 1987). The conditional auto-regressive model of `TESS` was applied to confirm the detection of contact zones in the geographic range of each plant species (Durand *et al.* 2009b; François and Durand 2010b). In `TESS` and `POPS`, we set the number of clusters to values less than 4, thus only retaining population clusters that corresponded to the most divergent clades in NJ trees. To evaluate whether or not the assignment of plant populations into well-defined geographic clusters could explain the high proportion of species with warm genetic clusters, we used random partitions of the study area to create geographically structured clusters (10,000 replicates).

Measures of contact zone shifts and intra-specific turnover.

To interpret ancestry distribution model predictions in terms of gene migration and population structure changes, we first measured the shift of contact zones between pairs of geographically adjacent clusters from projections on geographic maps. The magnitude of displacement was computed for temperatures increasing from 0.25 °C to 4 °C along a southwest-northeast transect paralleling the alpine axis. This was achieved by following the projections of contact zone points (sharing 50% of ancestry in each cluster), and monitoring displacements on the geographic map. In addition, we assessed the extent of intra-specific turnover by measuring the correlation between ancestry coefficients estimated using genetic data and current temperatures and those predicted for increased forecasted temperatures. The closer to 1 is this correlation, the smaller are expected

changes in population structure. While contact zone shifts between clusters are calculated pair-wise, intra-specific turnover can be viewed as a global measure of change in the population genetic structure of a given species.

Evidence supporting the existence of adaptive variation.

In our niche modeling assumptions, we implicitly assumed that plant populations were adapted to their local climatic conditions previous to climate warming. In particular, we supposed that genetic differentiation between clusters was — at least partly — explained by genetic barriers reflecting local adaptation to past and contemporary environments. We applied an outlier locus approach based on generalized linear models (GLMs, Joost *et al.* 2007) to evaluate the evidence for local adaptation to contemporary environmental conditions in each plant species. For each locus, we fitted a model of allele distribution via logistic regression of the binary alleles on latitude, and analyzed the clines corresponding to extreme values of the empirical distribution of z-statistics. Note that we used latitude because we were interested in identifying which clusters harbor appropriate genetic variation due to local adaptation (we did not examine which genes are targeted by natural selection). In this respect latitude was more robust than temperature which exhibits high levels of variability within small geographic areas. To overcome multiple testing issues, we controlled the false discovery rate at the level of 1% (Benjamini and Hochberg 1995). For all species with four or more outlier loci detected, we displayed the fitted allele frequency gradients corresponding to the four lowest P-values. We reported latitudes for which the gradients exhibited a point of change in curvature. To control for spatial auto-correlation in the sampling design of INTRABIODIV, we applied tests using generalized estimating equations (GEE; R Development Core Team 2011) as proposed in (Poncet *et al.* 2010).

Spatially explicit models of demographic history.

To test whether the clines observed at outlier loci could result from purely demographic processes, we performed computer simulations of a coalescent model of migration incorporating geographic information (SPLATCHE, Currat *et al.* 2004). The demographic history of plant populations was simulated using a model of range expansion defined on a lattice of demes reproducing the sampling design of the data used here. The source of expansion was located in regions of maximal genetic diversity identified as potential refugia during the ice age (Schönswetter *et al.* 2005). Genetic diversity was computed in each grid cell using Nei's estimator for AFLPs (Bonin *et al.* 2007). In simulations, the onset of expansion was either located in the southwest of the Alps (SW, 44 °N, 6 °E) or in the northeast of the Alps (NE, 47 °N, 12 °E). The simulation parameters were calibrated so that the Alps were colonized in less than 600 generations (Higgins and Richard-

son 1999). Within each cell, population size grew according to a logistic equation with carrying capacity equal to C = 50 effective individuals. Using cell locations and number of loci equal to those of the data used here, we tested differences between simulated and empirical data for two statistics. The first test statistic was defined by the correlation between ancestry proportions estimated from the genetic data and those predicted from the geographic and environmental covariates (without change). The second statistic corresponded to the amplitude of the latitudinal allele frequency cline estimated for the most extreme locus. The amplitude of a cline was computed as the difference between the minimum and maximum of the allele frequency cline within the latitudinal range of a species' distribution.

Results

Population genetic structure.

When used as an assignment method, `POPS` split the samples into geographically well-defined clusters which agreed with the results obtained from TESS, hierarchical clustering and previous analyses (Alvarez *et al.* 2009). One striking feature shared by 18 (90%) of the 20 plant species studied was the existence of "warm"genetic clusters (Figs. 4.3 and S1; Table 4.2). "Warm"clusters, as inferred by POPS, were characterized by temperatures significantly higher for the individuals assigned to this cluster than for individuals assigned to other clusters (two-sample t-tests, Table 4.2). To evaluate whether the assignment of individuals into well-defined geographic clusters could explain the high proportion of species with warm genetic clusters, we used a simulation approach based on artificial geographic clusters. Significant differences in average temperature among clusters were observed in 49% of the replicates. Thus, the high proportion of warm clusters observed in the study data set (90%) was not a mere outcome of the association between geography and climate.

For 15 species, the warm clusters were located in the southwest (SW) of the Alps, south of latitude 46 °N and west of longitude 8 °E in France and Italy (Table 4.2). Proportions of ancestry in warm clusters exhibited a gradual decrease in a north-eastern direction suggesting contact and gene flow with "colder"clusters (first rows in Fig. 2, Supplementary Movie SM1). The contact zones, associated with clines of admixture around 45.5 °N of latitude, reflected a potential intrusion of genes from lower latitude regions, and provided evidence that the geographic boundaries between warmer and colder populations were highly permeable. Other warm clusters were identified in the south of Austria (around 46.5 °N, 14 °E) for two species (*Saxifraga stellaris* and *Hypochaeris uniflora*). Two species (*Luzula alpinopilosa* and *Phyteuma hemisphaericum*) showed no evidence of differentiation with respect to temperature. For one species (*Dryas octopetala*), the warm cluster was

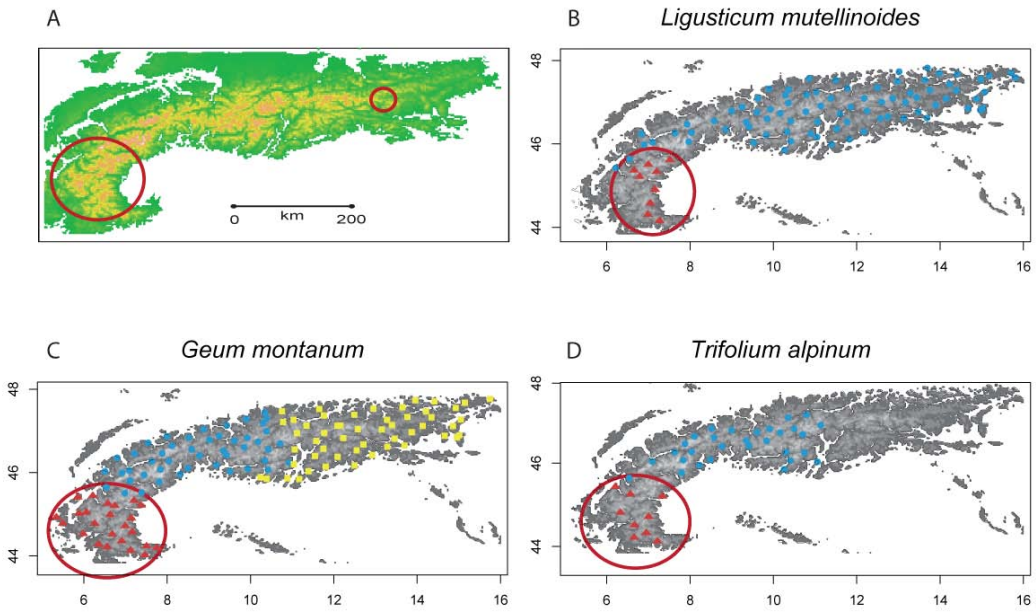


Figure 4.3: Population structure and location of warm genetic clusters in alpine plant species. A. Geographic map of the study area showing elevations > 1000 m.a.s.l and localization of warm clusters in the Southwest and East (circles). Population structure inferred by POPS for B. *Ligusticum mutellinoides*, C. *Geum montanum*, D. *Trifolium alpinum*. Sample locations in warm clusters are represented by red triangles. Population structure for all species is displayed in Figure S1.

spread over the western Alps at longitudes lower than 10 °E.

Model predictions of contact zone shifts.

The POPS ancestry distribution model enabled forecasts of genetic admixture under various climate change scenarios of similar magnitude as projected by the Intergovernmental Panel on Climate Change (2007). Starting from current average temperature values, changes in genetic admixture were computed over the alpine range for all individuals in all species and for temperatures increasing by steps of 0.25 °C up to 4 °C. Figure 4.4 shows model predictions of the movement of the contact zone between warm and cold clusters for three representative species (*Ligusticum mutellinoides*, *Geum montanum*, *Trifolium alpinum*). Among the 18 species that exhibited a warm cluster, 16 species were predicted to experience significant movement of contact zones in response to temperature increase (Supplementary Movie SM1). For all these species, the ancestry distribution models predicted that genes from warm clusters would invade colder clusters by progressively moving toward a north-eastern direction with increasing temperature (Fig. 4.4, Fig. 4.5A). One species, Hypochaeris uniflora, was predicted to behave in a different way as its warm cluster was located in the east, and thus the movement was predicted to prevail in a reverse direction, from east to the southwest. Two species with high altitudinal ranges, *Androsace obtusifolia* and *Cerastium uniflorum*, exhibited no sig-

nificant changes. In these two species, warm clusters were more weakly defined than in other species: They were significantly warmer according to average daily minimum temperature ($P < 0.05$ for both species) but not according to maximum temperature ($P = 0.06$ and $P = 0.45$, respectively).

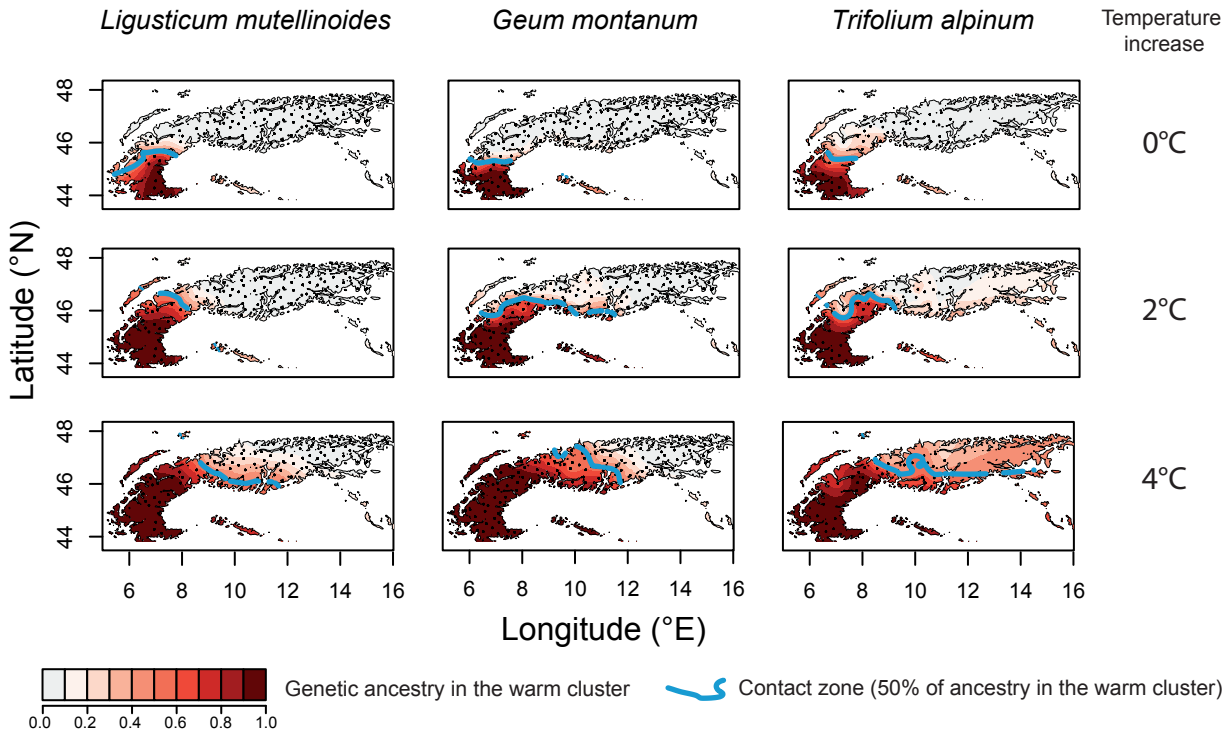


Figure 4.4: Predictions of contact zone movements in alpine plant species under increased temperature scenarios. The predictions were obtained from the POPS model for mean annual temperatures increasing by 0 °C, 2 °C and 4 °C. For each location on the maps, the color intensities represent the amounts of genetic ancestry (or individual admixture coefficients) in the warm clusters. The blue line represents the contact zone between warm and cold populations, defined as 50% of ancestry in each cluster. Maps are given for *Ligusticum mutellinoides*, *Geum montanum* and *Trifolium alpinum*. Predictions for all species are displayed in Supplementary Movie SM1.

To further quantify our predictions of gene migration under increased temperature, we estimated the displacement of the contact zones for species exhibiting warm clusters in the southwest. Figure 4.5A reports the magnitudes of contact zone shift estimated by tracking a contact point along a southwest-northeast transect. While a global trend of movement toward a north-eastern direction was observed, the magnitude of displacement was highly species-specific. For a temperature increase of 2 °C, contact zones were predicted to shift in the northeast direction by 92 km (minimum of 5 km, maximum of 212 km, SD = 68 km) and by 188 km (minimum of 11 km, maximum of 393 km, SD = 139 km) for an increase of 4 °C. For species *Ligusticum mutellinoides*, *Geum montanum*, *Trifolium alpinum*, we quantified the uncertainty of these POPS model predictions for a temperature increase of 2 °C (Figure 4.4), and obtained posterior predictive standard deviations equal to 78, 42

and 74 km respectively. Note that these estimates of cluster interface movement are only coarse summaries of model predictions: gene migration is expected to follow the complex topography of the alpine mountain range. A more accurate description of the forecasted movements of contact zones is given in the Supplementary Movie SM1.

Intra-specific turnover.

For each incremental increase in mean annual temperature, the extent of intra-specific turnover was measured by the correlation between current and predicted ancestry coefficients for future temperatures (Fig. 4.5B). For temperature increases of less than 2 °C, the correlation remained larger than 70% in 18 of the 20 species. Thus a modest impact of climate change on intra-specific structure was predicted for most species. For greater changes in temperature, histograms of correlation between current and predicted ancestry coefficients exhibited two modes, and species responses divided into two categories. For a temperature increase of 4 °C, correlations were greater than 60% for ten plant species. For those species, intra-specific turnover remained moderate. However, for the ten remaining species, the correlations dropped below 60%, meaning that the changes in genetic population structure were substantial for these species.

Evidence of adaptation in warm clusters.

Existence of populations already adapted to a warmer climate is a pre-condition for interpreting contact zone displacements as a consequence of gene migration to track shifting environments. To examine this hypothesis, we applied two models of allele distribution, respectively based on GLMs (Joost *et al.* 2007) and GEEs (Poncet *et al.* 2010), to detect loci for which allele frequency gradients exhibited strong correlations with latitudinal gradients. The detected outlier loci were likely to be associated with genes under selection by environmental factors linked to climate (Haldane 1948; Slatkin 1973; Coop *et al.* 2010; Manel *et al.* 2010a,b). GLMs and GEEs led to almost equal cline estimates, and we used the outlier loci of GLMs to identify geographic regions where adaptation to higher temperature could have occurred. More specifically, the models detected outlier loci exhibiting sharp latitudinal clines in 17 of 20 species including all species with warm clusters in the southwest of the Alps (Table 4.2; Figs. 4.6, S3). Figure 4.6 displays examples of "extreme"allele frequencies at four outlier loci in three representative species (examples for 17 species is given in Fig. S3). Sharp changes in allele frequency gradients were observed around latitudes 45-46 °N, which also correspond to the location of the contemporary contact zones observed between warm and cold clusters.

Table 4.2: **Table 2.** Maxima of genetic diversity, warm cluster locations and latitudinal clines for 20 Alpine plant species.

Species name	Maximum of genetic diversity [A]	Warm cluster [B]				Latitude of the most outlier cline [C]
		Localization of the warm cluster	Mean temperature within	Mean temperature outside	p-value	
Androsace obtusifolia	SW [1]	SW (45.4°N, 6.95°E)	-2.13	-3.24	0.00171	45°
Arabis alpina	SW	SW (44.9°N, 6.87°E)	2.60	-1.31	1.91e-16	45°/46.2°
Campanula barbata	N [1]	SW (45.5°N, 7.43°E)	-0.27	-1.08	0.00292	45.2°/46.6°
Cerastium uniflorum	SW	SW (45.9°N, 8.74°E)	-3.21	-3.81	0.0252	45°/46°
Dryas octopetala	NE [1]	W (45.7°N, 7.68°E)	-0.34	-1.42	3.37e-06	-
Gentiana nivalis	NE	SW (45.2°N, 7.22°E)	-1.22	-2.29	0.00346	45°/46.2°
Geum montanum	SW	SW (44.9°N, 7.05°E)	0.68	-1.61	1.64e-09	45°/46°
Geum reptans	NE	SW (45.3°N, 7.04°E)	-2.30	-3.71	1.78e-07	46.2°
Gypsophila repens	SW	SW (44.6°N, 6.79°E)	3.19	-0.88	2.64e-14	45°
Hedysarum hedysaroides	SW	SW (44.9°N, 7.95°E)	1.43	-1.80	1.07e-06	45°
Hypochaeris uniflora	E [1]	E (47.0°N, 14.4°E)	-0.61	-1.24	0.0211	-
Juncus trifidus	SW	SW (44.7°N, 7.06°E)	1.28	-2.23	1.22e-09	45°/46.5°
Ligusticum mutellinoides	NE	SW (45.0°N, 7.03°E)	-0.18	-2.10	0.000151	45°
Loiseleuria procumbens	NE	SW (44.7°N, 7.13°E)	0.65	-2.37	0.000138	45°/46.5°
Luzula alpinopilosa	SW	-	-	-	NS[3]	-
Phyteuma hemisphaericum	SW	-	-	-	NS	45.7°
Rhododendron ferrugineum	SW	SW (44.5°N, 6.76°E)	2.03	-1.09	1.93e-08	45°/46.5°
Saxifraga stellaris	- [2]	E (46.5°N, 13.4°E)	1.58	-0.99	1.54e-09	-
Sesleria caerulea	E	SW (45.4°N, 7.66°E)	-0.33	-1.71	0.00019	46°/47°
Trifolium alpinum	E	SW (44.9°N, 6.92°E)	0.14	-2.09	0.00172	45°

1: SW = southwest, N = north, NE = northeast, E = east. 2: '-' indicates absence of results. 3: NS stands for non significant p-value
A: Maxima of genetic diversity. B: Localization, mean temperature within and outside warm clusters, p-value for two-sample t-tests.
C: For each species, latitudes correspond to the points where allele frequency clines crossed the 50% threshold (4 most extreme P-values).

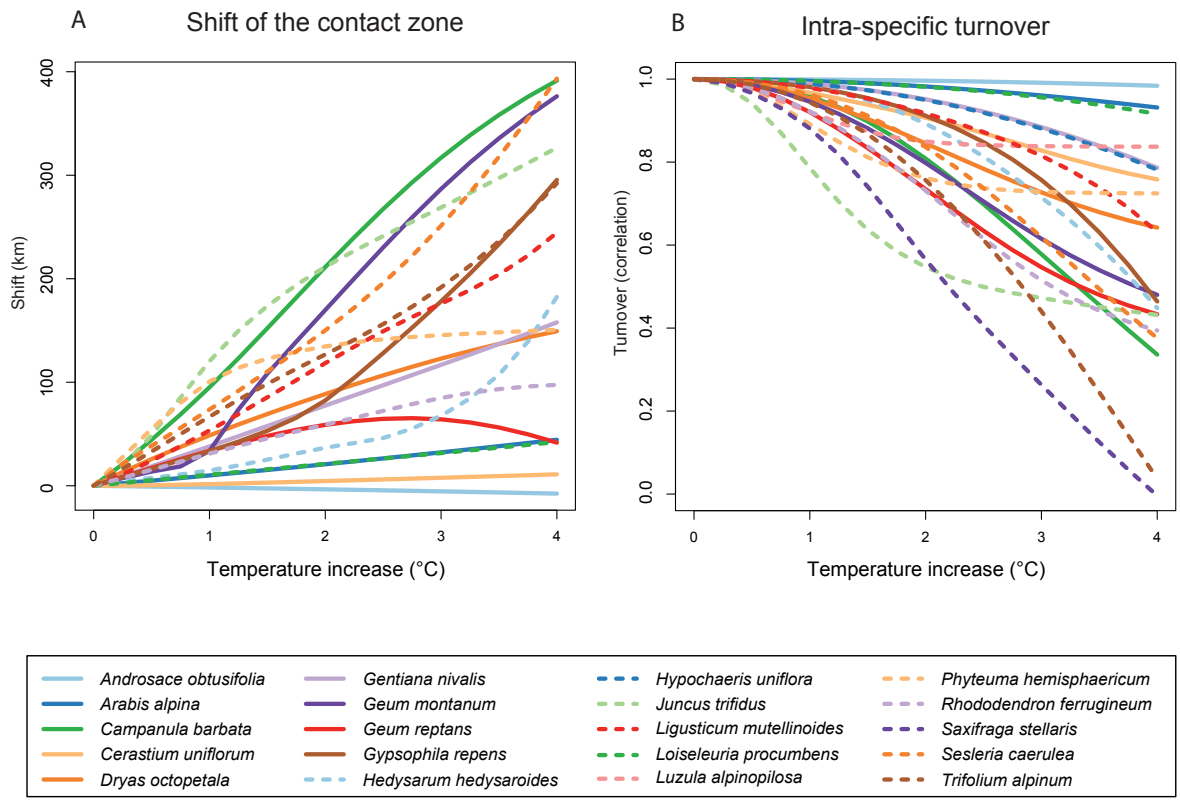


Figure 4.5: Predictions of shifts and intra-specific turnover in Alpine plant species under increased temperature scenarios. Ancestry coefficients and contact zone locations were predicted for temperatures increasing from 0 °C to 4 °C in steps of 0.25 °C (x-axes). A. Contact zone displacement: distance (km) measured along a south-western to north-eastern transect. B. Intra-specific turnover measured by the correlation between inferred and predicted ancestry coefficients for all individuals and all clusters.

Demographic history.

In order to identify the relative contribution of adaptation versus demographic history in shaping the contemporary population genetic structure of alpine plant species, we evaluated signatures of demographic events in our datasets. Alpine plant phylogeography is strongly influenced by post-glacial expansion after the ice age (Schönswetter *et al.* 2005). During range expansions, recurrent founder effects and genetic drift at the expansion front are expected to generate gradients in diversity decreasing from the source of expansion (Austerlitz *et al.* 2000; François *et al.* 2008). In agreement with previous observations (Thiel-Egenter *et al.* 2009), we found that the locations of areas of maximal diversity were species-specific (Fig. S2). This indicated that plant species probably followed distinct post-glacial colonization routes, associated with distinct ecological constraints during dispersal (Alvarez *et al.* 2009). Therefore the clusters observed in the south-western Alps for 18 species can hardly be explained by shared demographic history. This lack of association indicated that local adaptation contributed in shaping population genetic structure during the history of the species studied.

