Statistiques bayésiennes en génétique des populations : modèle à facteurs et processus gaussiens pour étudier la variation génétique neutre et adaptative

Nicolas Duforet-Frebourg

▶ **To cite this version:**

Nicolas Duforet-Frebourg. Statistiques bayésiennes en génétique des populations : modèle à facteurs et processus gaussiens pour étudier la variation génétique neutre et adaptative. Génétique des populations [q-bio.PE]. Université de Grenoble, 2014. Français. NNT : 2014GRENS015 . tel-01558168

HAL Id: tel-01558168
https://theses.hal.science/tel-01558168

Submitted on 7 Jul 2017

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

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

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

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

Arrêté ministériel : 7 août 2006

Présentée par

Nicolas Duforet-Frebourg

Thèse dirigée par **Michael Blum**

préparée au sein **du laboratoire TIMC-IMAG**
et de **l'Ecole doctorale de l'ingénierie pour la santé, la cognition et l'environnement**

Statistiques bayésiennes en génétique des populations : modèle à facteurs et processus gaussiens pour étudier la variation génétique neutre et adaptative

Thèse soutenue publiquement le **Jeudi 2 octobre 2014**,
devant le jury composé de :

Denis Allard
Directeur de recherche, INRA Avignon, Président
Christophe Ambroise
Professeur, Université d'Evry val d'Essone, Rapporteur
Renaud Vitalis
Directeur de recherche, INRA Montpellier, Rapporteur
Julien Mairal
Chargé de recherche, INRIA Montbonnot, Examinateur
Michael Blum
Chargé de recherche CNRS, Directeur de thèse

Résumé

Résumé Nous présentons dans cette thèse plusieurs travaux de statistiques bayésiennes appliquées à la génétique des populations. La génétique des populations a pour but d'expliquer les variations génétiques au sein d'une espèce, et d'inférer les processus ayant conduits à ces variations. Pour cela, des données génétiques massives sont utilisées et il y a un besoin grandissant de méthodes statistiques pour traiter ces données. Le travail de cette thèse s'inscrit dans cet effort de modélisation statistique pour répondre aux enjeux de la génétique des populations, et de la biologie de l'évolution. Nous nous intéressons tout particulièrement à la détection de traces d'adaptation locale dans les génomes, et à l'inférence des variations spatiales non stationnaires. Un modèle d'analyse factorielle bayésien est proposé pour détecter les traces d'adaptation locale. Nous comparons notre approche aux méthodes existantes, et démontrons qu'elle permet d'obtenir un plus faible taux de fausses découvertes. Nous présentons également un modèle bayésien basé sur des processus gaussiens pour caractériser les variations génétiques spatiales dans l'aire de répartition d'une espèce. Les performances de ces méthodes sont démontrées sur différents exemples issus de simulations ou de données. Plusieurs logiciels open source qui implémentent ces méthodes ont été développés pendant la thèse.

Mots-clés Statistiques bayésiennes, génétique des populations, adaptation locale, F_{st}, Non stationnarité.

Abstract In this thesis we present several works related to Bayesian statistics in population genetics. Population genetics aims at explaining genetic variation within natural species, and infer the different processes that lead to current genetic variation. Large scale genomic datasets are produced, and there is an increasing need of statistical methods to extract information from these datasets. My thesis work is part of this statistical modeling effort to answer to evolutionary biology and population genetic questions. We are interested in detecting footprints of local adaptation without, and infering non-stationary patterns of spatial variation. A Bayesian factor model is used to detect genes involved in local adaptation. We compare our factor model to existing methods, and show that it can reduce the false discovery rate. We also present a Bayesian model based on Gaussian processes to caracterize spatial genetic variations within species. The performances of these methods are tested on simulations and real datasets. Several open source software are available online.

Keywords Bayesian statistics, population genetics, local adaptation, F_{st}, Non-stationary process.

Table des matières

1 Introduction **9**

1.1 La génétique des populations 9

1.2 Applications 10

1.3 Données en génétique des populations 11

1.3.1 Les marqueurs génétiques 11

1.3.2 Jeux de données de référence 12

1.4 Objectif de la thèse 14

2 Modèles mécanistes pour détecter l'adaptation locale **17**

2.1 Modèles mécanistes et modèles phénoménologiques 17

2.2 Modèles d'évolution de populations et les F_{st} 18

2.2.1 Modèle de divergence et modèle en îles 18

2.2.2 La mesure de F_{st} et le test de Lewontin Krackauer 19

2.2.3 Scanner le génome avec les F_{st} 21

2.3 Modèles mécanistes pour l'adaptation 23

2.3.1 fdist 23

2.3.2 Bayescan 23

2.4 Vers un modèle phénoménologique 25

3 Un modèle phénoménologique pour détecter l'adaptation locale **27**

3.1 Analyse factorielle 27

3.1.1 Présentation du modèle 27

3.1.2 Lien avec l'Analyse en Composantes principales et applications en génétique des populations 29

3.1.3 Applications en génétique des populations 30

3.1.4 Version bayésienne du modèle 33

3.2 Détection de la sélection 34

3.2.1 Interprétations des *loadings* 34

3.2.2 Variables indicatrices z et modèle de Potts 36

3.2.3 MCMC et facteurs de Bayes 37

3.2.4 Application à un modèle de divergence 38

3.3 Vers une version rapide du logiciel 39

3.3.1 Analyse en composantes principales des données 39

3.3.2 Statistiques de l'ACP 40

3.3.3 Intérêts et limites . . . 41
3.3.4 Analyse du jeu de données du projet 1000 Génomes . . . 45

4 Un modèle pour la différenciation génétique locale 49
4.1 Le modèle d'isolement par la distance . . . 49
4.1.1 Modèle d'isolement par la distance . . . 49
4.1.2 Schéma d'isolement par la distance et non-stationnarité . . . 51
4.2 Non-stationnarité et processus gaussiens . . . 52
4.2.1 Processus Gaussiens . . . 52
4.2.2 Modèle paramétrique non-stationnaire . . . 53
4.3 La différenciation génétique locale . . . 55
4.3.1 Définition d'un voisinage . . . 55
4.3.2 Prédiction de la covariance . . . 56
4.4 Applications . . . 59
4.4.1 Exemple sur des simulations . . . 59
4.4.2 Exemple sur des données de plantes alpines . . . 59

5 Travaux annexes en génétique des populations 63
5.1 Construction d'haplotypes informatifs . . . 63
5.1.1 Le gain d'information pour l'assignation . . . 63
5.1.2 Un algorithme de combinaison . . . 64
5.2 Structure génétique de la population de l'Ouest de la France . . . 65
5.2.1 Données du projet DESIR-CavGens . . . 65
5.2.2 Analyse de structure . . . 65

6 Conclusions et perspectives 69
6.1 Conclusions . . . 69
6.2 Perspectives . . . 70

7 Appendices 87
7.1 Estimer les F_{st} . . . 87
7.2 Algorithme MCMC pour le modèle à facteurs . . . 89

8 Notes bibliographiques 91

8.1 Article A : *Nicolas Duforet-Frebourg, Eric Bazin, and Michael G.B. Blum Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model. Molecular Biology and Evolution, 2014, doi : 10.1093/molbev/msu182*. 91
8.2 Article B : *Nicolas Duforet-Frebourg and Michael G.B. Blum, Nonstationary patterns of isolation-by-distance, inferring measures of local genetic differentiation with Bayesian kriging. Evolution, 2014, 68 : 1110–1123. doi : 10.1111/evo.12342*. 91
8.3 Article C : *Nicolas Duforet-Frebourg, Lucie M. Gattepaille, Michael G.B. Blum and Mattias Jakobsson. HaploPOP : a software that improves population assignment by combining markers into haplotypes. En preparation* 91
8.4 Article D : *Matilde Karakachoff, Nicolas Duforet-Frebourg, et al., Fine scale human genetic structure in Western France. European journal of Human genetics, 2014, AOP, doi : 10.1038/ejhg.2014.175*. 91

1 Introduction

1.1 La génétique des populations

L'objectif premier de la génétique des populations est de décrire les variations génétiques entre individus, au sein d'une même espèce. Cette étude est justifiée car les génotypes des membres d'une même espèce comportent en général des différences. A un même emplacement physique sur un chromosome, l'ADN peut varier entre individus, et ces différentes versions de l'ADN sont appelées allèles. Les individus portent ainsi différents allèles, ce qui représente un ensemble de variations génétiques au sein de l'espèce. Ces variations émergent au fil des générations, de telle sorte que lorsque l'on étudie des données en génétique des populations, on n'observe que le résultat à un instant t de la superposition de différents processus aléatoires qui ont agit au fil des générations précédentes. En d'autres termes, une espèce évolue, et nous observons le résultat de cette évolution au temps présent.

Différents processus influencent cette évolution. Comme l'écrit Luigi Luca Cavalli-Sforza dans [Cavalli-Sforza, 2011], il existe "quatre piliers de l'évolution" qui sont les forces agissant sur l'évolution d'une espèce. Ces quatre processus sont les mutations, les migrations, la dérive génétique et la sélection naturelle. Les trois premiers sont des processus dits neutres. La variation génétique est dit neutre si elle n'affecte pas la capacité d'un organisme à survivre et à se reproduire, c'est à dire si elle n'affecte pas son adéquation à l'environnement [Holderegger et al., 2006]. Les mutations peuvent apparaître sur chaque nucléotide du matériel génétique avec un certain taux, appelé le taux de mutation. Les migrations se produisent lorsque les individus sont répartis en sous-groupes séparés par des barrières naturelles, telles qu'une mer, ou une chaîne de montagnes. Certains individus traversent occasionnellement ces barrières, et deviennent ainsi des migrants. La troisième force, la dérive génétique, est plus abstraite. Elle consiste en l'accumulation au fil des générations des changements imprévisibles dans les fréquences d'apparition des allèles [Gillespie, 2010]. Ces changements résultent de la formation aléatoire des gamètes. Cette dérive est d'autant plus complexe qu'elle est influencée par la démographie. Toutes les espèces peuvent au cours de leur histoire connaître des expansions, ou des réductions drastiques du nombre de leurs représentants. Ces changements de taille d'effectif ont des effets connus sur l'intensité de la dérive [Hartl and Clark, 1997]. Les variations dans la fréquence d'apparition d'un allèle sont d'autant plus grandes que le nombre d'individus dans la population est faible.

La quatrième grande force est la sélection naturelle. Contrairement aux autres, celle ci n'est pas neutre, et résulte d'une pression de l'environnement. Lorsque la sélection s'exerce, les individus les plus adaptés survivent en moyenne plus longtemps, ou ont une fertilité accrue. Ils ont ainsi

plus de chance de se reproduire. La sélection peut se décliner sous plusieurs formes. La sélection peut être dite positive, dans le cas où elle favorise les mutations et les nouveaux allèles. Elle peut être négative, ou purifiante, dans le cas où les individus porteurs de nouveaux variants seraient défavorisés. Un autre type de sélection qui retiendra notre attention est la sélection diversifiante où différents allèles sont favorisés dans différentes populations [Gaggiotti, 2013]. Ainsi certains allèles confèrent un avantage dans certaines populations d'une espèce, et leurs porteurs voient leur chance d'avoir une descendance et de transmettre ce gène augmentée. Lorsqu'une telle pression sélective s'exerce à long terme sur une population, les individus ont une plus grande adéquation à leur habitat, que des individus provenant d'autres habitats. On parle alors d'adaptation locale [Kawecki and Ebert, 2004]. L'habitat est défini par un ensemble de variables environnementales.

Ainsi, les variations génétiques entre les individus d'une population sont le résultat de forces évolutives et de processus démographiques. Le but de la génétique des populations est de retrouver les empreintes laissées par ces quatre piliers évolutifs sur la diversité génétique. Le travail de thèse s'inscrit dans cet effort de recherche.

1.2 Applications

La plupart des organismes connaissent des histoires démographiques complexes. Leur aire de répartition et le nombre d'individus qui composent leur population peut changer, produisant des schémas complexes qui laissent des traces dans les variations génétiques de leur population. La génétique des populations permet de reconstruire à l'aide de marqueurs génétiques les processus qui ont pu conduire une population d'organismes à ce qu'elle est aujourd'hui.

Néanmoins, s'il est aujourd'hui possible de retracer l'histoire démographique de la majeure partie des espèces, le travail de la génétique des populations se concentre majoritairement sur un faible nombre d'espèces. Comme John Gillespie l'écrit dans son livre "Population genetics and its great obsession grew out of this fascination with variation in species we love, not out of a desire to explain the origin of major evolutionary novelties" [Gillespie, 2010]. L'une de ces espèces dont nous aimons retracer l'évolution est l'espèce humaine. La reconstruction des processus démographiques et adaptatifs a par exemple permis de mieux comprendre comment l'homme moderne a colonisé l'ensemble de la planète. Nous donnons ci-dessous des exemples de processus neutres et adaptatifs qui ont façonné la variabilité génétique humaine.

La grande quantité de données disponibles a permis d'apporter des éléments de réponse, ou de formuler des hypothèses quant à l'origine de l'espèce humaine, telle que l'hypothèse out-of-Africa [Hammer et al., 1998] déclarant que les premiers individus *Homo sapiens* étaient localisés en Afrique de l'est ou du sud, avant de coloniser le reste du monde. Les données ont également per-

mis d'avancer des éléments de réponse quant à la parenté entre les hommes modernes et Néandertal [Green et al., 2010].

Les données de génétique des populations permettent aussi de détecter dans le génome des traces de l'adaptation locale. Un exemple d'adaptation locale peut être observé chez les individus de la population tibétaine. Les Tibétains se sont installés dans les hauts plateaux himalayens il y a plus de 5000 ans. A très haute altitude l'oxygène se fait plus rare et les individus sont soumis à des conditions d'hypoxie. Les individus de la population tibétaine possèdent des traits leur conférant un avantage en condition d'hypoxie tels qu'une fréquence respiratoire accrue et une plus forte pression sanguine. Les études génétiques ont permis de mettre en avant les différents facteurs de transcription qui confèrent ces capacités aux individus tibétains [Simonson et al., 2010] [Xu et al., 2011]. Ce résultat est l'un des exemples d'adaptation locale les plus récents mis en évidence chez l'espèce humaine.

1.3 Données en génétique des populations

1.3.1 Les marqueurs génétiques

Avec l'avènement des nouvelles technologies pour le séquençage, la quantité de données disponibles pour étudier la diversité génétique d'une espèce est abondante. Ces jeux de données sont constitués de marqueurs génétiques présents sur chaque individu des données. Nous appelons marqueur génétique un locus polymorphe. Il existe différents types de polymorphismes génétiques. Nous citons les plus utilisés dans les données.

Polymorphisme d'un unique nucléotide (SNP) Le polymorphisme le plus simple possible, est une différence d'un nucléotide. A une position physique donnée d'un chromosome, une base azotée, Adénine, Thymine, Cytosine ou Guanine, est remplacée par une autre. Ce changement est le résultat d'une mutation apparue aléatoirement chez un individu (Figure 1). Ce polymorphisme est le marqueur que nous rencontrons le plus souvent dans les données. Nous supposons toujours que ce polymorphisme n'a que deux versions possibles, c'est à dire qu'il est bi-allélique.

Insertions-deletions Les insertions-délétions, ou indel, sont des suppressions ou ajout d'un groupe de nucléotides dans la séquence. Ces insertions proviennent d'enchâssement de matériel génétique dans un chromosome. Les tailles de ces groupes de nucléotides varient d'une base à une région génomique entière. Les conséquences de ces insertions ou délétions sont variables, selon si elles touchent une région codant des protéines.

Microsatellites Un microsatellite est une chaîne de nucléotides qui comporte la répétition d'un mot. Ce mot est une chaîne d'au moins deux nucléotides. Le polymorphisme de ces marqueurs provient du nombre de répétitions du mot. C'est un marqueur qui peut posséder un grand nombre d'allèles. Dans ce cas, chaque allèle représente un nombre de répétitions différent.

Il existe d'autres polymorphismes tel que le polymorphisme de longueur des fragments amplifiés. Dans les données que nous considérons par la suite nous appliquons les méthodes statistiques développées à des marqueurs polymorphiques du type SNP.

Notons cependant que la quantité, et la qualité, des données varient énormément en fonction de l'espèce. Certaines espèces pour lesquelles on dispose d'importantes masses de données génomiques, appelées espèces modèles sont choisies pour être étudiées de façon plus approfondie. Ces espèces sont choisies selon des critères génétiques, expérimentaux, et économiques [Twyman, 2002]. Ces organismes doivent avoir un génome de taille relativement faible, comparée à la taille du génome humain. Elle doivent également offrir des facilités d'élevage en laboratoire tel que des générations courtes. L'objectif est de pouvoir comprendre les processus biologiques qui ont influencé l'évolution de cette espèce. Pour ces espèces modèles, les données couvrent l'intégralité du génome, et sont disponibles pour un grand nombre d'individus de l'espèce. En génétique des populations, plusieurs espèces, animales ou végétales sont des espèces modèles, *Arabidopsis thaliana*, *Drosophila melanogaster* ou encore *Mus musculus*.

1.3.2 Jeux de données de référence

Nous citons ici deux jeux de données de référence pour l'espèce humaine. Nous faisons une courte présentation de ces projets et par la suite nous les étudierons plus en détail.

Human genome diversity Panel Le Projet d'étude de la diversité du génome humain ; ou Human Genome Diversity Project (HGDP) a été initié à l'institut Morrison de l'université de Stanford [Weiss et al., 1997]. Ce projet est né du travail de Luigi Luca Cavalli-Sforza, avec pour but d'étudier les relations entre les diverses populations humaines. Les données regroupent un total de 1043 individus, provenant de 53 populations [Cann et al., 2002]. Ces 53 populations sont réparties sur les cinq continents (Figure 2), et ont été choisies pour leur caractère endogène. La description de ces données permet de comprendre plus en détail l'histoire et l'évolution de l'espèce humaine. L'un des premiers résultats mis en avant par ce jeu de données est la structure inter-continentale de l'espèce humaine [Rosenberg et al., 2002]. Pour cette étude, les auteurs utilisèrent un ensemble de 377 microsatellites. Ces marqueurs génétiques ont suffit à mettre en évidence la structure de

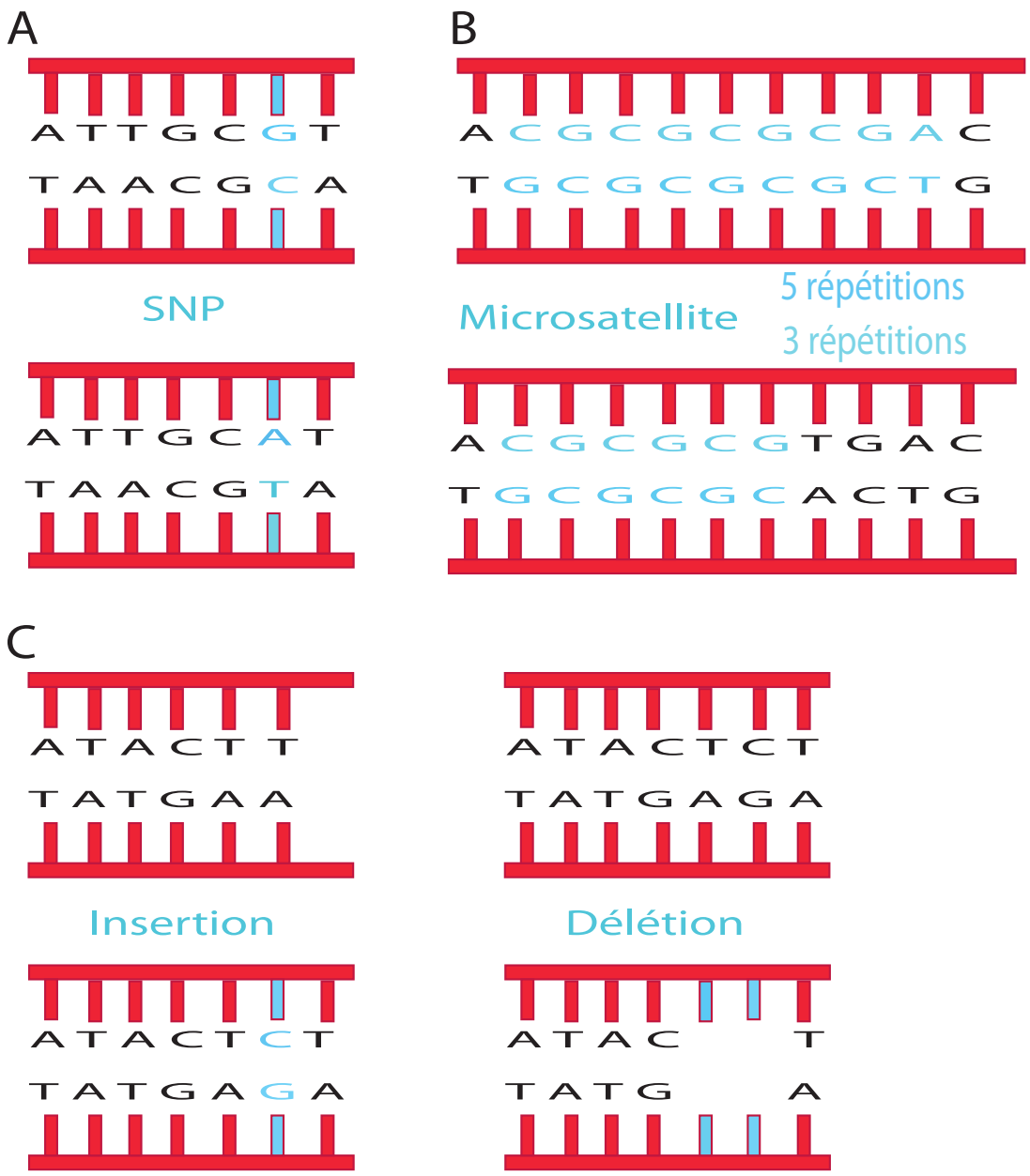


FIGURE 1 – panel A : Marqueur de type SNP. Ces marqueurs consistent en un changement d'une base sur les deux brins d'ADN. panel B : Marqueur de type microsatellite. Ce marqueur est un nombre de répétitions d'un motif, ici *CG*. Le polymorphisme vient du nombre de répétitions de ce motif. Deux allèles sont montrées, l'allèle contenant 5 répétitions, et l'allèle contenant 3 répétitions. panel C : Marqueurs de type insertion ou délétion. L'insertion est un ajout d'un certain nombre de nucléotides à un site donné. La délétion est la suppression d'un groupe de nucléotides.

population entre les différents continents [Rosenberg et al., 2002]. Ces données ont été utilisées avec un nombre croissant de marqueurs génétiques [Rosenberg et al., 2006]. Avec les avancées dans les technologies de génotypage, ce jeu de données a été exploité pour produire des données à plus grande échelle [Li et al., 2008]. Dans l'étude de Li et al., ces données sont analysées à l'échelle du génome entier avec plus de $650\ 000$ marqueurs du type SNP. La quantité de variables qui caractérisent les individus est augmentée d'un facteur 1000. Ces données de grande dimension permettent d'observer la même structure inter-continentale. Mais celles-ci sont utiles pour étudier d'autres questions que la structure inter-continentale des populations humaines. Un des intérêts est par exemple la détection de régions du génome impliquées dans des processus d'adaptation, comme nous le verrons dans la section 3.

Thousand genomes project Un autre jeu de données de référence plus récent est le projet 1000 génomes. Ce jeu de données regroupe les séquences de 1092 individus provenant de 14 populations différentes. Contrairement au HGDP, tous les continents ne sont pas représentés dans ce jeu de données. Les données contiennent également des populations issues d'un métissage inter continental, comme les individus afro-américains du sud ouest des États-Unis. Ce jeu n'a pas pour but de caractériser la diversité génétique des populations humaines, mais plutôt une compréhension approfondie du génome humain et de son association avec divers traits d'intérêt médical (voir *1000genomes.org*). Des données de séquençage sont utilisées pour détecter les variations les moins fréquentes du génome. Pour cela les séquences entières des individus ont été étudiées. Il en résulte un jeu de données contenant plus de $38\ 000\ 000$ marqueurs SNP, et $1\ 400\ 000$ d'insertions-délétions. Une première analyse de ces données a été réalisée pour démontrer tout le potentiel d'information des données [Consortium, 2012]. Nous utilisons les marqueurs SNP de ces données pour démontrer la capacité de la méthode proposée dans la section 3 à détecter des traces d'adaptation locale avec des données massives.

1.4 Objectif de la thèse

Le contexte de la thèse s'inscrit dans le développement de modèles statistiques pour traiter des données de génétique des populations à grande échelle telles que les données des jeux présentés dans la section 1.3.2. En une dizaine d'années, grâce aux technologies de séquençage, les données sont passées de l'ordre de la centaine de marqueurs par individu [Rosenberg et al., 2002], à des millions de marqueurs [Consortium, 2012]. De nombreuses réponses à des questions biologiques peuvent être apportées grâce à cet accroissement. Des questions telles que l'étude de la structure de population à une échelle géographique très fine [Humphreys et al., 2011] ou la recherche de traces

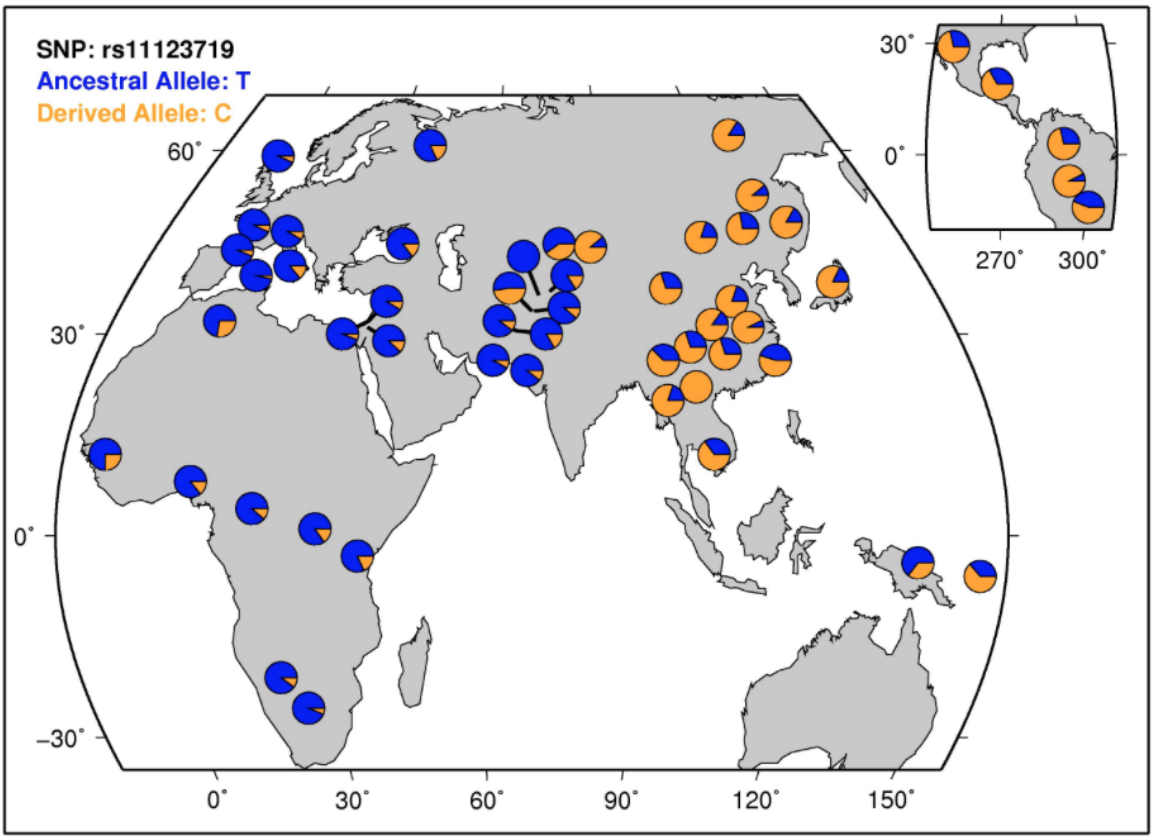


FIGURE 2 – Carte indiquant tous les lieux dans le monde où des données du Human Genome Diversity Project (HGDP) sont échantillonnées. La carte montre l'exemple de la variation de fréquence pour le marqueur SNP rs11123719.

d'adaptation biologique dans des régions du génome peuvent désormais avoir des réponses. Néanmoins, pour traiter ces questions, il y a un besoin grandissant de modèles statistiques et de moyens informatiques. Certaines méthodes statistiques communément utilisées ne sont plus applicables à cause des calculs nécessaires trop coûteux [Pritchard et al., 2000]. De plus, la plupart des méthodes classiques utilisées sont basées sur des modèles mécanistes. Nous définissons les modèles mécanistes dans la section 2. Ces méthodes font des hypothèses fortes sur les processus ayant engendrés les données. Avec la nature et la quantité des données analysées, nous pensons qu'il est utile de proposer des modèles avec moins d'hypothèses sur ces processus, et qui accordent plus d'importance à l'information contenue dans les données.