To test whether gradients in allele frequencies and high correlations between population genetic structure and environment could be explained by demographic history, we performed simulations of post-glacial re-colonization scenarios based on spatially explicit neutral coalescent models. In simulations, the source of expansion was either located in the southwest or the northeast of the Alps. For most species these two areas contained the maximum of their genetic diversity, and thus could correspond to putative glacial refugia. For the simulated data sets, the distribution of correlations between estimated admixture coefficients and those predicted from the current environmental covariates had their mean around 0.64 (SD = 0.13), whereas the actual mean was around 0.91 (SD = 0.04) for the 20 plant species studied here. The test based on those correlations was significant in 19 of the 20 species ($P < 0.01$ in all cases). For the demographic simulations, these results indicated that the environmental variables did not provide good explanations for the inferred population genetic structure. In contrast, the environmental variables provided high predictive power in the empirical plant data used here. In conclusion, neutral range expansion processes generated significantly lower levels of gene/environment associations than those observed in the data set.

Finally, we further examined the differences in amplitude between the observed allele frequency clines and those obtained from demographic simulations. Differences were significant in 15 (83%) of the 18 species exhibiting warm clusters when expansion started from the northeast. The differences were significant in 8 (44%) of the 18 species when expansion started from the southwest (Fig. S4). These results suggest that the gradients observed at outlier loci reflected the combined effects of neutral demographic processes and adaptation of plant genetic variation.

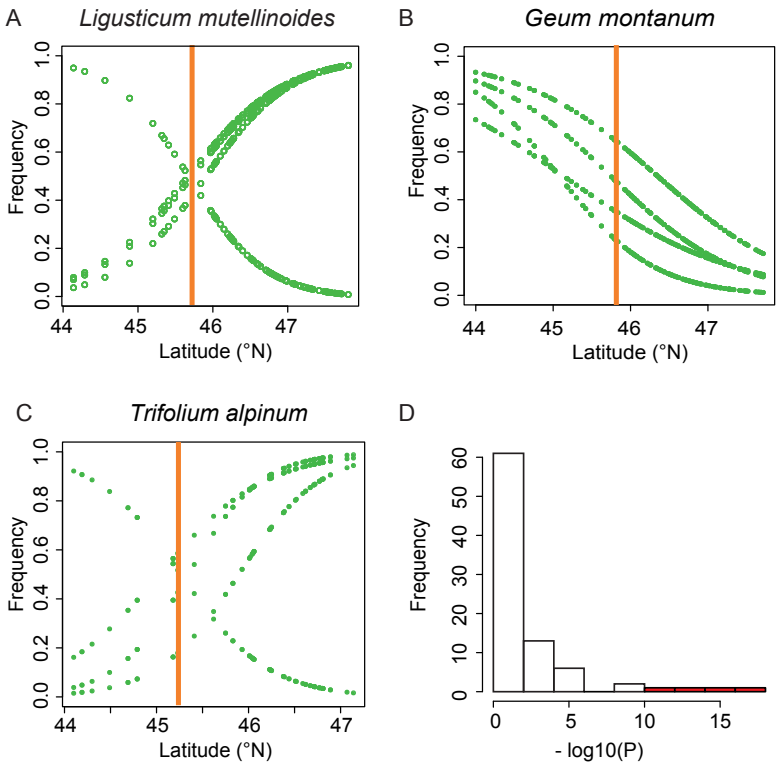


Figure 4.6: Latitudinal clines at outlier loci of alpine plant species. Allele frequency gradients at loci detected as outliers exemplified in the three species A. *Ligusticum mutellinoides*, B. *Geum montanum*, and C. *Trifolium alpinum*. Each dotted line represents one locus, the vertical bars correspond to the 50% frequency value. D. Histogram of P-values for *Geum montanum*. The shaded area in the histogram corresponds to the four loci displayed in panel B. Latitudinal clines for other species are displayed in Figure S3.

Discussion

We combined the philosophy of species distribution models with population genetic methods to predict the neutral component of the intra-specific response of alpine plants to global warming.

Modeling assumptions.

As with species distribution models, the results obtained with our ancestry distribution models rely on obvious simplifications. First, biotic interactions (*e.g.*, competition with other species) are supposed to be constant in space and time. This simplification is also present in species distribution models Guisan and Thuiller (2005), and implies that species are studied separately. Second, no adaptation to changing conditions occurs by new mutations during the period of environmental change. This assumption means that environmental change occurs within short timescales. In addition, while ancestry distribution models consider genotypes, phenotypes are not included in analyses. Thus the effects of phenotypic plasticity cannot be measured.

Standard species distribution models assume a global niche and niche conservatism over time (Gallien *et al.* 2010). By substituting species occurrence data with genetic data and applying ancestry distribution models, the global niche assumption is replaced by intra-specific niche variation. Though quantities computed by ancestry distribution models are summaries of allele frequencies at neutral markers, model predictions suppose that population structure is, at least partly, associated with alleles that played a role in adaptation to warmer conditions (Davis and Shaw 2001). For contiguous populations, adaptation to a warming climate will be triggered by gene flow from populations already adapted to areas of warm climate within the species range (Holderegger and Wagner 2008). Genes involved in this adaptation, allowing populations to keep track with environmental change, are probably not observable directly from the data used here, but are nevertheless expected to follow the same patterns of migration as those forecasted for neutral alleles (Bierne *et al.* 2011).

In summary, ancestry distribution models assume that gene migration represents the main component in the response of plant populations to environmental modifications. The future migration rates of neutral alleles are estimated through predicted changes in genetic admixture and by using geographic mapping of future genetic clusters. Because biotic interactions, phenotypic plasticity and new adaptive mutations might reduce the relative importance of gene flow, the model used here possibly leads to over-estimations of the rate at which genes may migrate (Aitken *et al.* 2008).

Interpretations of results.

In 18 of the 20 species studied here, populations in warmer environments showed higher predisposition to disperse out of their natural range in response to rapid climate change. For example, one of these species, *Rhododendron ferrugineum*, was partitioned into three genetic clusters. The warm cluster, having its center at around 44.5 °N, 6.76 °E, showed a difference of 3 °C in average annual minimum daily temperatures with the other two clusters. Forecasting the rate of gene migration from projections of the colonization edge of the warm cluster on a geographic map, yielded values around 0.9 km/year for 4 °C of temperature warming in the next century. Since ancestry distribution models favored a migration hypothesis, the magnitude of shift estimated for *Rhododendron ferrugineum* and for other species is likely to represent an upper bound of future rates of migration of neutral alleles. With the mean annual temperature rising by 2 °C, the plant populations studied showed a transversal shift within the range 5 - 212 km in the next 100 years. For all species, the migration process will create changes in population genetic structure that are species-specific and that could imply considerable changes in ecological interactions within plant communities (Engler *et al.* 2011).

A limitation of neutral population genetic models, however, is that they ignore the selective pressures encountered during plant migration. For example, two species, *Luzula alpinopilosa* and *Phyteuma hemisphaericum*, did not exhibit differentiated clusters with respect to climate. For these two species, contact zone shifts were more difficult to predict than for species exhibiting significantly warmer clusters, and the corresponding results could be interpreted as examples of population structure rather being shaped by demographic history or unknown environmental factors (soil, microenvironment) rather than by climatic gradients.

Phenotypic plasticity and rapid adaptation in a historical perspective.

As phenotypic plasticity and rapid adaptation are not included either in ancestry distribution models or in traditional species distribution models (Chevin *et al.* 2010), it is not yet possible to track the competitive and adaptive abilities of invaded gene pools and their subsequent capacity to reduce the level of introgression by the invading clusters. However, from a theoretical perspective, when assuming that central or northern alpine genotypes could quickly adapt to a changing climate and/or are plastic with respect to warmer conditions, it could be likely that the range expansion of what we describe here mainly as the northward expansion of south-western warm clusters would be more limited. Such limitation might be partly supported by the Pleistocene history of alpine plants, in which survival of gene pools during cold periods occurred all along the southern margin of the Alps (and not only in the southwest), followed by progressive re-colonization of de-

glaciated areas towards the North during warmer periods (Schönswetter *et al.* 2005). Evaluating whether this is a rational scenario would only be possible by modeling the past history of a species' gene pool backwards until the last glacial maximum, a task whose realistic implementation in our models is currently not possible given the lack of precise data on the location of realized glacial refugia for each of the 20 studied species (Alvarez *et al.* 2009).

Long-distance dispersal.

In plants, dispersal occurs through seeds, while gene flow and allele migration may occur both by seed and pollen dispersal. Despite evidence of adaptation to relatively warm environment in alpine plants, it is not assured that the migration rates of plant species could keep pace with fast rates of environmental change (Malcolm *et al.* 2002), particularly if northern gene pools show little phenotypic plasticity and/or slow local adaptation to warmer temperatures. Long-distance dispersal ability is thus seen as a key parameter for plant species to respond to climate change (Higgins and Richardson 1999). Although long-distance dispersal events might be rare and difficult to observe, they are known to have a strong impact on population structure and adaptation (Nathan 2006; Kunstler *et al.* 2007). A prominent role of wind dispersal has been recognized in alpine habitats (Tackenberg and Stöcklin 2008), and alpine species have thus a good chance to be dispersed by wind or animals over long distances. The required level of dispersal per generation, estimated within the range of 0.5 - 2.1 km by our ancestry distribution models, represents realistic rates for alpine species (Tackenberg and Stöcklin 2008), especially given that this estimates might be tempered by phenotypic plasticity or existing adaptation of resident populations.

Gene surfing.

Dispersal creates opportunities for moving populations to interbreed with resident populations at higher latitudes. The ancestry distribution models predict genetic admixture of plant populations from warm areas with populations from currently colder environments. Invasion by south-western populations certainly has the potential to increase local standing genetic variation and allowing for new adaptive potentials in invaded areas. We are however cautious about interpreting replacement of resident populations by invading ones for two reasons. First, local plants often perform significantly better than foreign plants at their site of origin (Leimu and Fischer 2008). Though better adapted to a warmer climate, there is no evidence that south-western populations could adapt swiftly to the conditions encountered in newly colonized environments (Walther *et al.* 2009). Ancestry distribution models accounted for differences in habitat between distinct clusters, and favored migration toward similar environments, but they did not

assess the adaptive potential of organisms in those environments. Second, during population movement, the genomes of invaders are theoretically predicted to be massively introgressed by resident alleles (Petit 2004; Jump and Penuelas 2005; Currat *et al.* 2008), which would slow down the migration process. On the other hand, invasions by south-western populations also have the potential to increase local standing genetic variation in invaded areas, which could balance the previous effect. Examining these questions would require further investigations using a higher density of genetic markers, and a more precise modeling of genetic drift.

Trailing and altitudinal edges.

The prediction of the presence of plant species at the edge of their distribution ranges is rendered difficult by the nature of the data used (individual genotypes were mainly sampled from species' distribution centers) and a lack of occurrence data for species in these areas. The predictions of population genetic models may also underestimate the risk of extinction at the trailing edge of the migration of plant populations (Hampe and Petit 2005) and at the upper altitudinal edge (Jump *et al.* 2009b). Drought and dieback events occurring at the trailing edge of shifting clusters have a high likelihood to distort population structure (Jump *et al.* 2009a). Since selection pressures acting on trailing edge populations may be intense, species harboring high levels of standing genetic variation in the southwest should have a greater chance to adapt than species with low levels of adaptive genetic diversity in this geographic area (Table 4.2). Examining these questions in further details will require coupling ancestry distribution models with more traditional species distribution models using absence-presence or abundance data and stratified niche sampling (Albert *et al.* 2010).

Conclusions

Rising global temperatures alter the distribution of plant species (Dukes and Mooney 1999; Walther *et al.* 2009; Pereira *et al.* 2010). Here, we predicted that global change would also create intra-specific turnover impacting both on intra-specific genetic variation within plant communities. Since migration within a species range would be followed by admixture with resident populations, local levels of genetic diversity might not experience a decrease, and some populations might even increase in genetic diversity. Although distributional changes are inevitable for most alpine plant species, intra-specific variation seems to be weakly affected by global warming at least considering a scenario in which the increase in the mean annual temperature is lower than 2 °C. For changes greater than 2 °C, however, our ancestry distribution models predicted higher levels of intra-specific turnover in many plant species, corresponding to low correlations between actual and

projected population genetic structure. Though the predictions of population genetic models do not account for rapid adaptation, phenotypic plasticity or biotic interactions, they agreed with the tolerance to temperature increase predicted by non-genetic observations (Theurillat and Guisan 2001). It is however important to keep in mind that anthropogenic changes in land use and land cover will surely affect migration patterns as climate change in the coming century.

Acknowlegments

We are grateful to the INTRABIODIV Consortium for providing us the plant data. We also thank Felix Gugerli, Renaud Vitalis, Etienne Klein for fruitful discussions, José Diniz-Filho, Victoria Sork and three anonymous reviewers for their useful comments. FJ and OF were supported by a CNRS grant on "Computational Landscape Genetics"and by a grant from "La Région Rhône-Alpes". SM was supported by the Institut Universitaire de France. NA and RH were funded by the Swiss National Science Foundation (Ambizione fellowship PZ00P3_126624 and Sinergia AVE CRSI33_127155/1), respectively. WT acknowledged support of the ANR EVORANGE (ANR-09-PEXT-011) project.

Supporting Materials http://membres-timc.imag.fr/Flora.Jay/forecastSM.zip

Figure S1. Population structure for 20 Alpine plant species. Sample locations in warm clusters are given in red.
Figure S2. Interpolated gradients of genetic diversity for 20 Alpine plant species.
Figure S3. Latitudinal clines at outlier loci for 20 Alpine plant species.
Figure S4. Null distribution of the shape parameter of the cline at the strongest outlier locus for 18 Alpine plant species under scenarios of range expansion from the NE and the SW.
Supplementary Movie SM1. Movement of contact zones for 18 Alpine plant species.

Chapitre 5

POPS : un logiciel pour la prédiction de la structure génétique des populations

Cet article présente le logiciel `POPS` (*Prediction Of Population Structure*). Nous détaillons les méthodes implémentées dans `POPS` et ses principales applications. Celles-ci sont illustrées à l'aide de l'analyse d'un jeu de données d'une espèce de plante alpine. Un manuel technique est disponible pour guider plus précisément l'utilisateur dans l'utilisation de `POPS` à partir de l'interface graphique ou de lignes de commande (Jay 2011).

Article C

F. Jay, O. François, E.Y. Durand, and M.G.B Blum. POPS: A software for prediction of population genetic structure using latent regression models. *Submitted*

Abstract

The software `POPS` performs inference of population genetic structure using genetic, geographic and environmental data. Based on a hierarchical Bayesian framework using latent regression models, `POPS` implements algorithms to estimate admixture proportions and cluster membership probabilities at the individual level, and the correlations between those quantities and environmental data. The models implemented in `POPS` improve the estimation of population genetic structure by using geographic and environmental information, and estimate the effects of environmental factors on population structure. Similarly to species distribution models, `POPS` defines ancestry distribution models, that allow their users to forecast individual cluster membership and admixture proportions under scenarios of environmental change. A typical use of `POPS` is to evaluate how the spatial population genetic structure of a given species would be modified by climate change. We illustrate the program features by applying it to molecular markers genotyped in an alpine plant species.

Keywords latent class regression models, mixture models, MCMC, population genetic structure, environmental covariates

5.1 Introduction

Ascertaining population genetic structure from multi-locus genotype data sets and identifying the environmental variables that correlate with this structure is important to several domains, for example, in population genetics, molecular ecology, landscape genetics, conservation genetics or genetic epidemiology (*e.g.*, Manel *et al.* 2003; Storfer *et al.* 2007; Balding 2006; Balkenhol *et al.* 2009; Segelbacher *et al.* 2010), and associations between population genetic structure and ecological variables have been frequently reported in the recent literature (for examples, see Duminil *et al.* 2007; Aitken *et al.* 2008; Sork *et al.* 2010; Lee and Mitchell-Olds 2011). Population genetic structure can be estimated by identifying *genetic clusters*, that are defined as genetically divergent groups of individuals commonly arising from isolation of populations, and by computing the probability of membership of individuals to the genetic clusters (Davies *et al.* 1999; Pritchard *et al.* 2000a). Because individual membership can be shared among several genetic clusters, an alternative objective is to infer individual *admixture proportions*, the relative contributions of distinct ancestral populations to an admixed genome. The first efforts to infer genetic clusters or individual admixture proportions using Bayesian modeling date to the computer programs `STRUCTURE` and `PARTITION` (Pritchard *et al.* 2000a; Dawson and Belkhir 2001). Although these algorithms and their derivatives are widely used to study the influence of landscape features on evolutionary processes, they do not incorporate information from environmental variables.

In this article, we present `POPS`, a software that implements Bayesian clustering algorithms based on genetic, geographic and environmental covariates. The idea of `POPS` is that individuals sharing similar environmental conditions and geographically close to each other are also likely to share genetic ancestry. The principle of `POPS` consists of jointly clustering individuals into genetic groups and estimating the effects of covariates on individual membership and admixture proportions. Identifying genetic clusters can be viewed as an instance of unsupervised learning based on multivariate categorical variables. `POPS` uses latent regression models that include covariates in models of latent responses (*e.g.* Bandeen-Roche *et al.* 1997; Chung *et al.* 2006; Linzer and Lewis 2011). Compared to programs based on genotypic information only, the benefit of including environmental variables has been shown from both simulated and biological data (Durand *et al.* 2009b; Jay *et al.* 2011a). In addition, fitting hidden regression models enables the users of `POPS` to predict cluster membership or admixture proportions from geographic and environmental variables. For example, in the context of climate change, `POPS` was used to forecast changes in population genetic structure of alpine plant species in response to temperature increase (Jay *et al.* 2011b).

`POPS` is implemented in the `C++` programming language and can be run from a command-line engine or a graphical user interface. The program takes input data files in a format

compatible with existing clustering algorithms like STRUCTURE (Pritchard *et al.* 2000a) and TESS (Chen *et al.* 2007). POPS returns textual and graphical results of inferred and predicted membership probabilities and admixture coefficients, estimated regression coefficients, allele frequencies in the estimated clusters, and values of criteria for model selection. The hidden regression models implemented in POPS differ according to whether or not the data set is assumed to contain admixed genotypes. In sections 2 and 3, we describe the hierarchical models used by POPS and their implementation with Markov Chain Monte Carlo (MCMC) algorithms. In section 4, we explain how to use the software from its graphical interface and from the command-line engine. In section 5, we illustrate the use of POPS on a particular example data set and show the main features of the program.

Ascertaining population genetic structure from multi-locus genotype data sets and identifying the environmental variables that correlate with this structure is important to several domains, for example, in population genetics, molecular ecology, landscape genetics, conservation genetics or genetic epidemiology (*e.g.*, Manel *et al.* 2003; Storfer *et al.* 2007; Balding 2006; Balkenhol *et al.* 2009; Segelbacher *et al.* 2010), and associations between population genetic structure and ecological variables have been frequently reported in the recent literature (for examples, see Duminil *et al.* 2007; Aitken *et al.* 2008; Sork *et al.* 2010; Lee and Mitchell-Olds 2011). Population genetic structure can be estimated by identifying *genetic clusters*, that are defined as genetically divergent groups of individuals commonly arising from isolation of populations, and by computing the probability of membership of individuals to the genetic clusters (Davies *et al.* 1999; Pritchard *et al.* 2000a). Because individual membership can be shared among several genetic clusters, an alternative objective is to infer individual *admixture proportions*, the relative contributions of distinct ancestral populations to an admixed genome. The first efforts to infer genetic clusters or individual admixture proportions using Bayesian modeling date to the computer programs STRUCTURE and PARTITION (Pritchard *et al.* 2000a; Dawson and Belkhir 2001). Although these algorithms and their derivatives are widely used to study the influence of landscape features on evolutionary processes, they do not incorporate information from environmental variables.

In this article, we present POPS, a software that implements Bayesian clustering algorithms based on genetic, geographic and environmental covariates. The idea of POPS is that individuals sharing similar environmental conditions and geographically close to each other are also likely to share genetic ancestry. The principle of POPS consists of jointly clustering individuals into genetic groups and estimating the effects of covariates on individual membership and admixture proportions. Identifying genetic clusters can be viewed as an instance of unsupervised learning based on multivariate categorical variables. POPS uses latent regression models that include covariates in models of latent responses (*e.g.* Bandeen-Roche *et al.* 1997; Chung *et al.* 2006; Linzer and Lewis 2011). Compared to programs based on genotypic information only, the benefit of including environmen-

tal variables has been shown from both simulated and biological data (Durand *et al.* 2009b; Jay *et al.* 2011a). In addition, fitting hidden regression models enables the users of POPS to predict cluster membership or admixture proportions from geographic and environmental variables. For example, in the context of climate change, POPS was used to forecast changes in population genetic structure of alpine plant species in response to temperature increase (Jay *et al.* 2011b).

POPS is implemented in the C++ programming language and can be run from a command-line engine or a graphical user interface. The program takes input data files in a format compatible with existing clustering algorithms like STRUCTURE (Pritchard *et al.* 2000a) and TESS (Chen *et al.* 2007). POPS returns textual and graphical results of inferred and predicted membership probabilities and admixture coefficients, estimated regression coefficients, allele frequencies in the estimated clusters, and values of criteria for model selection. The hidden regression models implemented in POPS differ according to whether or not the data set is assumed to contain admixed genotypes. In sections 2 and 3, we describe the hierarchical models used by POPS and their implementation with Markov Chain Monte Carlo (MCMC) algorithms. In section 4, we explain how to use the software from its graphical interface and from the command-line engine. In section 5, we illustrate the use of POPS on a particular example data set and show the main features of the program.

5.2 Models

We consider genetic data that consist of a matrix, X, with N rows and L columns where N is the number of individuals and L is the number of genetic markers or *loci*. Each genetic marker is coded as a categorical variable where each category corresponds to a given *allele*. In polyploid species there are $A \times N$ rows instead of N rows in the data matrix corresponding to the A chromosomes carried by each individual.

POPS uses models in which the membership of individuals to a genetic cluster or their admixture proportions are latent variables. The program implements two main classes of models according to whether we assume genetic admixture or not. A model without admixture assumes that each individual belongs to a single cluster, whereas a model with admixture assumes that each genome may originate from several populations.

The statistical models implemented in POPS can be viewed as extensions of the models implemented in STRUCTURE (Pritchard *et al.* 2000a). Previous extensions of STRUCTURE have improved its models by accounting for biological processes (Falush *et al.* 2003; Gao *et al.* 2007; Shringarpure and Xing 2009), by increasing computational performances using faster algorithms (Chen *et al.* 2006; Wu *et al.* 2006; Alexander *et al.* 2009), or by including geographical data (*e.g.* François *et al.* 2006; Corander *et al.* 2008; Durand *et al.* 2009b).

5.2.1 Models without admixture

Assuming that the number of genetic clusters is equal to K, the algorithm infers latent cluster labels $Z_i \in \{1, \ldots, K\}$ that correspond to the membership of each individual i, $i = 1, \ldots, N$. To infer the membership probabilities (or coefficients), `POPS` considers the allele frequencies $P = (p_{k\ell j})$, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$ and $j = 1, \ldots, J_\ell$. The entries of P correspond to the frequency of allele j at locus ℓ in population k, and J_ℓ is the number of distinct alleles observed at locus ℓ. Let us denote by A the ploidy of the studied organisms and by $\left(x_\ell^{(i,1)}, \ldots, x_\ell^{(i,A)}\right)$ the genotype of individual i at locus ℓ. We assume that a set of p covariates is available, and we denote by $\tilde{X}$ the $N \times (p+1)$ design matrix containing the value 1 in the first column and the p covariates in the next p columns. `POPS` computes the posterior distribution of the multidimensional parameters Z, P, and β

$$\Pr(Z, P, \beta | X, \tilde{X}) \propto \Pr(X|Z, P)\Pr(Z|\tilde{X}, \beta)\Pr(\beta)\Pr(P), \tag{5.1}$$

where $\Pr(P)$ denotes a prior distribution, β is a multidimensional matrix of regression coefficients and $\Pr(\beta)$ its prior distribution. In this statistical model, the likelihood is defined as the product of the following terms (Pritchard *et al.* 2000a)

$$\Pr(x^{(i)}|Z_i = k, P) = \prod_{\ell=1}^{L} \prod_{a=1}^{A} p_{k\ell x_\ell^{(i,a)}} \tag{5.2}$$

for all $i = 1, \ldots, N$. The prior distributions on the allele frequencies at each locus within each population are Dirichlet distributions

$$p_{k\ell \cdot} \sim \mathcal{D}(\lambda, \ldots, \lambda), \tag{5.3}$$

where λ is set to 1 by default. Note that other values of λ can be chosen by the users of the program.