L'objectif est de proposer des méthodes innovantes en génétique de populations pour détecter les zones du génome impliquées dans des processus d'adaptation locale (section 3), et de caractériser les variations génétiques spatiales à une échelle fine (section 4). Ces méthodes ont également pour but d'être applicables à des données contenant un grand nombre de marqueurs génétiques. Pour cela nous développerons des méthodes de statistiques bayésiennes basées sur différentes approches telles que l'analyse factorielle et les processus gaussiens. Un intérêt tout particulier sera porté à l'application de ces méthodes à divers jeux de données humaines telles que les données du Human Genome Diversity Project, ou du projet 1000 Génomes. Un aspect important de la thèse sera d'incorporer ces méthodes à la panoplie de méthodes à la disposition des biologistes. Pour cela nous avons développé des logiciels *open source* qui implémentent ces méthodes statistiques.

Dans la section 2 nous détaillons dans un premier temps les distinctions entre modèles mécanistes et phénoménologiques. Puis nous présentons les approches classiques pour détecter l'adaptation locale ainsi que certaines méthodes qui en dérivent. Nous verrons que ces méthodes sont issues de modèles explicites de génétique des populations, et que ce type d'approche peut être mis en difficulté. Dans la section 3, nous présentons une nouvelle approche pour apprendre conjointement la structure de population et détecter des traces d'adaptation locale. Cette méthode repose sur le modèle statistique de l'analyse factorielle. L'utilité de la méthode et du logiciel est mise en avant avec des simulations. Dans la section 4, nous proposons un modèle pour analyser la structure des populations dans un contexte spatial, et en particulier détecter les régions de l'aire de répartition d'une espèce où la différenciation génétique entre populations apparaît à plus courte échelle de distance. Là encore l'utilité de la méthode sera illustrée avec divers exemples issus de l'écologie moléculaire et de la génétique humaine. Dans la section 5, nous présentons d'autres travaux effectués pendant ma thèse qui concernent la caractérisation de la structure génétique des populations à une échelle fine et l'étude de la structure des populations de l'ouest de la France.

2 Modèles mécanistes pour détecter l'adaptation locale

Dans cette section, nous présentons les différences entre modèles mécanistes et modèles phénoménologiques. Cette distinction est importante car les méthodes adoptées pour déterminer les marqueurs impliqués dans l'adaptation locale sont souvent de nature mécanistes. Nous présentons la méthode basée sur les scans du génome à l'aide de mesures de F_{st} et le modèle sous-jacent. Nous détaillons ensuite deux méthodes basées sur ce modèle.

2.1 Modèles mécanistes et modèles phénoménologiques

Il est possible d'adopter différentes stratégies en biologie des populations en général pour décrire et expliquer les processus observés [Levins, 1966]. L'approche mécaniste consiste à vouloir décrire au plus proche le mécanisme naturel qui a conduit à l'observation des données. Un modèle mécaniste décrit la structure des données en modélisant les processus biologiques supposés avoir engendré les données [Hilborn, 1997]. Les paramètres de ce modèle ont donc une interprétation biologique directe. Le modèle résultant a pour avantage d'avoir des paramètres interprétables, et permet une bonne compréhension du modèle.

Néanmoins, une critique de cette approche est son manque de généralité. Une telle modélisation implique un ensemble d'hypothèses qui sont violées dans la majorité des cas. Dans les exemples pour détecter l'adaptation locale que nous citons dans les sections 2.3.1 et 2.3.2, les modélisations sont assez mécanistes. En effet la méthode implémentée dans le logiciel *fdist* repose sur l'hypothèse que la distribution des estimateurs dans le modèle neutre est produite par des populations qui évoluent selon un modèle en îles. La méthode implémentée dans le logiciel *Bayescan* suppose que les différentes populations évoluent selon un modèle de divergence instantanée comme détaillé dans la section 2.3.2. Ces hypothèses décrivent un mécanisme précis qui a pu conduire à la génération des données, mais qui sont violées dans la majorité des jeux de données. Le fait même de grouper les individus par population, comme dans les scans à F_{st} est une hypothèse. Les populations naturelles peuvent être structurées sans présenter de sous-populations nettes. Il est souvent préférable de travailler à l'échelle des individus [Manel and Holderegger, 2013].

Une autre stratégie est d'adopter une approche plus phénoménologique. Nous définissons un modèle phénoménologique comme une structure entre les variables du modèle qui a pour but de décrire au mieux les données [Hilborn, 1997]. Une telle approche a pour avantage d'être plus robuste car elle cherche à décrire les données en général et non pas juste à reconstruire un mécanisme. Néanmoins, un modèle phénoménologique est moins facile à interpréter puisqu'il ne renvoie pas à des paramètres biologiques, mais à des objets statistiques plus abstraits. L'analyse en compo-

santes principales, que nous détaillerons dans la section 3 est un exemple de modèle phénoménologique. Les composantes principales résultantes servent à représenter les individus dans les directions maximisant la variance.

Il est impossible de dire s'il vaut mieux adopter une stratégie ou une autre, et en général, les modèles ne sont jamais complètement mécanistes, car il y aurait trop de paramètres à prendre en compte, ni jamais complètement phénoménologiques puisque des paramètres biologiques sont souvent introduits. Dans la suite de cette section, nous présentons les méthodes classiques de scan à sélection qui supposent un modèle particulier de génétique des populations. Dans la section 3 nous présentons une méthode de scan à sélection qui ne repose pas sur des modèles mécanistes de génétique des populations mais sur un modèle phénoménologique.

2.2 Modèles d'évolution de populations et les F_{st}

2.2.1 Modèle de divergence et modèle en îles

Pour modéliser l'évolution de populations naturelles, différents scénarios ont été proposés. Ils reviennent à grouper les individus en un nombre de populations, finis ou infinis et à supposer une relation entre elles. Nous présentons ici les modèles les plus utilisés, le modèle de divergence instantanée, et le modèle en îles.

Le modèle de divergence instantanée est un modèle où l'on suppose que toutes les populations observées sont les descendantes d'une population source. Cette population source s'est divisée en un nombre n de populations à un instant donnée. Par la suite ces populations évoluent sans interaction les unes avec les autres (Figure 3). En d'autres termes, les seuls processus agissant sur chaque population seraient les mutations et la dérive génétique. Ce modèle a été étudié en détail par Nicholson et al. [Nicholson et al., 2002]. Si l'on considère un marqueur SNP j des données, avec $j = 1..l$, où l est le nombre de locus, soit $p_0^{(j)}$ la fréquence de ce marqueur dans la population ancestrale. La fréquence $p_i^{(j)}$ de ce marqueur dans une population i pour $i = 1..n$, suit une loi normale centrée en $p_0^{(j)}$ telle que

$$p(p_i^{(j)}) = \mathcal{N}(p_0^{(j)}, F_i(1 - p_0^{(j)})p_0^{(j)}) \tag{1}$$

où F_i est un paramètre d'intensité de la dérive dépendant de la population. Une forte dérive, due notamment à un faible nombre d'individus dans la population implique une plus forte variance des fréquences alléliques. Ce modèle permet d'estimer l'intensité de cette dérive. Dans les données, la fréquence ancestrale p_0 n'est pas observée, à l'exception de certaines données observées sur plusieurs générations [Krimbas and Tsakas, 1971].

Un autre modèle d'évolution est le modèle en îles qui a été proposé par Wright pour décrire les liens entre différentes sous-populations d'une même espèce. Il suppose que les n populations sont à l'équilibre démographique et échangent une certaine proportion de migrants à chaque génération (Figure 4). Les processus qui influencent l'évolution de la population sont les trois processus neutres : migration, dérive et mutation [Hudson, 1998]. L'intensité des migrations est ici un paramètre essentiel. Un trop fort tôt de migrations aurait un effet supérieur à celui de la dérive, et les fréquences alléliques seraient homogénéisées. Dans ce cas, les sous-populations n'auraient plus de sens et les variations génétiques entre les groupes seraient invisibles. Pour étudier la non-homogénéité de ces groupes, Wright introduit une quantité F_e [Wright, 1920] qui est un coefficient de corrélation entre les gamètes des individus d'une sous-population, relativement à la population totale. Nous revenons plus en détails sur ce coefficient dans la section 2.2.2 et dans l'annexe 7.1. Un lien fort existe entre l'intensité de la migration et le coefficient F_e [Slatkin, 1985].

Bien sur, ni l'un, ni l'autre de ces modèles ne reflètent exactement les processus ayant engendré les données. Un modèle d'isolement avec migration intégrant les deux scénarios a été étudié [Latter, 1973, Hey, 2010]. Dans les deux modèles, il est possible de quantifier le niveau de différenciation entre les populations à l'aide du coefficient de dérive F_i ou le coefficient de corrélation entre individus F_e. Les deux modèles ont en commun qu'ils supposent une structure de covariance identique entre les populations. Toutes les paires de populations selon ces modèles ont la même covariance entre leurs fréquences alléliques normalisées. Cette hypothèse peut être violée dans le cas où les données sont issues de processus démographiques plus complexes. Nous voyons dans la section 3 que dans ce cas, les méthodes basées sur ces modèles déclarent un grand nombre de locus neutres comme étant impliqués dans un processus d'adaptation locale.

2.2.2 La mesure de F_{st} et le test de Lewontin Krackauer

Les deux paramètres F_e et F_i de la section 2.2.1 sont des statistiques F dans le cadre du modèle de divergence ou du modèle en îles [Lewontin and Krakauer, 1973]. Il servent à quantifier la différenciation entre les sous-populations pour un marqueur donné. Historiquement, la statistique F est une statistique servant à quantifier le niveau d'hétérozigotie d'un marqueur dans une population, et la proximité de la fixation d'un allèle. Elle a été proposée par Sewall Wright pour estimer le manque d'hétérozygotes chez les bovins [Wright, 1920]. Cette statistique est fondamentale pour décrire la structure d'une population car elle permet de mesurer l'écart à l'équilibre théorique d'Hardy-Weinberg [Gillespie, 2010]. Cet équilibre établit que pour un marqueur de fréquence p dans une population où tous les individus se reproduisent aléatoirement entre eux, la proportion d'hétérozygotes est $2p(1-p)$. Cependant une population pouvant être divisée en groupes d'individus, comme

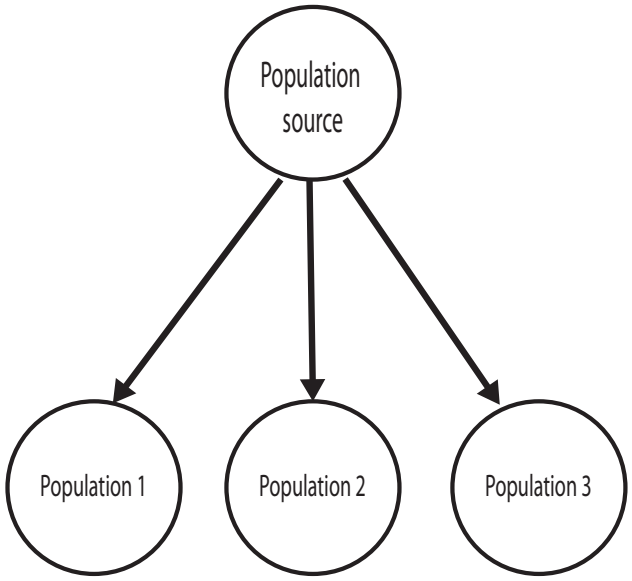


FIGURE 3 – Modèle de divergence instantanée. Toutes les populations observées dans la données sont supposées descendre d'une population source, ou population ancestrale.

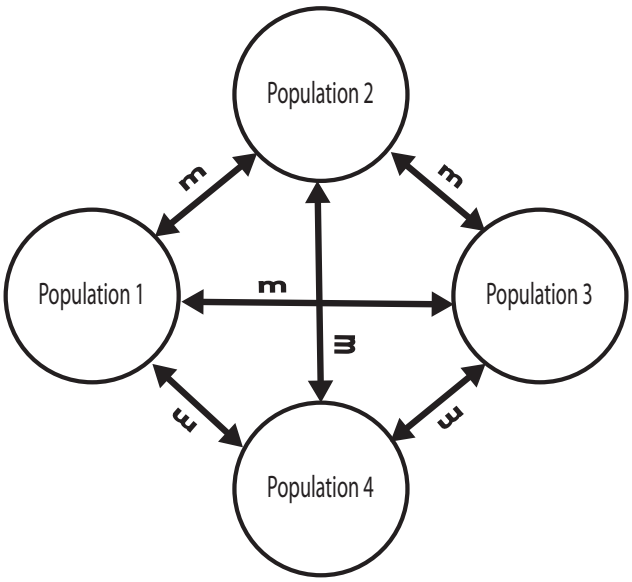


FIGURE 4 – Modèle en îles à 4 populations utilisé pour simuler des marqueurs neutres. Toutes les populations échangent à chaque génération une proportion m de migrants, représentés par une flèche.

nous le voyons dans la section 2.2.1, cette mesure est devenue un outil pour quantifier la différenciation entre sous-groupes d'individus. La définition de cette quantité est ambiguë [Balding, 2003], ce qui a conduit à différentes interprétations et estimations de ce coefficient. Dans la suite nous noterons cet indice F_{st}.

La F_{st} peut être estimée pour un nombre variable de sous-population, et un nombre variable de loci. L'estimateur proposé par Wright [Wright, 1943] pour un nombre n de sous-populations, et un marqueur bi-allélique de fréquence allélique f_i dans la sous-populations i de fréquence allélique $\bar{f}$ dans l'ensemble de la population est

$$\theta = \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})}{\bar{f}(1-\bar{f})} \tag{2}$$

Cet estimateur est un ratio de variances qui n'est pas robuste aux échantillonnages irréguliers, et un meilleur estimateur est détaillé dans l'annexe 7.1 [Weir and Cockerham, 1984].

Lewontin et Krackauer proposent d'utiliser cette mesure pour tester l'implication de marqueurs dans des processus adaptatifs dans les modèles de divergence, ou les modèles en îles [Lewontin and Krakauer, 1973 L'objectif de scanner le génome avec cette mesure est de calculer une mesure pour chaque marqueur génétique à disposition. L'intérêt de calculer une valeur par marqueur est de détecter les marqueurs ayant une forte F_{st}. Une forte valeur de l'estimateur pour un marqueur donné peut être interprétée comme un signal résultant d'une pression de sélection sur ce marqueur dans une des sous-populations considérées [Lewontin and Krakauer, 1973]. L'idée de Lewontin et Krackauer est qu'en l'absence de sélection, tous les processus évolutifs neutres, la dérive, les migrations ou les mutations ont un effet uniformément réparti sur le génome. Les variations causées par ces processus, sont appelées la structure neutre, ou encore structure de reproduction [Lewontin and Krakauer, 1973]. Ainsi, tous les marqueurs génétiques sélectivement neutres devraient avoir des valeurs de F_{st} issues de la même distribution. Dans le cas de présence de forces de sélection, la distribution des valeurs est sensée être déformée par la présence de certains marqueurs atypiques.

2.2.3 Scanner le génome avec les F_{st}

Les scans à F_{st} sont une méthode utile pour détecter des régions du génome susceptibles d'avoir été sous adaptation locale, en cherchant des locus outliers, i.e. ayant une valeur de F_{st} atypiquement forte. Ceux-ci ont déjà prouvé leur utilité avec de nombreux exemples. En particulier nous pouvons reconsidérer l'exemple de la section 1 et des individus tibétains. Il a été démontré que les tibétains ont une meilleure tolérance aux conditions d'hypoxie. Ils sont adaptés aux conditions de vie dans les hauts-plateaux himalayens. Ainsi pour détecter les marqueurs impliqués dans ce

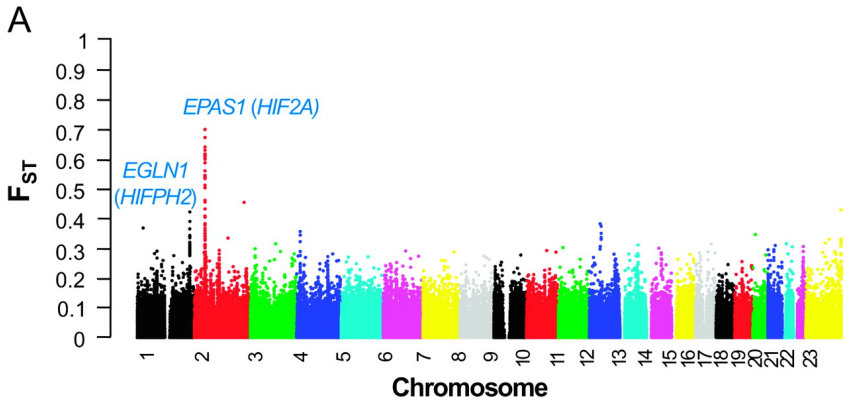


FIGURE 5 – Scan du génome à l'aide de F_{st} pour un groupe d'individus tibétains et un groupes d'individus Hans de Chine [Xu et al., 2011]. Les chromosomes 1 et 2 révèlent des régions enrichies en fortes F_{st}. Les régions EPAS1 et EGLN1 sont connues pour être activées en condition d'hypoxie.

processus d'adaptation, il est possible de considérer une deuxième sous population qui n'est pas adaptée à ces conditions, et de scanner le génome à la recherche de fortes mesures de F_{st}. Ces fortes mesures indiquent, entre les tibétains et la population non adaptée, une différenciation génétique extrêmement forte par rapport au reste du génome. Ce scan a été effectué [Xu et al., 2011] et en l'occurrence, la population choisie est la population des Hans de Chine. Le scan révèle de forts signaux sur les chromosomes 1 et 2, dans des régions contenant les gènes *EPAS1* et *EGLN1*, tous deux connus pour leur implication dans les facteurs d'activation en condition d'hypoxie. Notons ici que, si le scan est effectué pour les marqueurs un par un indépendamment, les estimateurs des F_{st} sont corrélés aux estimateurs des marqueurs voisins. Cela est du au processus de recombinaison qui échange des grandes quantités de matériel génétique entre chromosomes appareillés lors de la méiose. A cause de ces échanges, les sites proches les uns des autres ont des états alléliques corrélés. Cette corrélation est appelée le déséquilibre de liaison. Les scans montrent la présence de ce déséquilibre de liaison puisque même si potentiellement peu de marqueurs sont impliqués dans l'adaptation, des régions entières des chromosomes sont enrichies en fortes valeurs de F_{st}. Ce schéma de régions enrichies en fortes valeurs de F_{st} est attendu, et permet d'éviter de fausses découvertes. Un marqueur ayant une forte valeur de F_{st} au milieu d'une région non enrichie en forte F_{st} sera très probablement un faux positif.

2.3 Modèles mécanistes pour l'adaptation

Une façon de détecter les marqueurs subissant des pressions de sélection est de scanner le génome avec des estimateur de F_{st}. Cette méthode permet de révéler certaines régions génomiques dont les variations atypiques en comparaison du reste du génome semblent ne pas être expliquées seulement par la structure de reproduction. Nous détaillons par la suite plusieurs extensions de ces scans qui utilisent des modèles mécanistes de génétique des populations.

2.3.1 fdist

L'un des premiers modèles basé sur une approche mécaniste pour détecter les marqueurs sujets à des pressions de sélection a été proposé par Beaumont et Nichols [Beaumont and Nichols, 1996]. La méthode proposée est d'estimer les F_{st} pour un groupe de marqueurs, et de comparer ces valeurs à des valeurs simulées selon un modèle neutre. Le modèle neutre choisi par les auteurs est un modèle en île (Figure 4) où aucun des marqueurs ne subit de pression de sélection. Le but de cette approche est de pouvoir simuler un ensemble de valeurs de F_{st} pour des marqueurs neutres, et d'obtenir une distribution nulle des estimateurs de F_{st}. Les auteurs résument leur approche en quatre points :

- Estimer les F_{st} et l'hétérozigotie des marqueurs contenus dans les données.
- Estimer une F_{st} moyennée sur les marqueurs à l'aide de la formule (2).
- Simuler un grand nombre de marqueurs selon le modèle en île, et estimer leur hétérozigotie et F_{st}.
- Estimer les quantiles à 5% des F_{st} pour une hétérozigotie donnée.

La comparaison à l'hétérozigotie est importante car elle influe sur la valeur de l'estimateur de F_{st}. L'hétérozigotie est corrélée positivement à la F_{st} (Figure 6), et ainsi, les marqueurs à faible fréquence, ou faible hétérozigotie, ont une distribution de leur valeur de F_{st} modifiée [Jakobsson et al., 2013].

2.3.2 Bayescan

Un autre modèle mécaniste de génétique des populations utilisé est la cas où toutes les populations seraient issues d'une population source [Beaumont and Balding, 2004]. Dans le cas d'un marqueur génétique j, provenant d'une sous-population i, la mesure F_{st}^{ij} peut être estimée en utilisant la vraisemblance d'un modèle multinomial-Dirichlet, comme présenté dans [Beaumont and Balding, 2004]. Dans ce cas, F_{st}^{ij} est un paramètre qui peut être interprété comme proportionnel à la probabilité que deux individus aient un ancêtre commun dans la sous-population [Crow et al., 1970]. Pour détecter l'implication de certains marqueurs dans un processus d'adaptation locale, le modèle suivant a été

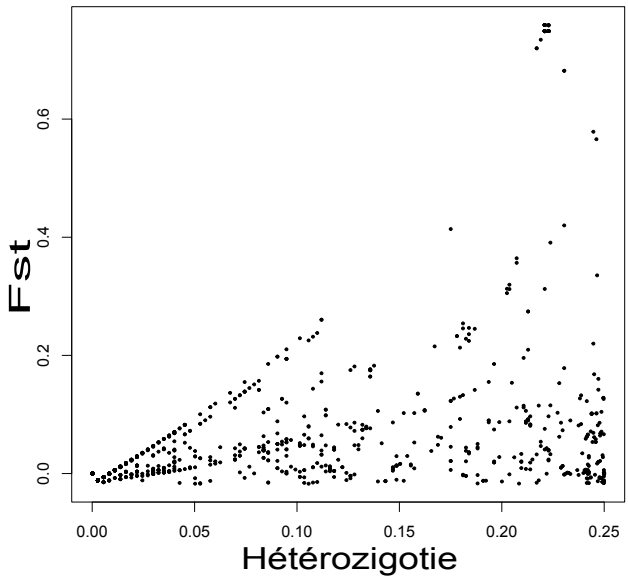


FIGURE 6 – Mesure de F_{st} en fonction de l'hétérozigotie d'un marqueur. la F_{st} est plus forte pour des marqueurs ayant une forte hétérozigotie. Ces valeurs ont été calculées pour des données simulées selon un modèle en île avec le logiciel *ms*.

proposé

$$\log(\frac{F_{st}^{ij}}{1-F_{st}^{ij}}) = \alpha_j + \beta_i + \gamma_{ij} \tag{3}$$

avec α_j un effet spécifique au marqueur j, et β_i un effet commun à l'ensemble de la population i. Le paramètre γ_{ij} est un paramètre de bruit [Beaumont and Balding, 2004, Foll and Gaggiotti, 2008, Riebler et al., 2008]. L'utilité de ce paramètre est discutée, et il pourra être ignoré. L'attente au vu de ce modèle est que les marqueurs n'ayant pas d'effet spécifique aient une valeur α_j nulle. Au contraire, un marqueur dont la valeur α_j est forte indique un fort effet non partagé par les autres marqueurs dans une population, rendant ce marqueur atypique. Deux interprétations sont proposées pour les valeurs atypiques de α_j. Une forte valeur positive indiquerait une pression de sélection liée à l'adaptation locale. Au contraire, une valeur fortement négative indiquerait une pression de sélection équilibrante, visant à homogénéiser les fréquences alléliques au sein des populations. Nous ne discutons pas de ce type de sélection, puisque notre objectif est de détecter les marqueurs génétiques impliqués dans des processus d'adaptation locale.

Le modèle bayésien est implémenté dans le logiciel *Bayescan*. Le logiciel implémente un algo-

rithme de méthode de Monte Carlo par chaîne de Markov avec des sauts réversibles qui permet l'estimation des lois à posteriori des différents paramètres du modèle. L'une des différences entre les approches de [Beaumont and Balding, 2004], [Riebler et al., 2008] et [Foll and Gaggiotti, 2008] provient de l'algorithme d'estimation des lois à posteriori. [Beaumont and Balding, 2004] proposent un prior gaussien centré et réduit pour les α_j, et mettent à jour ces paramètres à l'aide de pas Metropolis-Hastings. [Riebler et al., 2008] proposent eux une reparamétrisation du modèle $\eta_{ij} = \delta_j \alpha_j + \beta_i$ et estiment la distribution conditionnelle a posteriori des paramètres η_{ij} où δ_j est une variable indicatrice, indiquant un effet spécifique du locus j. L'implémentation du logiciel *Bayescan* est faite de sorte qu'au fil de la chaîne, des paramètres α_j peuvent être exclus et fixés à une valeur nulle. L'aspect saut réversible de l'algorithme exclut, ou inclut à chaque itération des paramètres α_j. La méthode de [Foll and Gaggiotti, 2008] ajoute de plus la possibilité d'utiliser la méthode avec des marqueurs dominants. Un marqueur est dit dominant s'il est impossible de distinguer entre hétérozygote et homozygote (voir *scienceboard.net*).

2.4 Vers un modèle phénoménologique

Les modèles basés sur les F-statistiques ont prouvé leur utilité notamment avec l'exemple de l'adaptation à l'altitude (Figure 5). Néanmoins leur utilisation n'est pas toujours bien appropriée, et nous montrons que dans certains cas, ces méthodes peuvent détecter un grand nombre de faux positifs (section 3). De plus, les données ne sont pas toujours adaptées au regroupement des individus en populations. En effet dans de nombreux exemples, l'échantillonnage des individus se fait continuellement dans l'ensemble de leur aire de répartition [Gugerli et al., 2008]. Dans ces conditions, les différences génétiques s'accumulent avec la distance selon un schéma d'isolement par la distance comme nous le présentons dans la section 4, plutôt que comme des populations distinctes. Le but de la section 3 est de présenter un modèle qui fait moins d'hypothèses quant au mécanisme qui a engendré les données. Une motivation importante est de tester la robustesse de cette nouvelle approche dans de nombreux exemples ou les hypothèses des approches classiques ne sont pas respectées.

3 Un modèle phénoménologique pour détecter l'adaptation locale

Dans cette section nous présentons un nouveau modèle pour détecter l'adaptation locale. Nous présentons dans un premier temps la famille des modèles à facteurs, son lien avec l'analyse en composantes principales et son utilité en génétique des populations. Par la suite nous donnons les détails du modèle et expliquons en quoi il permet de détecter les marqueurs ayant subi des pressions de sélection. Nous finissons par présenter une variante du modèle, ayant l'avantage d'être moins coûteuse en temps de calcul.

Une partie des résultats de cette section a fait l'objet d'une publication dans Molecular Biology and Evolution : *Nicolas Duforet-Frebourg, Eric Bazin, and Michael G.B. Blum Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model. Mol. Biol. Evol. 2014, doi :10.1093/molbev/msu182*
L'article se trouve dans l'annexe 8.1.

3.1 Analyse factorielle

3.1.1 Présentation du modèle

L'analyse factorielle a pour objectif de décrire et d'expliquer des données en grande dimension. Lorsque les données sont multidimensionnelles, les différentes variables qui les composent sont toujours corrélées à un certain degré. La génétique des populations est un très bon exemple pour illustrer cela. Comme dit précédemment les puces servant à génotyper les individus comportent des millions de marqueurs, et il est clair que les données ne vivent pas en si grande dimension. L'analyse factorielle vise à expliquer ces données par un nombre K de variables latentes, non observées, et construites à l'aide de combinaisons linéaires des variables originales [Mardia et al., 1979].

Dans le cas d'une matrice de données Y avec n lignes et p colonnes, l'analyse factorielle peut être vue comme un produit de deux matrices de faible rang K (Figure 7). Ces deux matrices, respectivement de dimensions $n \times K$ et $K \times p$, seront appelées la matrice des facteurs, et la matrice des loadings. Les loadings, que l'on pourrait traduire par poids, sont les contributions de chacune des p variables aux variables latentes. Les facteurs sont les projections des individus sur ces variables. Nous adoptons les notations suivantes : Y est la matrice de données à expliquer. U et V sont les deux matrices des facteurs et des loadings. Dans les données, n représente le nombre de réalisations, et p la dimension de ces données. Dans notre cas, n est un nombre d'individus et p est un nombre de marqueurs génétiques. Les données n'étant initialement pas centrées, on utilise une

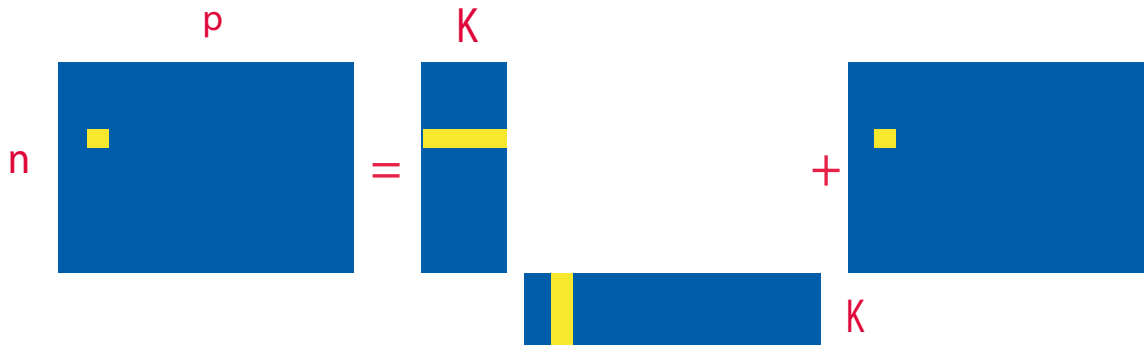


FIGURE 7 – Approximation de rang K d'une matrice de dimensions $n \times p$

matrice μ pour la moyenne. On appelle ϵ la matrice des résidus. Nous choisissons la formulation

$$Y = \mu + UV + \epsilon, \tag{4}$$

où l'on suppose que chaque colonne de V suit une loi normale $\mathcal{N}_K(0, I_K)$ et chaque colonne de ϵ suit une loi normale $\mathcal{N}_n(0, \Psi)$. Cette formulation est une formulation alternative aux formulations possibles données par [Bartholomew et al., 2011] ou [Jolliffe, 2005] où ce sont les lignes de U qui sont supposées suivre une loi normale $\mathcal{N}_K(0, I_K)$.