To include covariates, `POPS` considers *latent class regression* models (Dayton and Macready 1988; Hagenaars and McCutcheon 2002; Linzer and Lewis 2011). In these models, the membership labels $Z_1, \ldots, Z_N$ are regressed on the set of covariates using a multinomial probit model (Jay *et al.* 2011a). To emphasize the difference between geographic and environmental covariates, we denote by $\tilde{X}^{\mathrm{S}}$ the subset of spatial covariates and $\tilde{X}^{\mathrm{E}}$ the subset of environmental covariates. The probit model considers the vector of latent variables $C_i = (C_{i,1}, \ldots, C_{i,K-1})$ as response variables in $K-1$ regression equations (Albert and Chib 1993)

$$C_{i,k} = \tilde{X}_i^{\mathrm{E}} \beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}}) \beta_k^{\mathrm{S}} + \epsilon_{i,k}, \tag{5.4}$$

$$\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,K-1}) \sim \mathcal{N}(0, \mathrm{Id}),$$

where Id is the identity matrix, β_k^{E} and β_k^{S} are 2 column vectors of regression parameters in the k^{th} regression equation, and f is a polynomial function of degree inferior to 3. For example, to include a quadratic trend surface, f can be set to a polynomial function of degree 2. The value of Z_i is obtained from the value of C_i as follows

$$Z_i = \begin{cases} K & \text{if } \max_h C_{i,h} < 0 \\ k & \text{if } \max_h C_{i,h} > 0 \text{ and } \max_h C_{i,h} = C_{i,k}\,. \end{cases} \tag{5.5}$$

Since no regression equation is associated to the K^{th} cluster, the multinomial probit model is not symmetric with respect to the K clusters. The values of the regression parameters are then defined with respect to the K^{th} cluster, called the *reference* cluster.

5.2.2 Models with admixture

Models with admixture assume that the genome of an individual descents from K ancestral populations. To define models with admixture, we introduce a matrix, Q, storing the admixture proportions for all the individuals. Each element of the matrix, q_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, K$, corresponds to the proportion of individual i's genome that originates from cluster k. In admixture models, there is one cluster label $z_\ell^{(i,a)}$ for each allele copy at each locus, and Z is a matrix of size $N \times L \times A$. The model has 2 additional multidimensional parameters, and the posterior distribution of (Z, P, β, Q, α) is given by

$$\Pr(Z, P, \beta, Q, \alpha | X, \tilde{X}) \propto \Pr(X|Z, P)\Pr(Z|Q)\mathrm{P}(Q|\alpha)\Pr(\alpha|\tilde{X}, \beta)\Pr(\beta)\Pr(P). \tag{5.6}$$

As in model without admixture, the prior distribution on P is given by equation (5.3). The conditional distribution $\Pr(Z|Q)$ is given by

$$\Pr(z_\ell^{(i,a)} = k|Q) = q_{ik},\ i = 1, \ldots, N,\ a = 1, \ldots, A,\ \ell = 1, \ldots, L.$$

The admixture coefficients of each individual are assumed to follow a Dirichlet distribution

$$q_{i\cdot}|\alpha \sim \mathcal{D}(\alpha_{i1}, ..., \alpha_{iK}), \tag{5.7}$$

where $\alpha_{i\cdot}$ is proportional to the expected amount of individual admixture from each ancestral population.

In `POPS`, the covariates, $\tilde{X}$, are included in a regression model for the parameter α. The latent regression model was initially developed to include spatial information in the approach of Durand *et al.* (2009b). Denoting by $\tilde{X}_i^{\mathrm{S}}$ the set of spatial covariates, and by

$\tilde{X}_i^{\mathrm{E}}$ the environmental covariates, `POPS` considers the following log-normal model

$$\log(\alpha_{ik}) = \tilde{X}_i^{\mathrm{E}}\beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}} + y_{ik}, \tag{5.8}$$

where y_{ik}, is a zero-mean conditional autoregressive Gaussian model (CAR; Besag 1975; Ripley 1981; Vounatsou *et al.* 2000). In equation (5.8) the intermediate term represents a spatial trend and accounts for broad-scale spatial patterns, the last term corresponds to a spatially autocorrelated residual that accounts for a local spatial effect and the first term measures the effect of environmental covariates once spatial effects have been corrected (Lichstein *et al.* 2002). Note that this framework encompasses partial regression models and techniques based on uncorrelated combinations of environmental variables.

In the CAR model, the distribution of y_{ik} depends on the values of this variable for neighboring individuals. The model is conditionally Gaussian with mean

$$E[y_{ik}|y_{jk}, j \neq i] = \rho_k \sum_{j\sim i} w_{ij} y_{jk}, \tag{5.9}$$

and variance

$$Var(y_{ik}|y_{jk}, j \neq i) = \sigma_k^2, \tag{5.10}$$

where ρ_k is the magnitude of the spatial neighborhood effect in cluster k, w_{ij} are weights that determine the relative influence of j on i, and σ_k^2 is a variance parameter for the cluster k. The neighborhood is obtained from a Dirichlet tessellation built on individual spatial coordinates (François *et al.* 2006). The weights are functions of the great-circle distance between individuals i and j, d_{ij},

$$w_{ij} = \exp(\frac{d_{ij}}{\theta}), \tag{5.11}$$

where θ is equal to the mean value of great-circle distances between the individual locations.

5.2.3 Differences and similarities between models with and without admixture.

Models with and without admixture are biologically and statistically different. Each model underpins specific biological hypotheses, depending on whether or not the data contain admixed individuals. For example, clines of admixture proportions are expected when secondary contact between several genetically divergent populations occurs (Barton and Hewitt 1985). The choice of a particular statistical model is crucial as it can

impact the estimation of population genetic structure, especially of the admixture clines. François and Durand (2010b) showed that models without admixture are not robust to the presence of admixed individuals in the data, whereas admixture models are robust to an absence of admixture in the sample. On the other hand, when no individuals are admixed, models with admixture may find some artificial level of admixture for non-admixed individuals (Alexander and Lange 2011).

Statistically, models without admixture are based on a refinement of latent class models to include covariates. The principle of a latent class model is that the distribution of the observed data is a mixture of K class-specific distributions

$$\Pr(x^{(i)}) = \sum_{k=1}^{K} \Pr(Z_i = k)\Pr(x^{(i)}|Z_i = k). \tag{5.12}$$

Within each class, the loci are assumed to be independent and the distribution of allele counts at each locus is multinomial with the allelic frequency $p_{k\ell\cdot}$. In addition, we have

$$\Pr(x^{(i)} = (x_1^{(i,1)}, \ldots, x_L^{(i,A)})|P) = \sum_{k=1}^{K} \Pr(Z_i = k|P) \prod_{\ell=1}^{L} \prod_{a=1}^{A} \Pr(x_\ell^{(i,a)}|Z_i = k, P) \tag{5.13}$$

$$= \sum_{k=1}^{K} \Pr(Z_i = k) \prod_{\ell=1}^{L} \prod_{a=1}^{A} p_{k\ell x_\ell^{(i,a)}} \tag{5.14}$$

In `POPS` we are interested in modeling distributions of the following form

$$\Pr(x^{(i)}|P, \tilde{X}) = \sum_{k=1}^{K} \Pr(Z_i = k|\tilde{X})\Pr(x^{(i)}|Z_i = k, P, \tilde{X}).$$

When $\Pr(Z_i = k|\tilde{X}) = \Pr(Z_i = k)$, the model is called a *latent class regression model*, and is a special case of the *mixture regression models* (DeSarbo and Cron 1988; Wedel and DeSarbo 1994). These models have been implemented by Leisch (2004) in the `R` package **flexmix**.

If the covariates have no influence on the distribution of the data in a given class, *i.e.* $\Pr(X|Z_i = k, \tilde{X}) = \Pr(X|Z_i = k)$, the model is called a *concomitant-variable latent class model* (Dayton and Macready 1988) or a *latent class feed-forward model* (*LC feed-forward*; Vermunt and Magidson 2003). This model was implemented by Linzer and Lewis (2011) in the `R` package **poLCA**. To include the covariates $\tilde{X}$ in models without admixture, we considered a *LC feed-forward* model. The probabilities $\Pr(X|Z_i = k, \tilde{X})$ are then modeled with the probit regression model in equations (5.4) and (5.5).

Models with admixture do not fit into latent class models *stricto sensu*, because there is one latent class variable per individual and per marker. In the case of admixture models,

equation (5.12) extends to

$$\Pr(x^{(i)}) = \prod_{\ell}^{L} \prod_{a}^{A} \sum_{k=1}^{K} \Pr(z_l^{(i,a)} = k|q_{ik}) \Pr(q_{ik}) \Pr(x_l^{(i,a)}|z_l^{(i,a)} = k). \tag{5.15}$$

The inclusion of covariates in admixture models is similar to the inclusion of covariates in *LC feed-forward* models,

$$\Pr(x^{(i)}) = \prod_{\ell}^{L} \prod_{a}^{A} \sum_{k=1}^{K} \Pr(z_l^{(i,a)} = k|q_{ik}) \Pr(q_{ik}|\tilde{X}) \Pr(x_l^{(i,a)}|z_l^{(i,a)} = k), \tag{5.16}$$

where $\Pr(q_{ik}|\tilde{X})$ is obtained from equations (5.7) and (5.8). Thus, models with admixture are more similar to *LC feed-forward* models than to classical latent class regression models.

5.3 Inference and prediction

In this section we describe the Markov chain Monte Carlo algorithms implemented in `POPS` to perform parameter inference.

5.3.1 Models without admixture

To sample from the posterior distribution $\Pr(Z, P, C, \beta|X, \tilde{X})$, `POPS` implements an MCMC algorithm using Gibbs sampling updates.

UPDATING P. This step is the same as in the software `STRUCTURE` (PRITCHARD *et al.* 2000a). It is performed by simulating the allele frequencies as follows

$$p_{k\ell\cdot}|X, Z \sim \mathcal{D}(\lambda + n_{k\ell 1}, \ldots, \lambda + n_{k\ell J_\ell}), \tag{5.17}$$

where $n_{k\ell j}$ denotes the number of copies of allele j in population k at locus ℓ, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$, $j = 1, \ldots, J_\ell$.

UPDATING (C, Z). Since Z can be obtained from C in a deterministic fashion, Z and C are updated simultaneously. Using the Bayes formula, the joint conditional distribution of (C, Z) can be written as

$$\Pr(C, Z|\beta, P, X, \tilde{X}) \propto \Pr(X|\beta, P, Z)\Pr(C|\beta, \tilde{X})\Pr(Z|C)$$

For each i, (C_i, Z_i) is simulated using the following rejection algorithm.

- Step 1. Simulate the couple (C_i, Z_i) from the multinomial probit model by generating C_i from the regression equation (5.4) and determine $Z_i = k$ with the rule given in equation (5.5).

– Step 2. Accept the couple (C_i, Z_i) with probability

$$\frac{\Pr(X_i = x^{(i)}|P, Z_i = k)}{\max_k \Pr(X_i = x^{(i)}|P, Z_i = k)},$$

otherwise return to step 1. The probability $\Pr(X_i = x^{(i)}|P, Z_i = k)$ is computed as in equation (5.2).

UPDATING β. `POPS` uses a diffuse prior distribution on β, $\beta \sim \mathcal{N}(0, B^{-1})$, with $B = 0$. The Gibbs sampler proceeds by updating values of β using its conditional distribution (ALBERT and CHIB 1993)

$$\beta_k|C \sim \mathcal{N}(V\Omega^T C_{\cdot k},\ V), \tag{5.18}$$

where $V = (\Omega^T\Omega)^{-1}$, and Ω is the concatenation of $\tilde{X}^{\mathrm{E}}$ and $f(\tilde{X}^{\mathrm{S}})$.

5.3.2 Models with admixture

To sample from the posterior distribution in models with admixture, `POPS` uses a hybrid MCMC algorithm. Unless specified, updates are done using Gibbs samplers. Updates of P, Q and Z are similar to those used in the algorithm of `STRUCTURE` (PRITCHARD *et al.* 2000a). The other updates are similar to those used in the algorithm of `TESS` (DURAND *et al.* 2009b).

UPDATING P. Given X and Z, P is simulated according to an equation similar to equation (5.17) describing models without admixture. In models with admixture, individual cluster labels are simulated for each allele copy of each locus.

UPDATING Q. The admixture coefficients of individual i are sampled from a Dirichlet distribution

$$q_{i\cdot}|X, Z, \alpha \sim \mathcal{D}(\alpha_{i1} + m_{i1}, \ldots, \alpha_{iK} + m_{iK}) \tag{5.19}$$

where m_{ik} is the number of allele copies of individual i that originate from population k, $m_{ik} = \#((\ell, a)/z_\ell^{(i,a)} = k)$.

UPDATING Z. A cluster label $z_\ell^{(i,a)}$ is simulated independently for each i, a, ℓ from

$$\Pr(z_\ell^{(i,a)} = k|P, Q, X) = \frac{q_k^{(i)} \Pr(x_\ell^{(i,a)}|p_{k\ell x_\ell^{(i,a)}},\ z_\ell^{(i,a)} = k)}{\sum_{k'=1}^K q_{k'}^{(i)} \Pr(x_\ell^{(i,a)}|p_{k'\ell x_\ell^{(i,a)}}, z_\ell^{(i,a)} = k')}. \tag{5.20}$$

UPDATING (α, y). The parameters (α_{ik}, y_{ik}) are updated for each i, k using a Metropolis-

Hastings algorithm. For each $(i_0, k_0) \in \{1, \ldots, N\} \times \{1, \ldots, K\}$, a new value $y^*_{i_0 k_0}$ is sampled from the conditional density

$$y^*_{i_0 k_0} | \rho, \sigma^2, y_{ik}, i \neq i_0 \ \sim \ \mathcal{N}(\rho \sum_j w_{i_0 j} \ y_{j k_0}, \sigma^2) \tag{5.21}$$

$y^*_{i_0 k_0}$ is accepted and replaces $y_{i_0 k_0}$ with the probability

$$\min \left(1, \ \frac{\Gamma\left(\sum_{k=1}^K \alpha^*_{i_0 k}\right) \Gamma\left(\alpha_{i_0 k_0}\right)}{\Gamma\left(\sum_{k=1}^K \alpha_{i_0 k}\right) \Gamma\left(\alpha^*_{i_0 k_0}\right)} q_{i_0 k_0}^{\alpha^*_{i_0 k_0} - \alpha_{i_0 k_0}} \right) \tag{5.22}$$

where $\alpha^*_{i_0 k_0} = \exp\left(\tilde{X}^E_{i_0} \beta^E_{k_0} + f(\tilde{X}^S_{i_0}) \beta^S_{k_0} + y^*_{i_0 k_0}\right)$ and Γ denotes the Gamma function. If $y_{i_0 k_0}$ is updated to $y^*_{i_0 k_0}$ then $\alpha_{i_0 k_0}$ is updated to $\alpha^*_{i_0 k_0}$.

UPDATING β. POPS uses a diffuse prior distribution on β, $\beta \sim \mathcal{N}(0, B^{-1})$, with $B = 0$, so that the conditional posterior distribution on β is given by

$$\beta_k | \alpha, \rho_k, \sigma_k, \tilde{X} \ \sim \ \mathcal{N}(V \Omega^T (\mathrm{Id} - \rho_k W) \log(\alpha_k), \sigma_k^2 V) \tag{5.23}$$

where $V = \left(\Omega^T (\mathrm{Id} - \rho W) \Omega\right)^{-1}$, Id is the identity matrix, Ω is the concatenation of the matrices $\tilde{X}^{\mathrm{E}}$ and $f(\tilde{X}^{\mathrm{S}})$, and $W = (w_{ij})$.

UPDATING σ^2. POPS uses a Gamma distribution for the hyperprior parameter $\phi_k = 1/\sigma_k^2$ for each k in $\{1, \ldots, K\}$

$$\phi_k | y \sim \mathrm{Ga}\left(\frac{N}{2}, \frac{1}{2} \sum_{i=1}^N \sum_{j \sim i} w_{ij} \ y_{ik} \ y_{jk} \right), \tag{5.24}$$

where $\mathrm{Ga}(a, b)$ denotes the Gamma distribution with shape a and rate b.

UPDATING ρ. Let $e = (e_1, \ldots, e_N)$ the N eigenvalues of the weight matrix W, and e_{max} the largest eigenvalue. POPS uses a uniform hyperprior distribution in the range $(0, e_{max}^{-1})$ on ρ. POPS updates ρ_k independently for each k by using a Metropolis-Hastings step. A new value ρ^* of ρ_k is proposed from a Gaussian random walk with a fixed variance equal to 0.05. POPS rejects the proposed value if it is outside the range $(0, e_{max}^{-1})$. Otherwise, the program accepts it with the probability

$$\min \left(1, \ \prod_{i=1}^N \left(\frac{1 - \rho^* e_i}{1 - \rho_k e_i} \right)^{1/2} \ \exp\left(-\frac{1}{2\sigma_k^2} \ y_{ik} \sum_{j=1}^N w_{ij} \ y_{jk} \ (\rho_k - \rho^*) \right) \right). \tag{5.25}$$

5.3.3 Posterior predictive simulations and model selection

An important feature of POPS is that environmental and spatial covariates can be used for predicting cluster membership and admixture proportions given new covariate values. In models without admixture, the probabilities of membership for a new individual, or new values of the covariates, can be obtained by sampling latent cluster labels from equations (5.4) and (5.5), where the regression coefficients are sampled according to their posterior distribution. Similarly, in admixture models, admixture coefficients can be obtained from new covariates using equations (5.7) and (5.8).

To measure the ability of a set of covariates to predict membership or admixture coefficients, POPS computes a Pearson correlation coefficient. This coefficient measures the correlation between two vectors of membership or admixture coefficients. The first vector, of size $K \times N$, corresponds to the membership or admixture coefficients estimated from the posterior distribution and the second vector corresponds to the coefficients predicted from the spatial and environmental data. The use of correlation scores to compare models including different set of covariates has been studied by Jay *et al.* (2011a).

For each run, POPS also computes the Deviance Information Criterion (DIC; Spiegelhalter *et al.* 2002). The DIC is calculated as the sum of the expected deviance of a model and a penalty term corresponding to the effective number of parameters. The DIC has proven useful to discriminate between models with different priors or with distinct numbers of cluster (Durand *et al.* 2009b; François and Durand 2010b; Gao *et al.* 2011).

5.4 Using POPS

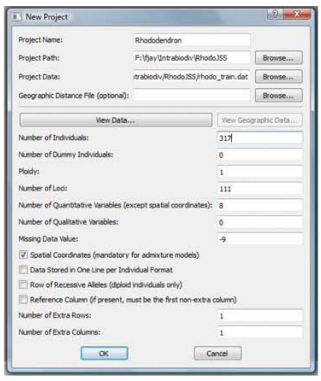
POPS is a free software implemented in the C++ programming language, and the graphical user interface uses the library Qt. In this section, we briefly describes its user interface, and we survey the main instructions of the command-line engine.

5.4.1 POPS graphical user interface (GUI)

To start an analysis using the POPS GUI, the user must create a project by loading a file containing genetic data, and possibly spatial and environmental data (Figure 5.1). The data format required by POPS is similar to STRUCTURE or TESS formats. The default format uses two rows to store the data of a diploid individual (one row for an haploid individual), but a special format that uses only one row can be specified. In addition to geographic or environmental data, there can be additional rows or columns present in the input file for informational or other purposes. Columns must be stored in a predefined order : 1) extra columns (optional, *e.g.* identifier for samples, population labels), 2) qualitative environmental covariates (optional), 3) quantitative environmental covariates

(optional), 4) longitude and latitude (mandatory in admixture models), 5) genetic markers (mandatory).

A



B

	1	2	3	4	5	6	7	8	9	10	11	12	
1	Ind	prcpavgy	tmaxavgty	tminavgty	asp200	slp200	topo	twi200s	dem200ibd	longitude	latitude	AAT_CAC_63.6	AAT
2	E21-1	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
3	E21-2	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
4	E21-4	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
5	E23-1	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	0	
6	E23-2	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	0	
7	E23-4	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	1	
8	E25-1	309.16	6.37	-0.57	102.46	10.38	261	-0.73	1739	12.33	47.71	0	
9	E25-2	309.16	6.37	-0.57	102.46	10.38	261	-0.73	1739	12.33	47.71	0	

Figure 5.1: A: Creation of a new project. The user has to input a datafile and data information. B: View of the datafile in `POPS`. Environmental data are highlighted in green, coordinates in blue, and genetic markers in white.

Users ought to specify the following parameters

- model with or without admixture,
- degree of the spatial trend surface (0, 1, 2, or 3),
- range of values for the maximal number of clusters K,
- number of runs to launch for each value of K,
- number of sweeps of each run (total number of steps and number of burn-in iterations).

Optional parameters can be specified, including hyperparameters of admixture models: the scale parameter θ, initial CAR variances (one value for all clusters), inference of the CAR variances $(\sigma_1, \ldots, \sigma_K)$ during the MCMC algorithm (see equations (5.10) and (5.11)). The interface of `POPS` is similar to the interface of `TESS` (Durand *et al.* 2009a), and users of this program should be able to familiarize with `POPS` swiftly.

5.4.2 Outputs

Each run of `POPS` produces textual results containing estimates of model parameters, graphical plots of estimated membership or admixture coefficients, the log-likelihood history of the run for convergence diagnosis, and values of the DIC for model selection (Spiegelhalter *et al.* 2002). Additionally, when geographic information is provided,

POPS displays *hard-clustering* assignments where each individual is assigned to the cluster for which the membership or admixture coefficient is maximal. Supplementary scripts are provided on the POPS webpage to display the membership probabilities or the admixture coefficients on geographic maps using the **fields** package for R (Furrer *et al.* 2009). Plots of regression coefficients history and barplots of membership/admixture coefficients predicted from the geographic or environmental covariates are also produced, and the correlation between inferred and predicted coefficients is computed.

Output results are accessible by double-clicking on corresponding items in the tree widget (Figure 5.2A). In addition, the "Summarize Runs" button allows the user to quickly access the summaries of all runs (K, trend degree, model used, DIC, average log-likelihood, correlation score). The window also provides a tool to export runs to CLUMPP, a software that averages the membership and admixture coefficients estimated from multiple MCMC runs (Jakobsson and Rosenberg 2007). Finally, the summary window can be used to run prediction. To predict membership or admixture coefficients for new values of covariates, the user needs to load a file containing the new data and choose runs that will be used for predictions. Textual and graphical results of predicted coefficients are computed for each selected run, and are accessible from the tree widget (Figure 5.2A). Prediction outputs from independent runs can also be exported to CLUMPP.

5.4.3 POPS command-line options

POPS is based on two command-line engines: one for the models without admixture, and the other for the models with admixture. We describe the main commands for both programs. When there are no options given to POPS, it will show its typical usage and exit. Here are the main options. They can be specified in any order.