Les hypothèses utilisées pour résoudre ce modèle sont $E[Y] = \mu, E[\epsilon] = 0, E[\epsilon^t \epsilon] = \Psi$ et $E[\epsilon^t V] = 0$ [Jolliffe, 2005]. En résumé, les hypothèses formulées sont que les résidus et produits UV soient de moyenne nulle, les résidus et les loadings ne soient pas corrélés, et que les erreurs soient de covariance Ψ une matrice diagonale. En faisant l'hypothèse de résidus gaussiens, et en notant Y_l la l^{eme} colonne des données, la vraisemblance conditionnellement aux loadings s'écrit ainsi ([Bartholomew et al., 2011])

$$p(Y_l|V) = \mathcal{N}(\mu_l + UV_l, \Psi), \tag{5}$$

où V_l est la l^{eme} colonne de V, et à l'aide des hypothèses énoncées

$$p(Y_l) = \mathcal{N}(\mu_l, U^t U + \Psi). \tag{6}$$

Le maximum de vraisemblance de U et Ψ peut être obtenu avec un algorithme d'Expectation Maximisation détaillé par Rubin et Thayer [Rubin and Thayer, 1982]. D'autres extensions de cet algorithme ont été développées pour des modèles de mélanges [Ghahramani et al., 1996].

Pour simplifier, nous supposerons que les données sont centrées et ainsi la matrice μ est nulle. Le vecteur Y_l, la l^{eme} colonne de Y, représentant les allèles des n individus pour le locus l s'exprime

de la manière suivante

$$Y_l = \sum_{k=1}^{K} U_k V_{kl} + \epsilon_l, l = 1 \dots p \tag{7}$$

où U_k est la k^{me} colonne de la matrice des facteurs U, V_{kl} est le coefficient (k, l) de la matrice des loadings V, et ϵ_l est la l^{me} colonne de ϵ. Cela permet d'exprimer la matrice Y comme le produit de deux matrices U et V de rang K.

3.1.2 Lien avec l'Analyse en Composantes principales et applications en génétique des populations

L'analyse en composante principale (ACP) est la méthode de référence pour représenter en très peu de dimensions des données multidimensionnelles [Jolliffe, 2005]. Cette méthode est habituellement vue comme la projection des réalisations composant les données dans des directions maximisant la variance [Hotelling, 1933]. Une vision alternative de l'ACP et de considérer le problème de minimisation de la distance entre les données et la reconstruction des données à partir des projections [Pearson, 1901]. Cette approche est aussi appelée Multi-Dimensionnal scaling (MDS). Dans le cas où l'on considère une distance euclidienne, ces deux problèmes sont des problèmes duaux et ont la même solution [Mardia et al., 1979].

L'ACP peut également être vue en termes de produit de matrices de rang faible. Quelque soit le point de vue adopté pour le problème, maximisation ou minimisation, l'ensemble des solutions est le même. Un moyen direct de calculer une solution pour l'ACP d'une matrice de données Y de rang r est de considérer la décomposition en valeurs singulières (DVS) de Y [Jolliffe, 2005]. La DVS est un résultat d'algèbre disant que toute matrice Y peut être écrite $Y = U\Lambda V$, où Λ est une matrice diagonale de rang r et U et V sont orthonormales. En permutant les valeurs de Λ pour que celles ci soient en ordre décroissant, et les colonnes de U et V de même, ces matrices contiennent les r premières composantes principales de Y. Ce résultat est important pour deux raisons. Premièrement il permet de calculer une solution de l'analyse en composantes principales sans calculer la matrice de covariance des données. C'est en général la solution choisie, comme dans la fonction `prcomp` de *R* [Core, 2012], qui utilise la SVD de la librairie *LAPACK* [Anderson et al., 1999]. Deuxièmement, ce résultat permet de voir l'ACP comme un produit matriciel, et montre son lien très fort avec les modèles à facteurs. Nous revenons plus en détail sur l'ACP classique dans la section 3.3.

Bishop et Tipping ont proposé une approche probabiliste de l'ACP [Tipping and Bishop, 1999] (voir aussi [Roweis, 1998]) qui présente de nombreux avantages par rapport à une ACP classique [Bishop, 2006]. Nous pouvons citer quelques atouts tels que

– la possibilité de résoudre le problème avec un algorithme EM, sans avoir à estimer la matrice

de covariance des données.

- la possibilité de traiter les données manquantes avec un algorithme de type EM
- la formulation de modèle de mélange pour l'ACP.
- la possibilité de faire un traitement bayésien du modèle.

Cette ACP probabiliste est très proche d'un modèle à facteurs puisque la seule différence entre les deux méthodes est dans l'hypothèse sur les résidus [Bishop, 2006]. En effet, Bishop et Tipping supposent que $\epsilon_l \approx \mathcal{N}_p(0, \sigma^2 I_n), l = 1..p$. Les résidus ont donc une covariance isotrope au lieu d'une covariance diagonale Ψ. Ainsi cette ACP probabiliste est en fait un cas particulier de l'analyse factorielle [Basilevsky, 2009].

En génétique des populations, les méthodes basées sur des modèles à facteurs sont apparues assez récemment [Engelhardt and Stephens, 2010, Frichot et al., 2014].

3.1.3 Applications en génétique des populations

L'ACP et l'analyse factorielle ont sensiblement le même objectif : décrire les données en faible dimension. Cet objectif les rend très appréciables en génétique des populations. L'ACP a cependant était préférée aux modèles à facteurs dans la littérature récente pour décrire la structure de population. Patterson et ses collègues [Patterson et al., 2006] ont contribué à cela avec le logiciel *smartpca*, qui permet de réaliser une analyse en composantes principales sur des jeux de données. Le logiciel réalise principalement une ACP, mais ajoute quelques fonctionnalités, telles que la suppression d'individus trop atypiques, ou l'estimation d'une valeur de K à l'aide du test de Tracy-Widom [Tracy and Widom, 1994]. De nombreuses études ont utilisé l'ACP ou MDS pour présenter des résultats à propos de la structure des populations [Novembre et al., 2008, Jakobsson et al., 2008]. La réduction en faible dimension a l'avantage de pouvoir représenter synthétiquement et élégamment la structure génétique (Figure 8).

Pour montrer l'utilité de l'analyse factorielle afin d'observer la structure de population, nous simulons un scénario dans lequel des individus issus de 4 populations de 1000 individus chacune, sont échantillonnés. L'histoire de ces populations se divise en deux phases. Dans un premier temps, une population ancestrale, commune aux 4 populations échantillonnées se divise en deux populations de 1000 individus de même taille que la population ancestrale. Ces deux populations évoluent alors selon un modèle de Wright-Fischer pendant un temps T, sans qu'aucune pression de sélection n'agisse. Le modèle de Wright-Fischer fait un ensemble d'hypothèses idéalisant l'évolution d'une population au fil des générations. Les hypothèses principales sont que la population est de taille constante, les générations sont non-chevauchantes, et il n'y a pas de recombinaison, ni de sélection [Hein et al., 2004]. Dans un deuxième temps, les deux populations se divisent à nouveau en 2

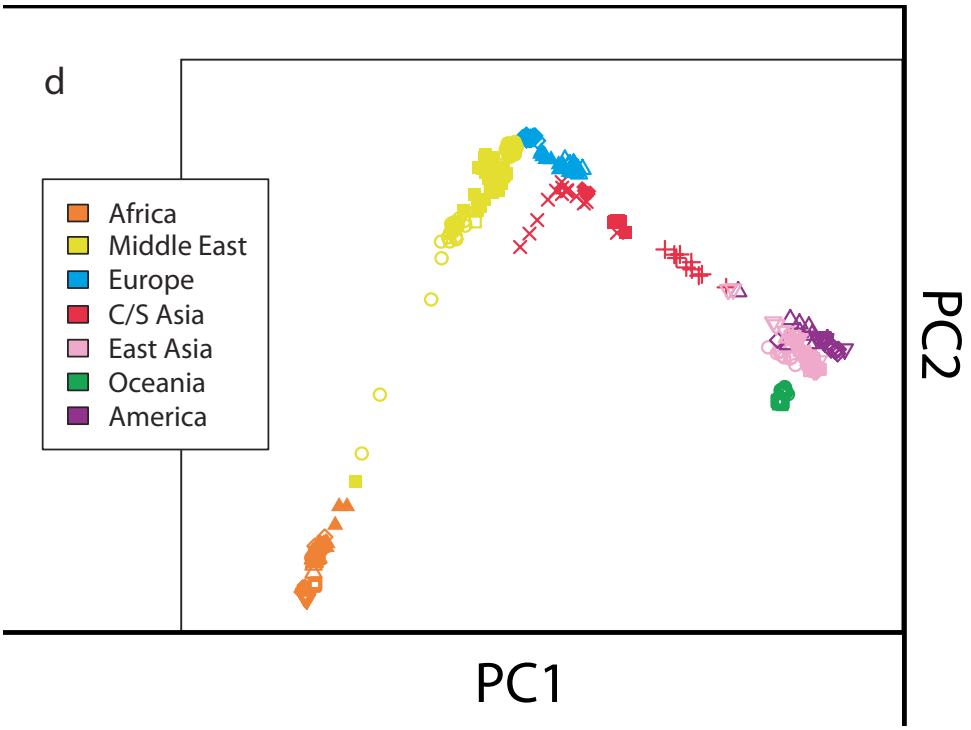


FIGURE 8 – Exemple d'utilisation de l'analyse en composantes principales en génétique des populations humaines. l'Analyse des données du projet HGDP [Jakobsson et al., 2008] montre que les deux premières composantes principales permettent d'observer la structure génétique entre les différents continents.

sous-populations de 1000 individus, notées (A_1, A_2) et (B_1, B_2). Les quatre populations résultantes évoluent pendant 100 générations, pour devenir les populations présentes dans les données. Durant ces 100 dernières générations, 100 marqueurs SNP dans chaque population sont soumis à des pressions de sélection. Les données sont simulées en utilisant un modèle de coalescent [Kingman, 1982] avec le logiciel *ms* [Hudson, 2002] pour la divergence neutre, et avec la librarie *SimuPOP* pour la sélection [Peng and Kimmel, 2005].

La structure de population est apprise par les facteurs latents contenus dans la matrice U, de l'équation (7) du modèle. Pour le choix de K, nous calculons l'erreur résiduelle au carré de l'équation (7). Plus la valeur de K augmente, plus la quantité de variance expliquée par le modèle augmente, et plus l'erreur résiduelle diminue. Cette mesure indique que le choix optimal pour le nombre de facteurs latents dans l'exemple est $K = 3$ (Figure 10), ce qui est attendu dans le cas de 4 populations [Patterson et al., 2006]. Le facteur associé à la plus grande variance distingue les populations (A_1, A_2) de (B_1, B_2), soit les populations séparées par la divergence la plus ancienne. Les deux facteurs suivants distinguent A_1 de A_2 et B_1 de B_2 (Figure 9).

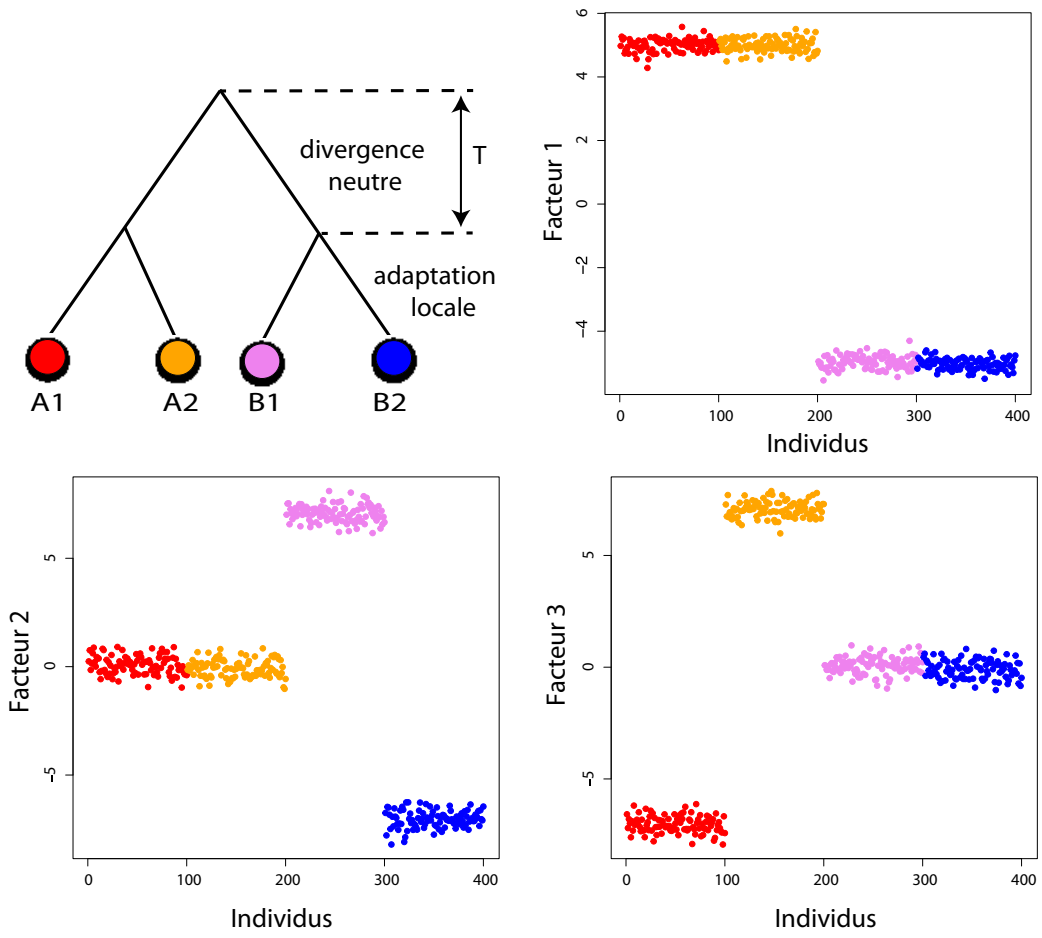


FIGURE 9 – Scénario utilisé pour la simulation. Le premier facteur, première colonne de la matrice U, sépare les individus des populations (A_1, A_2) et (B_1, B_2). Les deuxièmes et troisièmes facteurs séparent respectivement les individus des populations A_1 et A_2, et B_1 et B_2.

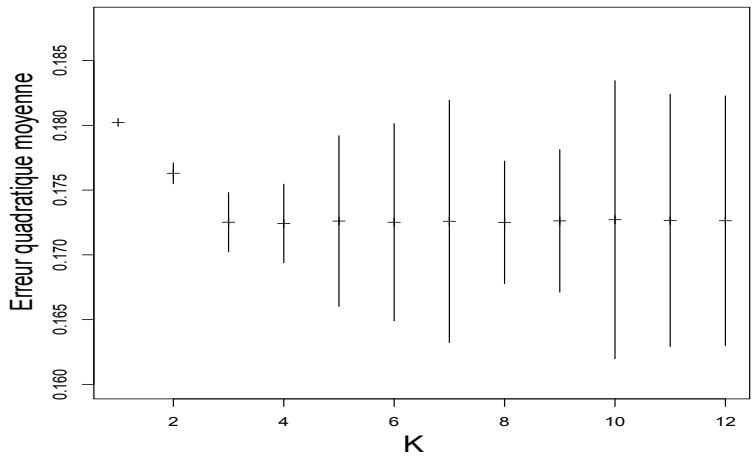


FIGURE 10 – erreur quadratique moyenne calculée à partir de l'équation (7), en fonction du nombre de facteurs utilisés pour les simulations selon le modèle de divergence à 4 populations. Les barres d'erreurs sont de taille 2 écarts types de l'erreur quadratique moyenne estimée sur 10 répétitions de l'algorithme MCMC. La valeur optimale de K est 3 comme attendu.

3.1.4 Version bayésienne du modèle

L'analyse factorielle est une méthode citée dans les manuels depuis des décennies [Harman, 1960, Tucker and MacCallum, 1997]. Cependant, plusieurs versions bayésiennes de cette méthode ont récemment été proposées [Marlin, 2003, Salakhutdinov and Mnih, 2008, Hofmann, 1999, Ding et al., 2011]. Ces modèles probabilistes ont plusieurs avantages, comme cités dans la section 3.1.2. L'un des avantages mis en avant est le fait de pouvoir appliquer la méthode avec des données manquantes [Srebro et al., 2003]. Un autre avantage que nous utiliserons tout particulièrement par la suite est la possibilité d'utiliser des algorithmes d'inférence bayésienne, et de construire un modèle hierarchique pour détecter l'adaptation locale. Nous détaillons ici la version probabiliste du modèle qui sera la base de notre modèle. L'idée de la version bayésienne est essentiellement de supposer que les lignes de U et colonnes de V suivent a priori une distribution normale $K-$variée.

$$p(U|\mu_U, \Sigma_U) = \prod_{i=1}^{n} \mathcal{N}(U_i; \mu_U, \Sigma_U) \tag{8}$$

$$p(V|\mu_V, \Sigma_V) = \prod_{l=1}^{p} \mathcal{N}(V_l; \mu_V, \Sigma_V) \tag{9}$$

Par la suite, les paramètres $\mu_V, \Sigma_V, \mu_U, \Sigma_U$ peuvent avoir différentes distributions a priori, selon le modèle. Un choix classique est une distribution jointe Normal-Wishart comme présentée dans le

modèle Bayesian Probabilistic matrix factorization de Salakhutdinov [Salakhutdinov and Mnih, 2008]. La loi a posteriori du modèle peut être approchée à l'aide d'algorithmes de type Monte Carlo par chaîne de Markov. Ces algorithmes regroupent des pas d'échantillonnage de Gibbs, pour lesquels il suffit de calculer les lois a posteriori conditionnelles des différents paramètres du modèle, ou des pas Metropolis-Hasting pour lesquels il faut proposer aléatoirement une valeur pour le paramètre et l'accepter selon un ratio d'acceptation. Cette flexibilité dans la modélisation va nous permettre de développer un modèle d'analyse factorielle bayésienne qui caractérise l'adaptation locale par détection d'outlier.

3.2 Détection de la sélection

3.2.1 Interprétations des *loadings*

A l'aide de l'analyse factorielle, il est possible de décrire les données avec un faible nombre de variables latentes. Dans l'exemple présenté précédemment, les données simulées correspondent à 4 populations, et les données peuvent être décrites à l'aide de 3 variables latentes, ou facteurs (Figure 9). Les variables sont des combinaisons linéaires des p variables présentes dans les données. Pour détecter des marqueurs génétiques subissant des pressions de sélection, nous nous intéressons aux marqueurs ayant une valeur atypiquement forte dans un vecteur de la matrice des loadings V. Une forte valeur V_{kl} pour un coefficient k du vecteur V_l implique que le marqueur l contribue fortement au k^{me} facteur latent. Cela revient à dire que le marqueur l est fortement différencié dans la direction k [Laloë et al., 2011]. Si l'on connaît les facteurs $U_1, \dots U_K$, cela peut être vu comme une régression multiple, où le vecteur Y_l est régressé par les K facteurs $U_1 \dots U_K$ (Figure 9). Les valeurs $Y_{1l} \dots Y_{Kl}$ sont les coefficients de régression. Puisque les facteurs représentent la structure de population [Engelhardt and Stephens, 2010], un fort coefficient de régression implique une forte corrélation avec la différenciation représentée par cette structure. Dans notre exemple de la section 3.1.3, une valeur atypiquement forte de V_{2l} signifie que le marqueur est atypiquement différencié entre la populations B_1 et B_2. Cela peut être interprété comme une trace d'adaptation locale dans une des deux populations.

Ainsi notre recherche de marqueurs impliqués dans des processus d'adaptation locale revient à rechercher des marqueurs outliers dans les matrices résultantes d'une analyse factorielle. Nous définissons un outlier, comme une valeur atypique au vu de la distribution des réalisations. La recherche d'une forte différenciation d'un marqueur rappelle l'approche de Lewontin et Krackauer [Lewontin and Krakauer, 1973]. Cette recherche correspond à l'hypothèse de travail de la génomique des population qui dit que la majeure partie des variations génétique est neutre [Luikart et al., 2003], et qu'une faible part des variations est influencée par des pressions de sélection. L'ACP a déjà été

proposée pour chercher les outliers dans des données en grande dimension [Filzmoser et al., 2008, Jackson and Chen, 2004]. Nous définissons le choix de modélisation pour chercher les outliers ci-dessous.

Pour la détection d'oultier dans la matrice V, nous utilisons un modèle d'inflation de la variance [Box and Tiao, 1968, Chaloner and Brant, 1988]. Nous définissons un vecteur de variables indicatrices aléatoires $z = (z_1 \dots z_p)$ à valeurs dans $\{0, \dots K\}$, qui indique pour chaque locus si Y_l est outlier dans la k^{ieme} composante $z_l = k$, ou non outlier $z_l = 0$. Dans le cas où $z_l = 0$, nous considérons que le locus neutre. Ainsi nous adoptons le prior suivant pour chacun des p vecteurs de V

$$p(V_l|z_l = 0, \Sigma_V) = \mathcal{N}_K(0, \Sigma_V), \tag{10}$$

$$p(V_l|z_l = k, \Sigma_V) = \mathcal{N}_K(0, C_k^2\Sigma_V) \tag{11}$$

où Σ_V est la matrice de covariance entre les loadings. Nous choisissons cette matrice diagonale avec comme termes diagonaux σ_k^2. La matrice C_k^2 est la matrice d'inflation de la variance. C'est une matrice diagonale, avec comme termes diagonaux, $(1, \dots, 1, c_k^2, \dots, 1)$ où la valeur non-unitaire est le k^{eme} élément. Chaque terme de variance est modélisé tel que $\sigma_k^2 = \sigma^2\rho_k^2$, où σ^2 est la variance des résidus de l'équation (7). Cette re-paramétrisation est utilisée en factorisation de matrice bayésienne pour supprimer des axes dont la variance serait trop faible par rapport au bruit. C'est à dire que si la variance de l'axe k, σ_k^2 devient faible devant la variance des résidus σ^2 le paramètre ρ_k^2 deviendra très faible, ce qui permettra l'élimination de cet axe [Oba et al., 2003]. Nous prenons des priors non-informatifs pour les variances $p(\sigma^2) \propto \frac{1}{\sigma^2}$ et $p(\rho_k^2) \propto \frac{1}{\rho_k^2}, k = 1 \dots K$. Pour les priors des paramètres d'inflation nous utilisons un prior uniforme entre deux valeurs réelles a et b. En pratique, nous prenons $a = 1$ de façon à ce que la variance augmentée soit toujours supérieure à la variance standard. La valeur de b importe peu, en pratique nous choisissons $b = 10$.

Dans le modèle d'analyse factorielle présenté dans l'équation (7), il faut gérer un problème d'identifiabilité des facteurs [West, 2003]. En effet, pour toute matrice orthonormale P $(K \times K)$, $UV = UP^tPV$. Pour simplifier l'identifiabilité des paramètres, nous prenons un prior

$$p(U) = \prod_{i=1}^{n} \mathcal{N}_K(U_i|0, I_K) \tag{12}$$

pour la matrice des scores U [Oba et al., 2003]. Ce prior ne rajoute pas de paramètres au modèle et simplifie l'identification des paramètres. La simplification du modèle en réduisant le nombre de paramètre est utile pour une meilleure estimation de la loi a posteriori. Nous verrons dans la section

3.2.3 que l'estimation à l'aide d'algorithme MCMC présente quelques difficultés dues à la grande dimension des paramètres. Nous proposons également une solution heuristique pour empêcher les rotations de la matrice U.

3.2.2 Variables indicatrices z et modèle de Potts

Nous avons détaillé le modèle bayésien hiérarchique, à l'exception des variables indicatrices z. Ces p variables prennent des valeurs entières entre 0 et K. Nous introduisons le paramètre π qui est la proportion d'outlier, ou de marqueurs sous sélection dans le jeu de données. Pour refléter la faible proportion de gènes sous sélection dans le génome, nous choisissons un prior uniforme sur le logarithme de π. Pour chaque variable z_l la probabilité a priori de ne pas être outlier est de $1-\pi$, et π/K d'être dans une des K classes, soit

$$p(z_l) = (1-\pi, \frac{\pi}{K}, \dots \frac{\pi}{K}). \tag{13}$$

Néanmoins la méthode développée est faite pour être appliquée à des données à grande échelle. En génétique, lorsque le nombre de marqueurs augmente, les marqueurs physiquement proches sont de plus en plus corrélés à cause du déséquilibre de liaison. Ainsi, on ne s'attend jamais à voir apparaître un seul marqueur sous sélection, mais plutôt une région entière du génome [Reich et al., 2001]. En terme de modélisation, nous choisissons de représenter cela dans le modèle hiérarchique. Le but est de lisser les résultats. Nous entendons par là, de faire en sorte qu'un marqueur avec un signal fort, mais isolé dans une région du génome où il n'y a pas d'autre signal fort, soit pénalisé car il est plus vraisemblablement un faux positif, qu'un marqueur impliqué dans un processus d'adaptation. L'incorporation de cette information dans les priors a été utilisée en traitement de l'image [Ding et al., 2011].

Au lieu de prendre le prior naïf présenté dans l'équation (13), nous considérons un modèle de Potts [Winkler, 2003]. Ce modèle est une généralisation du modèle d'Ising [Ising, 1925] dans le cas de plus de deux classes. Il est utilisé en segmentation d'image pour modéliser la corrélation entre les pixels [Celeux et al., 2003]. Nous considérons comme portée du champ, seulement les marqueurs voisins. Pour décrire l'intensité du champs, nous introduisons une constante β. Cette constante peut être interprétée comme l'inverse d'une température. Si β est faible, l'entropie est grande, et l'état d'un marqueur, sa classe, n'est pas corrélé à celui de ses voisins. Pour une marqueur l, conditionnellement à ses voisins

$$p(z_l = k | z_{l-1}, z_{l+1}) = \begin{cases} \frac{\alpha}{K} e^{\beta(1_{z_{l-1}=k} + 1_{z_{l+1}=k})} \text{ si } k > 0 \\ (1-\alpha) e^{\beta(1_{z_{l-1}=0} + 1_{z_{l+1}=0})} \text{ si } k = 0 \end{cases} \tag{14}$$

Sur l'ensemble du vecteur z

$$p(z_1, \dots z_p) \propto (1-\pi)^{p_0} \pi^{(p-p_0)} e^{\beta \sum_{i \sim j} 1_{z_i = z_j}} \tag{15}$$

où p_0 est le nombre de marqueurs dans la classe 0. Le choix de la constante β influence la régularité des résultats. Nous choisissons de ne pas l'estimer, et de laisser ce choix à l'utilisateur. Dans notre analyse du jeu de données HGDP, nous choisissons $\beta = 1$. Il est possible d'incorporer l'estimation de la constante β dans l'algorithme bayésien présenté dans la section 3.2.3 à l'aide d'un pas Metropolis-Hastings. Egalement, une étude comparative de [Zhou and Schmidler, 2009] présente l'efficacité de différentes stratégies quant à l'estimation de β.

3.2.3 MCMC et facteurs de Bayes

Les paramètres du modèle hiérarchique bayésien peuvent être estimés de différentes façons. Nous choisissons d'utiliser une méthode Monte Carlo par chaîne de Markov. Le but est d'estimer la loi a posteriori

$$p(U, V, \Sigma_V, Z, C^2, \pi, \sigma^2 | Y) \propto p(Y|U, V, \sigma^2) p(U) p(\sigma^2) p(V|\Sigma_V, Z, C^2) p(\Sigma_V) p(C^2) p(Z|\pi) p(\pi). \tag{16}$$

L'algorithme final est un algorithme hybride contenant des pas d'échantillonneur de Gibbs ainsi que des pas Metropolis-Hasting. Cet algorithme est détaillé dans l'annexe 7.2. Comme mentionné dans la section 3.1.3 deux exécutions de l'algorithme peuvent donner des matrices U de projection des individus différentes, à cause des rotations possibles entre les facteurs. Une solution possible sans étendre d'avantage le modèle est d'initialiser la matrice U à l'aide d'une solution. Dans les jeux de données où les rotations rendent les résultats entre deux exécutions de l'algorithme trop différents, nous choisissons d'initialiser U avec la solution de la décomposition en valeurs singulières. En théorie, l'initialisation d'un MCMC n'a plus d'importance lorsque l'on atteint la distribution stationnaire. Cependant, la distribution que nous considérons est en très grande dimension, puisqu'il y a plusieurs paramètres pour chacun des p locus. l'algorithme ne décrit pas la loi a posteriori complètement, et explore des minima locaux. Dans ce cas, l'initialisation de la chaîne a une importance.