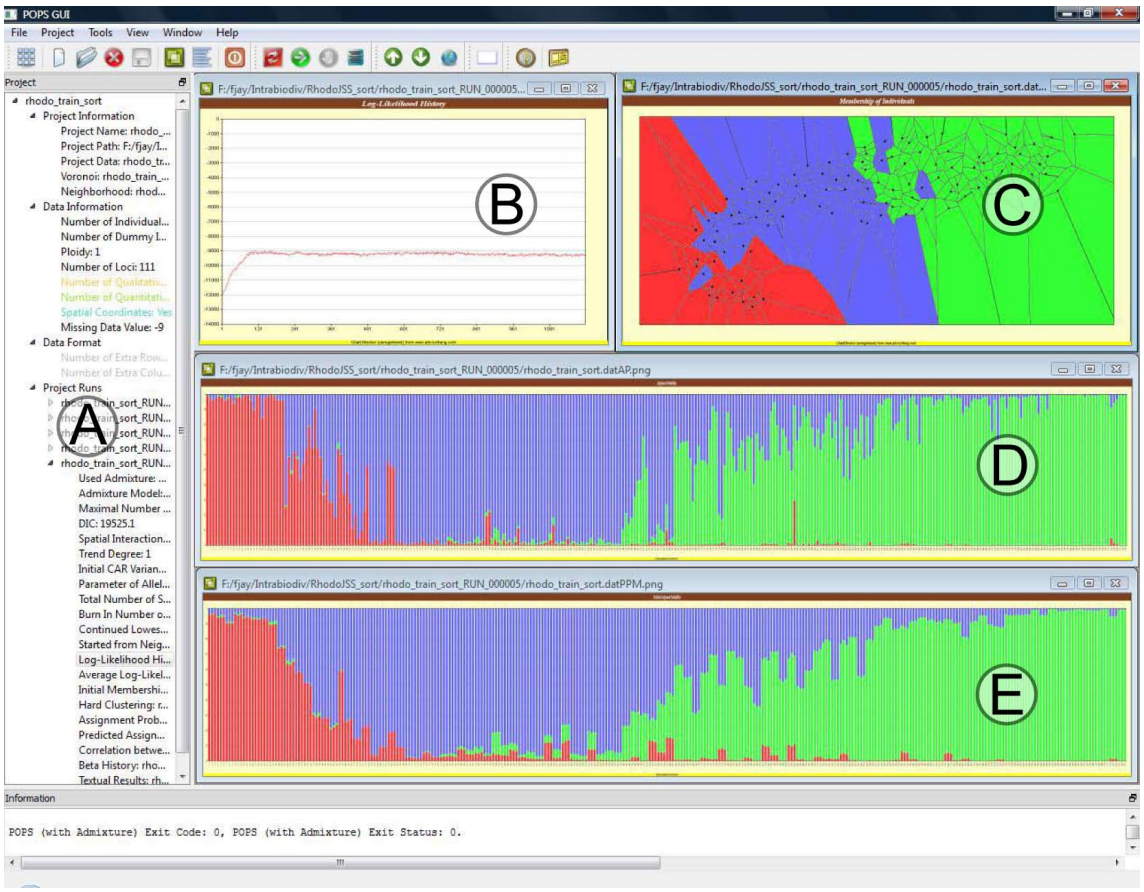


Figure 5.2: Analysis of a 317 *Rhododendron ferrugineum* plants dataset using `POPS` GUI. Graphical results are displayed for one run. A: a tree widget from which the user has access to project and data information, run options, textual and graphical outputs. All output results can be displayed by double-cliking on the appropriate items in the tree widget B: Log-likelihood history of the run. C,D,E: 3 graphical outputs, where colors correspond to clusters. C: "Hard clustering", each plant is assigned to one cluster (for which the admixture coefficient is maximal) and assignments are displayed on a map of the Alps where each black point represents a sample site. D: Admixture coefficients estimated for each sample. E: Admixture coefficients predicted from environmental data for each sample. In D and E, individuals are represented as vertical lines partitioned into segments corresponding to the fraction of their genomes assigned/predicted to belong to each genetic cluster.

Required Parameters

`-F` File Name of Input Data File
`-N` Number of Individuals
`-A` Ploidy (1 = Haploid, 2 = Diploid, ...)
`-L` Number of Loci
`-K` Maximal Number of Clusters
`-XL` Number of Qualitative Variables
`-X` Number of Quantitative Covariates (other than spatial coordinates)
`-T` Degree of Trend (-1: No spatial coordinates in datafile,
0: Spatial coordinates present but not used,
1,2,3: Degree 1, 2, or 3)
`-D` Parameter of Dirichlet Allele Frequency Model
`-S` Total Number of Sweeps of MCMC
`-B` Burn In Number of Sweeps of MCMC
`-P` Spatial Interaction Parameter (for admixture models only)

Optional Parameters

`-r` Number of Extra Rows in Data File
`-c` Number of Extra Columns in Data File
`-i` Folder Name of Input Data File (default: Current Folder)
`-o` Folder Name of Output Result Files (default: Current Folder)
`-orun` Suffix to Append to Output Result Files Names
e.g. a Run Number (`-orun1` or `-orun0001`) or a Specific Run Name (`-orunAdm002`)
`-sp` Special Data Format: 1 individual = 1 row (-spy: yes, -spn: no, default)
... Execute `pops|more` and `popsAdm|more` or see `POPS` manual for extra options

The command `pops` (or `pops.exe` on Windows system) runs models without admixture, whereas `popsAdm` (or `popsAdm.exe`) runs admixture models.

For example, suppose that the data set is stored in a file named "example.txt" in a folder "Example". This file is provided with the software. The data contain 268 haploid individuals (`-N268 -A1`) genotyped at 86 loci (`-L86`), without any environmental data (`-XL0 -X0`), no spatial coordinates (`-T-1`), 1 extra row `-r1` and 4 extra columns `-c4`. Assuming there are at most 3 clusters (`-K3`), we set the parameter of the Dirichlet allele frequency model to 1.0 (`-D1.0`). To run the MCMC algorithm for a total of 1,000 sweeps (`-S1000`) with the first 200 sweeps discarded as burn-in period (`-B200`), we use the following command

```
pops -Fexample.txt -N268 -A1 -XL0 -X0 -T-1 -L86 -K3 -D1.0 -S1000 -B200 -r1
-c4 -iExample -oExample -orun001
```

Output results will be stored in the directory "Example" (`-oExample`) and the string `"_RUN001_"` will be appended to the output names (`-orun001`).

Assume now that the 4 first columns of the file "example.txt" contain an extra column (identifier for samples, `-c1`), 1 quantitative covariate (temperature `-X1`), and 2 columns for longitude and latitude. To run `POPS` with the same run parameters as before but using the covariates, and setting the degree of the trend surface to 1 (`-T1`), the command is

```
pops -Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000 -B200 -r1
-c1
-iExample -oExample -orun002
```

To run the `POPS` admixture model using the same data and the same parameters as before, we additionally specify the default value of spatial interaction parameter (`-P0.6`)

```
popsAdm -Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000 -B200
-P0.6 -r1 -c1 -iExample -oExample -orun003
```

5.5 Examples

In this section, we illustrate several features of `POPS` with a re-analysis of a data set of 377 individuals from plant species *Rhododendron ferrugineum L.*, with each individual genotyped at 111 loci (INTRABIODIV database, Gugerli *et al.* 2008). *Rhododendron ferrugineum L.* is a small and evergreen shrub present in European mountains. With the genetic data, we consider also geographic and environmental covariates consisting of latitude, longitude, average minimum and maximum annual temperatures, average precipitations, and an additional set of topographic variables measured at each sampled site.

5.5.1 Estimating population genetic structure

A subset of 60 individuals are randomly chosen among the 377 samples to constitute a test set. We run `POPS` on the remaining 317 plants, and we report the results from a single run using the admixture model with $K = 3$ clusters. The run used 1,000 sweeps following a burn-in period of 200 sweeps. The log-likelihood function increases quickly during the run, and then reaches a stationary state (Figure 5.2B). Admixture coefficients are estimated for each individual and displayed in the graphical user interface in Figure 5.2D. Hard-clustering assignments are displayed on the tessellation built from the sample sites locations (Figure 5.2C). The map shows that the 3 inferred genetic clusters corresponds to three well-separated geographical regions, in the southwestern region (red cluster), central region (blue cluster) and northeastern region (green cluster). `R` scripts based on kriging methods allow us to display admixture coefficients spatially (Figure 5.3). Though substantial admixture occurs within contact zones between clusters, the results are con-

	southwestern cluster			northeastern cluster			central cluster		
	Estimate	95% C.I.	Signif.	Estimate	95% C.I.	Signif.	Estimate	95% C.I.	Signif.
intercept	71.12	[5.31,142.89]	**	-120.64	[-198.03,-38.6]	****	18.36	[-63.66,93.23]	-
precipitation	0.01	[0,0.02]	-	0	[-0.01,0.01]	-	0	[-0.01,0.01]	-
temp. max.	0.51	[-0.28,1.31]	-	1.09	[0.06,1.82]	**	-0.68	[-1.3,-0.22]	****
temp. min.	0.23	[-0.24,0.73]	-	0.19	[-0.33,0.69]	-	-0.39	[-0.89,0.09]	-
aspect	0	[-0.01,0]	-	0	[-0.01,0]	-	0	[0,0]	-
slope	0.01	[-0.06,0.06]	-	0.02	[-0.04,0.08]	-	0	[-0.05,0.05]	-
topo	0	[0,0]	-	0	[-0.01,0]	-	0	[-0.01,0]	-
soil humidity	0.21	[-0.7,0.84]	-	-0.19	[-1.01,0.53]	-	0.11	[-0.62,0.7]	-
altitude	0	[0,0.01]	-	0.01	[0,0.01]	***	-0.01	[-0.01,0]	****
longitude	0.03	[-0.26,0.29]	-	1.06	[0.68,1.47]	****	-0.53	[-0.89,-0.2]	****
latitude	-1.77	[-3.11,-0.56]	***	1.9	[0.42,3.5]	***	0.06	[-1.5,1.78]	-

Signif. codes: 100% '****' 99.9 % '***' 99% '**' 95% '*' 90% '-' 0%
'precipitation' stands for mean annual precipitation sum, 'temp. max.' for mean annual maximum temperature, 'temp. min.' for mean annual minimum temperature, and 'topo' for integrated topographic exposure map

Table 5.1: Estimates, 95% Bayesian confidence intervals, and significance levels for the regression coefficients in 3 clusters for an analysis on 317 *Rhododendron ferrugineum* plants.

sistent with hard-clustering assignments. The influence of each covariate is measured by their regression coefficients in the 3 clusters. Table 5.1 reports the posterior means of regression coefficients, their 95% Bayesian confidence interval, and their significance level. `POPS` detects an influence of average maximum daily temperature, altitude, longitude, and latitude on the admixture proportions. To evaluate the improvement on the estimation of population genetic structure obtained with models using environmental covariates, we additionally compute a DIC for models that do not use any environmental information. We find that the DIC decreases from 19565 to 19525 when adding the environmental covariates.

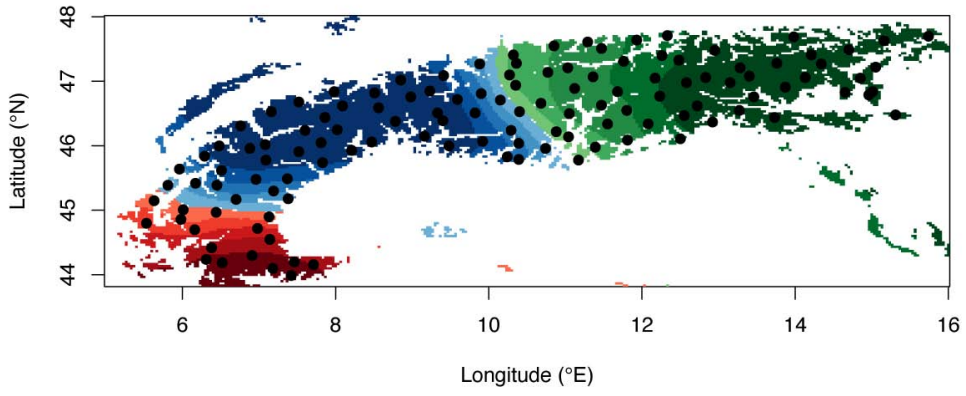


Figure 5.3: Admixture coefficients estimated for 317 *Rhododendron ferrugineum* plants are displayed spatially on a map of the European Alps. The map is computed using an R script based on kriging methods and provided with `POPS`. Only coefficients greater than 0.5 are displayed

5.5.2 Predicting population genetic structure based on covariate information

POPS can test if a set of covariates included in a model is useful to predict population genetic structure by computing the correlation between the estimated admixture coefficients and coefficients predicted from the geographic and environmental covariates only (Figures 5.2D and 5.2E). A correlation score of ≈ 0.96, reported in the tree widget, indicates that prediction from environmental variables is accurate. When we use POPS to predict admixture coefficients from the covariates of the 60 individuals contained in the test set (and not used for inference), the correlation score is equal to 0.85 (Figure 5.4).

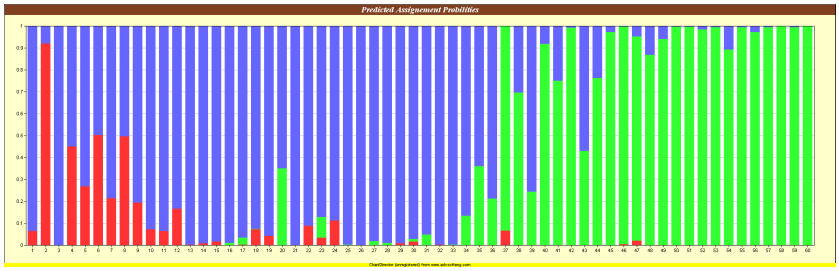


Figure 5.4: Admixture coefficients predicted for the 60 individuals contained in the test set. Admixture coefficients are computed for the 3 clusters inferred, using geographic, climatic and topographic information but no genetic data. The correlation between predicted admixture coefficients and coefficients estimated for the 60 test individuals using all environmental and genetic information is 0.85.

5.5.3 Forecasting population genetic structure under environmental changes

According to the Intergovernmental Panel on Climate Change, environmental conditions may change drastically during the coming century (Intergovernmental Panel on Climate Change 2007). Temperatures are predicted to rise by 1.8 to 4 °C, depending on the IPCC expert projections. Precipitations are also likely to increase in several regions. These changes are now acknowledged to have an impact on species distributions and there has been increasing evidence of species' range shifts due to climate change (e.g Parmesan and Yohe 2003). POPS provides a framework to investigate and forecast modifications in population genetic structure in response to environmental changes. We used POPS to forecast changes in population genetic structure of the species *Rhododendron ferrugineum* under a 2 °C temperature increase and a 40% augmentation in precipitation levels (see Jay *et al.* 2011b, for a more extensive study). Admixture coefficients computed with the projected climatic variables are displayed on Figure 5.5. A comparison with current admixture coefficients (Figure 5.2) provides evidence that the contact zones between the central and the southern clusters, and between the central and the northern

clusters, shift in the northward and westward directions respectively. Only a few sample sites are predicted to host individuals with a high admixture coefficient in the central cluster (blue-painted vertical lines in Figure 5.5A), and this cluster is not visible when the coefficients are displayed spatially (Figure 5.5B).

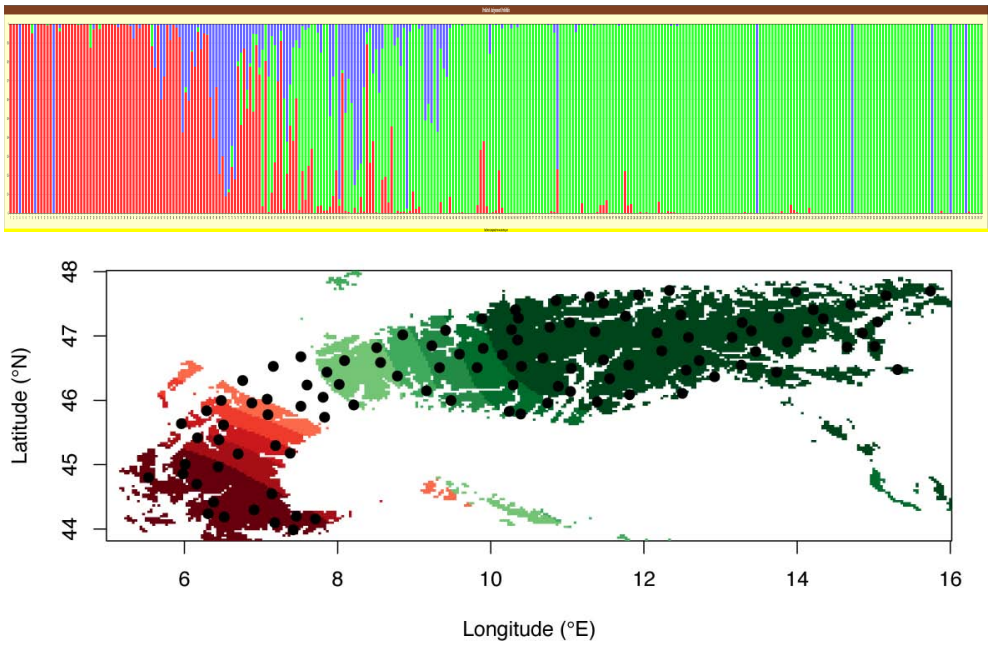


Figure 5.5: Admixture coefficients predicted for 317 individuals under a global warming scenario (2 °C temperature increase and 40% precipitation increase). A: Admixture coefficients are displayed on a bar chart. B: Admixture coefficients greater than 0.5 are displayed spatially using kriging methods.

5.6 Conclusion

POPS is a software to estimate population structure from individual multi-locus genotypes and individual covariates without assuming predefined populations. POPS jointly infers population genetic structure and the effect of environmental covariates on this structure. It can use environmental covariates to predict genetic structure for new individuals or for changing environmental conditions. POPS can be used either from a graphical user interface or from a command-line engine, both available at http://membres-timc.imag.fr/Olivier.Francois/pops.html

R scripts to post-process results and to display membership or admixture coefficients spatially are available from the POPS package.

Chapitre 6

Conclusion et perspectives

Nous avons développé une nouvelle méthode pour étudier les relations entre la structure génétique des populations et l'environnement. Cette méthode repose sur des modèles hiérarchiques bayésiens qui utilisent conjointement des données génétiques multilocus et des variables environnementales. Son premier avantage est que l'information environnementale peut permettre d'améliorer l'estimation de la structure génétique des populations par rapport à une méthode n'utilisant que les données génétiques. Ensuite, cette méthode nous a permis de nous pencher sur deux problèmes : d'une part, la détection des facteurs environnementaux liés à la structure génétique des populations ; d'autre part, la projection de la structure en cas de changement des conditions environnementales. Nous avons appliqué notre méthode à des données de génétique humaine pour évaluer le rôle de la géographie et des langages dans la structure génétique des populations, et à des données de plantes alpines pour étudier les conséquences du changement climatique sur la structure des populations.

Le premier problème, à savoir la recherche des facteurs environnementaux déterminant les variations génétiques, a été abordée dans de nombreuses études, et plusieurs méthodes ont déjà été développées. Certaines sont générales comme les tests de Mantel ou `GESTE` (Smouse *et al.* 1986; Foll and Gaggiotti 2006) ; d'autres dédiées à des facteurs spécifiques, comme les langages (Hunley and Long 2005). L'approche que nous avons présentée est différente, de par sa méthodologie et de par sa modélisation explicite de la structure génétique des populations. En particulier, elle ne nécessite pas de prédéfinir des populations. Son but n'est pas de remplacer les approches précédentes, mais de proposer un cadre d'étude nouveau. Appliquée à des questions connues, comme celle de la relation entre gènes et langages, elle permet d'apporter des éléments de réponse complémentaires.

Le deuxième problème concerne la projection de la structure génétique des populations. Elle s'inscrit dans la suite logique des modèles de projection de distribution d'espèces (SDMs), mais accorde de l'importance à une dimension majeure : la variation intraspéci-

fique. En combinant des méthodes de génétique des populations au principe des SDMs, l'approche que nous proposons se démarque clairement des approches classiques et devrait être à l'origine d'une nouvelle classe de modèles bioclimatiques.

Il faut garder à l'esprit que cette première méthode bénéficiera des nombreuses réflexions et avancées déjà réalisées dans le domaine des SDMs, à la fois d'un point de vue méthodologique et d'un point de vue biologique. D'un point de vue méthodologique, la multitude des modèles statistiques utilisés dans les SDMs pourrait inspirer des extensions de notre méthode (Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Heikkinen *et al.* 2006). De même, les discussions sur les stratégies de validation des SDMs seront profitables (voir, par exemple, Araújo *et al.* 2005). Enfin, comme le font les SDMs, il est nécessaire de pouvoir modéliser les données d'absence d'individus pour pouvoir projeter plus finement les distributions intraspécifiques. Notre méthode le permettra aisément et le problème sera plutôt de constituer des jeux de données suffisamment complets, c'est-à-dire contenant des données d'absence/présence, des données environnementales et des données génétiques. Le fort développement, ces dernières années, des études de génétique du paysage devrait néanmoins faciliter la tâche. D'un point de vue biologique, il s'agit surtout de supprimer, autant que possible, les hypothèses les moins réalistes de notre méthode. Une partie de ces hypothèses sont communes avec les SDMs et ont donc déjà été discutées (Guisan and Thuiller 2005; Sax *et al.* 2007; Jeschke and Strayer 2008). Il faudrait, entre autres, envisager la possibilité d'intégrer les données de dispersion, les phénotypes et les interactions biotiques.

Histoire démographique et adaptation locale. Les relations entre structure et environnement sont la conséquence de l'histoire démographique des populations et de phénomènes adaptatifs. Dans notre première étude, nous nous sommes intéressés à l'influence de la géographie et des traits culturels sur la structure génétique des populations humaines (Jay *et al.* 2011a). Les traits culturels participent à l'histoire démographique en provoquant des barrières potentielles aux flux de gènes. Nous n'avons envisagé le rôle de l'adaptation locale que dans notre deuxième étude, portant sur la réponse des plantes alpines au changement climatique (Jay *et al.* 2011b). Comme nous l'avons expliqué dans l'article, il est difficile de séparer la contribution de l'histoire démographique, dans les relations entre structure génétique et environnement, de la contribution de l'adaptation locale. Nous avons montré que la forme des gradients de fréquences alléliques ou que la corrélation entre la structure estimée et la structure prédite à partir des variables environnementales pouvaient apporter des éléments de réponse. Toutefois, c'est une question qui mérite d'être approfondie, et le développement de méthodes pour distinguer la contribution neutre de l'histoire démographique, de la contribution adaptative est aujourd'hui un défi (Joost *et al.* 2007; Coop *et al.* 2010; Manel *et al.* 2010a).

Données génétiques. Un aspect à prendre en compte lors du développement de méthodes en génétique des populations est le type de marqueurs disponibles. Comme nous l'avons signalé dans le chapitre 1, une très nette amélioration des technologies de séquençage a été réalisée ces dernières années et les méthodes statistiques actuelles ne sont pas toutes adaptées à des données de si grande dimension. C'est par exemple le cas pour `POPS`, dans son format actuel, du fait de son algorithme MCMC computationnellement lourd. Toutefois, il convient parfaitement à d'autres types de marqueurs, comme les microsatellites et les AFLPs, qui, malgré leur nombre réduit, sont souvent suffisamment informatifs pour détecter les grands traits de la structure génétique des populations ; d'autant plus que les espèces non-modèles bénéficient plus lentement des progrès du séquençage. L'évolution du type des marqueurs ne devrait donc pas entraîner le remplacement de telles méthodes, mais plutôt motiver l'émergence de questions auxquelles il n'était pas possible ou très difficile de répondre jusqu'à présent. Entre autres, les études de l'adaptation locale bénéficieront des jeux de données de SNPs, puisque ceux-ci offrent une meilleure couverture du génome et donc facilitent la détection des gènes potentiellement sous sélection (Manel *et al.* 2010a). Du point de vue démographique, ces jeux de données devraient permettre, grâce à la découverte des allèles rares, de s'intéresser à l'histoire de migrations bien plus récentes (en génétique humaine, voir Novembre and Ramachandran 2011).