D'autres solutions qu'un algorithme MCMC pourraient être envisagées. On pourrait essayer de calculer les estimateurs de maximum a posteriori (MAP). Si une formule explicite n'est pas obtenue, Il est possible d'utiliser des algorithmes de gradient. Une autre méthode est l'inférence bayésienne variationnelle. Cette méthode récente (2000) vise à trouver les paramètres à l'aide d'une transformation variationnelle [Jaakkola and Jordan, 2000].

Pour chaque marqueur il est possible d'estimer un facteur de Bayes. Ce facteur est le rapport entre les vraisemblances marginales de deux modèles. Dans notre cas, c'est le rapport de la vraisemblance du modèle outlier divisé par la vraisemblance du modèle non-outlier. Pour chaque marqueur l,

$$BF_l = \frac{p(V_l|z_l > 0, C^2, \Sigma_V)}{p(V_l|z_l = 0, \Sigma_V)} \tag{17}$$

L'interprétation directe de ce facteur de Bayes est que si la valeur est supérieure à 1, le modèle outlier est plus vraisemblable que le modèle neutre pour ce marqueur. Une table d'interprétation et de significativité de ces facteurs est proposée dans la littérature [Jeffreys, 1998, Kass and Raftery, 1995]. Cette quantité est estimée à chaque itération de l'algorithme MCMC. Nous faisons l'approximation que nous pouvons estimer BF_l par la moyenne des ratios des vraisemblances à chaque itération. En pratique un marqueur outlier a le plus souvent une forte probabilité d'être atypique pour un seul facteur, et rarement pour plusieurs. Le facteur de Bayes peut être approximé par vraisemblance du modèle outlier dans une classe k divisé par la vraisemblance du modèle non-outlier.

3.2.4 Application à un modèle de divergence

Nous reprenons l'exemple présenté dans la section 3.1.3. Les données de cette simulation comportent 4 populations ayant évolué selon une structure hiérarchique complexe, i.e différente du modèle de divergence instantanée. Les facteurs de Bayes estimés par la méthode sont comparés aux F_{st} et $q-$valeurs estimées par les logiciels *fdist* et *Bayescan*. Les performances de chaque méthode sont estimées en calculant le taux de fausses découvertes (FDR) pour une sensibilité donnée. Pour un seuil de facteur de Bayes donné, la sensibilité est la proportion de marqueurs étant sujet à de l'adaptation avec un facteur de Bayes supérieur à ce seuil. En notant les quantités VP (pour vrais positifs) le nombre de marqueurs impliqués dans l'adaptation locale avec un facteur de Bayes supérieur au seuil, FP (pour faux positifs) le nombre de marqueurs neutres avec un facteur de Bayes supérieur au seuil, FN (pour faux négatifs) le nombre de marqueurs impliqués dans l'adaptation locale avec un facteur de Bayes inférieur au seuil, alors la sensibilité vaut VP/(VP + FN). Le taux de fausses découvertes est la proportion de marqueurs neutres parmi les marqueurs détectés comme impliqués dans un processus d'adaptation à une sensibilité fixée. En d'autres termes, FDR = FP/(FP + VP). Dans les exemples nous prenons un seuil tel que la sensibilité vaut 50% (la moitié des marqueurs impliqués dans un processus d'adaptation sont détectés) et nous observons le taux de faux positifs associé à cette sensibilité.

En calculant le taux de fausses découvertes pour une sensibilité de 50% sur l'exemple de la section 3.1.3, il apparaît que le modèle basé sur l'analyse factorielle a toujours un FDR plus faible que les méthodes implémentées dans *Bayescan* et *fdist*. Cela reste vrai si l'on choisit d'autres valeurs

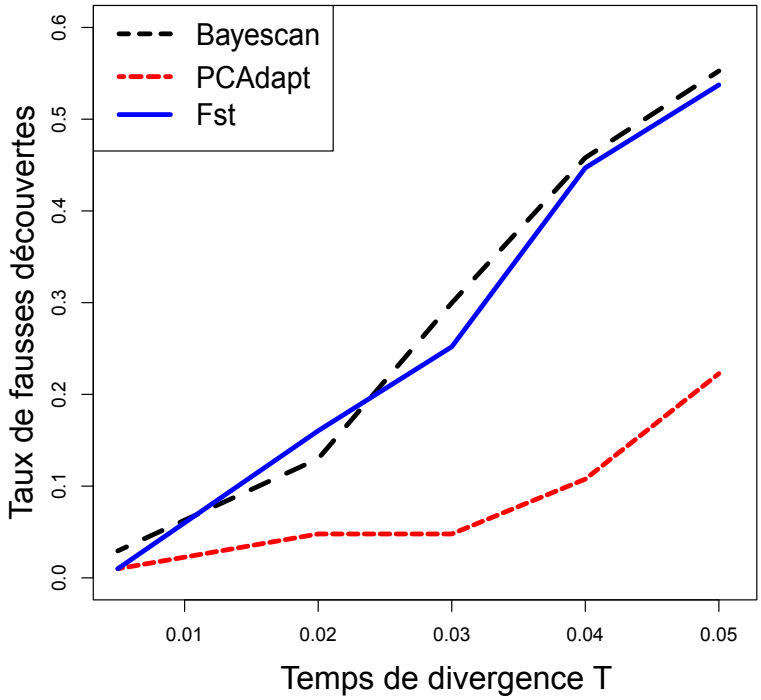


FIGURE 11 – Taux de fausses découvertes en fonction du temps de divergence neutre T. Le taux de fausses découvertes pour *Bayescan* et *fdist* augmente lorsque T augmente, ce qui correspond à une augmentation de l'écart au modèle supposé par les méthodes. Le taux de fausses découvertes augmente moins rapidement pour le modèle à facteurs.

pour la sensibilité telles que 25% ou 75%. Il apparait de plus que lorsque le temps de divergence neutre T augmente, le taux de fausses découvertes augmente (Figure 11). Cela peut s'expliquer par l'écart grandissant au modèle de divergence instantanée. En augmentant le temps de divergence neutre T, le processus démographique ayant engendré les données devient plus complexe, et les scénarios sont plus en contradiction avec les hypothèses des modèles des figures 3 et 4. Le modèle à facteurs est beaucoup plus robuste à l'augmentation de T. Cette tendance se vérifie avec d'autres seuils pour la sensibilité (25% et 75%).

3.3 Vers une version rapide du logiciel

3.3.1 Analyse en composantes principales des données

Nous avons détaillé précédemment les liens entre l'ACP et les modèles à facteurs. La proximité entre les deux modèles fait que s'il est possible d'interpréter les valeurs des loadings dans un modèle

à facteurs, il est possible de donner la même interprétation aux loadings de l'ACP. Dans le cas où nous considérons une matrice de données Y de dimensions $n \times p$, nous choisissons ici de centrer et réduire cette matrice. La matrice résultante notée $\tilde{Y}$, est une matrice dont les colonnes sont de moyenne nulle, et de variance unitaire. Pour réaliser cette analyse en composantes principales, Nous calculons la matrice de covariance $(n \times n)$ entre les individus des données

$$\Sigma = \tilde{Y}^t \tilde{Y} \tag{18}$$

et à la diagonaliser, de façon à ce que

$$\Sigma = U \Lambda^{2t} U \tag{19}$$

où Λ^2 est la matrice diagonale des valeurs propres, et U est la matrice des vecteurs propres. Ainsi les données peuvent être écrites

$$\tilde{Y} = U \Lambda V. \tag{20}$$

Avec cette écriture, U est la projection des individus dans les sous espaces propres, et il est possible d'estimer les loadings de la matrice V en utilisant l'équation $V = \Lambda^{-1t} U \tilde{Y}$. Par analogie avec l'analyse factorielle U est la matrice des facteurs, et V est la matrice des loadings. Il est donc possible de chercher des traces d'adaptation dans la matrice V.

Une approche plus classique de l'analyse en composantes principales serait de calculer et diagonaliser $\Sigma = \frac{1}{n-1} {}^t\tilde{Y}\tilde{Y}$. Dans ce cas la matrice Σ serait $(p \times p)$ et par conséquent, bien trop grande pour être diagonalisée. La diagonalisation est effectuée avec la fonction `dsyevr` de *Lapack* [Anderson et al., 1999]. Cette opération est très peu coûteuse car il est possible de ne calculer qu'un faible nombre K de vecteurs et valeurs propres avec cette opération. L'opération la plus coûteuse reste le calcul de la matrice de covariance avec un grand nombre p de marqueurs.

3.3.2 Statistiques de l'ACP

Pour calculer les statistiques pouvant indiquer l'adaptation locale, nous utilisons la matrice V. Nous avons uniquement besoin de calculer les valeurs impliquées dans les K premières directions. A partir de ces valeurs, nous étudions plusieurs statistiques pour détecter les marqueurs fortement différenciés pour un des K facteurs. Les statistiques suivantes sont étudiées :

$$S_1 = \sum_{k=1}^{K} (\frac{V_{kl}}{mad_k})^2 \quad (21)$$

$$S_2 = max_{k=1...K} (\frac{V_{kl}}{mad_k})^2 \quad (22)$$

$$S_3 = \sum_{k=1}^{K} \lambda_k^2 V_{kl}^2 \quad (23)$$

où $mad_k = median(|V_{kl} - median(V_l)|)$, et λ_k est la racine carrée de la k^{me} valeur propre, ou le k^{me} terme diagonal de Λ. La statistique S_1 est la somme des K loadings normalisés au carré d'un marqueur. Les marqueurs ayant de fortes contributions pour les facteurs ont donc de fortes statistiques S_1. La statistique S_2 ne prend que le plus fort des K loadings normalisés au carré. Comme mentionné dans la section 3.2.3, un marqueur impliqué dans un processus d'adaptation est souvent associé à un seul facteur. Notons que pour ces deux premières statistiques, la variance expliquée par les facteurs λ_k^2 n'intervient pas. La quantité $\lambda_k V_{kl}$ est la corrélation entre la variable l et le facteur k. Ainsi, dans le cas où la variable l est de variance unitaire, la statistique S_3 est la somme des coefficients de corrélation linéaire entre un marqueur et les facteurs. C'est donc la proportion de variance du l^{eme} locus expliquée par les K premières composantes principales. Dans le cas où $K+1$ populations sont présentes dans le jeu de données, cette statistique est équivalente à la F_{st}, en considérant la F_{st} comme une statistique d'analyse de la variance [Michalakis and Excoffier, 1996].

A partir de ces statistiques, nous obtenons des scores pour classer les marqueurs candidats. Nous étudions le résultat de ces classements sur le jeu de données simulées présenté dans la section 3.1.3. Nous observons les taux de fausses découvertes pour une sensibilité de 50% (Figure 12). La statistique S_1 donne des résultats similaires aux facteurs de Bayes de la section 3.2.3. Nous voyons que sur cet exemple cette statistique fait le moins de faux positifs. Cette tendance se confirme dans d'autres simulations (résultats non montrés). Il est intéressant de remarquer également que, comme prévu, S_3 donne les mêmes résultats qu'un scan à F_{st} (Figure 13).

3.3.3 Intérêts et limites

Cette approche, beaucoup plus simple méthodologiquement que le modèle bayésien d'analyse factorielle, a l'avantage d'être beaucoup moins coûteuse en temps de calcul que l'algorithme MCMC implémenté pour estimer les paramètres du modèle à facteurs et les facteurs de Bayes. Il nous est en particulier possible d'appliquer cette méthode sur les données du projet 1000 génomes [Consortium, 2012] comportant près de 38 millions de marqueurs SNP (section 3.3.4). Pour des jeux de données de faible taille ($p < 10000$) le temps de calcul de l'algorithme MCMC croit déjà fortement (Figure 14). Cette différence devient bien plus importante avec des données telles que

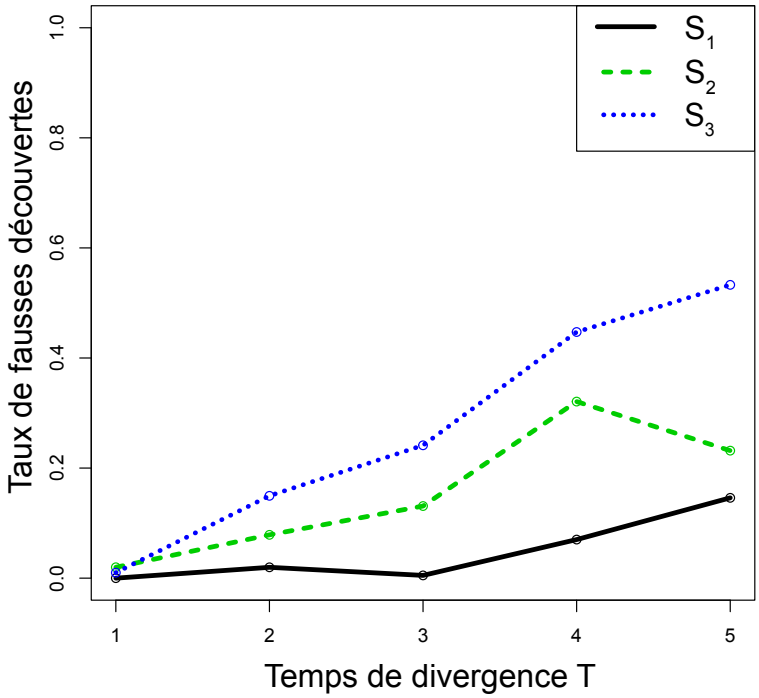


FIGURE 12 – Courbe du taux de fausses découvertes pour une sensibilité de 50%. Les trois courbes représentent les trois statistiques S_1, S_2 et S_3 présentées dans la section 3.3.2.

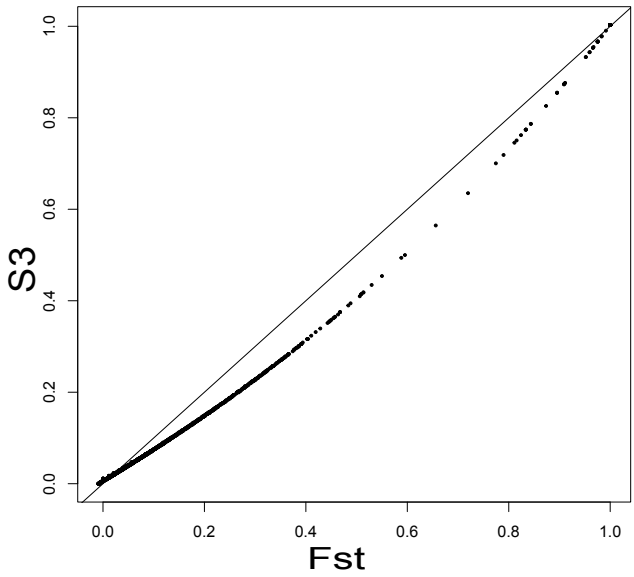


FIGURE 13 – F_{st} estimées à l'aide de l'équation (2) tracées en fonction de la statistique S_3. Ces valeurs ont été calculées sur une partie des marqueurs des données simulées de la section 3.1.3.

celles du projet 1000 génomes, et rend la version MCMC impossible à exécuter.

Néanmoins, cette approche peut s'avérer moins puissante et moins robuste au choix du nombre de facteurs. Nous pouvons observer cela en estimant le FDR sur les données simulées de la section 3.2.4. En fonction du nombre de facteurs, le taux de fausses découvertes varie fortement (Figure 15). Pour une valeur de K optimale, les résultats sont semblables. Lorsque la valeur de K devient trop grandes, les composantes principales considérées ne contiennent plus d'information mais essentiellement du bruit. Sommer les loadings normalisés de ces composantes principales masque les signaux potentiels de sélection. Pour $K = 3$, les deux méthodes ont un taux de fausses découvertes d'environ 5% pour une sensibilité de 50% (Figure 15). Pour $K = 6$, le modèle à facteurs bayésien a toujours un taux de fausses découvertes de 5%, alors que la version rapide basée sur l'ACP a un taux de fausses découvertes d'environ 40%.

Nous comparons les deux approches avec un autre jeu de données simulées à l'aide de *SFS_CODE* [Hernandez, 2008]. La simulation contient 179 individus séparés en 3 populations (données de [Fagny et al., 2014]). Le scénario de simulation reproduit l'expansion humaine depuis la sortie d'Afrique vers l'Europe et l'Asie à l'aide d'un modèle hiérarchique (Figure 16, Panel A). Dans ce jeu de données la valeur optimale de K est 2 puisqu'il y a trois continents à distinguer. La structure inter-continentale apparaît sur les deux premières composantes principales (Figure 16, Panel

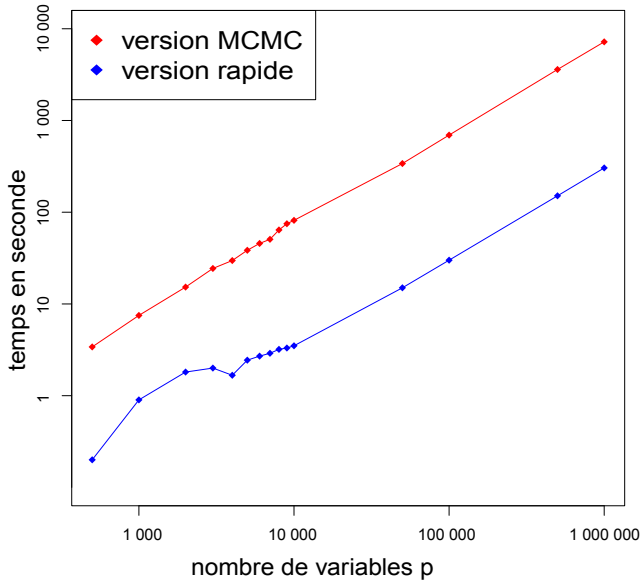


FIGURE 14 – Temps d'exécution en secondes de l'algorithme MCMC et de la version rapide en fonction du nombre de marqueur dans les données p. L'algorithme MCMC est lancé avec 100 itérations, en incluant la phase de chauffe.

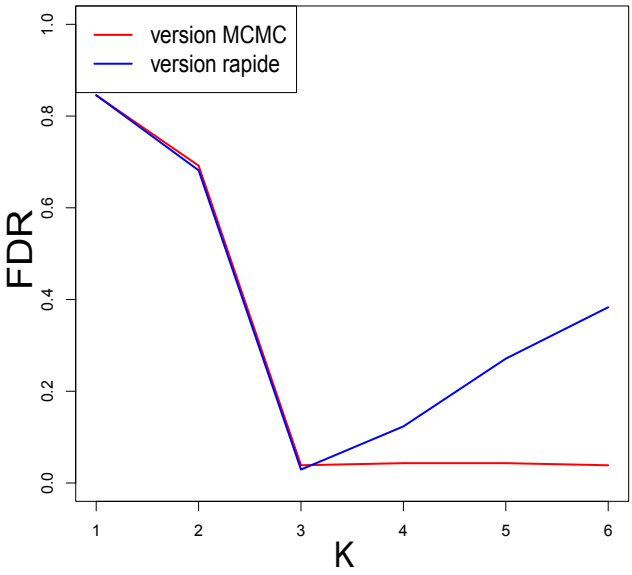


FIGURE 15 – Taux de fausses découvertes en fonction du nombre de facteurs K du modèle. Les FDR sont estimés sur les données simulées présentées dans la section 3.1.3

B).

Les performances du modèle à facteurs et de la statistique S_1 de l'analyse en composantes principales sont proches. La version ACP est avantageuse dans ce cas, car le jeu de données comporte plus de $1\,700\,000$ SNP, et le temps de calcul est beaucoup plus faible. L'exécution de 100 itérations du MCMC pour ce jeu de données prend $11\,468$ secondes, alors que le calcul des statistiques de l'analyse en composantes principales se fait en 344 secondes.

Une grande partie des marqueurs impliqués dans un processus d'adaptation sont détectés par les deux méthodes. Pour une valeur de K idéale, les résultats sont équivalents (Figure 16, Panel C). A toute sensibilité donnée, la précision (ou le taux de fausse découverte) est égale pour les deux méthodes. Cependant le choix de la valeur optimale de K est essentiel dans la version ACP. Le modèle bayésien est beaucoup plus robuste au choix de K (Figure 16, Panel C).

En conclusion, cette version rapide de la méthode basée sur l'analyse factorielle peut égaler les performances de la méthode bayésienne présentée dans la section 3.2. De plus, la nature de la méthode basée sur l'ACP la rend applicable à tout type de données, même les plus volumineuses. L'algorithme MCMC du modèle bayésien ne peut plus être utilisé lorsque p est de l'ordre du million, ce qui devient fréquent avec les données de séquençage. Pour les jeux de données de taille raisonnable ($p < 10^6$), le modèle bayésien reste une meilleure option. En effet, cet algorithme est plus robuste au choix de K. Cette robustesse peut être expliquée par la capacité du modèle bayésien à réduire les facteurs de faible variance à une variance nulle. Grâce à la paramétrisation $\sigma_k^2 = \sigma^2 \rho_k^2$ présentée dans la section 3.2, les facteurs de variance σ_k^2 faible devant σ^2, entraînent les valeurs de ρ_k^2 vers 0 [Oba et al., 2003] dans l'algorithme MCMC. Puisque ρ_k^2 est la variance mesurée d'un facteur, une variance faible entraîne les loadings correspondant à ce facteur vers 0.

La méthode est prévue pour une analyse exploratoire des données. Elle apprend la structure des populations, contrairement aux méthodes décrites dans la section 2, et donne des marqueurs candidats. Dans ces conditions, le choix de K est une question délicate, puisque il n'y a pas de connaissance des populations a priori. La robustesse au choix de K pour la détection de marqueurs impliqués dans des processus d'adaptation locale est donc un atout essentiel du modèle bayésien. Les deux méthodes sont implémentées dans le logiciel *PCAdapt*. Le logiciel est disponible à l'adresse `http://membres-timc.imag.fr/Nicolas.Duforet-Frebourg/PCAdapt.html`.

3.3.4 Analyse du jeu de données du projet 1000 Génomes

Nous utilisons la version rapide du logiciel *PCAdapt* pour analyser les données du projet 1000 Génomes [Consortium, 2012]. Dans ce jeu de données nous retenons 36 536 154 marqueurs SNP. Ces marqueurs sont tous de couverture faible. Nous choisissons de ne garder que ceux de couverture

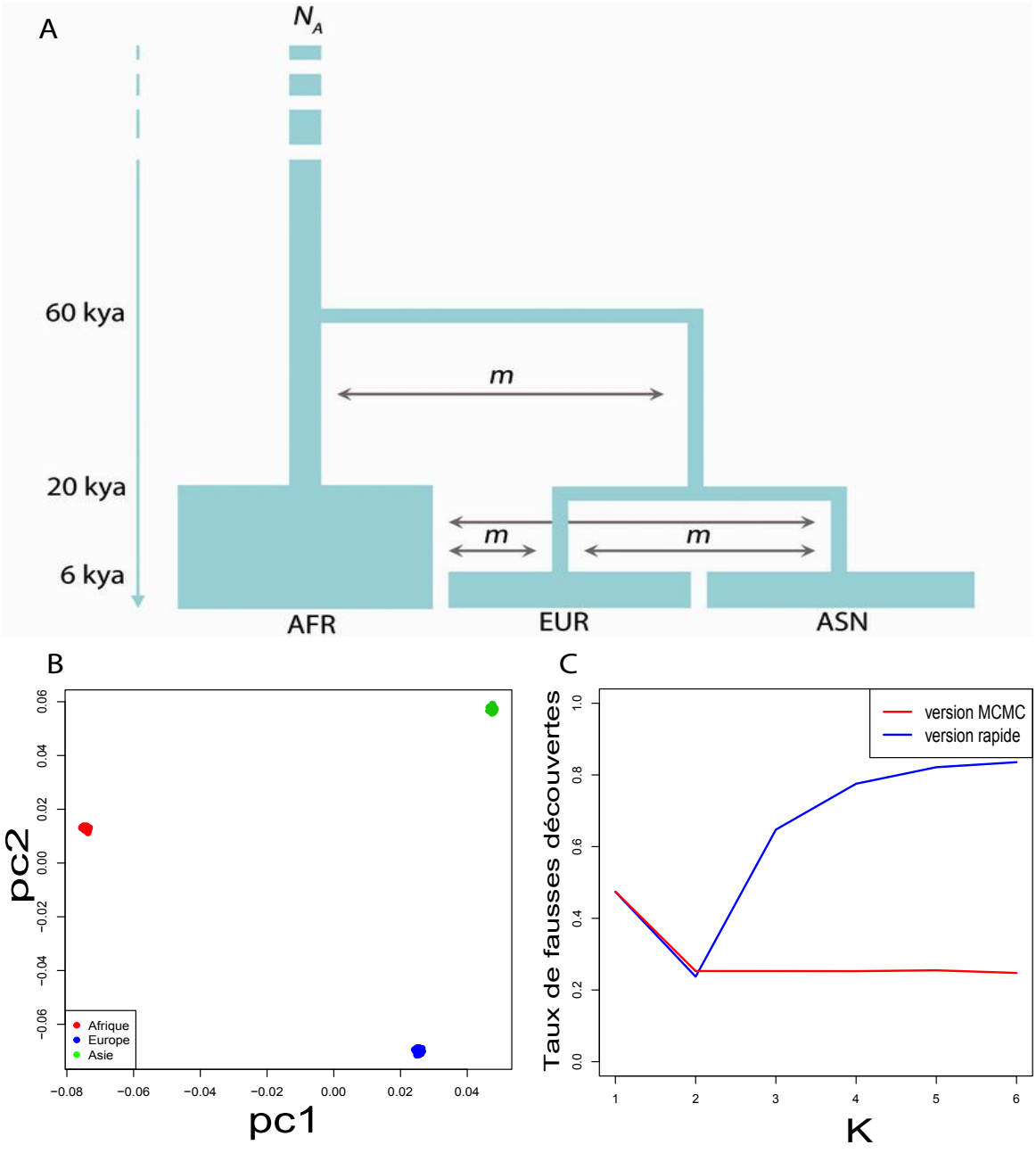


FIGURE 16 – Panel A : Scénario utilisé pour simuler les données les paramètres ont été calibrés pour reproduire au mieux l'expansion humaine hors d'Afrique [Fagny et al., 2014]. Panel B : Première et deuxième composantes principales du jeu de données simulées. Panel C : Taux de fausses découvertes en fonction de K pour les données simulées de la section 3.3.3.

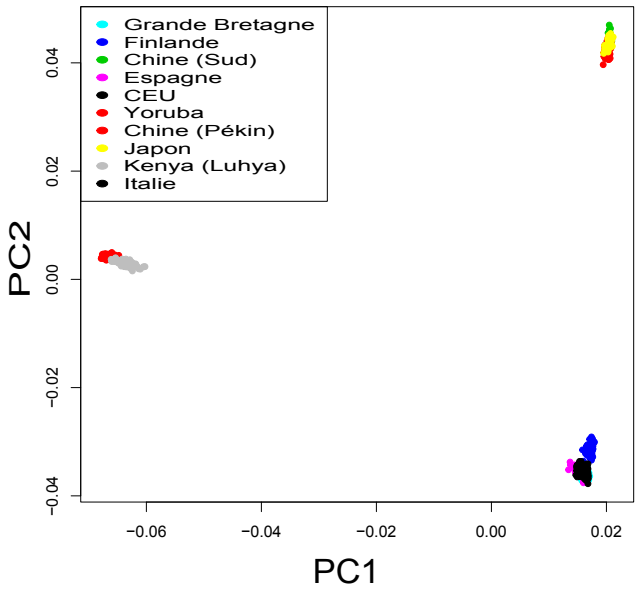


FIGURE 17 – Première et deuxième composantes principales pour les données du 1000 génomes.

faible pour avoir une couverture uniforme du génome. Parmi les 1092 individus, nous choisissons de n'en garder que 850. Nous excluons de l'analyse les individus des populations afro-américaines du sud ouest des États-unis, mexicaines, colombiennes et porto ricaines. L'avantage est de ne garder que des populations des continents africains, européens et asiatiques. Cela permet d'avoir une analyse de données très similaires aux données simulées présentées dans la section 3.3.3. Ainsi, les facteurs représenteront la différenciation inter-continentale, et les signaux de sélection seront faciles à interpréter.

La première composante principale sépare les individus d'Afrique du reste des individus. La deuxième composante principale sépare les individus européens des individus asiatiques (Figure 17). Des analyses d'enrichissements et d'ontologies sont nécessaires et sont en cours pour exploiter rigoureusement ces résultats. Néanmoins, nous pouvons observer les meilleurs marqueurs candidats pour les deux facteurs (Tables 1 et 2). Certains de ces gènes sont connus pour leur implication dans certains traits physiologiques. Le gène SLC24A5 est impliqué dans la couleur de peau en Eurasie [Canfield et al., 2013]. Le gène EDAR est lui associé à des traits tels que l'épaisseur des cheveux et la transpiration [Kamberov et al., 2013].

Chromosome	début	identifiant rs	description	distance
chr6	105883148	9486092	PREP prolyl endopeptidase	32148
chr6	105891509	1325421	PREP prolyl endopeptidase	40509
chr5	119745985	6862601	PRR16 proline-rich protein	54034
chr20	53252641	6014096	DOK5 docking protein	0
chr20	53253868	2208074	DOK5 docking protein	0
chr5	119786254	2973133	PRR16 proline-rich protein 16	13765
chr5	142079283	2879382	FGF1 fibroblast growth factor	1647
chr5	142079591	10072789	FGF1 fibroblast growth factor	1955
chr5	142080485	6879239	FGF1 fibroblast growth factor	2849
chr5	142080603	6875059	FGF1 fibroblast growth factor 1 isoform 1 precursor	2967

TABLE 1 – Table des 10 marqueurs ayant les plus fortes statistiques S_3 et associés à la composante principale 1. Cette composante représente la différenciation entre le continent africain et le reste des individus des données. Pour chacun des marqueurs la distance au gène le plus proche est indiquée dans la colonne distance. Les détails sur la fonction de ce gène sont précisés.

Chromosome	début	identifiant rs	description	distance
chr15	48426485	1426654	SLC24A5 sodium/potassium/calcium exchanger	0
chr15	48485927	2413887	CTXN2 cortexin-2	0
chr5	33951694	16891982	SLC45A2 membrane-associated transporter	0
chr6	2745353	6927195	MYLK4 myosin light chain kinase	0
chr15	48390957	12441154	SLC24A5 sodium/potassium/calcium exchanger	22212
chr5	33951117	35397	SLC45A2 membrane-associated transporter	0
chr15	48394587	57108441	SLC24A5 sodium/potassium/calcium exchanger	18582
chr2	109543884	922452	EDAR tumor necrosis factor receptor	0
chr15	48389925	12440301	SLC24A5 sodium/potassium/calcium exchanger	23244
chr15	48392166	1834640	SLC24A5 sodium/potassium/calcium exchanger	21003

TABLE 2 – Table des 10 marqueurs ayant les plus fortes statistiques S_3 et associés à la composante principale 2. Cette composante représente la différenciation entre les individus du continent européen et les individus du continent asiatique. Pour chacun des marqueurs la distance au gène le plus proche est indiquée dans la colonne distance. Les détails sur la fonction de ce gène sont précisés.

4 Un modèle pour la différenciation génétique locale

Une autre tâche importante de la génétique des populations est de caractériser la structure génétique au sein des individus. Nous présentons ici dans un premier temps un schéma de structure génétique spatiale, le schéma d'isolement par la distance. Puis nous introduisons un modèle non mécaniste basé sur des processus gaussiens pour caractériser les régions de forte différenciation locale au sein de l'aire de répartition d'une espèce. Ce modèle est en particulier utile pour la détection de barrières aux flux de gènes qui peuvent séparer les individus d'une même espèce.

Une partie des résultats de cette section a fait l'objet d'une publication dans Evolution :
Nicolas Duforet-Frebourg and Michael G.B. Blum, Nonstationary patterns of isolation-by-distance, inferring measures of local genetic differentiation with Bayesian kriging. Evolution, 2014, 68 : 1110–1123. doi : 10.1111/evo.12342.
L'article se trouve dans la section 8.2.

4.1 Le modèle d'isolement par la distance

4.1.1 Modèle d'isolement par la distance

En 1943, Wright introduit le modèle d'isolement par la distance [Wright, 1943]. Dans ce modèle, il suppose que la différenciation entre des sous-groupes d'individus augmente avec la distance géographique séparant les sous-groupes. Les descendants d'une génération d'individus sont supposés vivre et se reproduire près de leur lieu de naissance, ainsi les différences dans les fréquences d'apparition d'un allèle augmentent systématiquement avec la distance. La distance entre les sous-groupes d'individus au sein d'une espèce empêche donc cette espèce de former un ensemble génétiquement homogène, et induit de la structure de population. Cette structure a également été étudiée par Malécot [Malécot et al., 1948], en terme de changement de la corrélation entre les allèles des individus en fonction de la distance géographique.

Kimura propose un modèle de simulation [Kimura, 1953] pour décrire ce processus qu'il appellera le modèle de stepping-stone. Ce modèle est en fait un modèle en îles, comme présenté dans la section 2, avec la particularité de localiser géographiquement les îles. Grâce à cette localisation, il est possible de n'autoriser les migrations qu'entre sous-populations voisines (Figure 18). A l'aide de ce schéma de simulation, Kimura et Weiss [Kimura and Weiss, 1964] montrent théoriquement que la corrélation décroit en fonction de la distance. En considérant deux sous-populations séparées par k sous-populations, la corrélation r entre les fréquences alléliques des deux sous-populations

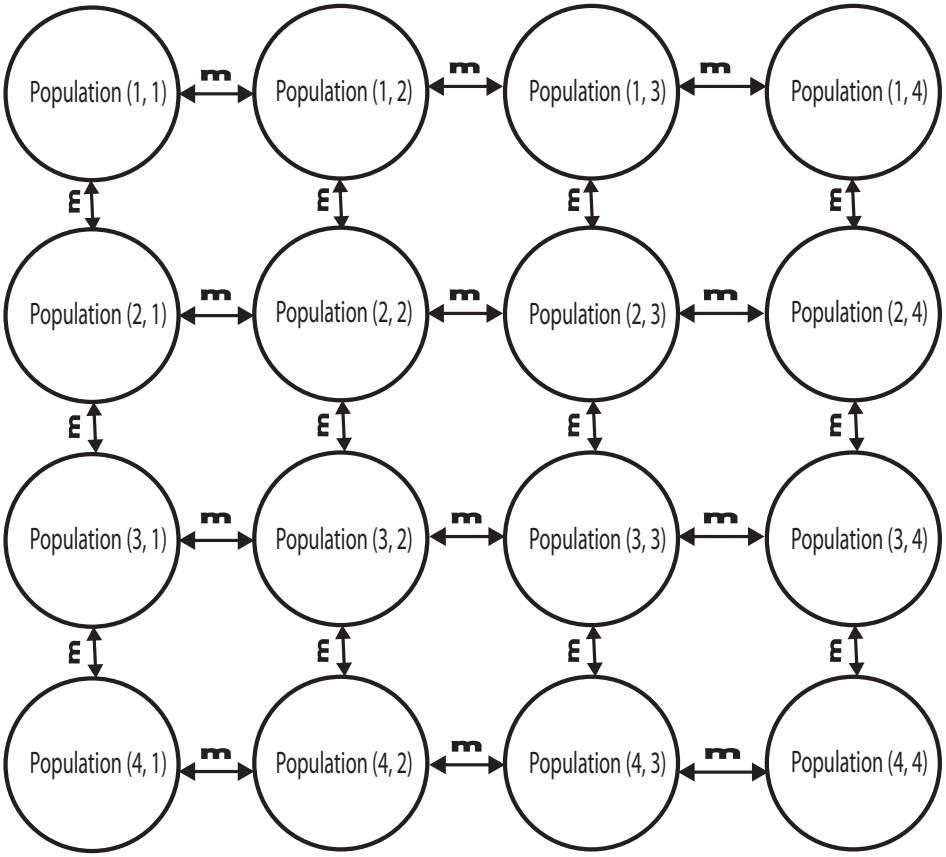


FIGURE 18 – Modèle de stepping-stone en 2 dimensions. Les populations voisines échangent à chaque génération une proportion m de migrant, représenté par une flèche. Par défaut cette proportion est constante pour toutes les paires de sous-populations.

en présence d'une infinité de sous-populations est donnée par

$$r(k) \approx e^{-\sqrt{2\alpha}k} \tag{24}$$

dans le cas d'un modèle unidimensionnel. En deux dimensions, 2 populations séparées par k_1 sous-populations dans la première dimension et k_2 sous-populations dans la seconde dimension, ont une corrélation

$$r(k_1, k_2) \approx \frac{e^{-\sqrt{4\alpha}\sqrt{k_1^2+k_2^2}}}{\sqrt{k_1^2+k_2^2}} \tag{25}$$

où α est un paramètre dépendant des taux de migration du modèle.

D'autres développements théoriques sur ce modèle ont par la suite était apportés [Slatkin, 1993, Rousset, 1997].

4.1.2 Schéma d'isolement par la distance et non-stationnarité

Le modèle de stepping-stone est un modèle de simulation composé de sous-populations. Dans ces simulations, il est donc possible de prédire la corrélation entre les fréquences alléliques de deux populations, à condition de connaître les taux de migrations entre les sous-populations. Dans les données, les hypothèses du modèle ne sont pas forcément respectées. Les individus peuvent être répartis uniformément sans former de sous-populations évidentes, comme c'est le cas en génétique du paysage [Manel et al., 2003]. Néanmoins, l'augmentation de la différenciation génétique avec la distance, ou respectivement la décroissance de la corrélation entre les fréquences alléliques, reste un résultat attendu. C'est ce que nous appelerons le schéma d'isolement par la distance. Ce schéma est attendu dans la majorité des populations naturelles, y compris les populations humaines [Sharbel et al., 2000, Ramachandran et al., 2005, Hardy et al., 2006, Hellberg, 2009].

Nous pouvons considérer l'exemple de l'espèce *Phyteuma hemisphericum*. Cette plante alpine est l'une des 28 espèces étudiées dans le projet Intrabiodiv [Gugerli et al., 2008]. Ce projet a pour but d'étudier la diversité génétique de diverses espèces de plantes alpines. Pour cela, des individus des différentes espèces ont été échantillonnés sur l'ensemble de la chaîne des Alpes. Ainsi, nous disposons avec *Phyteuma hemisphericum* d'une espèce échantillonnée sur l'ensemble de son aire de répartition. En calculant les corrélations entre fréquences alléliques, et en les comparant aux distances géographiques, il est donc possible d'observer le schéma d'isolement par la distance (Figure 19).

Le schéma d'isolement par la distance peut souvent cacher des processus démographiques complexes qui ont pour conséquence des variations dans la décroissance de cette corrélation à une échelle donnée. Dans le cas de régions avec des migrations plus faibles [Slatkin, 1985, McRae and Beier, 2007], cette diminution serait plus importante entre les populations dans ces régions. En analysant la diminution de la corrélation en fonction de la distance entre les individus de *Phyteuma hemisphericum*, des variations spatiales apparaissent (Figure 19). Les fréquences alléliques des populations proches dans la partie est des Alpes sont en moyenne plus corrélées que les fréquences alléliques des populations de la partie ouest. Nous appellerons ces différences dans la décroissance des corrélations des fréquences alléliques la non-stationnarité du schéma d'isolement par la distance. Nous faisons ici référence à la non-stationnarité faible, ou encore non-stationnarité de second ordre d'un processus. Si les fréquences alléliques sont les valeurs du processus, la covariance entre des paires de fréquences également espacées varie en fonction de l'espace.

Cette non-stationnarité spatiale peut être due à différents facteurs, tels que des variations de taille dans les sous-populations [Serrouya et al., 2012], des régions favorisant les flux de gènes, [Arnaud, 2003, Munshi-South, 2012], ou encore des barrières naturelles telles qu'une mer, ou une

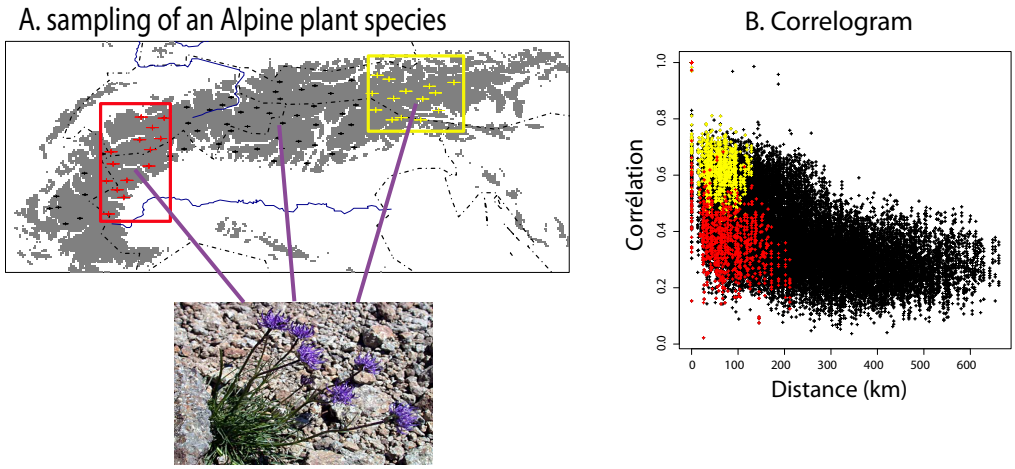


FIGURE 19 – Panel A : Schéma d'échantillonnage de la plante *Phyteuma hemisphericum* sur l'arc alpin. Panel B : Schéma d'isolement par la distance. Les corrélations sont calculées pour chaque paire de populations à l'aide des fréquences alléliques de marqueurs AFLP. Les paires comprenant deux populations dans le secteur d'échantillonnage en jaune (respectivement en rouge), sont coloriées en jaune (respectivement rouge). En moyenne, les populations proches géographiquement dans le secteur jaune (rouge) sont plus (moins) corrélées génétiquement.

chaîne de montagnes [Castella et al., 2000, Epps et al., 2005]. La détection de barrières en particulier a fait l'objet de développements méthodologiques [Barbujani et al., 1989, Manni et al., 2004, Cercueil et al., 2007]. Le but ici est de proposer un modèle permettant de caractériser la non-stationnarité du schéma d'isolement par la distance. Ce modèle permettra de détecter les barrières au flux de gène, mais également des variations dues à d'autres processus démographiques.

4.2 Non-stationnarité et processus gaussiens

Pour caractériser la non-stationnarité du schéma d'isolement par la distance, nous choisissons d'utiliser des processus gaussiens. Nous décrivons ici ces processus, leur utilité, et une approche paramétrique pour estimer les covariances non-stationnaires.

4.2.1 Processus Gaussiens

Dans cette section, nous notons n le nombre de populations échantillonnées dans les données, et p le nombre de marqueurs génétiques dans les données. Nous notons X_j le vecteur des n fréquences alléliques d'un marqueur j. Nous supposons que ce vecteur, est la réalisation d'un processus gaus-

sien [Rasmussen, 2006], i.e.

$$p(X_j) = \mathcal{N}(m, \Psi), \tag{26}$$

où m est une fréquence allélique moyenne, et Ψ est la matrice de covariance du processus.

Nous nous intéressons plus particulièrement à la non-stationnarité de la covariance Ψ. Notons d_{ij} la distance entre les populations i et j. Sous l'hypothèse de stationnarité de second ordre, la covariance entre i et j est une fonction décroissante qui dépend uniquement de d_{ij}. La matrice de covariance pourrait s'écrire pour tout $i, j = 1 \dots n, \Psi_{i,j} = f(d_{i,j})$. Cette hypothèse est utilisée en krigeage dit ordinaire ou en krigeage simple [Cressie, 1993]. La fonction f serait une fonction décrivant la décroissance de la covariance avec la distance spatiale, aussi appelée corrélogramme. Plusieurs familles de fonctions peuvent être utilisées pour le corrélogramme, telles que des fonctions exponentielles, gaussiennes, sphériques, ou de Matèrn [McBratney and Webster, 1986]. Le krigeage permet d'estimer la réalisation d'un processus à un lieu donné de l'espace. Dans sa formulation originale [Matheron, 1962, Cressie, 1993], le krigeage ne fait aucune hypothèse gaussienne. Ici nous présentons une formulation du krigeage en termes de processus gaussiens [Bishop, 2006].

4.2.2 Modèle paramétrique non-stationnaire

Plusieurs modèles ont été proposés pour la modélisation de processus spatiaux non-stationnaires [Higdon et al., 1999, Nott and Dunsmuir, 2002, Paciorek and Schervish, 2006]. Un modèle paramétrique pour la fonction f, définissant la covariance a été proposé par [Paciorek and Schervish, 2006]. Ce modèle repose sur la convolution de noyaux gaussiens en différents sites de l'espace. Considérons x, un site des données, K_x est une densité gaussienne multivariée de covariance Σ_x, qui décrit localement la covariance du processus. Paciorek et Schervish montrent que

$$\Psi(x_i, x_j) = \sigma^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left|\frac{\Sigma_i + \Sigma_j}{2}\right|^{\frac{-1}{2}} e^{-t(x_i - x_j)(\frac{\Sigma_i + \Sigma_j}{2})^{-1}(x_i - x_j)}, i, j = 1 \dots n, \tag{27}$$

où σ^2 est un paramètre d'amplitude et chaque matrice Σ_i informe sur l'intensité de la décroissance dans chaque direction de l'espace en un site x_i. Il est démontré que Ψ est une fonction de covariance valide [Paciorek and Schervish, 2006].

Pour notre étude, nous choisissons un modèle isotrope non-stationnaire. L'intensité de la décroissance est la même dans toutes les directions, ainsi $\Sigma_i = r_i I_2, i = 1 \dots n$, avec I_2 la matrice identité en dimension 2. Les paramètres r_i peuvent être vus comme des paramètres de portée. Une forte valeur de r_i implique une forte covariance à une distance donnée. Dans le cas de *Phyteuma hemisphericum* (figure 19), les sites en jaunes auraient de fortes valeurs de r_i alors que les sites en rouges auraient de faibles valeurs de r_i. Pour l'estimation des différents paramètres du modèle,

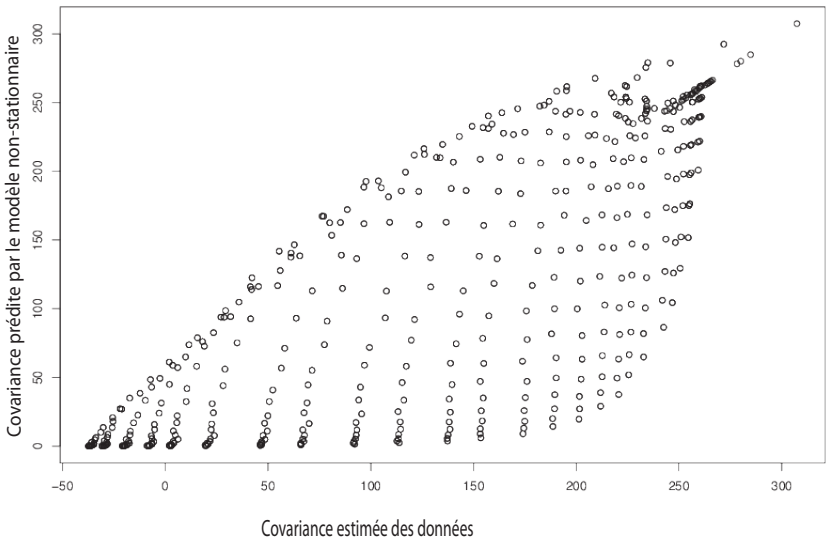


FIGURE 20 – Covariance estimée dans les données en fonction de la covariance prédite par le modèle paramétrique de [Paciorek and Schervish, 2006]. La covariance est estimée sur les simulations de gradient de migration, présentées dans la section 4.4.

[Paciorek and Schervish, 2006] proposent un algorithme MCMC en utilisant des priors gaussiens sur les r_i. Une extension propose un algorithme de gradient pour déterminer le maximum a posteriori des différents paramètres [Plagemann et al., 2008].

Cette approche permet à partir d'un modèle de covariance paramétrique d'apprendre un ensemble de paramètres interprétables notés r_i, qui indique pour chaque site où des données sont disponibles, quelle est la portée de la covariance. En d'autres termes, ces paramètres indiquent l'intensité de la décroissance dans le schéma d'isolement par la distance. Néanmoins, nous avons observé dans divers jeux de données simulées avec des modèles réalistes, que ce modèle ne permet pas de reproduire la covariance estimée dans les données (Figure 20). Ce modèle à l'inconvénient de ne pas être assez flexible. Nous choisissons donc d'estimer la non-stationnarité du schéma d'isolement par la distance par une méthode non-paramétrique. Dans l'article publié dans Evolution (section 8.2), nous utilisons le modèle paramétrique de [Paciorek and Schervish, 2006] uniquement pour simuler des données non-stationnaires. Cela nous permet de comparer la corrélation non-stationnaire à courte distance donnée par le modèle paramétrique et la différenciation génétique locale que nous estimons avec la méthode détaillée ci-dessous.

4.3 La différenciation génétique locale

Pour définir la différenciation génétique locale, nous choisissons de définir un voisinage autour de chaque site où des données sont échantillonnées, et d'estimer la covariance moyenne entre les données du site et le voisinage. A partir de la matrice de covariance entre les sites échantillonnés, l'objectif est donc d'estimer la covariance entre sites échantillonnés et le voisinage de ces sites. A la différence du krigeage où l'on cherche généralement à faire des prédictions sur le moment d'ordre 1, ici nous visons à faire des prédictions sur le moment d'ordre 2. Cela nous permet par la suite d'estimer la covariance, ou la F_{st}, entre une population et son voisinage qui se trouve à une distance fixée.

4.3.1 Définition d'un voisinage

Nous considérons toujours le vecteur de fréquences alléliques du marqueur j, X_j échantillonnées sur n sites de l'aire de répartition. Pour chacun de ces n sites, nous définissons un voisinage constitué de n_v sites à une distance d du site échantillonné. Le nombre de voisins n_v constituant le voisinage influe sur la précision des estimations. La distance d ne doit pas être trop grande pour que l'estimateur reste local. Il faudra en général choisir $d < min_{i,j \in 1...n} d_{i,j}$. Nous notons Y_j le vecteur des fréquences alléliques aux sites constituant les voisinages. Ces données sont non observées. Elles représentent les fréquences alléliques du marqueur j dans les n_v populations fictives avoisinant les n sites des données. Nous supposons que

$$(\mathbf{X_j}, \mathbf{Y_j}) \rightsquigarrow \mathcal{N}((\mathbf{m_x}, \mathbf{m_y}), \mathbf{\Psi}), \tag{28}$$

où (m_x, m_y) est un vecteur fréquence allélique et Ψ la matrice de covariance d'un processus gaussien $(n+n \times n_v)$-varié. Une approche de krigeage traditionnelle serait d'estimer Y_j en chaque point du voisinage. Cependant, nous choisissons à la place d'estimer la structure de covariance entre X et Y. Nous écrivons la structure de covariance

$$\mathbf{\Psi} = \begin{pmatrix} \mathbf{\Psi_{xx}} & \mathbf{\Psi_{xy}} \\ \mathbf{\Psi_{xy}}^{\mathbf{T}} & \mathbf{\Psi_{yy}} \end{pmatrix}, \tag{29}$$

où Ψ_{xx} (respectivement Ψ_{yy}) est la covariance de X (respectivement Y)

4.3.2 Prédiction de la covariance

A partir des équations (28) et (29), il est possible [Brown et al., 1994] d'écrire :

$$Y - m_y =^t \Psi_{xy}\Psi_{xx}^{-1}(X - m_x) + \epsilon \tag{30}$$

où ϵ est un résidu gaussien, indépendant de X. Par conséquent la covariance entre X et Y que nous notons Σ_{xy} est donnée par

$$\begin{aligned}
\Sigma_{xy} &= Cov(^t\Psi_{xy}\Psi_{xx}^{-1}(X - m_x) + \epsilon, X - m_x) &(31)\\
&= Cov(^t\Psi_{xy}\Psi_{xx}^{-1}(X - m_x), X - m_x) &(32)\\
&= {}^t\Psi_{xy}\Psi_{xx}^{-1}Var(X) &(33)
\end{aligned}$$

où $Var(X) = (X - m_x)^t(X - m_x)$ est p fois la matrice de variance mesurée entre les fréquences alléliques des n sites échantillonnés.

Pour pouvoir estimer la covariance entre X et Y, il est nécessaire de définir Ψ. Nous utilisons un modèle de corrélogramme exponentiel. Pour tout $i, j = 1 \ldots n \times n_v$, notons d_{ij} la distance entre le site i et j, alors

$$\Psi_{ij} = \frac{(1 - \alpha) + \alpha e^{-d_{ij}/r} + \lambda 1_{i=j}}{1 + \lambda} \tag{34}$$

où trois paramètres sont associées à cette fonction, appelée corrélogramme. Le paramètre r est un paramètre de portée, de manière similaire aux r_i présentés dans la section 4.2.2. Le paramètre α mesure l'amplitude, qui décrit jusqu'à quel seuil la covariance décroit (Figure 21). Enfin λ est un paramètre qui sert à prévenir d'éventuels problèmes numériques. Dans le cas par exemple où des sites échantillonnés seraient trop proches, la matrice Ψ_{xx} deviendrait numériquement impossible à inverser.

En statistiques spatiales, les différents paramètres du corrélogramme sont habituellement estimés [Clark, 1979]. Dans notre cas, estimer un triplet (α, r, λ) pour définir Ψ n'est pas une option pertinente. L'estimation de la covariance entre X et Y dépend fortement de la valeur choisie (Figure 22). Cela est en partie due à la nature non-stationnaire des données considérées. Il parait impossible de trouver un jeu de paramètres tel que le corrélogramme stationnaire de l'équation (34) s'adapte convenablement au schéma d'isolement par la distance. Si c'était le cas, il ne serait pas nécessaire d'estimer la différenciation génétique locale. Nous choisissons un point de vue bayésien et intégrons les prédictions de Σ_{xy} selon $p(\alpha, r, \lambda|X)$. Pour écrire la loi a posteriori des paramètres, nous utilisons la vraisemblance Wishart comme détaillée par Schmidt et O'Hagan [Schmidt and O'Hagan, 2003]. Nous utilisons ensuite un algorithme MCMC avec des pas Metro-

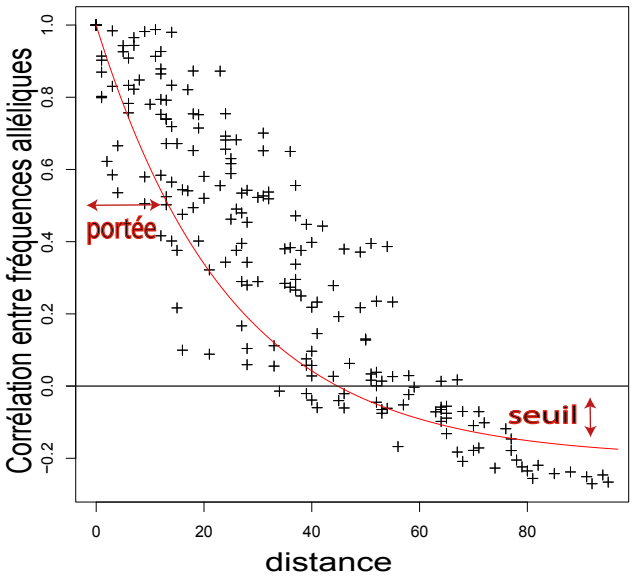


FIGURE 21 – Corrélations entre fréquences alléliques expliquées par la distance. La décroissance est modélisée en utilisant un corrélogramme exponentiel en rouge, ayant une portée (range) et un seuil (sill) estimés.

polis Hastings et d'échantillonneur de Gibbs comme détaillé dans l'article de la section 8.2.

Ce corrélogramme est un modèle standard en krigeage ordinaire. Il suppose que le processus des fréquences alléliques est stationnaire au second ordre. Paradoxalement, la non-stationnarité de l'isolement par la distance est estimée à l'aide d'un modèle stationnaire pour Ψ. En faisant cela, nous supposons que toute la non-stationnarité de Σ_{xy} vient des données, et non pas du modèle paramétrique de Ψ comme dans [Paciorek and Schervish, 2006].

A partir de l'estimateur de l'équation (33), il est possible d'estimer la covariance entre X et Y directement à partir de la matrice de covariance empirique des fréquences alléliques notées $Var(X)$. Ainsi pour chaque population, nous obtenons la covariance à d kilomètres de cette population. En remplaçant la matrice de covariance empirique par tout autre matrice de similarité, nous pouvons construire un estimateur similaire. En particulier, en utilisant la matrice de coefficient $i, j = 1 \dots n, 1 - F_{st}^{ij}$, où F_{st}^{ij} est la différenciation génétique entre les populations i et j sur l'ensemble des marqueurs (section 7.1). Nous pouvons construire l'estimateur de différenciation locale qui correspond à la F_{st} à d kilomètres de chaque population.

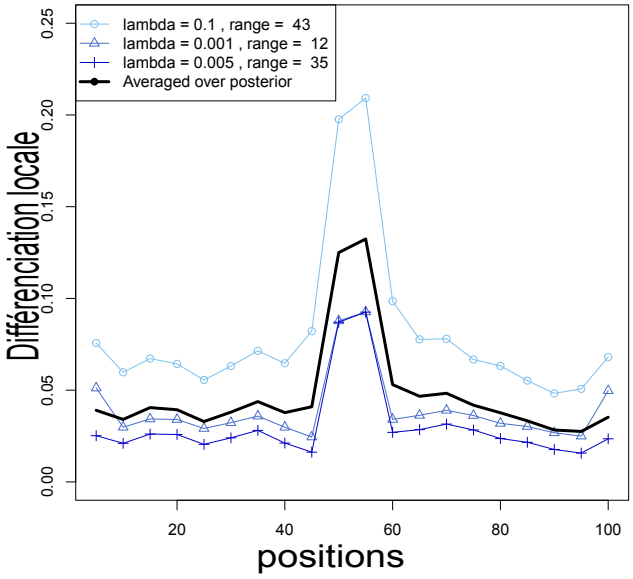


FIGURE 22 – Prédiction de la différentiation génétique locale pour un modèle de stepping stone unidimensionnel. Les migrations entre populations voisines sont constantes, excepté entre les populations 50 et 51, où il n'y a plus de migration depuis plusieurs générations. Les courbes bleues représentent la prédiction de la différenciation génétique locale, ou la corrélation entre fréquences alléliques à une distance $d = 0.1$, pour différents triplets de paramètres (α, r, λ). La courbe noire représente la prédiction de la différenciation génétique locale en intégrant selon la loi à posteriori des paramètres.