Articles publiés ou soumis au cours de la thèse

- Jay, F., O. François, E.Y. Durand, and M.G.B Blum. POPS : A software for prediction of population genetic structure using latent regression models. *Submitted*

- Jay, F., S. Manel, N. Alvarez, E.Y. Durand, W. Thuiller, R. Holderegger, P. Taberlet, O. François. Forecasting changes in population genetic structure of Alpine plants in response to global warming. *Submitted*

- Jay, F., O. François, and M.G.B. Blum. Predictions of native American population structure using linguistic covariates in a hidden regression framework. *PLoS ONE*, 6, 2011.

- Durand, E., F. Jay, O.E. Gaggiotti, and O. François. Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution* 26 : 1963-1973, 2009.

Bibliographie

Aitchison, J. and J. Bennett, 1970 Polychotomous quantal response by maximum indicant. Biometrika **57** : 253.

Aitken, S., S. Yeaman, J. Holliday, T. Wang and S. Curtis-McLane, 2008 Adaptation, migration or extirpation : climate change outcomes for tree populations. Evolutionary Applications **1** : 95–111.

Akaike, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19** : 716–723.

Albert, C., N. Yoccoz, T. Edwards Jr, C. Graham, N. Zimmermann *et al.*, 2010 Sampling in ecology and evolution – bridging the gap between theory and practice. Ecography **33** : 1028–1037.

Albert, J. H. and S. Chib, 1993 Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88** : 669–679.

Alexander, D. and K. Lange, 2011 Enhancements to the admixture algorithm for individual ancestry estimation. BMC bioinformatics **12** : 246.

Alexander, D., J. Novembre and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome research **19** : 1655.

Alvarez, N., C. Thiel-Egenter, A. Tribsch, R. Holderegger, S. Manel *et al.*, 2009 History or ecology ? substrate type as a major driver of patial genetic structure in alpine plants. Ecology letters **12** : 632–640.

Ando, T., 2007 Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. Biometrika **94** : 443.

Araújo, M., R. Pearson, W. Thuiller and M. Erhard, 2005 Validation of species–climate impact models under climate change. Global Change Biology **11** : 1504–1513.

Atkins, K. and J. Travis, 2010 Local adaptation and the evolution of species' ranges under climate change. Journal of theoretical biology **266** : 449–457.

Austerlitz, F., S. Mariette, N. Machon, P. Gouyon and B. Godelle, 2000 Effects of colonization processes on genetic diversity : differences between annual plants and tree species. Genetics **154** : 1309.

Avise, J., 1992 Molecular population structure and the biogeographic history of a regional fauna : a case history with lessons for conservation biology. Oikos : 62–76.

Ayub, Q., A. Mansoor, M. Ismail, S. Khaliq, A. Mohyuddin *et al.*, 2003 Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. Am. J. Phys. Anthropol. **122** : 259–68.

Balding, D., 2003 Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology **63** : 221–230.

Balding, D., 2006 A tutorial on statistical methods for population association studies. Nature Reviews Genetics **7** : 781–792.

Balding, D. and R. 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.

Balkenhol, N., L. Waits and R. Dezzani, 2009 Statistical approaches in landscape genetics : an evaluation of methods for linking landscape and genetic data. Ecography **32** : 818–830.

Bamshad, M., S. Wooding, W. Watkins, C. Ostler, M. A. Batzer *et al.*, 2003 Human population genetic structure and inference of group membership. Am. J. Hum. Genet. **72** : 578–89.

Bandeen-Roche, K., D. Miglioretti, S. Zeger and P. Rathouz, 1997 Latent variable regression for multiple discrete outcomes. J. Am. Stat. Assoc. **92** : 1375–1386.

Barrantes, R., P. E. Smouse, H. W. Mohrenweiser, H. Gershowitz, J. Azofeifa *et al.*, 1990 Microevolution in lower Central America : genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am. J. Hum. Genet. **46** : 63–84.

Barton, N. and G. Hewitt, 1985 Analysis of hybrid zones. Annual review of Ecology and Systematics : 113–148.

Belle, E. and G. Barbujani, 2007 Worldwide analysis of multiple microsatellites : language diversity has a detectable in influence on DNA diversity. Am. J. Phys. Anthropol. **133** : 1137–1146.

Bellwood, P. S., 2005 *The First Farmers : The Origins of Agricultural Societies*. Oxford : Blackwell.

Benjamini, Y. and Y. Hochberg, 1995 Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) : 289–300.

Bernal, M., B. Sinues, I. Johansson, R. McLellan, A. Wennerholm *et al.*, 1999 Ten percent of north spanish individuals carry duplicated or triplicated cyp2d6 genes associated with ultrarapid metabolism of debrisoquine. Pharmacogenetics and Genomics **9** : 657.

Besag, J., 1975 Statistical analysis of non-lattice data. The Statistician **24** : 179–195.

Bierne, N., J. Welch, E. Loire, F. Bonhomme and P. David, 2011 The coupling hypothesis : why genome scans may fail to map local adaptation genes. Molecular Ecology .

Blei, D. M., A. Y. Ng and M. I. Jordan, 2003 Latent dirichlet allocation. J. Mach. Learn. Res. **3** : 993–1022.

Blum, M. and M. Jakobsson, 2011 Deep divergences of human gene trees and models of human origins. Molecular Biology and Evolution **28** : 889.

Bolnick, D., B. Shook, L. Campbell and I. Goddard, 2004 Problematic use of Greenberg's linguistic classification of the Americas in studies of Native American genetic variation. Am. J. Hum. Genet. **75** : 519–522.

Bonin, A., D. Ehrich and S. Manel, 2007 Statistical analysis of amplified fragment length polymorphism data : a toolbox for molecular ecologists and evolutionists. Molecular Ecology **16** : 3737–3758.

Borcard, D., P. Legendre and P. Drapeau, 1992 Partialling out the spatial component of ecological variation. Ecology **73** : 1045–1055.

Bowcock, A., A. Ruiz-Linares and J. Tomfohrde, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368** : 455–457.

Bryc, K., A. Auton, M. Nelson, J. Oksenberg, S. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in west africans and african americans. Proceedings of the National Academy of Sciences **107** : 786.

Busby, J., 1991 Bioclim : a bioclimate analysis and prediction system. Plant Protection Quarterly **6** : 8.

Campbell, L., 1987 Review of language in the Americas by Joseph H. Greenberg. Language **64** : 591–615.

Campbell, L., 1997 *American Indian Languages : The Historical Linguistics of Native America*. Oxford University Press, New York.

Campbell, L., 2006 Long-range comparison : methodological disputes. In K. Brown, editor, *Encyclopedia of Language and Linguistics*. Oxford : Elsevier, 2nd edition, 324–331.

Cardon, L. and J. Bell, 2001 Association study designs for complex diseases. Nature Reviews Genetics **2** : 91–99.

Carpenter, G., A. Gillison and J. Winter, 1993 Domain : a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation **2** : 667–680.

Cavalli-Sforza, L., I. Barrai and A. Edwards, 1964 Analysis of human evolution under random genetic drift. In *Cold Spring Harbor symposia on quantitative biology*, volume 29. Cold Spring Harbor Laboratory Press, 9.

Cavalli-Sforza, L. and A. Edwards, 1967 Phylogenetic analysis. models and estimation procedures. American Journal of Human Genetics **19** : 233.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes*. Princeton University Press.

Cavalli-Sforza, L. L., E. Minch and J. L. Mountain, 1992 Coevolution of genes and languages revisited. Proc. Natl. Acad. Sci. USA **89** : 5620–5624.

Cavalli-Sforza, L. L. and A. Piazza, 1975 Analysis of evolution : evolutionary rates, independence and treeness. Theor. Popul. Biol. **8** : 127 – 165.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi and J. Mountain, 1988 Reconstruction of human evolution : bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA **85** : 6002–6006.

Celeux, G., F. Forbes, C. Robert and Titterington, 2006 Deviance information criteria for missing data models. Bayesian Analysis **1** : 651–706.

Chakraborty, R., R. Blanco, F. Rothhammer and L. E, 1976 Genetic variability in Chilean Indian populations and its association with geography, language, and culture. Soc. Biol. **23** : 73–81.

Chen, C., E. Durand, F. Forbes and O. François, 2007 Bayesian clustering algorithms ascertaining spatial population structure : a new computer program and a comparison study. Mol. Ecol. Notes **7** : 747–756.

Chen, C., F. Forbes and O. François, 2006 fastruct : model-based clustering made faster. Molecular Ecology Notes **6** : 980–983.

Chevin, L., R. Lande and G. Mace, 2010 Adaptation, plasticity, and extinction in a changing environment : towards a predictive theory. PLoS Biology **8** : e1000357.

Chung, H., B. Flaherty and J. Schafer, 2006 Latent class logistic regression : application to marijuana use and attitudes among high-school seniors. J. R. Stat. Soc. Ser. A **169** : 723–743.

Colonna, V., A. Boattini, C. Guardiano, G. Longobardi, D. Pettener *et al.*, 2010 Long-range comparisons between genes and languages based on syntactic differences. Hum. Hered. **In press**.

Coop, G., D. Witonsky, A. Di Rienzo and J. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics **185** : 1411.

Corander, J., J. Sirén and E. Arjas, 2008 Bayesian spatial modeling of genetic population structure. Comput. Stat. **23** : 111–129.

Corander, J., P. Waldmann and M. J. Sillanpaa, 2003 Bayesian analysis of genetic differentiation between populations. Genetics **163** : 367–374.

Crandall, K., O. Bininda-Emonds, G. Mace and R. Wayne, 2000 Considering evolutionary processes in conservation biology. Trends in Ecology and Evolution **15** : 290–295.

Cressie, N., 1993 *Statistics for spatial data*. Wiley, New York.

Currat, M., N. Ray and L. Excoffier, 2004 Splatche : a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes **4** : 139–142.

Currat, M., M. Ruedi, R. Petit and L. Excoffier, 2008 The hidden side of invasions : massive introgression by local genes. Evolution **62** : 1908–1920.

Daganzo, C., 1979 *Multinomial probit : the theory and its application to demand forecasting*. Academic Press.

Darwin, C., 1859 On the origin of species.

Das, A., S. Mohanty and W. Stephan, 2004 Inferring the population structure and demography of drosophila ananassae from multilocus data. Genetics **168** : 1975.

Davies, N., F. Villablanca and G. Roderick, 1999 Determining the source of individuals : multilocus genotyping in nonequilibrium population genetics. Trends in Ecology & Evolution **14** : 17–21.

Davis, M. and R. Shaw, 2001 Range shifts and adaptive responses to quaternary climate change. Science **292** : 673.

Davis, M., R. Shaw and J. Etterson, 2005 Evolutionary responses to changing climate. Ecology **86** : 1704–1714.

Dawson, K. J. and K. Belkhir, 2001 A bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. **78** : 59–77.

Dayton, C. M. and G. B. Macready, 1988 Concomitant-variable latent-class models. J. Am. Stat. Assoc. **83** : 173–178.

DeSarbo, W. and W. Cron, 1988 A maximum likelihood methodology for clusterwise linear regression. J. Classif. **5** : 249–282.

Doi, H., M. Takahashi and I. Katano, 2010 Genetic diversity increases regional variation in phenological dates in response to climate change. Global Change Biology **16** : 373–379.

Dugoujon, J. M., S. Hazout, F. Loirat, B. Mourrieras, B. Crouau-Roy *et al.*, 2004 GM haplotype diversity of 82 populations over the world suggests a centrifugal model of human migrations. Am. J. Phys. Anthropol. **125** : 175–192.

Dukes, J. and H. Mooney, 1999 Does global change increase the success of biological invaders ? Trends in Ecology & Evolution **14** : 135–139.

Duminil, J., S. Fineschi, A. Hampe, P. Jordano, D. Salvini *et al.*, 2007 Can population genetic structure be predicted from life-history traits ? American Naturalist : 662–672.

Durand, E., C. Chen and O. François, 2009a *Tess version 2.3 - Reference Manual*.

Durand, E., F. Jay, O. E. Gaggiotti and O. François, 2009b Spatial inference of admixture proportions and secondary contact zones. Mol. Biol. Evol. **26** : 1963–1973.

Edwards, A. and L. Cavalli-Sforza, 1965 A method for cluster analysis. Biometrics : 362–375.

Engelhardt, B. and M. Stephens, 2010 Analysis of population structure : A unifying framework and novel methods based on sparse factor analysis. PLoS Genetics **6** : e1001117.

Engler, R., C. Randin, W. Thuiller, S. Dullinger, N. Zimmermann *et al.*, 2011 21st century climate change threatens mountain flora unequally across europe. Global Change Biology .

Epps, C., P. Palsbøll, J. Wehausen, G. Roderick, I. Ramey *et al.*, 2005 Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. Ecology Letters **8** : 1029–1038.

Evanno, G., S. Regnaut and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure : a simulation study. Molecular Ecology **14** : 2611–2620.

Excoffier, L., R. Harding, R. Sokal, B. Pellegrini and A. Sanchez-Mazas, 1991 Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum. Biol. **63** : 273–297.

Falush, D., M. Stephens and J. Pritchard, 2003 Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. Genetics **164** : 1567.

Falush, D., M. Stephens and J. Pritchard, 2007 Inference of population structure using multilocus genotype data : dominant markers and null alleles. Molecular Ecology Notes **7** : 574–578.

Felsenstein, J., 1989 Phylogeny inference package (version 3.2). Cladistics **5** : 164–166.

Fisher, R., 1915 Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika **10** : 507–521.

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

Fraley, C. and A. Raftery, 1998 How many clusters ? which clustering method ? answers via model-based cluster analysis. The computer journal **41** : 578.

François, O., S. Ancelet and G. Guillot, 2006 Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population Genetics. Genetics **174** : 805–816.

François, O. and E. Durand, 2010a Spatially explicit Bayesian clustering models in population genetics. Mol. Ecol. Resour. **10** : 773–784.

François, O. and E. Durand, 2010b Spatially explicit Bayesian clustering models in population genetics. Mol. Ecol. Resour. **10** : 773–784.

François, O., M. Blum, M. Jakobsson and N. Rosenberg, 2008 Demographic history of european populations of arabidopsis thaliana. PLoS genetics **4** : e1000075.

Furrer, R., D. Nychka and S. Sain, 2009 ***fields*** *: Tools for spatial data.* R package version 5.02.

Gaggiotti, O., S. Brooks, W. Amos and J. Harwood, 2004 Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. Molecular Ecology **13** : 811–825.

Gallien, L., T. M
"unkem
"uller, C. Albert, I. Boulangeat and W. Thuiller, 2010 Predicting potential distributions of invasive species : where to go from here ? Diversity and Distributions **16** : 331–342.

Gao, H., K. Bryc and C. Bustamante, 2011 On identifying the optimal number of population clusters via the deviance information criterion. PloS one **6** : e21014.

Gao, H., S. Williamson and C. Bustamante, 2007 A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics **176** : 1635.

Gelman, A., 2004 *Bayesian data analysis*. CRC press.

Good, I., 1952 Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological) : 107–114.

Gordon, R. G., 2005 *Ethnologue : Languages of the World*. Dallas : SIL International, fifteenth edition.

Grabherr, G. and M. Gottfried, 1994 Climate effects on mountain plants. Nature **369** : 448.

Gray, R. and Q. Atkinson, 2003 Language-tree divergence times support the anatolian theory of indo-european origin. Nature **426** : 435–439.

Green, P., 1995 Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika **82** : 711.

Greenberg, J., 1987 *Language in the Americas*. Stanford : Stanford University Press.

Greenberg, J., C. I. Turner and S. Zegura, 1986 The settlement of the Americas : a comparison of the linguistic, dental, and genetic evidence. Curr. Anthropol. **27** : 477–97.

Gugerli, F., T. Englisch, H. Niklfeld, A. Tribsch, Z. Mirek *et al.*, 2008 Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation - a project synopsis. Perspectives in Plant Ecology, Evolution and Systematics **10** : 259 – 281.

Guillot, G., A. Estoup, F. Mortier and J. Cosson, 2005 A spatial statistical model for landscape genetics. Genetics **170** : 1261.

Guisan, A. and W. Thuiller, 2005 Predicting species distribution : offering more than simple habitat models. Ecology letters **8** : 993–1009.

Guisan, A. and N. Zimmermann, 2000 Predictive habitat distribution models in ecology. Ecological modelling **135** : 147–186.

Hagenaars, J. and A. McCutcheon, 2002 *Applied latent class analysis*. Cambridge University Press.

Haldane, J., 1948 The theory of a cline. Journal of Genetics **48** : 277–284.

Hampe, A., 2004 Bioclimate envelope models : what they detect and what they hide. Global Ecology and Biogeography **13** : 469–471.

Hampe, A. and R. Petit, 2005 Conserving biodiversity under climate change : the rear edge matters. Ecology Letters **8** : 461–467.

Handley, L. J. L., A. Manica, J. Goudet and F. Balloux, 2007 Going the distance : human population genetics in a clinal world. Trends Genet. **23** : 432 – 439.

Hasselblad, V., 1966 Estimation of parameters for a mixture of normal distributions. Technometrics : 431–444.

Hastie, T. and R. Tibshirani, 1990 *Generalized additive models*. Chapman & Hall/CRC.

Hastie, T., R. Tibshirani and J. H. Friedman, 2009 *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. New York : Springer-Verlag, second edition.

Hausman, J. and D. Wise, 1978 A conditional probit model for qualitative choice : Discrete decisions recognizing interdependence and heterogeneous preferences. Econometrica : Journal of the Econometric Society : 403–426.

Heath, S., I. Gut, P. Brennan, J. McKay, V. Bencko *et al.*, 2008 Investigation of the fine structure of european populations with applications to disease association studies. European Journal of Human Genetics **16** : 1413–1429.

Hedrick, P., M. Ginevan and E. Ewing, 1976 Genetic polymorphism in heterogeneous environments. Annual Review of Ecology and Systematics **7** : 1–32.

Heggarty, P., W. Maguire and A. McMahon, 2010 Splits or waves ? Trees or webs ? How divergence measures and network analysis can unravel language histories. Philosophical Transactions of the Royal Society B : Biological Sciences **365** : 3829–3843.

Heikkinen, R., M. Luoto, M. Araújo, R. Virkkala, W. Thuiller *et al.*, 2006 Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography **30** : 751.

Henn, B., C. Gignoux, M. Jobin, J. Granka, J. Macpherson *et al.*, 2011 Hunter-gatherer genomic diversity suggests a southern african origin for modern humans. Proceedings of the National Academy of Sciences **108** : 5154.

Higgins, S. and D. Richardson, 1999 Predicting plant migration rates in a changing world : the role of long-distance dispersal. American Naturalist **153** : 464–475.

Hoggart, C., E. Parra, M. Shriver, C. Bonilla, R. Kittles *et al.*, 2003 Control of confounding of genetic associations in stratified populations. The American Journal of Human Genetics **72** : 1492–1504.

Holderegger, R. and H. Wagner, 2008 Landscape genetics. BioScience : 199–207.

Hubisz, M., D. Falush, M. Stephens and J. Pritchard, 2009 Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources **9** : 1322–1332.

Huelsenbeck, J. P. and P. Andolfatto, 2007 Inference of Population Structure Under a Dirichlet Process Model. Genetics **175** : 1787–1802.

Hunley, K., M. Dunn, E. Lindström, G. Reesink, A. Terrill *et al.*, 2008 Genetic and linguistic coevolution in northern island Melanesia. PLoS Genet. **4** : e1000239.

Hunley, K. and J. C. Long, 2005 Gene flow across linguistic boundaries in Native North American populations. Proc. Natl. Acad. Sci. USA **102** : 1312–1317.

Hunley, K. L., G. S. Cabana, D. A. Merriwether and J. C. Long, 2007 A formal test of linguistic and genetic coevolution in native Central and South America. Am. J. Phys. Anthropol. **132** : 622–631.

Huntley, B., 1995 Plant species' response to climate change : implications for the conservation of european birds. Ibis **137** : S127–S138.

Hutchinson, G., 1957 Concluding remarks. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 22. Cold Spring Harbor Laboratory Press, 415–427.

Intergovernmental Panel on Climate Change, 2007 *Climate change 2007 : the physical science basis*. Cambridge University Press.

Jakobsson, M. and N. A. Rosenberg, 2007 CLUMPP : a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics **23** : 1801–1806.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451** : 998–1003.

Jay, F., 2011 *POPS short tutorial*.

Jay, F., E. Durand, O. François and M. Blum, 2011 Pops : A software for prediction of population genetic structure using latent regression models. *Submitted*.

Jay, F., O. François and M. Blum, 2011a Predictions of native american population structure using linguistic covariates in a hidden regression framework. PLoS ONE **6**.

Jay, F., S. Manel, N. Alvarez, E. Durand, W. Thuiller *et al.*, 2011b Forecasting changes in population genetic structure of alpine plants in response to global warming. submitted .

Jeschke, J. and D. Strayer, 2008 Usefulness of bioclimatic models for studying climate change and invasive species. Annals of the New York Academy of Sciences **1134** : 1–24.

Joost, S., A. Bonin, M. Bruford, L. Després, C. Conord *et al.*, 2007 A spatial analysis method (sam) to detect candidate loci for selection : towards a landscape genomics approach to adaptation. Molecular Ecology **16** : 3955.

Jump, A., R. Marchant and J. Peŋuelas, 2009a Environmental change and the option value of genetic diversity. Trends in plant science **14** : 51–58.

Jump, A., C. Mátyás and J. Peñuelas, 2009b The altitude-for-latitude disparity in the range retractions of woody species. Trends in ecology & evolution **24** : 694–701.

Jump, A. and J. Penuelas, 2005 Running to stand still : adaptation and the response of plants to rapid climate change. Ecology Letters **8** : 1010–1020.

Kullback, S. and R. Leibler, 1951 On information and sufficiency. The Annals of Mathematical Statistics **22** : 79–86.

Kumar, S., K. Tamura and M. Nei, 2004 Mega3 : integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in bioinformatics **5** : 150.

Kunstler, G., W. Thuiller, T. Curt, M. Bouchaud, R. Jouvie *et al.*, 2007 Fagus sylvatica l. recruitment across a fragmented mediterranean landscape, importance of long distance effective dispersal, abiotic conditions and biotic interactions. Diversity and distributions **13** : 799–807.

Lao, O., T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in europe. Current Biology **18** : 1241–1248.

Lazarsfeld, P., 1950 The logical and mathematical foundation of latent structure analysis. In *Measurement and Prediction*. Princeton University Press, 362–472.

Lazarsfeld, P., 1954 A conceptual introduction to latent structure analysis. Mathematical thinking in the social sciences : 349–387.

Leclerc, E., Y. Mailhot, M. Mingelbier and L. Bernatchez, 2008 The landscape genetics of yellow perch (perca flavescens) in a large fluvial ecosystem. Molecular ecology **17** : 1702–1717.

Lee, C. and T. Mitchell-Olds, 2011 Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. Molecular Ecology .

Lee, S., F. Zou and F. Wright, 2010 Convergence and prediction of principal component scores in high-dimensional settings. Annals of statistics **38** : 3605.

Legendre, P., 2000 Comparison of permutation methods for the partial correlation and partial mantel tests. Journal of Statistical Computation and Simulation **67** : 37–74.

Leimu, R. and M. Fischer, 2008 A meta-analysis of local adaptation in plants. PLoS One **3** : e4010.

Leisch, F., 2004 Flexmix : a general framework for finite mixture models and latent class regression in R. J. Stat. Softw. **11** : 1–18.

Lewin, R., 1988 American Indian language dispute. Science **242** : 1632–1633.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319** : 1100–1104.

Lichstein, J. W., T. R. Simons, S. A. Shriner and K. E. Franzreb, 2002 Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. **72** : 445–463.