4.4 Applications

4.4.1 Exemple sur des simulations

Pour observer les résultats obtenus à l'aide des mesures de différenciation génétique locale, nous créons des cartes de différentiation sur des jeux de données simulées selon différents scénarios. Les simulations sont faites avec le logiciel *ms* [Hudson, 2002] pour 100 populations composées de 1000 individus. Les populations sont disposées sur une grille (10×10) ou chaque population échange un certain nombre de migrants avec les populations voisines. Nous considérons deux scénarios possibles (Figure 23). Dans le premier, deux barrières aux flux de gènes apparaissent, de sorte qu'à partir d'un certain temps, les populations de part et d'autre de cette barrière n'échangent plus de migrants. On s'attend alors à voir la différenciation génétique locale augmenter le long des barrières puisque la corrélation entre les populations de part et d'autre de la barrière va diminuer au fil des générations. Dans un deuxième scénario, nous supposons que les échanges de migrants ne sont pas homogènes sur la grille. Le taux de migration entre populations voisines augmente selon un gradient allant du sud-ouest vers le nord-est. Les taux de migration formant un cline le long de ce gradient, l'attente est de voir la différenciation génétique locale diminuer selon ce gradient. Les cartes obtenues correspondent aux attentes (Figure 24), et permettent de visualiser synthétiquement les variations génétiques dans l'aire de répartition des individus.

4.4.2 Exemple sur des données de plantes alpines

Nous reprenons ici l'exemple de l'espèce *Phyteuma hemisphericum*, échantillonnée sur l'ensemble de l'arc alpin. La carte de différenciation génétique (Figure 25) locale révèle une forte différenciation génétique locale dans la partie ouest des Alpes, et beaucoup plus faible dans la partie Est. Cela est cohérent avec l'analyse du schéma d'isolement par la distance faite dans la section 4.1.2, où les fréquences alléliques des populations à l'ouest sont en moyenne plus corrélées aux populations voisines, que les fréquences alléliques des populations à l'est (Figure 19).

Un schéma de différenciation locale non-stationnaire est courant pour un grand nombre de plantes alpines. Une des explications proposées est la présence de vallées coupant les flux de gènes comme le val d'Aoste ou la vallée de l'Adige [Thiel-Egenter et al., 2011]. Les glaciations du Pléistocène qui ont fragmenté l'aire de répartition des plantes alpines sont une explication. Avec la fin de la période glaciaire, de nombreuses espèces ont re-colonisé la chaîne des Alpes. Un tel processus démographique laisse des traces telles qu'une forte différentiation génétique entre les individus provenant de refuges glaciaires différents [Barton and Hewitt, 1985]. Ainsi, les schémas non-stationnaires pourraient être expliqués par les processus démographiques de re-colonisation

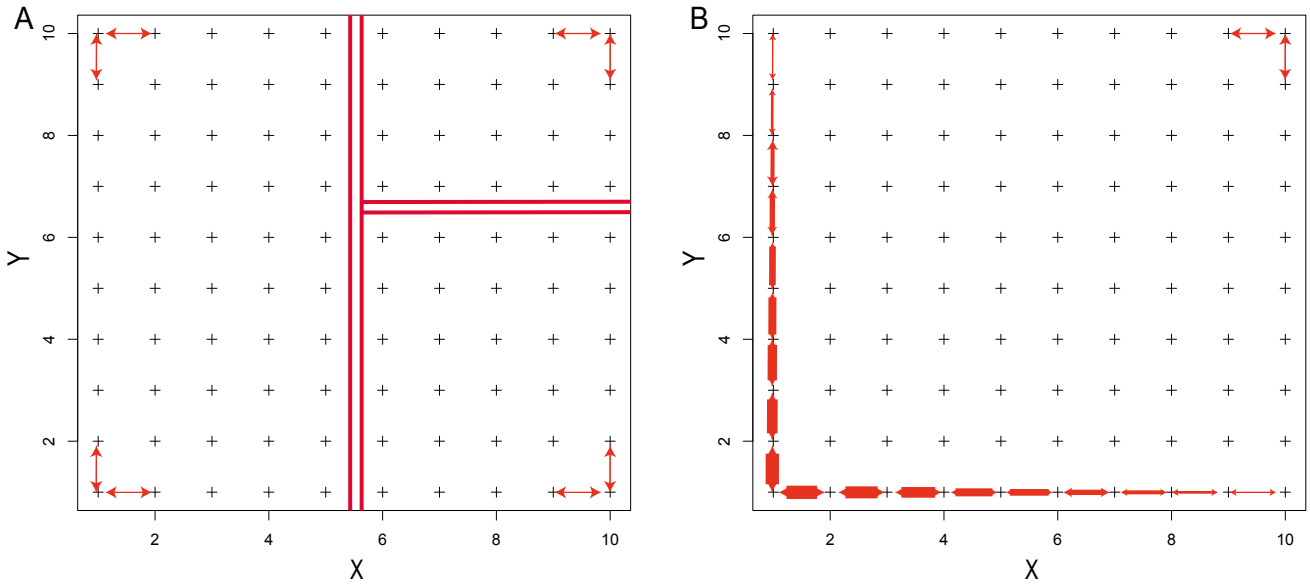


FIGURE 23 – Panel A : Modèle de stepping-stone en 2 dimensions contenant 10×10 populations. Les migrations sont constantes, dans l'ensemble de l'aire de répartition. Tous les flux de migration ne sont pas représentés sur le schéma. Deux barrières bloquent les flux de migrant à $X = 5$ et $Y = 6$. Panel B : Modèle de stepping-stone en 2 dimensions contenant 10×10 populations. Les flux de migrant représentés par des flèches diminuent selon un gradient partant du sud ouest, vers le nord est. Plus les flèches sont fines, plus les migrations sont faibles. Tous les flux de migration ne sont pas représentés sur le schéma.

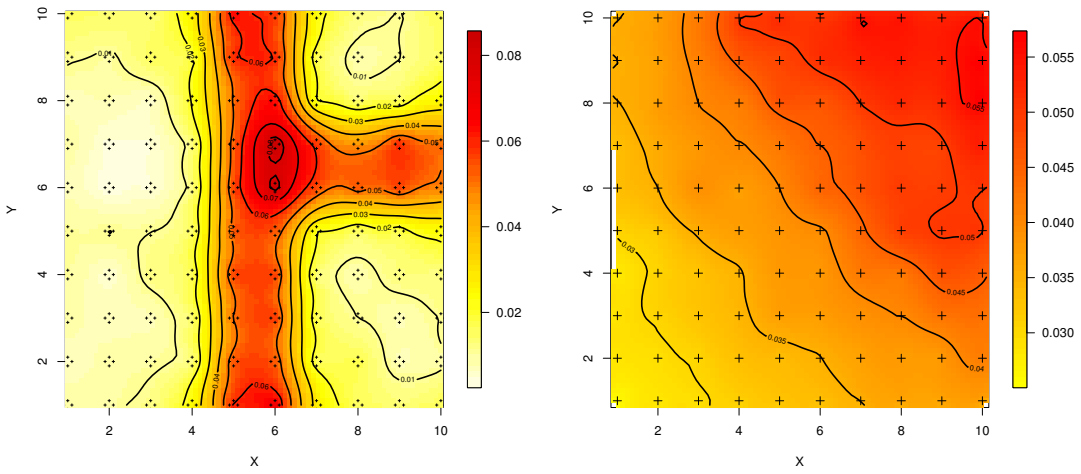


FIGURE 24 – Cartes de différenciation locale pour les deux scénarios présentés Figure 23. Les grandeurs mesurées à l'aide de l'échelle sont des F_{st} à une distance euclidienne de 0.1.

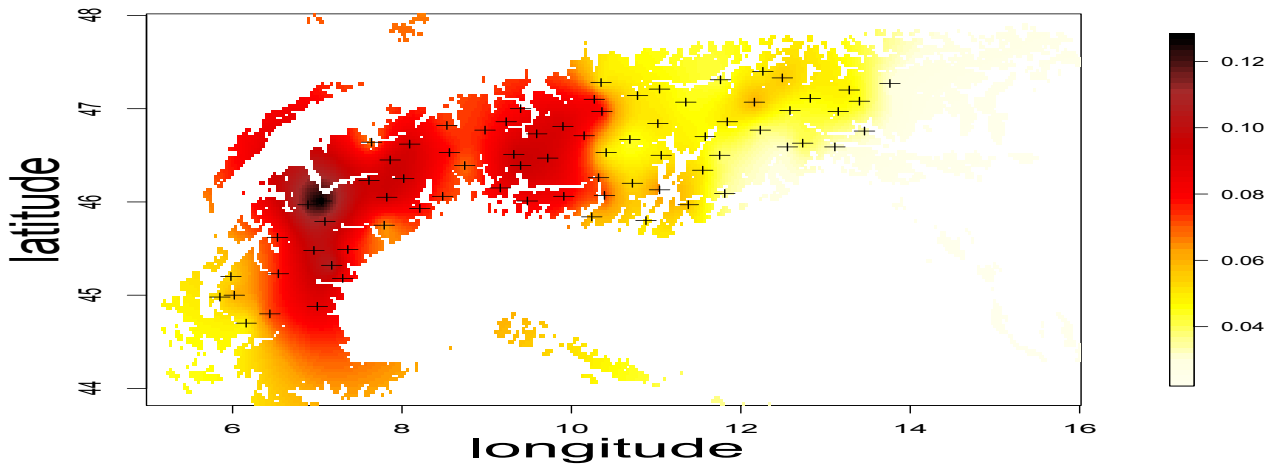


FIGURE 25 – Carte de différenciation génétique locale pour *Phyteuma hemisphericum*. La différenciation est estimée avec la corrélation entre les fréquences alléliques des marqueurs à chaque site. Les voisins sont des populations fictives situées à 8km.

des Alpes [Pawlowski, 1970, Schönswetter et al., 2005].

5 Travaux annexes en génétique des populations

Dans cette section, j'évoque deux projets sur lesquels j'ai travaillé durant ma thèse. Les deux projets sont en rapport avec l'analyse de la structure génétique des populations. Nous présentons d'abord une méthode pour construire des haplotypes permettant de détecter une différenciation qui n'est pas détectée sur les données brutes. Dans un deuxième temps nous présentons une analyse de données en génétique humaine à une échelle géographique fine. Cette exemple est l'occasion d'appliquer la méthode d'estimation de la différenciation génétique locale (section 4).

5.1 Construction d'haplotypes informatifs

Les résultats présentés dans cette section sont présentés dans l'article :
Nicolas Duforet-Frebourg, Lucie M. Gattepaille, Michael G.B. Blum and Mattias Jakobsson. HaploPOP : a software that improves population assignment by combining markers into haplotypes. (en preparation)
L'article se trouve dans la section 8.3

5.1.1 Le gain d'information pour l'assignation

L'inférence de l'ascendance d'un individu, et d'une population dont il est originaire est un problème central en génétique des populations. La caractérisation de la structure génétique à l'aide de données multivariées a souvent été ramenée à l'inférence de coefficients de métissage pour les individus. Ces coefficients de métissage décrivent pour chaque individu la proportion de génome issue de K populations inconnues. Des méthodes pour résoudre ce problème ont été proposées comme celle implémentée dans le logiciel *structure* [Pritchard et al., 2000], ou d'autres plus récentes [Alexander et al., 2009, Bouaziz et al., 2012, Frichot et al., 2014]. Certains exemples en génétique humaine ont montré l'utilité de ces approches, et leur capacité à inférer les populations pour un ensemble d'individus [Rosenberg et al., 2002]. Les marqueurs génétiques contenus dans les données apportent différentes quantités d'information quant à l'origine des individus dans ces K populations. Cette quantité d'information dépend des fréquences alléliques de ce marqueur dans chacune des populations. Cette information peut être quantifiée en utilisant l'entropie de Shannon [Shannon, 1949]. Cette quantité à été dérivée dans le cas d'un marqueur génétique dont la fréquence allélique est observée dans K populations [Rosenberg et al., 2003]. L'information pour l'assignation est une quantité nulle pour un marqueur dont les fréquences alléliques seraient les mêmes dans toutes les populations. En effet un tel marqueur n'aurait aucun pouvoir discriminant entre les populations, et n'apporte aucune information pour assigner un individu à une population.

Au contraire, un marqueur dont un allèle serait exclusif à une population aurait un grand pouvoir discriminant.

Les marqueurs génétiques d'un jeu de données apportent une certaine quantité d'information, au sens de [Rosenberg et al., 2003] pour assigner les individus à des populations sources. Il a été prouvé de plus que combiner ces marqueurs en haplotype peut accroître la quantité d'information dans les données [Gattepaille and Jakobsson, 2012]. Nous appellerons un haplotype toute combinaison de marqueurs des données. L'idée importante est que la présence de déséquilibre de liaison fait que les états alléliques des marqueurs d'une même région du génome vont être corrélés. Ainsi, l'information de deux marqueurs combinés n'est pas la somme des informations des deux marqueurs. Elle peut être plus grande ou plus faible. C'est pour savoir s'il est plus informatif de combiner deux marqueurs que [Gattepaille and Jakobsson, 2012] ont introduit la statistique de gain d'information pour l'assignation. Lorsque cette statistique est positive pour une paire de marqueurs, il est plus informatif de considérer l'haplotype résultant de la combinaison des deux marqueurs.

5.1.2 Un algorithme de combinaison

Lorsque j'ai commencé à travailler sur ce projet, en stage de master à l'université d'Uppsala en Suède, la statistique de gain avait déjà été introduite. Le problème était de transformer un jeu de données de p marqueurs en un jeu d'haplotypes informatifs en utilisant cette statistique de gain. Nous avons vu précédemment que cette statistique permet de dire s'il est plus informatif de considérer deux marqueurs indépendamment, ou de considérer l'haplotype résultant de leur combinaison. Cela requiert de connaître les K populations présentes dans les données. Néanmoins, s'il est possible de savoir s'il est plus informatif de combiner deux marqueurs entre eux pour accroître la quantité d'information dans les données, le problème de transformer un jeu de données entier en un jeu d'haplotypes dont l'information pour l'assignation aux populations sources serait optimale est un problème de combinatoire délicat. Le problème de regrouper les p marqueurs en haplotypes revient à trouver la meilleure partition de ces marqueurs au sens de [Rosenberg et al., 2003]. Le nombre de partitions d'un ensemble est un nombre qui rend impossible l'exploration de toutes les solutions [Rota, 1964].

Nous avons donc proposé un algorithme explorant les combinaisons de marqueurs en haplotypes en un temps polynomial. Nous proposons pour cela un algorithme glouton par région du génome, où la taille des régions est au choix de l'utilisateur. Le coût de cet algorithme augmente proportionnellement au cube du nombre de marqueurs dans une région (section 8.3). Nous démontrons l'efficacité de cet algorithme sur une partie du jeu de données POPRES [Nelson et al., 2008]. En particulier, nous démontrons qu'une analyse en composante principale ne permet pas de distin-

guer les individus des populations portugaises et espagnoles avec les marqueurs du jeu de données. Lorsque les haplotypes sont construits selon l'algorithme proposé, l'assignation de ces individus est amélioré. L'algorithme proposé est implémenté dans le logiciel *HaploPop*.

5.2 Structure génétique de la population de l'Ouest de la France

Les résultats de cette section font l'objet d'une soumission en révisions mineures dans European journal of Human genetics :
Matilde Karakachoff, Nicolas Duforet-Frebourg, et al. (2014), Fine scale human genetic structure in Western France. European journal of Human genetics, (minor revision).
L'article se trouve dans la section 8.4.

5.2.1 Données du projet DESIR-CavGens

Nous analysons les données génétiques échantillonnées sur un ensemble de 1684 individus. Les individus proviennent de deux études précédemment réalisées, les études D.E.S.I.R. et CavGens qui ont été menées dans les Pays de la Loire. Les individus sont génotypés sur une puce Illumina HumanHap300, contenant plus de $300\ 000$ marqueurs de type SNP. En plus des données génétiques, une information sur le lieu de naissance de chaque individu est disponible. Pour supposer que les données génétiques d'un individu sont corrélées à son lieu de naissance, il est préférable de contrôler que les quatre grand-parents sont bien nés dans un lieu proche. Néanmoins, nous ferons l'hypothèse que ces lieux de naissance sont informatifs. Les individus couvrent l'ensemble de la France, mais sont essentiellement originaires des régions de l'Ouest : Pays de la Loire, Bretagne, Normandie et Poitou-Charentes (Figure 26). Nous essayons de démontrer que même à une échelle géographique si fine, les variations génétiques humaines ne sont pas aléatoires et sont structurées. Une telle preuve de structure à une échelle fine impliquerait la présence de variants rares [Kristiansson et al., 2008]. Ces variants ont un intérêt médical car ils peuvent participer au déterminisme génétique de certaines pathologies [Cirulli and Goldstein, 2010, McClellan and King, 2010].

5.2.2 Analyse de structure

Depuis le travail initial de [Menozzi et al., 1978], il a été démontré qu'à l'échelle de l'Europe ou de n'importe quel autre continent, la variation génétique est structurée chez les humains [Novembre et al., 2008, Jay et al., 2011]. De même à l'intérieur des pays européens, les preuves de structure de populations sont nombreuses [T O'Dushlaine et al., 2010, Salmela et al., 2011, Humphreys et al., 201
Il est donc envisageable d'observer de la structure à l'échelle d'une partie de la France. Le contexte

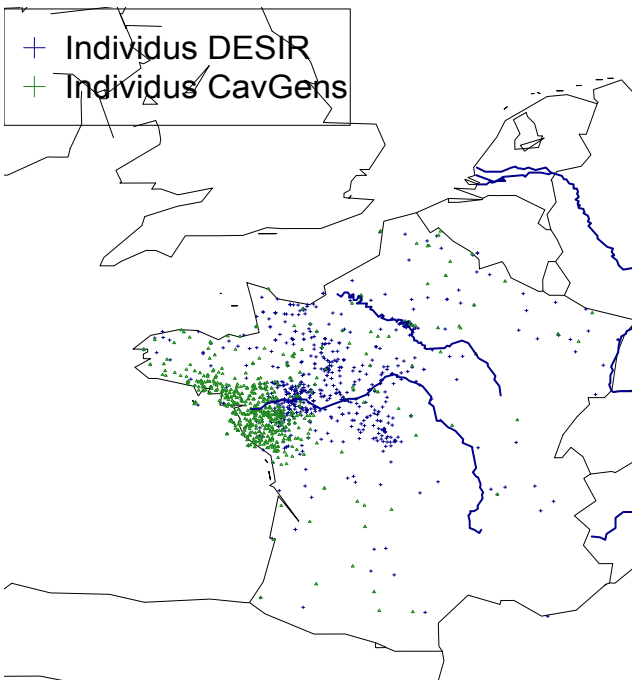


FIGURE 26 – Lieux de naissances des individus échantillonnés dans les jeux de données RAC et DESIR. L'échantillonnage couvre essentiellement les régions de l'ouest de la France.

spatial des données nous permet de chercher des traces du schéma d'isolement par la distance. Nous appliquons la méthode développée dans la section 4, et nous pouvons observer que la différenciation génétique locale est plus forte en Bretagne et en Vendée (Figure 27). Nous observons de plus un accroissement du déséquilibre de liaison moyen en Bretagne. Ce résultat est attendu, et reflète la taille efficace de population plus faible à une extrémité d'une aire de répartition [Wilkins and Wakeley, 2002]. D'autres méthodes classiques telles que l'analyse en composantes principales [Patterson et al., 2006] et l'inférence de coefficients d'ascendance avec *Admixture* [Alexander et al., 2009] confirment la présence de structure spatiale (Figure 28). De plus, nous fusionnons les données CavGens et DESIR avec le jeu de données POPRES [Nelson et al., 2008]. Nous retrouvons le résultat établissant que la structure de population est corrélée à la géographie à l'aide d'une analyse en composantes principales [Novembre et al., 2008]. Nous observons de plus que les individus nés dans les départements bretons se distinguent du reste du jeu de données par leur proximité avec les individus irlandais du jeu de données POPRES. Cette proximité est mesurée à l'aide de mesure d'identité par état allélique [Purcell et al., 2007], et de F_{st} sur des régions de 50 marqueurs consécutifs. Ce résultat peut refléter une ascendance commune récente entre bretons et irlandais.

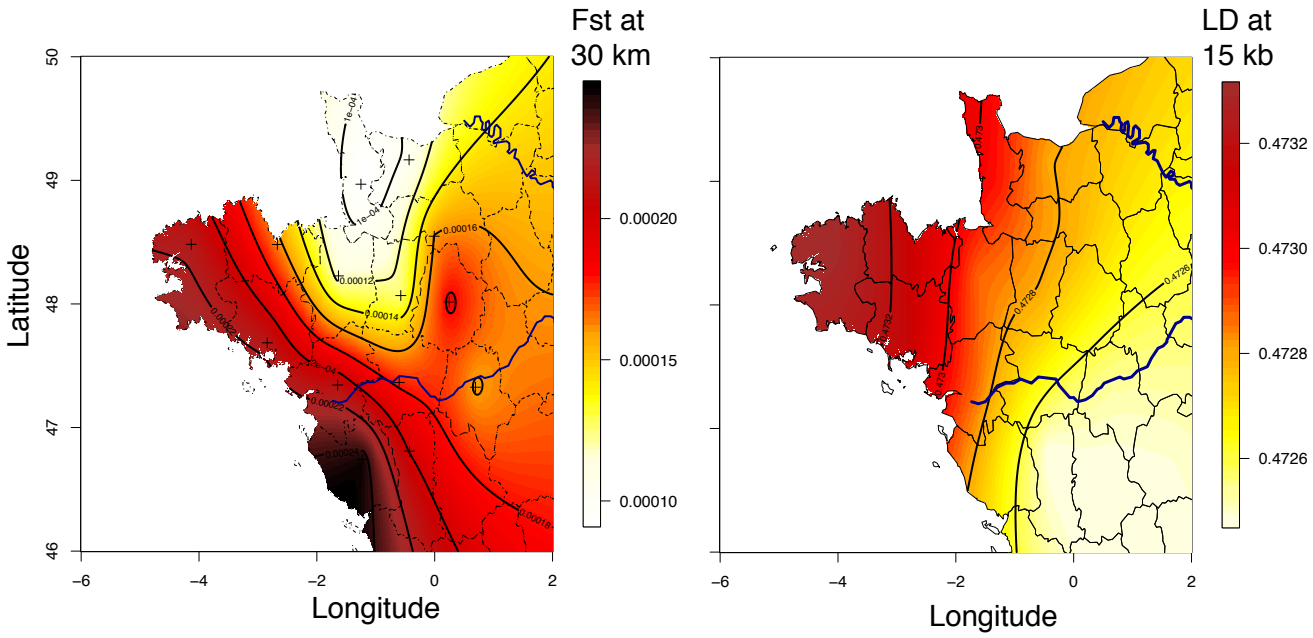


FIGURE 27 – Panel A : Carte de différenciation génétique locale des individus des jeux de données DESIR et CavGens. Les estimateurs de différenciation génétique locale sont des F_{st} à 30 kilomètres de chaque département. Panel B : Carte du déséquilibre de liaison moyen à 15kb par département.

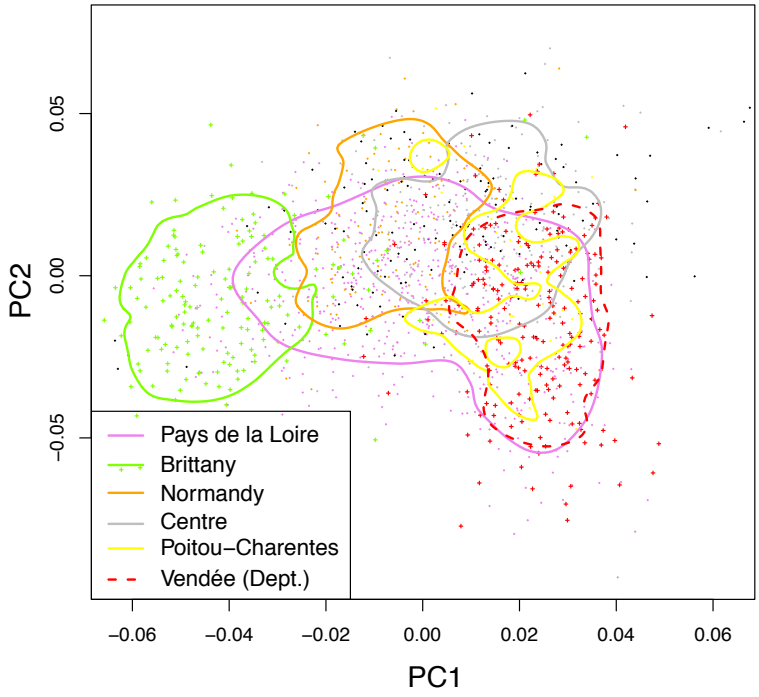


FIGURE 28 – Première et deuxième composante principale des individus des jeux de données DESIR et CavGens. Les individus sont coloriés en fonction de leur appartenance aux 5 régions les plus représentées dans les données. Les enveloppes contiennent 75% des individus d'une région. Les individus du département vendéen sont en rouge, pour être mis en évidence bien qu'ils fassent partie de la région Pays de la Loire.

6 Conclusions et perspectives

6.1 Conclusions

Dans ce manuscrit, nous avons présenté les enjeux de la génétique des populations. Ces enjeux regroupent un grand nombre d'applications possibles, incluant la reconstruction de la démographie des espèces et la détection de processus adaptatifs. La génétique des populations est une science qui aujourd'hui est transformée fondamentalement par le changement dans la nature et la quantité des données. Une part importante des efforts de recherche en génétique depuis le début des années 2000 concerne le traitement informatique et statistique des données massives. Le travail de cette thèse s'inscrit dans cet effort de modélisation statistique pour répondre aux enjeux de la génétique des populations, et de la biologie de l'évolution. Pour cela nous avons adopté un point de vue phénoménologique pour la modélisation statistique. Nous détacher des approches classiques souvent mécanistes est un moyen de proposer des méthodes innovantes qui montrent un taux de fausses découvertes plus faibles sur plusieurs jeux de données. Ces modèles et logiciels s'incorporent dans la panoplie des méthodes disponibles pour le traitement de données génétiques en grande dimension.

Une partie importante de mon travail de thèse est de proposer un modèle innovant et non supervisé pour détecter les marqueurs génétiques impliqués dans le processus d'adaptation locale. L'état de l'art dans le domaine de la détection de l'adaptation locale a été fait dans la section 2. Cet état de l'art est partiel, mais rappelle que l'approche classique, proposée par Lewontin et Krackauer [Lewontin and Krakauer, 1973], a posé les fondements des méthodes basées sur les scans à F_{st}. Par la suite de nombreux modèles ont développé l'idée de scanner le génome à l'aide de statistiques F [Beaumont and Nichols, 1996, Vitalis et al., 2003, Beaumont and Balding, 2004, Antao et al., 2008, Foll and Gaggiotti, 2008, Bonhomme et al., 2010, Fariello et al., 2013]. Nous n'avons pas détaillé toutes les méthodes, et nous sommes concentrés sur deux d'entre elles, implémentées dans les logiciels *Bayescan* et *fdist*. Ces deux méthodes nous servent de référence pour assurer les bons résultats du modèle que nous présentons dans la section 3. D'autres approches existent pour détecter l'adaptation locale basées sur des méthodes de corrélation entre données génétiques et variables environnementales [Coop et al., 2010, Frichot et al., 2013]. Le modèle que nous proposons pour détecter l'adaptation locale est basé sur une analyse factorielle, ce qui lui permet d'apprendre la structure de population, et de calculer des facteurs de Bayes pour chaque marqueur. Nous démontrons l'utilité de ce modèle sur des données réelles et simulées (section 8.1). Une approche rapide de la méthode a également été proposée. Cette autre méthode a l'avantage de passer à l'échelle des grands jeux de données, mais n'est pas robuste au choix du nombre de facteurs K. L'algorithme détaillé dans la section 7.2, ainsi que la version rapide sont implémentés dans le logiciel *PCAdapt*.

Une autre question adressée dans cette thèse est l'étude des variations génétiques dans l'espace. Il a été montré que certains processus évolutionnaires complexes peuvent entraîner des schémas d'isolement par la distance non-stationnaires. Pour quantifier la non-stationnarité résultante nous proposons un modèle de krigeage bayésien et introduisons la différenciation génétique locale. Cette mesure permet d'observer des schémas de non-stationnarité pour des données simulées selon un scénario complexe ou sur des données réelles dans la section 8.2. L'algorithme pour estimer la différenciation génétique locale est implémenté dans le logiciel *LocalDiff*. Un test d'hypothèse de non-stationnarité est présenté dans l'article de la section 8.2. Cette procédure teste l'hypothèse nulle de stationnarité du schéma d'isolement par la distance, sachant un schéma d'échantillonnage donné. Nous montrons par la suite sa capacité à détecter de la structure de population à une échelle géographique fine.

D'autres travaux sont présentés dans ce manuscrit et font partie du travail de thèse. Un logiciel implémentant une stratégie de construction d'haplotypes informatifs au sens de [Rosenberg et al., 2003] a été développé. Une analyse fine de la structure génétique dans l'ouest de la France a également été réalisée, et montre que les variations génétiques dans les régions de l'Ouest sont structurées. En particulier, il a été montré que la différenciation génétique à l'intérieur de la région Pays de la Loire est plus forte entre la Vendée et les autres départements. Un autre résultat intéressant est la proximité génétique entre les individus échantillonnés en Bretagne et les individus irlandais du jeu de données POPRES impliquant une ascendance commune entre les irlandais et les bretons des données DESIR et CavGens. Cette analyse de données a également été l'occasion d'utiliser la mesure de différenciation génétique locale (section 4).

6.2 Perspectives

Comme nous l'avons mentionné dans la conclusion, de nombreuses méthodes existent pour détecter les gènes impliqués dans l'adaptation locale. Les méthodes peuvent être basées sur des modèles à tendance mécanistes ou à tendance phénoménologique. Il n'y a pas une meilleure option entre les deux approches. Un grand nombre de modèles mécanistes ont été développés pour étendre le test de Lewontin-Krackauer, et ont prouvé leur utilité. Nous montrons qu'une approche phénoménologique peut diminuer le nombre de fausses découvertes. Les approches corrélatives dont nous avons parlé dans la conclusion cherche des corrélations entre données génétiques et variables environnementales. Elles peuvent être mécanistes ou phénoménologiques, et elles permettent de détecter des régions du génome fortement corrélées à l'environnement [Villemereuil et al., 2014]. Toute cette panoplie de méthodes donne un large choix quant à l'approche à adopter pour détecter des régions du génome impliquées dans des processus d'adaptation. Il est intéressant de voir que

pour un même jeu de données, les résultats sont rarement identiques pour deux méthodes différentes. Une approche intéressante serait de ne plus se restreindre à une méthode, mais d'observer les résultats de toutes les méthodes et de chercher des résultats récurrents. Cela serait un moyen d'intégrer les résultats sur les différentes méthodes, et ainsi de réduire la quantité de faux positifs [Grossman et al., 2013].

Nous avons présenté les possibilités de modélisation avec une approche bayésienne dans la section 3.2.3. Cette flexibilité offre la possibilité de modifier le modèle de la section 3 pour inclure d'autres informations. Une extension possible serait de modifier la structure de covariance des résidus du modèle (7). Comme mentionné dans la section 4, les génotypes des individus d'un jeu de données sont corrélés à la distance géographique séparant les individus. Cette corrélation spatiale apparaît naturellement dans les facteurs de l'analyse en composantes principales [Novembre and Stephens, 2008]. Cette corrélation peut avoir pour conséquence une inflation des facteurs de Bayes, et causer de faux positifs. Un moyen de prendre cela en compte serait de supposer la matrice Ψ non diagonale, de la même façon que dans la section 4.2.2, et l'équation (34). Une approche similaire a été proposée par [Frichot et al., 2012]. Une autre extension possible du modèle serait une autre définition de la notion d'outlier. Il serait possible de dire qu'un marqueur peut être outlier dans plusieurs directions. Un marqueur pourrait être atypique dans les directions (k_1, k_2) au lieu de k_1 ou k_2. Cela rajouterait de nombreuses classes possibles pour les marqueurs. En effet il y aurait 2^K classes possibles au lieu de $K + 1$. Néanmoins, la valeur de K étant faible en pratique, la combinatoire ne serait pas un vrai problème. Une telle approche pourrait permettre d'accroître les facteurs de Bayes des marqueurs faiblement corrélés à plusieurs composantes. Ainsi, cette extension pourrait permettre de détecter des cas d'évolution parallèle où des individus appartenant à des populations différentes voient le même allèle favorisé [Foll et al., 2014].

Cependant, avec l'accroissement de la taille des jeux de données, il est nécessaire de travailler avec des algorithmes performants. Les approches MCMC ont du mal à passer à l'échelle des données génomiques de nouvelle génération. Une extension intéressante serait d'utiliser des méthodes récentes d'inférence bayésienne comme les méthodes d'inférence variationnelle [Jaakkola and Jordan, 2000] pour estimer les paramètres du modèle. Une telle extension permettrait de conserver les avantages de la modélisation bayésienne, et d'estimer les paramètres du modèle avec un algorithme plus rapide qu'un algorithme MCMC. Le logiciel *structure* utilisant un algorithme MCMC a récemment été amélioré avec l'utilisation de telles méthodes [Raj et al., 2014].

Une approche alternative à été proposée dans la section 3.3 pour les jeux de données à grande échelle. Néanmoins cette solution est moins performante que le modèle bayésien original, en particulier quant à la robustesse des résultats au choix de K. Pour prévenir l'augmentation du taux de

fausses découvertes avec un mauvais choix de K, il serait intéressant de développer une approche permettant d'obtenir une quantification du taux de faux positifs. L'un des désavantages des statistiques S_1, S_2 et S_3, est qu'elles donnent un classement des marqueurs, et une liste de candidat, mais pas de mesure de significativité pour ces candidats.

Un test d'hypothèse peut être envisagé pour donner une significativité à chaque marqueur, et estimer un taux de fausses découvertes [Storey, 2002]. Une idée serait d'utiliser les tests de permutations. Ce test consisterait à prendre un nombre m de marqueurs $(m << p)$, et de permuter aléatoirement chacun de ces marqueurs [Chung and Storey, 2013]. Une analyse en composante principale sur les données, incluant les marqueurs permutés, permettrait alors de calculer les statistiques sur ces marqueurs, supposés être neutres. répéter cette opération de nombreuses fois donnerait une distribution des statistiques selon l'hypothèse de neutralité, et permettrait de calculer une p-valeur pour chaque marqueur des données. Néanmoins, une telle approche n'assure pas un faible taux de fausses découvertes. l'hypothèse nulle n'est pas strictement la neutralité sélective du marqueur, mais plutôt une corrélation nulle du marqueur avec une composante principale. En génétique des populations, un grand nombre de marqueurs neutres sont faiblement corrélés à la structure représentée par chaque composante. Ainsi toutes ces variables auraient une statistique naturellement plus élevée que les statistiques des variables permutées alors qu'elle n'ont pas été impliquées dans des processus adaptatifs.

En conclusion, les modèles phénoménologiques ont un grand potentiel pour la génétique des populations, et en particulier la détection de processus adaptatifs. L'analyse factorielle permet d'apprendre conjointement la structure génétique de population et d'identifier des marqueurs candidats à l'adaptation locale. La difficulté dans l'utilisation de modèles phénoménologiques est la capacité d'interprétation des résultats du modèle. Cependant, des efforts de compréhension ont été fait en particulier pour l'interprétation de l'analyse en composantes principales en génétique des populations [McVean, 2009, Engelhardt and Stephens, 2010]. Une fois la difficulté de l'interprétation maîtrisée, la significativité des résultats est une autre difficulté. Il est possible d'estimer des facteurs de Bayes ou des p-valeurs à l'aide de permutations par exemple. Néanmoins, ces valeurs donnent une significativité par rapport à un modèle nul qui n'est pas explicitement l'hypothèse de neutralité. L'hypothèse nulle va en général être une absence de corrélation entre deux paramètres, mais une corrélation pourrait cacher d'autres processus que de l'adaptation biologique.

Les possibilités d'extensions sont très nombreuses, et des modélisations plus phénoménologiques, en plus des approches mécanistes classiques, peuvent apporter de l'information complémentaire en biologie de l'évolution.

Références

[Alexander et al., 2009] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9) :1655–1664.

[Anderson et al., 1999] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammerling, S., McKenney, A., et al. (1999). *LAPACK Users' guide*, volume 9. Siam.

[Antao et al., 2008] Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., and Luikart, G. (2008). Lositan : a workbench to detect molecular adaptation based on a fst-outlier method. *BMC bioinformatics*, 9(1) :323.

[Arnaud, 2003] Arnaud, J. (2003). Metapopulation genetic structure and migration pathways in the land snail *helix aspersa* : influence of landscape heterogeneity. *Landscape Ecology*, 18(3) :333–346.

[Balding, 2003] Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology*, 63(3) :221–230.

[Barbujani et al., 1989] Barbujani, G., Oden, N., and Sokal, R. (1989). Detecting regions of abrupt change in maps of biological variables. *Systematic Biology*, 38(4) :376–389.

[Bartholomew et al., 2011] Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis : a unified approach*, volume 899. John Wiley & Sons.

[Barton and Hewitt, 1985] Barton, N. H. and Hewitt, G. (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*, pages 113–148.

[Basilevsky, 2009] Basilevsky, A. T. (2009). *Statistical factor analysis and related methods : theory and applications*, volume 418. John Wiley & Sons.

[Beaumont and Balding, 2004] Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, 13(4) :969–980.

[Beaumont and Nichols, 1996] Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 263(1377) :1619–1626.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

[Bonhomme et al., 2010] Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and SanCristobal, M. (2010). Detecting selection in population trees : the lewontin and krakauer test extended. *Genetics*, 186(1) :241–262.

[Bouaziz et al., 2012] Bouaziz, M., Paccard, C., Guedj, M., and Ambroise, C. (2012). Ships : Spectral hierarchical clustering for the inference of population structure in genetic studies. *PloS one*, 7(10) :e45685.

[Box and Tiao, 1968] Box, G. E. and Tiao, G. C. (1968). A bayesian approach to some outlier problems. *Biometrika*, 55(1) :119–129.

[Brown et al., 1994] Brown, P., Le, N., and Zidek, J. (1994). Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, 22(4) :489–509.

[Canfield et al., 2013] Canfield, V. A., Berg, A., Peckins, S., Wentzel, S. M., Ang, K. C., Oppenheimer, S., and Cheng, K. C. (2013). Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3 : Genes| Genomes| Genetics*, 3(11) :2059–2067.

[Cann et al., 2002] Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. *Science (New York, NY)*, 296(5566) :261.

[Castella et al., 2000] Castella, V., Ruedi, M., Excoffier, L., Ibanez, C., Arlettaz, R., and Hausser, J. (2000). Is the Gibraltar Strait a barrier to gene flow for the bat Myotis myotis (Chiroptera : Vespertilionidae) ? *Molecular Ecology*, 9(11) :1761–1772.

[Cavalli-Sforza, 2011] Cavalli-Sforza, L. (2011). *Aventure de l'espèce humaine (L') : De la génétique des populations à l'évolution culturelle*. Odile Jacob.

[Celeux et al., 2003] Celeux, G., Forbes, F., and Peyrard, N. (2003). Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern recognition*, 36(1) :131–144.

[Cercueil et al., 2007] Cercueil, A., François, O., and Manel, S. (2007). The genetical bandwidth mapping : a spatial and graphical representation of population genetic structure based on the Wombling method. *Theoretical Population Biology*, 71(3) :332–341.

[Chaloner and Brant, 1988] Chaloner, K. and Brant, R. (1988). A bayesian approach to outlier detection and residual analysis. *Biometrika*, 75(4) :651–659.

[Chesser, 1983] Chesser, R. K. (1983). Genetic variability within and among populations of the black-tailed prairie dog. *Evolution*, pages 320–331.

[Chung and Storey, 2013] Chung, N. C. and Storey, J. D. (2013). Statistical significance of variables driving systematic variation. *arXiv preprint arXiv :1308.6013*.

[Cirulli and Goldstein, 2010] Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6) :415–425.

[Clark, 1979] Clark, I. (1979). *Practical geostatistics*, volume 3. Applied Science Publishers London.

[Consortium, 2012] Consortium, . G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65.

[Coop et al., 2010] Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4) :1411–1423.

[Core, 2012] Core, R. (2012). R : A language and environment for statistical computing.

[Cressie, 1993] Cressie, N. (1993). Statistics for spatial data : Wiley series in probability and statistics.

[Crow et al., 1970] Crow, J. F., Kimura, M., et al. (1970). An introduction to population genetics theory. *An introduction to population genetics theory.*

[Ding et al., 2011] Ding, X., He, L., and Carin, L. (2011). Bayesian robust principal component analysis. *Image Processing, IEEE Transactions on*, 20(12) :3419–3430.

[Engelhardt and Stephens, 2010] Engelhardt, B. E. and Stephens, M. (2010). Analysis of population structure : a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics*, 6(9) :e1001117.

[Epps et al., 2005] Epps, C., Palsbøll, P., Wehausen, J., Roderick, G., Ramey II, R., and McCullough, D. (2005). Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. *Ecology Letters*, 8(10) :1029–1038.

[Fagny et al., 2014] Fagny, M., Patin, E., Enard, D., Barreiro, L. B., Quintana-Murci, L., and Laval, G. (2014). Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular biology and evolution*, page msu118.

[Fariello et al., 2013] Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3) :929–941.

[Filzmoser et al., 2008] Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3) :1694–1711.

[Foll and Gaggiotti, 2008] Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers : a bayesian perspective. *Genetics*, 180(2) :977–993.

[Foll et al., 2014] Foll, M., Gaggiotti, O. E., Daub, J. T., and Excoffier, L. (2014). Hierarchical bayesian model of population structure reveals convergent adaptation to high altitude in human populations. *arXiv preprint arXiv :1402.4348*.

[Frichot et al., 2014] Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4) :973–983.

[Frichot et al., 2012] Frichot, E., Schoville, S., Bouchard, G., and François, O. (2012). Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in genetics*, 3.

[Frichot et al., 2013] Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution*, 30(7) :1687–1699.

[Gaggiotti, 2013] Gaggiotti, O. (2013). Statistical software and methods for population genetics. *Oral communication*.

[Gattepaille and Jakobsson, 2012] Gattepaille, L. M. and Jakobsson, M. (2012). Combining markers into haplotypes can improve population structure inference. *Genetics*, 190(1) :159–174.

[Ghahramani et al., 1996] Ghahramani, Z., Hinton, G. E., et al. (1996). The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

[Gillespie, 2010] Gillespie, J. H. (2010). *Population genetics : a concise guide*. JHU Press.

[Green et al., 2010] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A draft sequence of the neandertal genome. *science*, 328(5979) :710–722.

[Grossman et al., 2013] Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., et al. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4) :703–713.

[Gugerli et al., 2008] Gugerli, F., Englisch, T., Niklfeld, H., Tribsch, A., Mirek, Z., Ronikier, M., Zimmermann, N., Holderegger, R., and Taberlet, P. (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(4) :259–281.

[Hammer et al., 1998] Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C., Templeton, A. R., and Zegura, S. L. (1998). Out of africa and back again : nested cladistic analysis of human y chromosome variation. *Molecular Biology and Evolution*, 15(4) :427–441.

[Hardy et al., 2006] Hardy, O. J., Maggia, L., Bandou, E., Breyne, P., Caron, H., Chevallier, M. H., Doligez, A., Dutech, C., Kremer, A., Latouche-hallé, C., et al. (2006). Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species. *Molecular Ecology*, 15(2) :559–571.

[Harman, 1960] Harman, H. H. (1960). Modern factor analysis.

[Hartl and Clark, 1997] Hartl, D. L. and Clark, A. G. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland.

[Hein et al., 2004] Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution : a primer in coalescent theory*. Oxford university press.

[Hellberg, 2009] Hellberg, M. (2009). Gene flow and isolation among populations of marine animals. *Annual Review of Ecology and Systematics*, 40 :291–310.

[Hernandez, 2008] Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23) :2786–2787.

[Hey, 2010] Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular biology and evolution*, 27(4) :905–920.

[Hiebert and Hamrick, 1983] Hiebert, R. D. and Hamrick, J. L. (1983). Patterns and levels of genetic variation in great basin bristlecone pine, pinus longaeva. *Evolution*, pages 302–310.

[Higdon et al., 1999] Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian statistics*, 6(1) :761–768.

[Hilborn, 1997] Hilborn, R. (1997). a., and m. mangel. 1997. the ecological detective : Confronting models with data.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

[Holderegger et al., 2006] Holderegger, R., Kamm, U., and Gugerli, F. (2006). Adaptive vs. neutral genetic diversity : implications for landscape genetics. *Landscape Ecology*, 21(6) :797–807.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417.

[Hudson, 1998] Hudson, R. (1998). Island models and the coalescent process. *Molecular Ecology*, 7(4) :413–418.

[Hudson, 2002] Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2) :337–338.

[Humphreys et al., 2011] Humphreys, K., Grankvist, A., Leu, M., Hall, P., Liu, J., Ripatti, S., Rehnström, K., Groop, L., Klareskog, L., Ding, B., et al. (2011). The genetic structure of the Swedish population. *PLOS ONE*, 6(8) :e22547.

[Ising, 1925] Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1) :253–258.

[Jaakkola and Jordan, 2000] Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1) :25–37.

[Jackson and Chen, 2004] Jackson, D. A. and Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics*, 15(2) :129–139.

[Jakobsson et al., 2013] Jakobsson, M., Edge, M. D., and Rosenberg, N. A. (2013). The relationship between fst and the frequency of the most frequent allele. *Genetics*, 193(2) :515–528.

[Jakobsson et al., 2008] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181) :998–1003.

[Jay et al., 2011] Jay, F., François, O., and Blum, M. G. B. (2011). Predictions of native American population structure using linguistic covariates in a hidden regression framework. *PLOS ONE*, 6(1) :e16227.

[Jeffreys, 1998] Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

[Jolliffe, 2005] Jolliffe, I. (2005). *Principal component analysis*. Wiley Online Library.

[Kamberov et al., 2013] Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., et al. (2013). Modeling recent human evolution in mice by expression of a selected edar variant. *Cell*, 152(4) :691–702.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430) :773–795.

[Kawecki and Ebert, 2004] Kawecki, T. J. and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology letters*, 7(12) :1225–1241.

[Kimura, 1953] Kimura, M. (1953). "stepping stone" model of population. *Ann. Rept. Nat. Inst. Genetics, Japan*, 3 :62–63.

[Kimura and Weiss, 1964] Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4) :561.

[Kingman, 1982] Kingman, J. F. (1982). On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43.

[Krimbas and Tsakas, 1971] Krimbas, C. B. and Tsakas, S. (1971). The genetics of dacus oleae. v. changes of esterase polymorphism in a natural population following insecticide control-selection or drift ? *Evolution*, pages 454–460.

[Kristiansson et al., 2008] Kristiansson, K., Naukkarinen, J., Peltonen, L., et al. (2008). Isolated populations and complex disease gene identification. *Genome Biol*, 9(8) :109.

[Laloë et al., 2011] Laloë, D., Gautier, M., et al. (2011). On the genetic interpretation of between-group pca on snp data.

[Latter, 1973] Latter, B. (1973). The island model of population differentiation : a general solution. *Genetics*, 73(1) :147–157.

[Levins, 1966] Levins, R. (1966). The strategy of model building in population biology. *American scientist*, 54(4) :421–431.

[Lewontin and Krakauer, 1973] Lewontin, R. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1) :175–195.

[Li et al., 2008] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866) :1100–1104.

[Luikart et al., 2003] Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics : from genotyping to genome typing. *Nature Reviews Genetics*, 4(12) :981–994.

[Malécot et al., 1948] Malécot, G. et al. (1948). Mathematics of heredity. *Les mathematiques de l'heredite.*

[Manel and Holderegger, 2013] Manel, S. and Holderegger, R. (2013). Ten years of landscape genetics. *Trends in ecology & evolution*, 28(10) :614–621.

[Manel et al., 2003] Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics : combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4) :189–197.

[Manni et al., 2004] Manni, F., Guerard, E., Heyer, E., et al. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation : how barriers can be detected by using Monmonier's algorithm. *Human Biology*, 76(2) :173–190.

[Mardia et al., 1979] Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate analysis. *Probability and Mathematical Statistics, London : Academic Press, 1979*, 1.

[Marlin, 2003] Marlin, B. M. (2003). Modeling user rating profiles for collaborative filtering. In *NIPS*.

[Matheron, 1962] Matheron, G. (1962). *Traité de géostatistique appliquée. 1 (1962)*, volume 1. Editions Technip.

[McBratney and Webster, 1986] McBratney, A. and Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, 37(4) :617–639.

[McClellan and King, 2010] McClellan, J. and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, 141(2) :210–217.

[McRae and Beier, 2007] McRae, B. and Beier, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, 104 :19885–19890.

[McVean, 2009] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10) :e1000686.

[Menozzi et al., 1978] Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358) :786–792.

[Michalakis and Excoffier, 1996] Michalakis, Y. and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, 142(3) :1061–1064.

[Munshi-South, 2012] Munshi-South, J. (2012). Urban landscape genetics : canopy cover predicts gene flow between white-footed mouse (Peromyscus leucopus) populations in New York City. *Molecular Ecology*, 21 :1360–1378.

[Nelson et al., 2008] Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., et al. (2008). The population reference sample, popres : a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3) :347–358.

[Nicholson et al., 2002] Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(4) :695–715.

[Nott and Dunsmuir, 2002] Nott, D. J. and Dunsmuir, W. T. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, 89(4) :819–829.

[Novembre et al., 2008] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. *Nature*, 456(7218) :98–101.

[Novembre and Stephens, 2008] Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5) :646–649.

[Oba et al., 2003] Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i., and Ishii, S. (2003). A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16) :2088–2096.

[Paciorek and Schervish, 2006] Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5) :483–506.

[Patterson et al., 2006] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12) :e190.

[Pawlowski, 1970] Pawlowski, B. (1970). Remarques sur l'endémisme dans la flore des alpes et des carpates. *Plant Ecology*, 21(4) :181–243.

[Pearson, 1901] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572.

[Peng and Kimmel, 2005] Peng, B. and Kimmel, M. (2005). simupop : a forward-time population genetics simulation environment. *Bioinformatics*, 21(18) :3686–3687.

[Plagemann et al., 2008] Plagemann, C., Kersting, K., and Burgard, W. (2008). Nonstationary gaussian process regression using point estimates of local smoothness. In *Machine learning and knowledge discovery in databases*, pages 204–219. Springer.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–959.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink : a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3) :559–575.

[Raj et al., 2014] Raj, A., Stephens, M., and Pritchard, J. K. (2014). Variational inference of population structure in large snp datasets. *Genetics*, pages genetics–114.

[Ramachandran et al., 2005] Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44) :15942–15947.

[Rasmussen, 2006] Rasmussen, C. E. (2006). Gaussian processes for machine learning.

[Reich et al., 2001] Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834) :199–204.

[Riebler et al., 2008] Riebler, A., Held, L., and Stephan, W. (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178(3) :1817–1829.

[Rosenberg et al., 2003] Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, 73(6) :1402–1422.

[Rosenberg et al., 2006] Rosenberg, N. A., Mahajan, S., Gonzalez-Quevedo, C., Blum, M. G., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L., Zapata, G., et al. (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from india. *PLoS genetics*, 2(12) :e215.

[Rosenberg et al., 2002] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602) :2381–2385.

[Rota, 1964] Rota, G.-C. (1964). The number of partitions of a set. *American Mathematical Monthly*, pages 498–504.

[Rousset, 1997] Rousset, F. (1997). Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, 145(4) :1219–1228.

[Roweis, 1998] Roweis, S. (1998). Em algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632.

[Rubin and Thayer, 1982] Rubin, D. B. and Thayer, D. T. (1982). Em algorithms for ml factor analysis. *Psychometrika*, 47(1) :69–76.

[Salakhutdinov and Mnih, 2008] Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM.

[Salmela et al., 2011] Salmela, E., Lappalainen, T., Liu, J., Sistonen, P., Andersen, P. M., Schreiber, S., Savontaus, M.-L., Czene, K., Lahermo, P., Hall, P., et al. (2011). Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. *PLoS One*, 6(2) :e16747.

[Schmidt and O'Hagan, 2003] Schmidt, A. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 65(3) :743–758.

[Schönswetter et al., 2005] Schönswetter, P., Stehlik, I., Holderegger, R., and Tribsch, A. (2005). Molecular evidence for glacial refugia of mountain plants in the European Alps. *Molecular Ecology*, 14(11) :3547–3555.

[Serrouya et al., 2012] Serrouya, R., Paetkau, D., Mclellan, B. N., Boutin, S., Campbell, M., and Jenkins, D. A. (2012). Population size and major valleys explain microsatellite variation better than taxonomic units for caribou in western Canada. *Molecular Ecology*, 21 :2588–2601.

[Shannon, 1949] Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1) :10–21.

[Sharbel et al., 2000] Sharbel, T., Haubold, B., and Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana* : biogeography and postglacial colonization of Europe. *Molecular Ecology*, 9(12) :2109–2118.

[Simonson et al., 2010] Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., et al. (2010). Genetic evidence for high-altitude adaptation in tibet. *Science*, 329(5987) :72–75.

[Slatkin, 1985] Slatkin, M. (1985). Gene flow in natural populations. *Annual review of ecology and systematics*, pages 393–430.

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

[Srebro et al., 2003] Srebro, N., Jaakkola, T., et al. (2003). Weighted low-rank approximations. In *ICML*, volume 3, pages 720–727.

[Storey, 2002] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :479–498.

[T O'Dushlaine et al., 2010] T O'Dushlaine, C., Morris, D., Moskvina, V., Kirov, G., Gill, M., Corvin, A., Wilson, J. F., Cavalleri, G. L., Consortium, I. S., et al. (2010). Population structure and genome-wide patterns of variation in ireland and britain. *European Journal of Human Genetics*, 18(11) :1248–1254.

[Thiel-Egenter et al., 2011] Thiel-Egenter, C., Alvarez, N., Holderegger, R., Tribsch, A., Englisch, T., Wohlgemuth, T., Colli, L., Gaudeul, M., Gielly, L., Jogan, N., et al. (2011). Break zones in the distributions of alleles and species in Alpine plants. *Journal of Biogeography*, 38(4) :772–782.

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3) :611–622.

[Tracy and Widom, 1994] Tracy, C. A. and Widom, H. (1994). Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1) :151–174.

[Tucker and MacCallum, 1997] Tucker, L. R. and MacCallum, R. C. (1997). Exploratory factor analysis. *Unpublished manuscript, Ohio State University, Columbus*.

[Twyman, 2002] Twyman, R. (2002). What are model organisms. *Wellcome trust*.

[Villemereuil et al., 2014] Villemereuil, P., Frichot, É., Bazin, É., François, O., and Gaggiotti, O. E. (2014). Genome scan methods against more complex models : when and how much should we trust them ? *Molecular ecology*, 23(8) :2006–2019.

[Vitalis et al., 2003] Vitalis, R., Dawson, K., Boursot, P., and Belkhir, K. (2003). Detsel 1.0 : a computer program to detect markers responding to selection. *Journal of Heredity*, 94(5) :429–431.

[Weir and Cockerham, 1984] Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *evolution*, pages 1358–1370.

[Weir et al., 1990] Weir, B. S. et al. (1990). *Genetic data analysis. Methods for discrete population genetic data.* Sinauer Associates, Inc. Publishers.

[Weiss et al., 1997] Weiss, K. M., Cavalli-Sforza, L. L., Dunston, G. M., Feldman, M., Greely, H. T., Kidd, K. K., King, M., Moore, J. A., Szathmary, E., Twinn, C. M., et al. (1997). Proposed model ethical protocol for collecting dna samples. *Houston law review/University of Houston*, 33(5) :1431.

[West, 2003] West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm.

[Wilkins and Wakeley, 2002] Wilkins, J. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. *Genetics*, 161(2) :873–888.

[Winkler, 2003] Winkler, G. (2003). *Image analysis, random fields and Markov chain Monte Carlo methods : a mathematical introduction*, volume 27. Springer.

[Wright, 1920] Wright, S. (1920). *Principles of livestock breeding*. Number 905. US Department of Agriculture.

[Wright, 1943] Wright, S. (1943). Isolation by distance. *Genetics*, 28(2) :114.

[Xu et al., 2011] Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., Yang, L., Pan, X., Wang, J., Shen, Y., et al. (2011). A genome-wide search for signals of high-altitude adaptation in tibetans. *Molecular biology and evolution*, 28(2) :1003–1011.

[Zhou and Schmidler, 2009] Zhou, X. and Schmidler, S. C. (2009). Bayesian parameter estimation in ising and potts models : A comparative study with applications to protein modeling. Technical report, Technical report, Duke University.