Linzer, D. and J. Lewis, 2011 polca : An r package for polytomous variable latent class analysis. Journal of Statistical Software .

Malcolm, J., A. Markham, R. Neilson and M. Garaci, 2002 Estimated migration rates under scenarios of global climate change. Journal of Biogeography **29** : 835–849.

Manel, S., J. Dias and S. Ormerod, 1999 Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions : a case study with a himalayan river bird. Ecological Modelling **120** : 337–347.

Manel, S., S. Joost, B. Epperson, R. Holderegger, A. Storfer *et al.*, 2010a Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Molecular Ecology .

Manel, S., B. Poncet, P. Legendre, F. Gugerli and R. Holderegger, 2010b Common factors drive adaptive genetic variation at different spatial scales in arabis alpina. Molecular Ecology .

Manel, S., M. K. Schwartz, G. Luikart and P. Taberlet, 2003 Landscape genetics : combining landscape ecology and population genetics. Trends in Ecology and Evolution **18** : 189 – 197.

Manica, A., F. Prugnolle and F. Balloux, 2005 Geography is a better determinant of human genetic differentiation than ethnicity. Hum. Genet. **118** : 366–371.

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer. Res. **27** : 209–220.

Marchini, J., L. Cardon, M. Phillips and P. Donnelly, 2004 The effects of human population structure on large genetic association studies. Nature genetics **36** : 512–517.

Marra, M., E. Smith, J. Shulmeister and R. Leschen, 2004 Late quaternary climate change in the awatere valley, south island, new zealand using a sine model with a maximum likelihood envelope on fossil beetle data. Quaternary Science Reviews **23** : 1637–1650.

McCulloch, R. and P. Rossi, 1994 An exact likelihood analysis of the multinomial probit model. Journal of Econometrics **64** : 207–240.

Melton, P., I. Briceno, A. Gomez, E. Devor, J. Bernal *et al.*, 2007 Biological relationship between Central and South American Chibchan speaking populations : evidence from mtDNA. Am. J. Phys. Anthropol. **133** : 753–770.

Menozzi, P., A. Piazza and L. Cavalli-Sforza, 1978 Synthetic maps of human gene frequencies in europeans. Science **201** : 786.

Moritz, C., 1994 Defining'evolutionarily significant units' for conservation. Trends in ecology and evolution **9** : 373–374.

Murillo, F., F. Rothhammer and L. E, 1977 The Chipaya of Bolivia : dermatoglyphics and ethnic relationships. Am. J. Phys. Anthropol. **46** : 45–50.

Nathan, R., 2006 Long-distance dispersal of plants. Science **313** : 786.

Nei, M., F. Tajima and Y. Tateno, 1983 Accuracy of estimated phylogenetic trees from molecular data. Journal of Molecular Evolution **19** : 153–170.

Nelder, J. and R. Wedderburn, 1972 Generalized linear models. Journal of the Royal Statistical Society. Series A (General) : 370–384.

Noelli, F., 2008 The Tupi expansion. In H. Silverman and W. H. Isbell, editors, *The Handbook of South American Archaeology*. Springer : New York, 400–401.

Noelli, F. S., 1998 The Tupi : explaining origin and expansions in terms of archaeology and of historical linguistics. Antiquity **72** : 648–663.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature **456** : 98–101.

Novembre, J. and S. Ramachandran, 2011 Perspectives on human population structure at the cusp of the sequencing era. Annual review of genomics and human genetics **12** : 1–30.

Parmesan, C. and G. Yohe, 2003 A globally coherent fingerprint of climate change impacts across natural systems. Nature **421** : 37–42.

Patterson, N., A. L. Price and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. **2** : e190.

Pearson, K., 1894 Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A **185** : 71–110.

Pearson, R. and T. Dawson, 2003 Predicting the impacts of climate change on the distribution of species : are bioclimate envelope models useful ? Global Ecology and Biogeography **12** : 361–371.

Pella, J. and M. Masuda, 2006 The gibbs and split merge sampler for population mixture analysis from genetic data with incomplete baselines. Canadian Journal of Fisheries and Aquatic Sciences **63** : 576–596.

Peñuelas, J. and M. Boada, 2003 A global change-induced biome shift in the montseny mountains (ne spain). Global Change Biology **9** : 131–140.

Pereira, H., P. Leadley, V. Proença, R. Alkemade, J. Scharlemann *et al.*, 2010 Scenarios for global biodiversity in the 21st century. Science **330** : 1496.

Peterson, A., 2003a Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology : 419–433.

Peterson, A., 2003b Projected climate change effects on rocky mountain and great plains birds : generalities of biodiversity consequences. Global Change Biology **9** : 647–655.

Peterson, A., J. Soberón and V. Sánchez-Cordero, 1999 Conservatism of ecological niches in evolutionary time. Science **285** : 1265.

Peterson, A. and D. Vieglais, 2001 Predicting species invasions using ecological niche modeling : new approaches from bioinformatics attack a pressing problem. BioScience **51** : 363–371.

Petit, R., 2004 Biological invasions at the gene level. Diversity and Distributions **10** : 159–165.

Poncet, B., D. Herrmann, F. Gugerli, P. Taberlet, R. Holderegger *et al.*, 2010 Tracking genes of ecological relevance using a genome scan in two independent regional population samples of. Molecular Ecology **19** : 2896–2907.

Premo, L. S. and J.-J. Hublin, 2009 Culture, population structure, and low genetic diversity in Pleistocene hominins. Proc. Natl. Acad. Sci. USA **106** : 33–37.

Price, A., N. Patterson, R. Plenge, M. Weinblatt, N. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature **38** : 904–909.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000a Inference of population structure using multilocus genotype data. Genetics **155** : 945–959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly, 2000b Association mapping in structured populations. Am. J. Hum. Genet. **67** : 170–181.

Prugnolle, F., A. Manica and F. Balloux, 2005 Geography predicts neutral genetic diversity of human populations. Curr. Biol. **15** : R159–R160.

R Development Core Team, 2011 *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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. Natl. Acad. Sci. USA **102** : 15942–15947.

Rannala, B. and J. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences **94** : 9197.

Reynolds, J., B. Weir and C. Cockerham, 1983 Estimation of the coancestry coefficient : basis for a short-term genetic distance. Genetics **105** : 767.

Richardson, B., G. Rehfeldt and M. Kim, 2009 Congruent climate-related genecological responses from molecular markers and quantitative traits for western white pine (pinus monticola). International Journal of Plant Sciences **170** : 1120–1131.

Ripley, B., 1981 *Spatial statistics*. Wiley, New York.

Ripley, B. D., 1996 *Pattern Recognition and Neural Networks*. Cambridge University Press.

Risch, N., E. Burchard, E. Ziv and H. Tang, 2002 Categorization of humans in biomedical research : genes, race and disease. Genome Biol **3** : 1–12.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science **298** : 2381–2385.

Ruhlen, M., 1987 *A Guide to the World's Languages : Classification*, volume 1. Stanford Univ Pr.

Ruhlen, M., 1991 *A Guide to the World's Languages. Volume 1 : Classification*. Stanford University Press.

Saitou, N. and M. Nei, 1987 The neighbor-joining method : a new method for reconstructing phylogenetic trees. Molecular biology and evolution **4** : 406.

Salamin, N., R. W
"uest, S. Lavergne, W. Thuiller and P. Pearman, 2010 Assessing rapid evolution in a changing environment. Trends in Ecology & Evolution .

Salzano, F. M., J. V. Neel, H. Gershowitz and E. C. Migliazza, 1977 Intra and intertribal genetic variation within a linguistic group : the Ge-speaking indians of Brazil. Am. J. Phys. Anthropol. **42** : 337–347.

Sax, D., J. Stachowicz, J. Brown, J. Bruno, M. Dawson *et al.*, 2007 Ecological and evolutionary insights from species invasions. Trends in Ecology & Evolution **22** : 465–471.

Scherrer, D. and C. Körner, 2011 Topographically controlled thermal-habitat differentiation buffers alpine plant diversity against climate warming. Journal of Biogeography .

Schönswetter, P., I. Stehlik, R. Holderegger and A. Tribsch, 2005 Molecular evidence for glacial refugia of mountain plants in the european alps. Molecular Ecology **14** : 3547–3555.

Schwartz, R., 2001 Racial profiling in medical research. New England Journal of Medicine **344** : 1392–1393.

Schwarz, G., 1978 Estimating the dimension of a model. The annals of statistics : 461–464.

Scott, A. and M. Symons, 1971 Note : On the edwards and cavalli-sforza method of cluster analysis. Biometrics : 217–219.

Segelbacher, G., S. Cushman, B. Epperson, M. Fortin, O. François *et al.*, 2010 Applications of landscape genetics in conservation biology : concepts and challenges. Conservation Genetics **11** : 375–385.

Serre, D. and S. Pääbo, 2004 Evidence for gradients of human genetic diversity within and among continents. Genome Research **14** : 1679.

Shringarpure, S. and E. P. Xing, 2009 mStruct : inference of population structure in light of both genetic admixing and allele mutations. Genetics **182** : 575–593.

Slatkin, M., 1973 Gene flow and selection in a cline. Genetics **75** : 733.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139** : 457.

Smouse, P. E., J. C. Long and R. R. Sokal, 1986 Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Biol. **35** : 627–632.

Smyth, P., 2000 Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. **10** : 63–72.

Sokal, R., N. Oden and B. Thomson, 1988 Genetic changes across language boundaries in europe. American Journal of Physical Anthropology **76** : 337–361.

Sork, V., F. Davis, R. Westfall, A. Flint, M. Ikegami *et al.*, 2010 Gene movement and genetic association with regional climate gradients in california valley oak (quercus lobata née) in the face of climate change. Molecular Ecology **19** : 3806–3823.

Spiegelhalter, S., N. Best, B. Carlin and A. Linde, 2002 Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B **64** : 583–639.

Spielman, R. S., E. C. Migliazza and J. V. Neel, 1974 Regional linguistic and genetic differences among Yanomama indians. Science **184** : 637–644.

Spuhler, J., 1972 Genetic, linguistic and geographical distances in Native North America. In H. J. Weiner J, editor, *The Assessment of Population Affinities in Man*. Oxford : Clarendon Press, 73–95.

Spuhler, J., 1979 Genetic distance, trees, and maps of North American Indians. In H. A. Laughlin WS, editor, *The First Americans : Origins, Affinities, and Adaptations*. New York : Gustav Fischer, 135–183.

Stanton, M. and C. Galen, 1997 Life on the edge : adaptation versus environmentally mediated gene flow in the snow buttercup, ranunculus adoneus. American Naturalist : 143–178.

Stinson, K., 2004 Natural selection favors rapid reproductive phenology in potentilla pulcherrima (rosaceae) at opposite ends of a subalpine snowmelt gradient. American Journal of Botany **91** : 531.

Storfer, A., M. Murphy, J. Evans, C. Goldberg, S. Robinson *et al.*, 2007 Putting the landscape in landscape genetics. Heredity **98** : 128–142.

Sturm, M., C. Racine and K. Tape, 2001 Climate change : increasing shrub abundance in the arctic. Nature **411** : 546–547.

Suits, D. B., 1957 Use of dummy variables in regression equations. J. Am. Stat. Assoc. **52** : 548–551.

Tackenberg, O. and J. Stöcklin, 2008 Wind dispersal of alpine plant species : a comparison with lowland species. Journal of Vegetation Science **19** : 109–118.

Tang, H., J. Peng, P. Wang and N. Risch, 2005 Estimation of individual admixture : analytical and study design considerations. Genetic epidemiology **28** : 289–301.

Theurillat, J. and A. Guisan, 2001 Potential impact of climate change on vegetation in the european alps : a review. Climatic change **50** : 77–109.

Thiel-Egenter, C., N. Alvarez, R. Holderegger, A. Tribsch, T. Englisch *et al.*, 2011 Break zones in the distributions of alleles and species in alpine plants. Journal of biogeography **38** : 772–782.

Thiel-Egenter, C., F. Gugerli, N. Alvarez, S. Brodbeck, E. Cieślak *et al.*, 2009 Effects of species traits on the genetic diversity of high-mountain plants : a multi-species study across the alps and the carpathians. Global Ecology and Biogeography **18** : 78–87.

Thomas, E., 1966 Mathematical models for the clustered firing of single cortical neurones. The British journal of mathematical and statistical psychology **19** : 151.

Thomassen, H., Z. Cheviron, A. Freedman, R. Harrigan, R. Wayne *et al.*, 2010 Spatial modelling and landscape-level approaches for visualizing intra-specific variation. Molecular ecology **19** : 3532–3548.

Thornton, P., S. Running and M. White, 1997 Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology **190** : 214–251.

Thuiller, W., 2003 Biomod-optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology **9** : 1353–1362.

Thuiller, W., C. Albert, M. Araújo, P. Berry, M. Cabeza *et al.*, 2008 Predicting global change impacts on plant species' distributions : future challenges. Perspectives in Plant Ecology, Evolution and Systematics **9** : 137–152.

Thuiller, W., S. Lavorel, M. Araújo, M. Sykes and I. Prentice, 2005 Climate change threats to plant diversity in europe. Proceedings of the National Academy of Sciences of the United States of America **102** : 8245.

Tishkoff, S., F. Reed, F. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of africans and african americans. Science **324** : 1035.

Tishkoff, S. A. and K. K. Kidd, 2004 Implications of biogeography of human populations for 'race' and medicine. Nature Genetics **36** : S21–S27.

Turesson, G., 1925 The plant species in relation to habitat and climate. Hereditas **6** : 147–236.

Vermunt, J. and J. Magidson, 2003 Latent class models for classification. Computational Statistics & Data Analysis **41** : 531–537.

Voronoï, G., 1908 Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. Journal für die reine und angewandte Mathematik (Crelles Journal) **1908** : 198–287.

Vounatsou, P., T. Smith and A. Gelfand, 2000 Spatial modelling of multinomial data with latent structure : an application to geographical mapping of human gene and haplotype frequencies. Biostatistics **1** : 177–189.

Walker, P. and K. Cocks, 1991 Habitat : a procedure for modelling a disjoint environmental envelope for a plant or animal species. Global Ecology and Biogeography Letters : 108–118.

Walther, B., M. Wisz and C. Rahbek, 2004 Known and predicted african winter distributions and habitat use of the endangered basra reed warbler (acrocephalus griseldis) and the near-threatened cinereous bunting (emberiza cineracea). Journal of Ornithology **145** : 287–299.

Walther, G., 2003 Plants in a warmer world. Perspectives in Plant Ecology, Evolution and Systematics **6** : 169–185.

Walther, G., E. Post, P. Convey, A. Menzel, C. Parmesan *et al.*, 2002 Ecological responses to recent climate change. Nature **416** : 389–395.

Walther, G., A. Roques, P. Hulme, M. Sykes, P. Pysek *et al.*, 2009 Alien species in a warmer world : risks and opportunities. Trends in Ecology & Evolution **24** : 686–693.

Wang, R., S. Farrona, C. Vincent, A. Joecker, H. Schoof *et al.*, 2009 Pep1 regulates perennial flowering in arabis alpina. Nature **459** : 423–427.

Wang, S., C. M. Lewis, Jr., M. Jakobsson, S. Ramachandran, N. Ray *et al.*, 2007 Genetic variation and population structure in Native Americans. PLoS Genet. **3** : e185.

Ward, R. H., A. Redd, D. Valencia, B. Frazier and S. Pääbo, 1993 Genetic and linguistic differentiation in the Americas. Proc. Natl. Acad. Sci. USA **90** : 10663–10667.

Wedel, M. and W. DeSarbo, 1994 A review of recent developments in latent class regression models. Advanced methods of marketing research : 352–388.

Wilson, J., M. Weale, A. Smith, F. Gratrix, B. Fletcher *et al.*, 2001 Population genetic structure of variable drug response. nature genetics **29** : 265–269.

Witkowski, S. and C. Brown, 1981 Mesoamerican historical linguistics and distant genetic relationship. Am. Anthropol. **83** : 905–911.

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

Wright, S., 1949 The genetical structure of populations. Annals of Human Genetics **15** : 323–354.

Wu, B., N. Liu and H. Zhao, 2006 Psmix : an r package for population structure inference via maximum likelihood method. BMC bioinformatics **7** : 317.

Yee, T. and N. Mitchell, 1991 Generalized additive models in plant ecology. Journal of vegetation science : 587–602.

Zhang, Y., 2008 Tree-guided bayesian inference of population structures. Bioinformatics **24** : 965.

Zhou, X., X. Wang and E. Dougherty, 2006 Multi-class cancer classification using multinomial probit regression with bayesian gene selection. IEE Proceedings-Systems Biology **153** : 70–76.

Zhu, X., S. Zhang, H. Zhao and R. Cooper, 2002 Association mapping, using a mixture model for complex traits. Genetic epidemiology **23** : 181–196.

Annexe A

Spatial inference of admixture proportions and secondary contact zones

Durand, E., F. Jay, O.E. Gaggiotti, and O. François. Spatial inference of admixture proportions and secondary contact zones. Molecular Biology and Evolution 26 : 1963-1973, 2009.

Spatial Inference of Admixture Proportions and Secondary Contact Zones

Eric Durand, Flora Jay,* Oscar E. Gaggiotti,† and Olivier François**

*Faculty of Medicine, Laboratoire des Techniques de 1'Ingénierie Médicale et de la Complexité, University Joseph Fourier, Grenoble IT, Group of Mathematical Biology, La Tronche, France; and †Laboratoire d'Ecologie Alpine, Unité Mixte de Recherche Centre National de la Recherche Scientifique 5553, University Joseph Fourier, Grenoble, France

Genetic admixture of distinct gene pools is the consequence of complex spatiotemporal processes that could have involved massive migration and local mating during the history of a species. However, current methods for estimating individual admixture proportions lack the incorporation of such a piece of information. Here, we extend Bayesian clustering algorithms by including global trend surfaces and spatial autocorrelation in the prior distribution on individual admixture coefficients. We test our algorithm by using spatially explicit and realistic coalescent simulations of colonization followed by secondary contact. By coupling our multiscale spatial analyses with a Bayesian evaluation of model complexity and fit, we show that the algorithm provides a correct description of smooth clinal variation, while still detecting zones of sharp variation when they are present in the data. We also apply our approach to understand the population structure of the killifish, *Fundulus heteroclitus*, for which the algorithm uncovers a presumed contact zone in the Atlantic coast of North America.

Introduction

Biological data based on geographic surveys often display global trends and spatial autocorrelation (Sokal and Oden 1978). Spatial autocorrelation is the correlation of a geographic variable with itself but at a certain distance apart. This phenomenon complicates the analysis of spatial patterns by creating departure from the standard independence hypothesis (Slatkin and Arter 1991; Epperson and Li 1996). This pattern may be driven by endogenous factors like dispersal limitation or by exogenous factors like an important environmental determinant that is spatially structured and that implies spatial structuring in the observed variable. It is widely acknowledged that underestimating autocorrelation in ecological data can bias inference from statistical models (Lichstein et al. 2002; Dormann 2007).

Traditional spatial statistical analyses take these points into account by decomposing the spatial variation of a response variable, z, into global and local effects

$$z = m(x) + y,$$

where x are the two-dimensional spatial coordinates. The first term, $m(x)$, is a trend surface—often defined as a first-order polynomial, $m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$—capturing regional or long-range variation. The second term, y, is a spatially autocorrelated residual that represents short-range variation. This approach is sometimes called "universal kriging" (Ripley 1988) or spatial trend analysis (Bocquet-Appel and Sokal 1989).

An important question that could greatly benefit from a more precise modeling of spatial patterns is the study of genetic admixture. The demography of natural populations is the result of phases of expansion, contraction and migration, or local mating that can produce shifting patches of genotypes. In such conditions, populations isolated for a long time may be brought into contact in a certain area, leading to the genetic admixture of different gene pools (Chakraborty 1986). Admixture is particularly pervasive in humans because migratory movements have brought together peoples from different origins (Cavalli-Sforza et al 1994), and its precise assessment is important for association studies that are susceptible to biases due to population structure (Pritchard et al. 2000; Yu et al. 2006). In addition, admixture between populations originating in different continents can be exploited to detect disease susceptibility loci at which risk alleles are distributed differentially between these populations (Chakraborty and Weiss 1988; Reich and Patterson 2005; Smith and O'Brien 2005).

Under natural conditions, admixture is known to happen in secondary contact zones, and it may generate Hardy–Weinberg and linkage disequilibrium at unlinked loci (Barton and Hewitt 1985; Durrett et al. 2000). These zones are places where the hybrid offspring of the interbreeding populations are present and where their allele frequencies form a cline (Endler 1977; Barton and Gale 1993). Secondary contact or hybrid zones have often been described as the consequence of post-Pleistocene recolonization of landmasses after the ice retreat (Taberlet et al. 1998). Detecting and identifying the relative contributions of these refugia to current populations are of paramount interest to the reconstruction of the demographic history of many organisms (Hewitt 2000).

Many admixture models compute population coefficients, considering hybrid genes as proportionally inherited from two or more populations that are thought of as being the relicts of some parental populations. The quantities being estimated, the admixture coefficients, are the respective contributions of the parental populations to the hybrid gene pools. Several approaches to estimating these proportions in populations have been proposed during the last few decades, including least-squares regression (Roberts and Hiorns 1965), maximum likelihood (Long 1991), estimation of coalescence times (Bertorelle and Excoffier 1998), Markov chain Monte Carlo (MCMC) algorithms or likelihood-based methods (Chikhi et al. 2001; Wang 2003), and approximate Bayesian computation (Excoffier et al. 2005). Regarding the estimation of admixture proportions in individuals, current methods are based on computer-intensive programs like STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), ADMIXMAP (Hoggart et al. 2004), INSTRUCT (Gao et al. 2007), LAMP (Sankararaman et al. 2008). Spatial models have

Key words: admixture, Bayesian inference, spatial trends, spatial autocorrelation, secondary contact zones.

E-mail: olivier.francois@imag.fr.

Mol. Biol. Evol. 26(9):1963–1973. 2009
doi:10.1093/molbev/msp106
Advance Access publication May 21, 2009

been implemented in TESS (Chen et al. 2007) and BAPS (Corander et al. 2008). Recent examples of the use of individual-based Bayesian clustering algorithms are for the genetic analysis of hybridization between two species of lemurs (Pastorini et al. 2009), the inference of a strong subdivision between two subpopulations of the lepidopteran *Chilo suppressalis* in China (Meng et al. 2008), the demographic history of European population of the model plant *Arabidopsis thaliana* (François et al. 2008), or the recolonization of the Swiss Alps by the Valais shrew *Sorex antinorii* (Yannic et al. 2008). Principal component analysis (PCA) may provide concurrent means to estimate admixture proportions, and spatial versions of PCA might also be relevant to this framework (Patterson et al. 2006; Jombart et al. 2008).

In this study, we extended the hierarchical Bayesian algorithm implemented in TESS in order to include spatial prior distributions on the individual admixture proportions, and we assessed the abilities of this approach to detect the admixture in secondary contact zones. The proposed approach adopts a formulation similar to universal kriging in which a response variable—here admixture proportion—can be modeled as the sum of two components: a trend surface plus a Gaussian autoregressive residual term (Besag 1975; Ripley 1981; Cressie 1993). The trend surface and the residual terms attempt to capture the broad-scale and fine-scale patterns that may be expected under migration or local isolation-by-distance processes (Bocquet-Appel and Sokal 1989).

The objective of the proposed algorithm is to improve the inference of admixture proportions when admixture proportions are variable across space. The inference method is tested on synthetic data obtained from simple models and from spatially explicit scenarios simulating secondary contact and mimicking realistic migration routes for a species that colonized Europe from two glacial refugia. We measure the relative fit of spatial and nonspatial models in terms of statistical information criteria, and we display their posterior spatial predictions using a two-dimensional graphical method. The approach is applied to analyzing an hypothesized contact zone in the marine species *Fundulus heteroclitus*, with individuals genotyped at eight microsatellite loci in 15 samples along the east coast of North America (Adams et al. 2006).

Materials and Methods

A Spatial Prior for Admixture Proportions

We consider N individuals genotyped at L loci, and we assume that their geographic coordinates were recorded at the sampling locations. Individuals can be either diploid or haploid. As in the algorithm underlying STRUCTURE (Pritchard et al. 2000), we assume that the individuals represent a mixture from at most $K_{\max}$ unobserved clusters and a matrix denotes the admixture proportions for all the individuals. Each element of the matrix, q_{ik}, is the proportion of individual is genome that originated from cluster k.

We perform inference of population structure in a Bayesian framework by incorporating individual geographic covariates in the prior distributions on the admixture coefficients. More specifically, we assume a Dirichlet distribution on the q_{ik}s for each individual i,

$$q_{i\cdot} \sim \mathscr{D}(\alpha_{i1}, \ldots, \alpha_{iK_{\max}}), \quad (1)$$

where α_{ik} is proportional to the average admixture coefficient, $\mathrm{E}(q_{ik})$. The novelty is that we consider a log-normal model for the α_{ik}, viewed as unobserved response variables

$$\log(\alpha_{i\cdot}) = f(x_i)^T \beta_{\cdot} + y_{i\cdot}, \quad (2)$$

where x_i represents a two-dimensional vector of spatial covariates for i, for example, latitude and longitude. Log-linear regressions of the average admixture levels on the spatial covariates are performed in each of the $K_{\max}$ clusters. The definition of the two terms appearing in the right-hand side of equation (2) is given hereafter.

The hidden regression model described in equation (2) is similar to universal kriging (Ripley 1981; Cressie 1993), and it can be separated into two components. The first component, $m = f(x_i)^T\beta$, represents the mean response, and it is modeled as a (possibly) nonlinear trend surface. Although this was not stated explicitly, latent regression models that may incorporate trend surfaces were previously considered by Gaggiotti et al. (2004), Foll and Gaggiotti (2006), and Faubet and Gaggiotti (2008) who studied population divergence measures and recent migration rates. We limited our further analyses to linear trend surfaces, but the proposed method is valid for arbitrary polynomial shapes, and our computer program allows the use of quadratic or cubic models. The second component, y_i, represents a zero-mean spatially autocorrelated random variable. This term is a conditional auto-regressive (CAR) Gaussian model (Besag 1975; Vounatsou et al. 2000). In the CAR model, the conditional expectation of y_i, given the response at all other locations, is a weighted sum of the mean-centered coefficients at neighboring locations

$$E(y_i \mid y_j \text{ at other locations}) = \rho \sum_{j \text{ neighbor of } i} w_{ij} y_j, \quad (3)$$

where ρ is a parameter that determines the magnitude of the spatial neighborhood effect and w_{ij} are weights that determine the relative influence of location j on location i. The CAR model is mathematically defined as a Gaussian random field, and it may represent the locally structured part of the variation. To better account for local mating, we defined neighbors from the Dirichlet tessellation (François et al. 2006), and we used an exponential covariance matrix to model the decay of correlation with geographic distance

$$w_{ij} = \exp(-d_{ij}/\theta), \quad (4)$$

where d_{ij} is the great-circle distance between the sites i and j and θ is a scale parameter that may be related to the intensity of gene dispersal. More specifically, the expression (3) for y_i implies the covariance matrix $\Lambda = \sigma^2(\mathrm{Id} - \rho W)^{-1}$, where W is an $N \times N$ matrix with zeros on the diagonal and the neighbor weights (w_{ij}) in the off-diagonal positions, Id is the identity matrix, and σ^2 is the variance of the CAR. Equation (3) underlines that ρ and θ are not simultaneously identifiable parameters and that estimates should focus on the product ρW. In practice, we set θ equal to the mean value of great-circle distances between the individual locations, and ρ is estimated from the data. We further refer

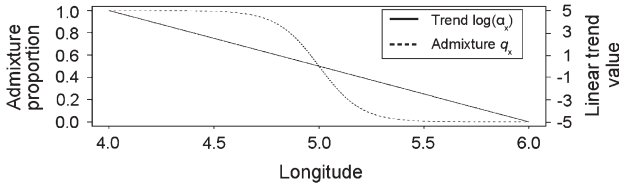


FIG. 1.—Expected admixture proportions q_x as a function of space under a longitudinal linear model, $\log(\alpha_x) = 25 - 5x$, in a simulation of the spatial prior distribution using two clusters. The results for the second cluster are symmetric with respect to the middle of the area.

to the model defined in equation (2) as the full regression model. A model without the CAR component is termed a trend model.

To give correct interpretations of linear trend surfaces, one should keep in mind that the assumption is not that the admixture proportions vary linearly in space. In fact, the model assumes that the q_{ik}s have sigmoidal shapes across space, mirroring theoretical predictions for allele frequency curves in hybrid zones (Barton and Hewitt 1985). To give an illustration of the shape of the admixture proportions under a linear trend model, we simulated realizations of the prior model using two clusters. Assuming dependence on the longitude, x, we parameterized the trend surface as $m_1 = a - bx$ in cluster 1 and as $m_2 = -a + bx$ in cluster 2 ($a = 25, b = 5$), and we sampled individuals along the longitudinal gradient ($x \in [4,6]$). A rough approximation of the average admixture proportion in cluster 1 at longitude x can then be given by $q_{x,1} = 1/(1+\exp(-2(a+bx)))$, which can be represented by a sigmoid curve. Figure 1 shows that the curve of the expected admixture proportions indeed varies spatially with a sigmoidal shape, staying almost constant in each cluster and decreasing sharply at the boundary between two clusters. Simulations with three adjacent clusters displayed similar patterns, with admixture coefficients showing stable values over large regions and varying substantially at their boundaries.

Implementation Details

Our Bayesian model was implemented as a hybrid MCMC algorithm, following Gelman et al. (2004) for the priors on regression models and Metropolis–Hastings rules for the CAR model (supplementary supporting text ST1, Supplementary Material online). For the parameter ρ, we used a noninformative prior over the interval $(0, 1/\lambda_{\max})$, where $\lambda_{\max}$ is the largest eigenvalue of W, and we implemented Metropolis–Hastings updates. An important feature of the hidden regression approach was the possibility to display posterior predictive maps of admixture coefficients. These maps can show the predictions of admixture proportions for an individual at an arbitrary geographic location, adding useful information to the standard unidimensional bar chart representations.

Because the model specified in equation (2) is not the unique way to define a spatially explicit prior for admixture, we implemented variants of the above Bayesian approach. One alternative is to use a multinomial logit regression model for the admixture proportions, q_{ik}, instead of the lognormal model for their average proportional values, α_{ik}. Another alternative is to use a convolution Gaussian prior with two variance parameters, τ^2 and σ^2, as defined by Besag et al. (1991) and used by Mollié (1996) in an epidemiological context (BYM model; supplementary supporting text ST1, Supplementary Material online). The CAR and BYM models are close to each other. Both of them were implemented in the TESS computer program and were used in the subsequent data analyses. They generally led to similar results. For the BYM model, we used noninformative priors on variance parameters, and updates of these parameters were performed according to a Gibbs sampling algorithm.

Model Choice

Following Pritchard et al. (2000), we suggest performing analyses of population structure for a range of values of $K_{\max}$. When choosing the number of population, we need to account for the fact that including a trend surface implies a regularization of the number of observed clusters so that the actual number of clusters, K, may be less than the number specified by the mixture model (François et al. 2006). To decide which $K_{\max}$ (and K) may provide the best fit to the genetic data, we used the deviance information criterion (DIC; Spiegelhalter et al. 2002). The DIC was computed along MCMC runs as the average model deviance plus a penalty term, p_D, that counts the effective number of parameters in a model. To select the number of clusters, the program was run for a range of values of $K_{\max}$, and we considered the values for which the DIC first reached a plateau, like it is usually done for STRUCTURE with the logarithm of evidence (Evanno et al. 2005). The DIC was also useful for selecting among a nonspatial prior, a trend-only prior or the full model (trend plus CAR prior). It allowed us to assess the presence of clines or clusters and to measure the relative importance of large-scale and local effects. In this case, the focus of model selection shifted to choosing the best regression model, and we utilized a conditional version of the DIC based on the average residuals of the hidden regression model (Celeux et al. 2006).

Simulations of Recent Admixture of Two Parental Populations

In a first series of experiments, we simulated spatial genetic data mimicking the instantaneous admixture of two weakly differentiated parental populations. The parental populations were assumed to be in migration/drift equilibrium, and genotypes for $n = 400$ diploid individuals were obtained from structured coalescent simulations with two islands, constant levels of gene flow and constant mutation rate (infinite allele model, $4\mu N_e = 1$). We controlled the simulations by varying the effective migration rate $M = 4mN_e$ between 4 and 12 so that the F_{ST} of the parental gene pool varied in the range $[0.02, 0.05]$ (estimated with HIERFSTAT [Goudet 2005]). To create a spatial framework, the individual locations were randomly generated with Gaussian distributions around two centroids put at distance 2 on a longitudinal axis (standard deviation [SD] = 1). The genotype of each individual at each of L

loci was built as follows. For each individual and each locus, we computed the distance d_1 (d_2) to the left (right) centroid, and we assumed that each allele originated in the first (second) parental population with probability $d_2/(d_1+d_2)$ ($d_1/(d_1+d_2)$). We used $L = 100$ loci. This simulation of individual levels of admixture was similar to the ones classically used in studies of population samples (Chikhi et al. 2001; Griebeler et al. 2006). The simulation imposed a longitudinal trend to the genetic data, with individuals at lower longitude sharing more alleles with the first parental population than with the second one. Spatial autocorrelation was neglected in this simulation process.

Simulations of Contact Zones in Europe

In a second series of simulations, we used spatially explicit simulations to generate synthetic population genetic data following secondary contact. Simulations were performed using SPLATCHE (Currat et al. 2004), a computer program that allows incorporation of geographic and environmental information in the migration scenario. The simulation of the demographic phase occurred in a two-dimensional nonequilibrium stepping-stone model defined on a lattice of ~25,000 cells (or demes) covering Europe. Each deme represented a surface of ~450 km^2 and exchanged migrants with its four neighbors at rate m. Topographic information was imported from a geographical information system, and it was encoded into distinct friction values for each cell. In these simulations, measures of genetic differentiation at neutral loci increased with geographic distance. Population sizes grew logistically at rate r in each deme and saturated at their carrying capacity, C. The three parameters r, m, and C determined the speed of the wave of advance. In our study, the growth rate was set to $r = 0.6$, the migration rate ranged between [0.2, 0.9], and carrying capacities were set either to $C = 100$ or to $C = 1,000$ in each deme. With the tested parameter settings, Europe was colonized in less than 600 generations.

The dynamics were started from an ancestral population of effective size $N_e = 1,000$ individuals. After an initial divergence phase of about 300–500 generations, populations started to colonize Europe from two distant southern foci, one in the Iberian peninsula and the other one in Turkey. Secondary contact occurred in Central Europe, in an area close to Germany. We used a friction map that made migration toward mountainous areas more difficult, and water masses were impossible to cross. We added two isthmi that connected the British Isles to France and Scandinavia to Denmark. We used two values for the total number of generations, $T = 1,000$ and $T = 2,500$. The genetic data were simulated as short tandem repeats at either $L = 10$ or $L = 100$ neutral loci according to the stepwise mutation model. We used a mutation rate of 5×10^{-4} per locus and generation, and we sampled 60 populations at random locations in Europe containing either 3 or 20 individuals per sample. Combining all the simulation parameters, we generated a total of 16 data sets.

Simulations of Equilibrium Stepping-Stone Models

In a third series of experiments, we used EASYPOP (Balloux 2001) to generate spatial genetic data sets under an equilibrium model of isolation by distance. Under this scenario, theory shows that measures of genetic differentiation at neutral loci increase with geographic distance due to the well-known process of accumulation of local genetic differences under geographically restricted dispersal (Wright 1943). Allele frequencies vary across the region, but they do not exhibit regional shapes. Equilibrium stepping-stone simulations are examples of data that do not correspond well to Bayesian clustering model assumptions. In absence of a reasonable number of source populations, the inferred value of the number of clusters and the corresponding allele frequencies in each cluster can be rather arbitrary (Pritchard et al. 2003).

The simulation took place in a two-dimensional stepping-stone model defined on a 10 by 10 lattice. We generated data sets for 60 populations of diploid individuals genotyped at 10 microsatellite loci. The mutation rate was set to $\mu = 5 \times 10^{-4}$, and the migration rate, m, was varied in the interval $[0.3, 0.9]$. Then we created two data sets by randomly resampling three individuals in each population. The presence of long-range isolation by distance was assessed by regressing the pairwise differentiation measures $F_{ST}/(1-F_{ST})$ on the geographic distances.

Application to *F. heteroclitus* Data

The mummichog *F. heteroclitus* is a small killifish. Its habitat ranges from northern Florida to the Gulf of St Lawrence along the eastern coast of North America. It has been shown that *F. heteroclitus* exhibited a steep latitudinal cline using allozymes, mtDNA, and microsatellite markers (Power et al. 1991; Adams et al. 2006). Several hypotheses for this clinal variation have been proposed, including secondary contact between two divergent populations or a northward expansion from a southern refugium after the last glacial age. Using 731 diploid individuals genotyped at eight microsatellite loci, Adams et al. (2006) showed that a pure northward expansion might not explain the observed nuclear pattern of variation, and they suggested an alternative model of postglacial colonization.

MCMC Runs

We studied a total of 22 simulated data sets plus one biological example. The scale parameter θ was set to 1 in the first four data sets (recent admixture) and to $\theta = 1,000$ in the other ones (contact zones). In the scenarios of recent admixture and the equilibrium isolation-by-distance simulations, we present results for the CAR model (similar results were obtained with the BYM model). In secondary contact simulations and for the killifish, we used the CAR and BYM models. Results were almost identical for both models, and we reported results for the second one.

For each data set, we investigated which of a nonspatial, a linear trend, or a full model provided the best fit. These analyses were performed for values of K_{max} ranging from 2 to 7. MCMC algorithms were run for a length of 50,000 sweeps with burn-in periods of 40,000 sweeps. For each data set and each model, we ran the algorithm 100 times, retained the 10 runs with the best DICs, and averaged admixture estimates using CLUMPP (Jakobsson

Table 1
Conditional DIC for Data Sets Simulating Recent Admixture of Two Populations

	No Covariate			Longitudinal Trend			Linear Trend Surface		
F_{ST}	Min	Mean	SD	Min	Mean	SD	Min	Mean	SD
0.03	414.7	419.8	3.87	417.4	422.7	4.31	419.4	424.8	3.41
0.04	416.0	422.9	4.98	**362.0***	369.1	4.94	367.5	398.4	5.72
0.05	387.7	397.7	4.77	**366.4***	379.5	3.72	373.6	383.4	3.22

NOTE.—Min, minimum; SD, standard deviation.

Table 2
Regression Table for a Data Set Simulating Recent Admixture of Two Populations

	Cluster 1		Cluster 2	
	Estimate	95% CI	Estimate	95% CI
Intercept	−25.13	(−33.57 to − 16.44)	23.86	(14.07–34.02)
latitude	1.03	(−0.56 to 2.71)	0.34	(−1.41 to 1.97)
longitude	4.58	(3.18–5.92)	−4.51	(−6.01 to −3.13)

NOTE.—CI, Credibility interval.

and Rosenberg 2007). As the full analysis required 41,400 runs, we put restriction on some computations when the results were obvious (scenarios 1–6). Runs were performed using an upgraded version of the program TESS (Chen et al. 2007) on a cluster of computers.

Results

Recent Admixture of Two Parental Populations

For $K_{max} = 2$ and $F_{ST} \geqslant 0.04$, the smooth longitudinal cline created in the simulated data was uncovered by the spatial algorithms (supplementary fig. S1A, Supplementary Material online). Note that the F_{ST} values were computed before creating admixture and that these numbers were likely to overestimate the true levels of differentiation in the data. Using a conditional version of the DIC for the hidden regression, we evaluated the fit of the nonspatial (trend of degree 0), longitudinal trend (trend of degree 1), and both longitudinal and latitudinal trend (trend of degree 1) models in table 1 (no autocorrelation term). Minimum values were computed over 100 runs (Min) and averages over the 10 best runs (mean and SDs). The best values are bolded and marked with a star. Values for $F_{ST} = 0.02$ were similar to those reported for $F_{ST} = 0.03$. The nonspatial algorithm was unable to obtain correct estimates of the admixture proportions when $F_{ST} = 0.04$. The clustering algorithms failed to uncover the cline at $F_{ST} \leqslant 0.03$. There was a steep decrease of DICs when the cline was detected, shifting from values around 420 to values around 370. In the latter case, the DIC analysis selected the longitudinal trend model (DIC = 362–366) in agreement with the synthetic data generation process. The correlation between the estimated admixture proportions and their true values was also highest for the longitudinal trend model ($r = 0.97$, $P < 10^{-10}$), indicating that the cline was almost perfectly reconstructed by the algorithm. Similar results were obtained for $K_{max} = 3$ and 4 for which $K = 2$ effective clusters were actually detected when $F_{ST} \geqslant 0.04$. We also obtained slightly better performances for these data sets when we used a multinomial logit regression model for the admixture proportions, uncovering the cline at $F_{ST} = 0.03$ (supplementary fig. S1B, Supplementary Material online).

The strength of the spatial effect was measured by the regression coefficients. Table 2 presents these coefficients for the trend model and for the scenario with $F_{ST} = 0.05$. As expected, there was a clear effect of longitude on the admixture proportions. Latitude, on the other hand, had no detectable influence because the credibility interval of its regression coefficient included zero (supplementary fig. S2, Supplementary Material online). Finally, the symmetric role of the two parental populations was reflected by regression estimates that were approximately symmetric for each cluster.

Contact Zones in Europe

The levels of differentiation in the 16 simulated data sets ranged from 0.02 to 0.28. The highest F_{ST}s were observed for the smaller migration rate, number of generations, and carrying capacities. In accordance with classical models, the F_{ST} decreased when one of these parameters increased. In all data sets, longitudinal clines separating the western and eastern part of the continent were inferred as soon as we set $K_{max} \geqslant 2$. These patterns clearly exhibited a contact zone localized in central Europe.

Separate DIC analyses were ran for the BYM model and for the small (180 individuals and 10 loci) and large (1,200 individuals and 100 loci) data sets (supplementary-figs. S3 and S4, Supplementary Material online). When the small data sets were used to compare the nonspatial, trend, and full models using $K_{max} = 3$, the relative differences in the DIC were in favor of the inclusion of spatial covariates. The DIC selected the full regression model $7/8$ times and the trend model for $1/8$ data sets (fig. 2*A*). For $K_{max} = 5$, the spatial models outperformed the nonspatial models, except for one data set (fig. 2*B*). The full model was also selected more often (5/8) than the trend model (2/8). For these data, the effective number of cluster varied between 2 and 4, with the lowest Ks found in data sets with small F_{ST}s. Figure 2*C–D* details the DIC analysis for two data sets (labels 1 and 8). Similar conclusions were reached for the big data sets, but the trend model was selected more often (3/8) than the full model (1/8) as more loci, and larger samples were used.

The main features observed in the spatial population structure analyses are illustrated in figure 3*A–B*, considering one particular data set with demographic parameters $T = 1,000$, $C = 100$, $m = 0.3$, total sample size $n = 1,200$, and $L = 100$ loci. For these data, the DIC selected the linear trend model (DIC = 181,456) and a value of $K_{max} \approx 5$ (label 6; supplementary fig. S4, Supplementary Material online). For $K_{max} = 2$, the admixture estimates exhibited a clinal pattern in central Europe. For $K_{max} = 3$, a clear separation was identified in Scandinavia, a pattern that was observed in a majority of the simulations. For $K_{max} = 4$, a small cluster—particular to the studied data—was found in the northeast of Europe (blue cluster). Setting $K_{max} \geqslant 5$ did not modify the estimates of the admixture proportions significantly. Figure 3*B* displays a posterior map of predicted admixture levels in Europe. The hidden regression model

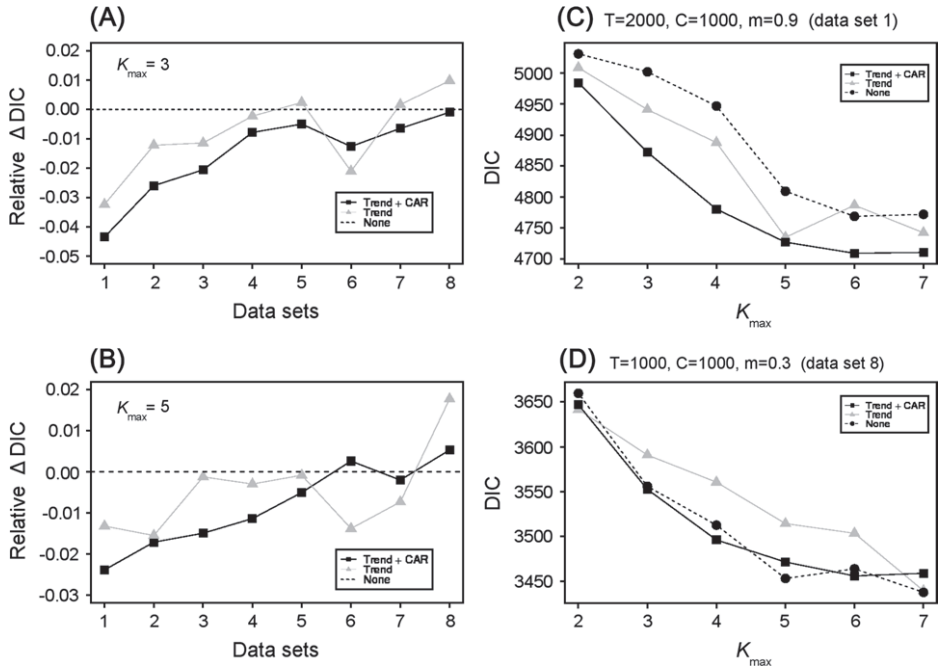


FIG. 2.—Bayesian model choice for secondary contact scenarios. (*A–B*) All simulations (180 individuals and 10 loci). Data sets are labeled 1–8. The plots represent relative differences in DIC for two hidden regression models. The reference model is the model without covariate (nonspatial model, dashed lines). (*C*) DIC as a function of $K_{\max}$ for one simulated data set ($T = 2,000$, $m = 0.9$, $C = 1,000$) for three models. (*D*) DIC as a function of $K_{\max}$ for one simulated data set ($T = 1,000$, $m = 0.3$, $C = 1,000$) for three models.

predicted a long and narrow contact zone (in red pixels) consistent with the shape of hybrid zones observed in many species (Barton and Hewitt 1985).

Equilibrium Stepping-Stone Simulations

For the data set with the largest migration rate, the pairwise F_{ST}s ranged from 0.0004 to 0.11, and the mean differentiation was equal to 0.042 (SD = 0.019). The extent of long-range isolation by distance was assessed in supplementary figure S5 (Supplementary Material online). With the smallest value of the migration rate, the levels of differentiation ranged from 0.0004 to 0.0037. Varying $K_{\max}$ between 2 and 9, we used the DIC to compare the nonspatial models, CAR models (trend of degree 0, $\rho > 0$), trend models (trend of degree 1, $\rho = 0$), and full models (trend of degree 1, $\rho > 0$). For the largest value of the migration rate m, the DIC analysis revealed that the values $K_{\max} = 4$ and 5 received the highest support and that no model performed better than the nonspatial models (supplementary fig. S6, Supplementary Material online). No cluster was effectively discovered. The results for the smallest value of m were similar to those obtained for the largest value. With more extensive sampling (20 individuals in each population) and more genetic data (100 microsatellite neutral markers), again no model performed better than the nonspatial models. We obtained four clusters, located in the corners of the study area that were not subsets of those obtained with $K_{\max} = 3$, suggesting that they might correspond to mathematical artifacts.

Application to *F. heteroclitus*

Using the same scheme as for the simulated data, we fitted nonspatial, linear trend, and full hidden regression models to the *Fundulus* data (BYM model). The linear trend model obtained the best DICs for values of $K_{\max}$ in $[2,7]$ (fig. 4). Increasing $K_{\max}$ above 3 did not lead to a significant decrease in DICs, and the clustering results remained unchanged, suggesting that the effective number of cluster could be estimated as $K = 3$ (fig. 3*C–D*). The best models detected a cline separating the northern and southern populations and grouped two isolated samples to the south to the study area. The posterior predictive map localized the cline to the east to New Jersey (in red pixels, fig. 3*D*) and agreed with the findings of Adams et al. (2006).

Discussion

We proposed a Bayesian algorithm to estimate individual admixture proportions by incorporating spatial trends and spatial autoregressive processes in the prior distribution on these coefficients. The priors were defined as hidden regression models with autocorrelated residuals including spatial effects at multiple scales. Although spatial autoregressive models have been known for a long time in the statistical literature, they have been considered in population genetics only recently (Vounatsou et al. 2000; Wasser et al. 2004). The new algorithms extend a previous work by François et al. (2006) who implemented a hidden Markov random field in a model without admixture. The

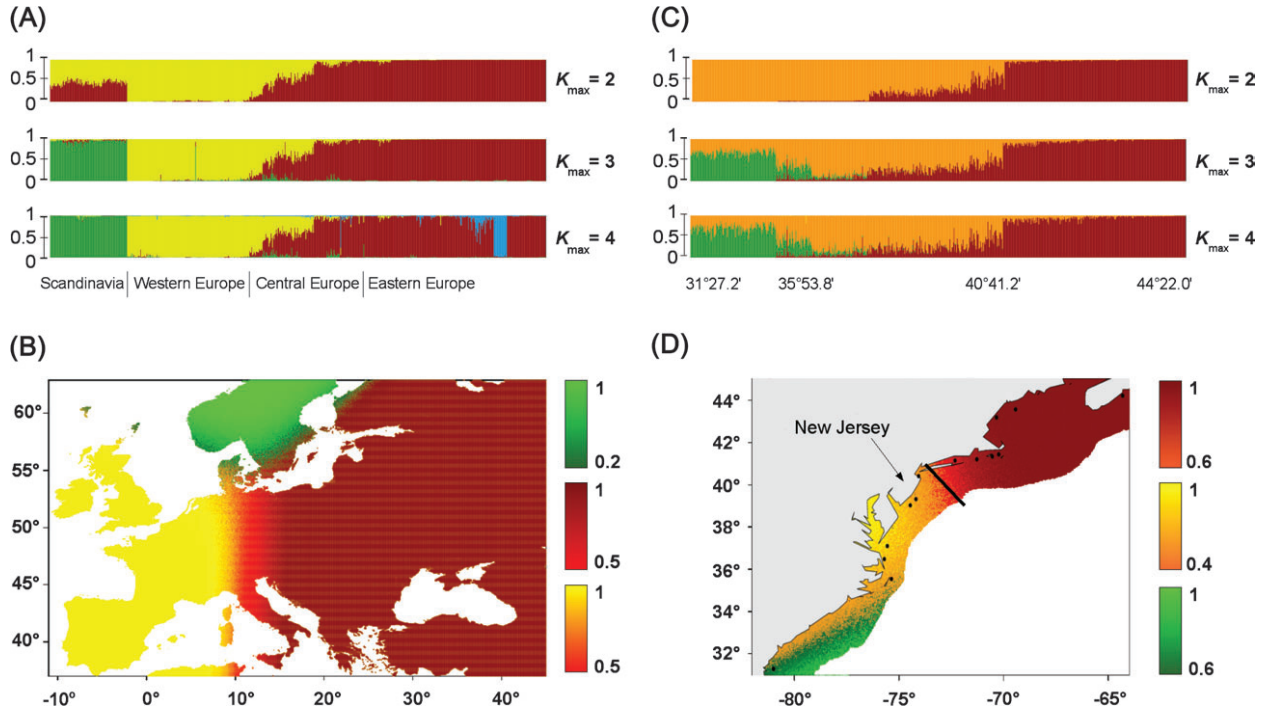


FIG. 3.—Posterior estimates of admixture proportions and predictive maps for selected models. (*A–B*) Range expansion from two refugia ($T = 1,000$, $m = 0.3$, $C = 100$, 1,200 individuals, $L = 100$ loci). These results are representative of a majority of the data sets. The contact zone is highlighted in red pixels. (*C–D*) *Fundulus heteroclitus.* In (*C*), the individuals are sorted by latitude. The cline at latitude 40°41.2′ (red pixels, black line) corresponds to the observation of Adams et al. (2006).

results of our simulation study indicate that our method can outperform those that ignore spatial information, especially when genetic information is not extensive. For example, this is the case in nonmodel species for which extensive genomic data sets are not yet available.

Regression of Admixture Proportions

Regression of admixture coefficients has received much attention in population genetics in recent years. For example, regression was previously utilized to examine the relationships between admixture and geographic distance in Europeans. This was done in order to support the hypothesis of a large contribution of the Neolithic farmers to the current European gene pool. Surveys of admixture clines in this context uncovered an approximate linear relationship between admixture proportions and distance to a putative eastern origin (Chikhi et al. 2001; Dupanloup et al. 2004; Belle et al. 2006) or a true eastern origin when simulations were used (Currat and Excoffier 2005). Because they assumed statistically uncorrelated residuals, regressions of posterior estimates might differ from those obtained by our approach in a drastic way. In our approach, the regression is part of the modeling process. Polynomial trend surfaces may account for clines in all directions, and autocorrelated residuals may account for isolation by distance. Including spatial information in the prior distribution on the admixture proportions can also provide posterior estimates that have been corrected for genealogical correlation between individuals. This is achieved in a rather natural fashion using the hierarchical Bayesian approach (Gelman et al. 2004).

Model Selection and DIC

An important intrinsic feature of imposing spatially structured priors was the possibility for the MCMC algorithm to eliminate a number of spurious clusters automatically. When we input a maximum of K_{max} clusters to the model, the effective number of clusters in the data may be a smaller value, K. In this case, the DIC sometimes selects models in which K_{max} is greater than K. An explanation may be the variability in estimated DICs. Theory pre-

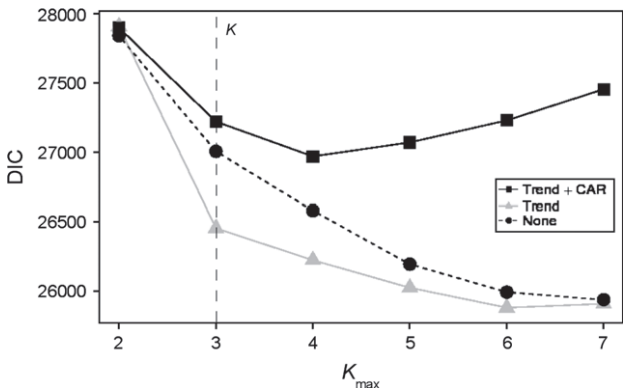


FIG. 4.—DIC as a function of K_{max} for *Fundulus heteroclitus*. The vertical dashed line corresponds to estimated effective number of clusters $K = 3$ obtained from the linear trend model.

dicts that errors in information criterion comparisons are of order $\sqrt{n}$, where n is the number of observations (Ripley 1996). We suspect that the constant term in this large-sample approximation could be rather big, especially in complex hierarchical models as implemented in this study. In the killifish example, models with $K_{\max} = 4$ and 5 clusters were given smaller values of the DIC than models with $K_{\max} = 3$ clusters. Nevertheless, it was obvious from the direct inspection of the posterior estimates that the effective number of populations was equal to 3 in the selected models. It is possible that the DIC decrease—around 100 units—may not be large enough to justify a choice of a model with a larger number of clusters. Note that although the DIC is widely acknowledged to be a useful measure, it does not always lead to choosing the best model (Brooks 2002).

Simulation Analyses

In the simulations of recent admixture, a given level of admixture was assigned to each individual according to a pure longitudinal trend model. These simulations were an approximation of more complex spatially explicit processes, for which we neglected spatial autocorrelation. The DIC analysis selected the correct covariate, and the observed number of clusters in the data agreed with two parental populations. The posterior estimates of the admixture coefficients exhibited a longitudinal clinal shape, as we expected. In secondary contact simulations, the best models were obtained when we included both the trend and autocorrelation terms in the statistical model. The estimated trends were apparent in the prediction maps, and they were oriented along a longitudinal axis. They were visible for $K_{\max} \geqslant 2$, and they captured the signature of the simultaneous range expansion from the two refugia. The inclusion of autocorrelation in the best model was not a surprising result as sampling was dense enough to observe the short-range effects that are inherent to the stepping-stone simulation (the average distance between nearest samples was around 300 km). The prediction maps for the admixture proportion described and highlighted the areas where the hybrids resided. These hybrid zones conformed to their theoretical predictions (Barton and Hewitt 1985). In some runs, more than three clusters were actually found, especially when we used the larger number of loci and the larger sample sizes. Only the continental cline and the northern cluster were consistently present in all runs. The additional clusters were often located in the northeast or in the British Isles and might have resulted from drift or localized founder effects within the main cline. Such founder effects were more frequent when the Baltic sea was crossed, leading us to observe a Scandinavian cluster more frequently.

One potential source of misleading interpretations is with data sets arising from homogeneous short-range migration process across time and space. Such data clearly violate the spatial admixture model assumptions. The formulation of the admixture model accounts for short-range isolation-by-distance effects by the way of the autocorrelated residuals and for regional effects by means of the latent regression model and the trend surface. Under an equilibrium stepping-stone model, we expect a long-range isolation-by-distance pattern. Because there are no regional effects, the trend surface is not useful, and genetic variation is partitioned over artificial clusters like for other Bayesian clustering algorithms. In addition, we observe that the estimated clusters are inconsistent over increasing values of $K_{\max}$. In contrast, the reason why it works well in the case of a secondary contact zone is that, in this case, variation is more structured and exhibits regional trends. Regional effects are well taken into account by the latent regression, which makes clusters easier to identify than in pure equilibrium situations. The residual autoregressive term can improve the admixture model by taking care of the short-range isolation by distance. Note that the goal of the proposed algorithm differs from detecting isolation by distance. For an approach able to separate the effects of isolation by distance from migration and to give an estimate of the scale at which each process operates see Bocquet-Appel and Sokal (1989).

Secondary Contact Hypothesis for the Killifish

The killifish *F. heteroclitus* has served as a model for understanding the local adaptation to variable environments (Avise 2004). This species is known to exhibit latitudinal clinal variation in a number of physiological traits, and patterns at mitochondrial and nuclear DNA loci have suggested a complex history of spatially variable selection and secondary contact, with an abrupt genetic transition between northern and southern populations (Adams et al. 2006). The spatial population structure analysis inferred a cline that separated the northern and southern populations. Adams et al. (2006) suggested that this cline was the result of recolonization of the whole current habitat from unfrozen water at the end of the last glacial age, creating a secondary contact zone between northern and southern populations. The best model did not include spatial autocorrelation effects. An explanation may be the use of population samples, which perhaps removed some local aspects of variation. We think that including spatial autocorrelation would have been more useful if individual sampling had been performed uniformly within the study area. A third cluster corresponded to the two southernmost samples of killifish. Because these two samples were geographically isolated from the rest, it was difficult to decide whether the smooth variation observed to the south of the area could be attributed to isolation by distance, that is, an artificial cluster, or to historical patterns of migration. In any case, coupling Bayesian clustering methods with additional demographic analyses seems always necessary as secondary contact and isolation by distance in an irregular sampling design might produce confounding signals.

Clines and Clusters

The methods presented in this study have the potential to detect coexisting clines and clusters through the inferred variation of admixture proportions (for a related discussion on clustering algorithms, see Rosenberg et al. [2005]). This was emphasized by the analysis of simulations of range expansion from two refugia. In these spatially explicit simulations, the algorithm detected a contact zone at the same time as it found clusters in the north of

Europe and elsewhere. In general, it might be difficult to distinguish between clines and clusters without a good spatial coverage of the study area. In this case, a DIC analysis will provide an assessment of the relative contribution of clines and clusters to the posterior estimates of the admixture coefficients. For example, a nonspatial analysis for the killifish data suggested the existence of four clusters partitioning the southern cline, but a spatial analysis coupled to a DIC evaluation indicated that a cline merging two clusters better explained the data.

Comparisons with Simpler Methods

Relationships between Bayesian clustering algorithms and PCA have been emphasized by Patterson et al. (2006) who considered a model of genetic structure in which populations have diverged from an ancestral population recently. If the model assumes K populations, PCA is then expected to have $K-1$ significant components under the Tracy–Widom theory (Patterson et al. 2006). Applying PCA to the killifish, the cline and the southern genetic cluster were visible in the first and in the third eigenvectors (PC1 and PC3; supplementary fig. S7, Supplementary Material online). In this example, the patterns found in PC1 and PC3 match those computed by the Bayesian clustering program. In simulations of recent admixture, the tests were significant for PC1 only, and this axis of variation clearly captured clinal variation at the contact zone. This was to be expected because the informative panel F_{ST} was low and the theory could be expected to perform very well. In contrast, the Tracy–Widom theory yielded more than 15 significant axes of variation ($P < 0.01$) in some simulations of contact zones in Europe (supplementary fig. S8, Supplementary Material online). For these components, the genetic meaning was hard to interpret. This happened in situations where the informative panel F_{ST} was high (>0.10) and the Tracy–Widom theory less valid. In this case, the Bayesian algorithm was more robust as it always detected no more than five clusters and provided interpretable values for the admixture proportions. Nevertheless, the first PCs always included the cline and clusters found by the Bayesian clustering algorithm, and we believe that the two methods are useful complementary exploratory tools.

Concluding Remarks

Bayesian algorithms for inference of population structure have traditionally focused on finding clusters, whereas less efforts have been devoted to detect clinal variation. To provide a better description of the relative contribution of clines and clusters, we coupled a multiscale spatial admixture analysis with a Bayesian assessment of model complexity and fit. This approach reduces the number of spurious clusters when the underlying variation is mainly clinal, while still detecting zones of small genetic discontinuities. Our new algorithm provides more accurate estimates of the admixture proportions compared with standard nonspatial methods, and this suggests its use when studying the spatial population structure, secondary contact zones, and when correcting for the population structure in phenotype–genotype association studies.

Supplementary Material

Supplementary supporting text ST1 and figures S1A, S1B, S2–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Stephanie Adams for communicating the *Fundulus* data and to Nicolas Ray for providing us a recent version of SPLATCHE. We also thank Michael GB Blum, Nick Patterson, Jonathan K Pritchard, and an anonymous referee for their comments. Simulations were run on the UJF-CIMENT cluster of computers (http://healthphy.grenoble.cnrs.fr/). OF was supported by grant ANR BLAN06-3-146282 MAEV.

Literature Cited

Adams SM, Lindmeier JB, Duvernell DD. 2006. Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. Mol Ecol. 15:1109–1123.

Avise JC. 2004. Molecular markers, natural history, and evolution. 2nd ed. Sunderland (MA): Sinauer Associates.

Balloux F. 2001. EASYPOP (version 1.7): a computer program for the simulation of population genetics. J Hered. 92:301–302.

Barton NH, Gale KS. 1993. Genetic analysis of hybrid zones. In: Harrison RG, editor. Hybrid zones and the evolutionary process. Oxford: Oxford University Press. p. 13–45.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. Annu Rev Ecol Syst. 16:113–148.

Belle EMS, Landry P-A, Barbujani G. 2006. Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proc R Soc Lond B Biol Sci. 273:1595–1602.

Bertorelle G, Excoffier L. 1998. Inferring admixture proportions from molecular data. Mol Biol Evol. 15:1298–1311.

Besag J. 1975. Statistical analysis of non-lattice data. Statistician. 24:179–195.

Besag J, York J, Mollié A. 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). Ann I Stat Math. 43:1–59.

Bocquet-Appel JP, Sokal RR. 1989. Spatial autocorrelation analysis of trend residuals in biological data. Syst Zool. 38(4):333–341.

Brooks SP. 2002. Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde. J R Stat Soc Ser B. 64:616–639.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton (NJ): Princeton University Press.

Celeux G, Forbes F, Robert CP, Titterington DM. 2006. Deviance information criteria for missing data models. Bayesian Anal. 1:651–674.

Chakraborty R. 1986. Gene admixture in human populations: models and predictions. Yearb Phys Anthropol. 29:1–43.

Chakraborty R, Weiss KM. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA. 85:9119–9123.

Chen C, Durand E, Forbes F, François O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes. 7:747–756.

Chikhi L, Bruford MW, Beaumont MA. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics. 158:1347–1362.

Corander J, Sirén J, Arjas E. 2008. Bayesian spatial modeling of genetic population structure. Comput Stat. 23:111–129.

Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.

Currat M, Excoffier L. 2005. The effect of the Neolithic expansion on European molecular diversity. Proc R Soc Lond B Biol Sci. 272:679–688.

Currat M, Ray N, Excoffier L. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Mol Ecol Notes. 4:139–142.

Dormann CF. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Glob Ecol Biogeogr. 16:129–138.

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. 2004. Estimating the impact of prehistoric admixture on the Europeans' genome. Mol Biol Evol. 21:1361–1372.

Durrett R, Buttel L, Harrison R. 2000. Spatial models for hybrid zones. Heredity. 84:9–19.

Endler JA. 1977. Geographic variation, speciation, and clines. Princeton (NJ): Princeton University Press.

Epperson BK, Li T. 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proc Natl Acad Sci USA. 93:10528–10532.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14:2611–2620.

Excoffier L, Estoup A, Cornuet JM. 2005. Bayesian analysis of an admixture model with mutations and arbitrary linked markers. Genetics. 169:1727–1738.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567–1587.

Faubet P, Gaggiotti OE. 2008. A new Bayesian method to identify the environmental factors that influence recent migration. Genetics. 178:1491–1504.

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

François O, Ancelet S, Guillot G. 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics. 174:805–816.

François O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic history of European populations of *Arabidopsis thaliana*. PLoS Genet. 4(5):e1000075.

Gaggiotti OE, Brooks SP, Amos W, Harwoods J. 2004. Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. Mol Ecol. 13:811–825.

Gao HS, Williamson S, Bustamante CD. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics. 176:1635–1651.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian data analysis. Boca Raton (FL): Chapman and Hall/CRC Press.

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

Griebeler EM, Müller JC, Seitz A. 2006. Spatial genetic patterns generated by two admixing genetic lineages: a simulation study. Conserv Genet. 7:753–766.

Hewitt G. 2000. The genetic legacy of the quaternary ice ages. Nature. 405:907–913.

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. 2004. Design and analysis of admixture mapping studies. Am J Hum Genet. 74:965–978.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 23:1801–1806.

Jombart T, Devillard S, Dufour A-B, Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity. 101:92–103.

Lichstein JW, Simons TR, Shriner SA, Franzreb KE. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol Monogr. 72:445–463.

Long JC. 1991. The genetic structure of admixed populations. Genetics. 127:417–428.

Meng XF, Shi M, Chen XX. 2008. Population genetic structure of *Chilo suppressalis* (Walker) (Lepidoptera: Crambidae): strong subdivision in China inferred from microsatellite markers and mtDNA gene sequences. Mol Ecol. 17:2880–2897.

Mollié A. 1996. Bayesian mapping of disease. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. London: Chapman & Hall p.359–379.

Pastorini J, Zaramody A, Curtis DJ, Nievergelt CM, Mundy NI 2009. Genetic analysis of hybridization and introgression between wild mongoose and brown lemurs. BMC Evol Biol. 9(1):32.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Power DA, Lauerman T, Crawford DL, DiMichele L. 1991. Genetic mechanisms for adapting to a changing environment. Annu Rev Genet. 25:629–659.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Pritchard JK, Wen X, Falush D. 2003. Documentation for structure software: version 2.3. Department of Human Genetics, University of Chicago.

Reich D, Patterson N. 2005. Will admixture mapping work to find disease genes? Philos Trans R Soc Lond B Biol Sci. 360:1605–1607.

Ripley BD. 1981. Spatial statistics. New York: Wiley.

Ripley BD. 1996. Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Roberts DF, Hiorns RW. 1965. Methods of analysis of the genetic composition of a hybrid population. Hum Biol. 37: 38–43.

Rosenberg NA, Saurabh S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the influence of human population structure. PLoS Genet. 1:660–671.

Sankararaman S, Kimmel G, Halperin E, Jordan MI. 2008. On the inference of ancestries in admixed populations. Genome Res. 18:668–675.

Slatkin M, Arter HE. 1991. Spatial autocorrelation methods in population genetics. Am Nat. 138:499.

Smith MW, O'Brien SJ. 2005. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet. 6:623–632.

Sokal RR, Oden NL. 1978. Spatial autocorrelation in biology. I. Methodology. Biol J Linn Soc. 10:199–228.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit (with discussion). J R Stat Soc Ser B. 64:583–639.

Taberlet P, Fumafalli L, Wust-Saucy AG, Cosson J-F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol. 7:453–464.

Vounatsou P, Smith T, Gelfand AE. 2000. Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. Biostatistics. 1:177–189.

Wang J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics. 164:747–765.

Wasser SK, Shedlock AM, Comstock K, Ostrander EA, Mutayoba B, Stephens M. 2004. Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. Proc Natl Acad Sci USA. 101:14847–14852.

Wright S. 1943. Isolation by distance. Genetics. 28:139–156.

Yannic G, Basset P, Hausser J. 2008. Phylogeography and recolonization of the Swiss Alps by the Valais shrew (*Sorex antinorii*), inferred with autosomal and sex-specific markers. Mol Ecol. 17:4118–4133.

Yu J, Pressoir G, Briggs WH, et al. (12 co-authors). 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203–208.

Jonathan Pritchard, Associate Editor

Accepted May 13, 2009

Titre : Méthodes bayésiennes en génétique des populations : relations entre structure génétique des populations et environnement.

Résumé : Nous présentons une nouvelle méthode pour étudier les relations entre la structure génétique des populations et l'environnement. Cette méthode repose sur des modèles hiérarchiques bayésiens qui utilisent conjointement des données génétiques multi-locus et des données spatiales, environnementales et/ou culturelles. Elle permet d'estimer la structure génétique des populations, d'évaluer ses liens avec des covariables non génétiques, et de projeter la structure génétique des populations en fonction de ces covariables. Dans un premier temps, nous avons appliqué notre approche à des données de génétique humaine pour évaluer le rôle de la géographie et des langages dans la structure génétique de populations amérindiennes. Dans un deuxième temps, nous avons étudié la structure génétique des populations pour 20 espèces de plantes alpines, et nous avons projeté les modifications intraspécifiques qui pourront être causées par le réchauffement climatique.

Mots-clés : structure génétique des populations, covariables environnementales, modèles bayésiens hiérarchiques, modèles à classes latentes, MCMC, modèles bioclimatiques.

Title: Bayesian methods for population genetics: relationships between population genetic structure and environment.

Abstract: We introduce a new method to study the relationships between population genetic structure and environment. This method is based on Bayesian hierarchical models which use both multi-loci genetic data, and spatial, environmental, and/or cultural data. Our method provides the inference of population genetic structure, the evaluation of the relationships between the structure and non-genetic covariates, and the prediction of population genetic structure based on these covariates. We present two applications of our Bayesian method. First, we used human genetic data to evaluate the role of geography and languages in shaping Native American population structure. Second, we studied the population genetic structure of 20 Alpine plant species and we forecasted intra-specific changes in response to global warming.

Keywords: population genetic structure, environmental covariates, Bayesian hierarchical models, latent class models, MCMC, bioclimatic models.