7 Appendices

7.1 Estimer les F_{st}

La F_{st} peut être estimée pour un nombre variable de sous-populations, et un nombre variable de loci. L'estimateur proposé par Wright [Wright, 1943] pour un nombre N de sous-populations, et un marqueur biallélique de fréquence allélique f_i dans la sous-population i et de fréquence allélique $\bar{f}$ dans l'ensemble de la population est

$$\theta = \frac{\frac{1}{N}\sum_{i=1}^{N}(f_i - \bar{f})}{\bar{f}(1-\bar{f})} = \frac{\sigma_s^2}{\bar{f}(1-\bar{f})} \tag{35}$$

Au vu des ambiguïtés dans le calcul des estimateurs de F_{st} dans la littérature [Chesser, 1983, Hiebert and Hamrick, 1983] et des approximations que comporte cet estimateur, Weir et Cockerham proposent un estimateur plus robuste basé sur une analyse de variance [Weir and Cockerham, 1984]. En notant a, la composante de variance inter population, b la composante de variance inter individus intra populationnelle, et c la composante de variance entre les gamètes d'un individu, l'estimateur s'écrit

$$\theta_{WC} = \frac{a}{a+b+c} \tag{36}$$

$$a = \frac{\bar{n}}{n_c}\left(\sum_{i=1}^{N}\frac{n_i(f_i-\bar{f})^2}{(N-1)\bar{n}} - \frac{1}{\bar{n}-1}\left(\bar{f}(1-\bar{f}) - \frac{N-1}{N}\sum_{i=1}^{N}\frac{n_i(f_i-\bar{f})^2}{(N-1)\bar{n}} - \frac{\bar{h}}{4}\right)\right) \tag{37}$$

$$b = \frac{\bar{n}}{\bar{n}-1}\left(\bar{f}(1-\bar{f}) - \frac{N-1}{N}\sum_{i=1}^{N}\frac{n_i(f_i-\bar{f})^2}{(N-1)\bar{n}} - \frac{2\bar{n}-1}{4\bar{n}}\bar{h}\right) \tag{38}$$

$$c = \frac{\bar{h}}{2} \tag{39}$$

où n_i est l'effectif de la i^{eme} sous population, $\bar{n}$ est la moyenne des effectifs et $n_c = (N\bar{n} - \sum_{i=1}^{N}\frac{n_i^2}{N\bar{n}})/(N-1)$ [Weir et al., 1990]. $\bar{h}_i$ est l'hétérozigotie moyenne dans la population i. Nous notons au passage qu'avec les hypothèses suivantes :

- Tous les échantillons sont de même effectif : $n_i = \bar{n}, i = 1 \ldots N$
- Les données contiennent un grand nombre de sous population : $\frac{1}{N} \approx 0$.
- Les échantillons contiennent un grand nombre d'individus : $\frac{1}{\bar{n}} \approx 0$.

alors l'estimateur 36 se simplifie en l'estimateur 2.

Dans la section 4, nous utilisons de nouveau l'estimateur de F_{st} de l'équation (36). Cependant, nous l'utilisons pour évaluer le niveau de différenciation entre plusieurs populations en utilisant tous les p loci contenus dans les données, au lieu de scanner les marqueurs du génome un à un.

Dans ce cas, l'estimateur que nous utilisons est :

$$\theta_{WC} = \frac{\sum_{l=1}^{p} a_l}{\sum_{l=1}^{p} a_l + b_l + c_l} \tag{40}$$

comme conseillé dans [Weir and Cockerham, 1984].

7.2 Algorithme MCMC pour le modèle à facteurs

Nous mentionnons dans la section 3.2.3 que nous estimons la distribution a posteriori des paramètres du modèle par un algorithme MCMC. Cet algorithme repose sur des pas d'échantillonneur de Gibbs pour les loi conditionnelles, et des pas Metropolis-Hastings pour deux des paramètres.

Nous modélisons la variance des coefficients de la matrice V, $\sigma_k^2 = \sigma^2\rho_k^2$. Par la suite, nous considérons $\Sigma_V = \sigma^2 \times \mathrm{Diag}(\rho_1^2, \cdots, \rho_K^2)$ où $\mathrm{Diag}(\rho_1^2, \cdots, \rho_K^2)$ est une matrice diagonale avec $(\rho_1^2, \cdots, \rho_K^2)$ comme termes diagonaux. $\Sigma_{V,k0}$ est identique à Σ_V excepté le k_0^{me} élément de la diagonale qui vaut $\sigma^2 c_{k_0}^2 \rho_{k_0}^2$ au lieu de $\sigma^2\rho_{k_0}^2$. Pour une valeur donnée de K, nous décrivons une itération de l'algorithme MCMC

–
$$\mathbf{V}_\ell \leftarrow \mathcal{N}(\mathbf{m}_\ell, \Sigma_\ell),\ \ell = 1, \ldots, p,$$
avec $\mathbf{m}_\ell = (\sigma^2\Sigma_{V,k_0}^{-1} + \mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}\mathbf{Y}_\ell)$ et $\Sigma_\ell = (\Sigma_{V,k_0}^{-1} + \frac{1}{\sigma^2}\mathbf{U}^T\mathbf{U})^{-1}$.

–
$$\rho_k^2 \leftarrow IG(\frac{p}{2}, \frac{1}{2\sigma^2}\sum_{\ell=1}^{p}\frac{V_{k\ell}^2}{(c_k^2)^{\mathbb{1}_{z_\ell=k}}}),$$
où $\mathbb{1}$ est la fonction indicatrice et IG est la distribution inverse-gamma.

–
$$\sigma^2 \leftarrow IG(\frac{(n+K)p}{2}, \frac{1}{2}\sum_{\ell=1}^{p}(\sum_{i=1}^{n}|Y_{i,\ell} - U_i V_\ell|^2 + \sum_{k=1}^{K}\frac{V_{k\ell}^2}{\rho_k^2(c_k^2)^{\mathbb{1}_{z_\ell=k}}})),$$

–
$$\mathbf{U}_i \leftarrow \mathcal{N}(\mathbf{m}_i, \Sigma),\ i = 1, \ldots, n,$$
avec $\mathbf{m}_i = (I_K + \frac{1}{\sigma^2}\mathbf{V}\mathbf{V}^T)^{-1}\frac{1}{\sigma^2}\mathbf{V}\mathbf{Y}_i$ et $\Sigma = (\frac{1}{\sigma^2}\mathbf{V}\mathbf{V}^T + I_K)^{-1}$. Les $\mathbf{U}_i$ sont des vecteurs de dimension K et sont les vecteurs lignes de la matrice $\mathbf{U}$. Notons que dans la section 3, nous parlons des colonnes de $\mathbf{U}$ avec la notation $\mathbf{U}_i$.

–
$$z_\ell = k \text{ avec une probabilité } p(z_\ell = k|\pi, V_l, \Sigma_V, c_k^2),\ k = 1, \ldots, K$$
avec
$$p(z_\ell = k|\pi, V_l, \Sigma_V, c_k^2) \propto \frac{\pi}{K}e^{\beta\sum_{i\sim\ell}\mathbb{1}_{z_i,z_\ell}}\frac{1}{c}e^{\frac{V_{k\ell}^2}{2\sigma^2\rho_k^2}(1-1/c_k^2)}, \tag{41}$$
et $p(z_\ell = 0|\pi, V_l, \Sigma_V, c_k^2) \propto (1-\pi)e^{\beta\sum_{i\sim\ell}\mathbb{1}_{z_i,z_\ell}}$ pour $k = 0$.

– $(c_1^2, \ldots, c_K^2)$ sont mis à jour à l'aide d'un pas Metropolis-Hastings. Le vecteur proposé est tiré aléatoirement suivant le noyau de proposition
$$(c_1^{2*}, \ldots, c_K^{2*}) \leftarrow \mathcal{N}((c_1^2, \ldots, c_K^2), I_K)$$

et chaque coefficient c_k^{2*} est accepté avec une probabilité

$$min(1, \frac{p(c_k^{2*}) \prod_{l \in \mathcal{L}_k} \frac{1}{c_k} e^{\frac{V_{kl}^2}{2c_k^{2*}\rho_k^2\sigma^2}}}{p(c_k^2) \prod_{l \in \mathcal{L}_k} \frac{1}{c_k^*} e^{\frac{V_{kl}^2}{2c_k^2\rho_k^2\sigma^2}}}),$$

où $\mathcal{L}_k$ est l'ensemble des marqueurs outliers selon le k^{me} facteur. Nous choisissons comme distribution a priori $p(c_k^2) = \mathcal{U}(1, 10)$.

– $\log_{10}(\pi)$ est mis à jour avec un pas Metropolis-Hastings, et un noyau de proposition

$$\log_{10}(\pi)^* \leftarrow \mathcal{N}(\log_{10}(\pi), .5)$$

et la valeur $\log_{10}(\pi^*)$ est acceptée avec une probabilité

$$min(1, \frac{p(\pi^*)\pi^{*(p-p_0)}(1-\pi^*)^{p_0}}{p(\pi)\pi^{(p-p_0)}(1-\pi)^{p_0}})$$

où p_0 est le nombre de marqueurs non outliers. Nous choisissons comme distribution a priori $p(log_{10}(\pi)) = \mathcal{U}(-4, -1)$.

Nous observons que l'algorithme converge rapidement, et supposons qu'une chaîne de 400 itérations suffit à donner une estimation des facteurs de Bayes.

8 Notes bibliographiques

Les articles suivants ont été écrits pendant la thèse :

8.1 Article A : *Nicolas Duforet-Frebourg, Eric Bazin, and Michael G.B. Blum Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model. Molecular Biology and Evolution, 2014, doi : 10.1093/molbev/msu182*.

8.2 Article B : *Nicolas Duforet-Frebourg and Michael G.B. Blum, Nonstationary patterns of isolation-by-distance, inferring measures of local genetic differentiation with Bayesian kriging. Evolution, 2014, 68 : 1110–1123. doi : 10.1111/evo.12342*.

8.3 Article C : *Nicolas Duforet-Frebourg, Lucie M. Gattepaille, Michael G.B. Blum and Mattias Jakobsson. HaploPOP : a software that improves population assignment by combining markers into haplotypes. En preparation*

8.4 Article D : *Matilde Karakachoff, Nicolas Duforet-Frebourg, et al., Fine scale human genetic structure in Western France. European journal of Human genetics, 2014, AOP, doi : 10.1038/ejhg.2014.175*.

MBE Advance Access published June 24, 2014

Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model

Nicolas Duforet-Frebourg,[1] Eric Bazin,[2] and Michael G.B. Blum*,[1]

[1]Laboratoire TIMC-IMAG, UMR 5525, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France

[2]Laboratoire d'Ecologie Alpine, UMR 5553, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France

***Corresponding author:** E-mail: michael.blum@imag.fr.

Associate editor: John Novembre

Abstract

There is a considerable impetus in population genomics to pinpoint loci involved in local adaptation. A powerful approach to find genomic regions subject to local adaptation is to genotype numerous molecular markers and look for outlier loci. One of the most common approaches for selection scans is based on statistics that measure population differentiation such as F_{ST}. However, there are important caveats with approaches related to F_{ST} because they require grouping individuals into populations and they additionally assume a particular model of population structure. Here, we implement a more flexible individual-based approach based on Bayesian factor models. Factor models capture population structure with latent variables called factors, which can describe clustering of individuals into populations or isolation-by-distance patterns. Using hierarchical Bayesian modeling, we both infer population structure and identify outlier loci that are candidates for local adaptation. In order to identify outlier loci, the hierarchical factor model searches for loci that are atypically related to population structure as measured by the latent factors. In a model of population divergence, we show that it can achieve a 2-fold or more reduction of false discovery rate compared with the software *BayeScan* or with an F_{ST} approach. We show that our software can handle large data sets by analyzing the single nucleotide polymorphisms of the Human Genome Diversity Project. The Bayesian factor model is implemented in the open-source PCAdapt software.

***Key words:* F_{ST}, population structure, landscape genetics, population genomics, selection scans.**

Introduction

With the development of sequencing and genotyping technologies, there is a considerable impetus in population genomics to pinpoint loci involved in local adaptation (Akey et al. 2002; Bonin et al. 2006). A working hypothesis of population genomics is that most loci have neutral patterns of variation that are similarly affected by demographic processes, whereas loci targeted by natural selection have atypical patterns (Luikart et al. 2003). Although this working hypothesis is fiercely debated (Sella et al. 2009), it has led to a protocol for finding genomic regions subject to local adaptation which consists of genotyping numerous molecular markers and looking for outlier loci. Different regions of the genome are expected to exhibit highly variable levels of genetic differentiation between populations ranging from genomic regions exhibiting little differentiation to regions where genetic divergence is extremely pronounced. Although high levels of differentiation can be explained by various causes, adaptation of individuals to their local environment is a prominent explanation such that patterns of differentiation for adaptive loci exceed neutral expectations (Nosil and Buerkle 2010). Measures of genetic differentiation between populations such as F_{ST} have been commonly used to find outlier loci, although there are many alternative approaches (Oleksyk et al. 2010). A proof of concept for approaches based on genetic differentiation was provided when studying human adaptation to altitude; the most differentiated variants between a Tibetan population living in a hypoxic environment and a lowland Han Chinese population were found in hypoxia-inducible transcription factors (Yi et al. 2010; Xu et al. 2011).

Genome scans based on F_{ST} were proposed by Lewontin and Krakauer (1973) and have been considerably expanded since (Beaumont and Nichols 1996; Vitalis et al. 2001; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler et al. 2008; Guo et al. 2009; Bazin et al. 2010; Bonhomme et al. 2010; Gompert and Buerkle 2011; Fariello et al. 2013). They are not limited to two populations as in the adaptation-to-altitude example and can be used with multiple populations. One possibility is to compute an overall F_{ST} measure of genetic differentiation and to determine a threshold at which the null hypothesis of neutral evolution can be rejected (Beaumont and Nichols 1996). Another possibility is to adopt a model-based perspective by implementing the multinomial-Dirichlet model or *F* model, which is parameterized by population-specific *F* statistics (Beaumont and Balding 2004). The *F* statistics can be interpreted as measures of divergence from a common immigrant gene pool (Wright 1931) or as a divergence from an initial and hypothetical ancestral population (Nicholson et al. 2002). The Bayesian approach for distinguishing between neutral or adaptive evolution offers the opportunity to assign a probability to each of the two evolutionary models at each locus (Foll and Gaggiotti 2008; Riebler et al. 2008). In the following, approaches based

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

Article

on F_{ST} or on population-specific F statistics are referred as genome scans based on F statistics. There are many software-implementing genome scans based on F statistics (e.g., *BayeScan*, *DetSel*, *fdist2*, and *Lositan*), and they contribute to the popularity of this approach in population genomics (Beaumont and Nichols 1996; Vitalis et al. 2003; Antao et al. 2008; Foll and Gaggiotti 2008).

However, a major issue with genome scans based on F statistics is that they can generate a high rate of false positives for both biological and statistical reasons (Bierne et al. 2013). Here, we propose to address the statistical and computational problems that arise with F statistics. The first problem arises because F statistics have been derived under Wright's F model of population subdivision, which assumes a particular covariance structure for gene frequencies among populations (Bierne et al. 2013; Fourcade et al. 2013). When spatial structure departs from Wright's island model of population subdivision, genome scans based on F statistics produce many false positives (Bierne et al. 2013). Alternative statistical measures that account for population structure have recently been proposed (Bonhomme et al. 2010; Günther and Coop 2013). A second potential problem concerns the computational burden of some Bayesian approaches, which can become an obstacle with a large number of single nucleotide polymorphisms (SNPs) (Lange et al. 2014). The last intrinsic problem of genome scans based on F statistics is that individuals should be grouped into populations. However, it has been advocated in landscape genetics to rather work at the scale of individuals because it avoids potential bias in identifying populations in advance and it offers the opportunity to conduct studies at a finer scale (Manel and Holderegger 2013).

To tackle the aforementioned problems, we propose a statistical method based on a Bayesian factor model (West 2003) to pick outlier loci involved in local adaptation. With factor models, we seek to jointly determine population structure and outlier loci. Factor models are strongly related to principal component analysis (PCA) because they both approximate the matrix of individual genotypes by a product of two lower-rank matrices, albeit using different constraints and priors for the lower-rank matrices (Engelhardt and Stephens 2010). One of the two matrices encodes population structure using latent factors, whereas the second matrix measures to what extent each individual SNP is related to the pattern of population structure. The proposed factor model seeks for loci that are atypically related to population structure. To show the potential of factor models for genome scans, we consider two examples. First, we consider a model of population divergence. In this example, we compare false discovery rates (FDRs) obtained from the proposed factor model with *BayeScan* and with a genome scan based on F_{ST}. The second example is a model of isolation-by-distance with selection. It is an instance of how factor models can be used to detect local adaptation when it would be arbitrary to group individuals into populations. Finally, we analyze the HGDP human data set (Li et al. 2008) to provide an example of how factor models can be used to detect local adaptation with a large number of SNPs.

New Approaches

We denote the $n \times p$ matrix of centered allele counts by $\mathbf{Y}$, where n is the number of individuals and p is the number of loci. The elements $Y_{i\ell}$, $i = 1,\ldots,n$, $\ell = 1,\ldots,p$ correspond to the centered allele counts of the ith individual at locus ℓ. Before centering, the allele counts belong to {0, 1} or {0, 1, 2} for haploid and diploid species, respectively. After centering, each column of the matrix $\mathbf{Y}$ has a mean of 0.

Factor models assume that the matrix of column-centered genotypes $\mathbf{Y}$ can be written as a product of two lower-rank matrices

$$\mathbf{Y} = \mathbf{UV} + \epsilon, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are of dimension $(n \times K)$ and $(K \times p)$, respectively, K is an hyper parameter that is much smaller than n and p, and ϵ is the matrix of residuals. In the following, the column vectors of $\mathbf{U}$ referred to as "factors" or "latent factors" are denoted by $\mathbf{U}_1, \ldots, \mathbf{U}_K$. Factor models assume that the vector—of size n—of centered allele counts $\mathbf{Y}_\ell$ can be obtained as follows:

$$\mathbf{Y}_\ell = \sum_{k=1}^{K} \mathbf{U}_k V_{k\ell} + \epsilon_\ell, \ \ell = 1, \ldots, p, \tag{2}$$

where ϵ_ℓ is a vector containing n independent Gaussian residuals of variance σ^2 and $V_{k\ell}$ are the elements of the matrix $\mathbf{V}$. Assuming that the K factors are known, then the elements $V_{k\ell}$ of the matrix $\mathbf{V}$ are the regression coefficients—sometimes called *factor loadings*—obtained after regressing the vector of centered allele counts $\mathbf{Y}_\ell$ by the K factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$. As candidates for local adaptation, we consider the loci ℓ that have large (in absolute value) regression coefficients $V_{k\ell}$ for one of the factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$. In factor models, the K factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$ are in fact unknown and have to be estimated; they are parameters of the model and represent population structure (Engelhardt and Stephens 2010). In our proposed framework, outlier loci are loci that are excessively related to population structure, the latter being measured by the K latent factors. After statistical inference, the factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$ are ordered by decreasing variances $\sigma_1^2 > \ldots > \sigma_K^2$, where σ_k^2 measures the variance of the regression coefficients $V_{k\ell}$, $\ell = 1\ldots p$, for the kth factor.

To provide a concrete example of how factors represent population structure, we consider a model of population divergence. We assume that an initial population splits into two populations A and B that diverged according to neutral evolution. The initial neutral divergence of duration T is followed by two concomitant splits where each daughter population A and B splits into two subpopulations (A_1, A_2) and (B_1, B_2). In contrast to the initial divergence which is purely neutral, the second phase of divergence between populations assumes some local adaptation with a small proportion of SNPs conferring selective advantage (fig. 1). We fit the factor models with $K = 3$ and we display the three factors in figure 1. The first factor discriminates individuals according to the initial split and the second and third factors discriminate individuals according to the subsequent splits which separate

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

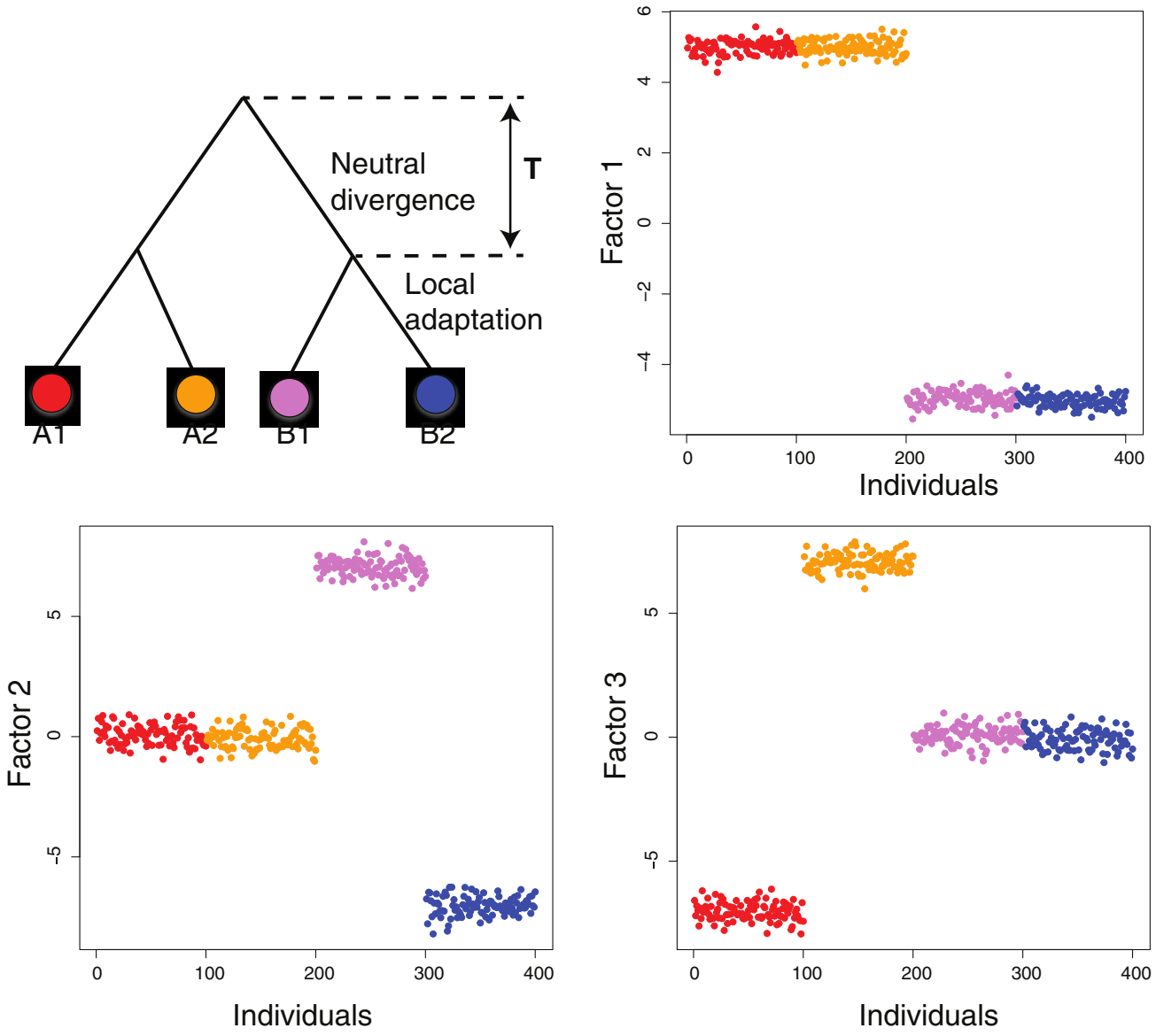


FIG. 1. Values of the $K = 3$ factors for a population divergence model with four populations. The upper-left panel shows the model of population divergence. The other panels show the values of the first three factors and each dot corresponds to one individual. As candidates for local adaptation, the factor model with $K = 3$ looks for SNPs whose variation is atypically well explained by one of the three factors. For the simulations, the effective population size is $N_e = 1{,}000$ diploid individuals in each population, 50 individuals are sampled in each population, and the neutral divergence time is $T = 200$ generations and is twice as long as the second phase during which there is adaptation. We assume that 400 SNPs, among a total of 10,000 SNPs, can confer selective advantage. The 400 adaptive SNPs are split into four sets of 100 SNPs, each set corresponding to adaptation in one of the four evolutionary lineages. We consider a selection coefficient $s = 0.1$ for homozygotes carrying two adaptive alleles and $s = 0.05$ for heterozygotes.

subpopulation B_1 from B_2 (second factor) and subpopulation A_1 from A_2 (third factor).

We now specify how we measure the degree of outlyingness for each locus using a Bayesian criterion. To account for outlier and non-outlier loci, we assume that, at a given locus ℓ, the vector of regression coefficients $\mathbf{V}_\ell = (V_{1\ell},\ldots,V_{K\ell})$ comes from a mixture of two different distributions. We introduce a vector $\mathbf{z}$ of indicator variables $(z_1,\ldots,z_p)$ whose elements are equal to 0 for non-outlier loci and take values $1,\ldots,K$ for outlier loci. For any locus ℓ, either outlier or nonoutlier, we assume that the vector $\mathbf{V}_\ell = (V_{1\ell},\ldots,V_{K\ell})$ is composed of independent Gaussian random variables. The model for non-outlier loci is a product of Gaussian distributions

$$V_{k\ell} \mid z_\ell = 0 \sim \mathcal{N}(0,\sigma_k^2),\ k = 1,\ldots,K, \qquad (3)$$

where $\mathcal{N}(m,\sigma^2)$ denotes the Gaussian distribution with mean m and variance σ^2. To model outlier loci, we consider a variance–inflation model which assumes an inflated variance to account for outlier loci (Box and Tiao 1968; Devlin and Roeder 1999). The model for outlier loci is itself a mixture model with K components of equal weights, where the kth component assumes an inflated variance for the kth regression coefficient but not for the other ones. Denoting the variance–inflation parameter for factor k by c_k^2 ($c_k^2 > 1$), the k_0th component of the mixture model for outlier assumes a product of Gaussian distributions

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

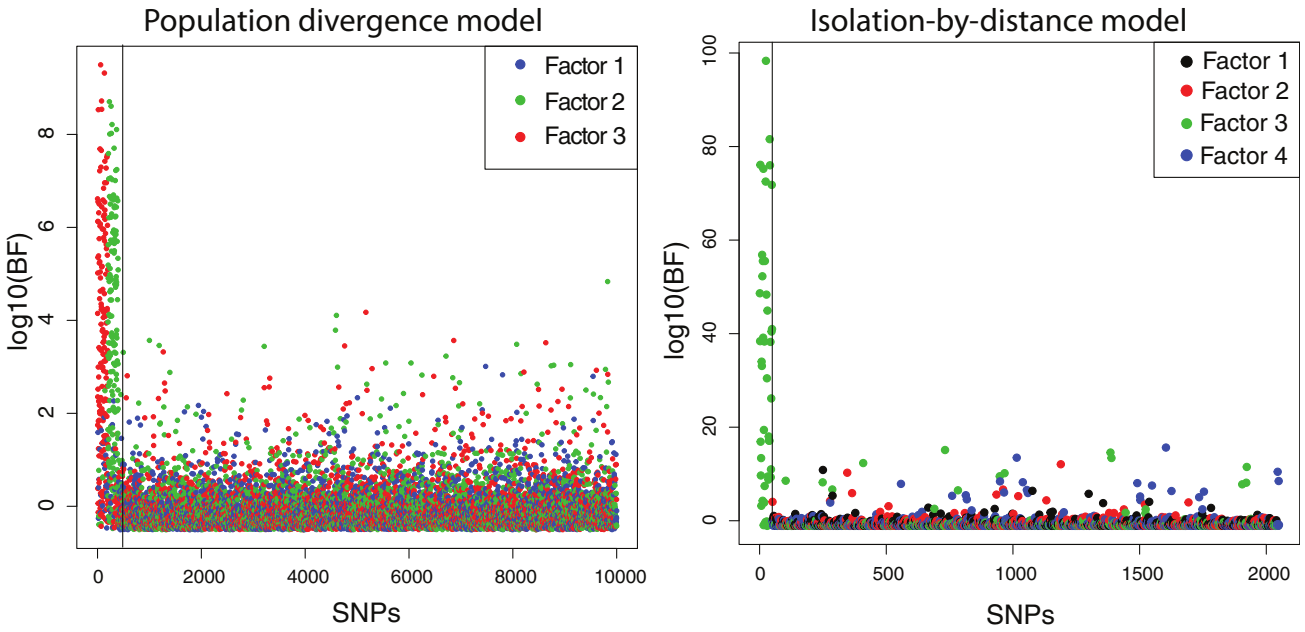


FIG. 2. Bayes factors for the two different simulation examples. The SNPs under selection are located on the left-hand side of the vertical bar.

$$V_{k\ell} \mid z_\ell = k \sim \mathcal{N}(0, c_k^2\sigma_k^2),\ k = k_0$$
$$V_{k\ell} \mid z_\ell = k \sim \mathcal{N}(0, \sigma_k^2),\ k \neq k_0 \quad (4)$$

The model for outliers has been chosen for the sake of interpretability. Each outlier locus can be related to one of the K factors because outlier loci should be atypically explained by one of the K factors. To measure the strength of evidence for outlyingness for each locus, we compute the Bayes factor of the outlier model against the non-outlier model. If a locus is considered as an outlier, the factor with which there is an atypical correlation is found by computing the posterior probabilities of each of the K components of the outlier mixture model. To account for linkage disequilibrium, we additionally consider a Potts model that encourages outlier loci to be clustered in the genome (see Materials and Methods).

When fitting the factor model with $K = 3$ to data simulated under the scenario of population divergence depicted in figure 1, the outlier model of equation (4) assumes three different types of outlier loci: loci that have large genetic differentiation when comparing the pair of subpopulations (A_1,A_2) with the pair (B_1,B_2) (large values of $|V_{1\ell}|$), loci that have large genetic differentiation when comparing subpopulation B_1 with B_2 (large values of $|V_{2\ell}|$), and loci that have large genetic differentiation when comparing subpopulation A_1 with A_2 (large values of $|V_{3\ell}|$). Because the simulation assumes that the initial period of divergence is purely neutral, the first types of outliers (large values of $|V_{1\ell}|$) are in fact false positives.

Results

Simulation Study

Population Divergence Model

The first simulation study investigates to what extent factor models better account for population structure than methods based on F statistics. We consider the model of population divergence depicted in figure 1. An initial neutral divergence is followed by adaptive divergence where 4% of the 10,000 simulated SNPs are involved in local adaptation. The set of adaptive SNPs is split in four equal parts and each subset of SNP confers a selective advantage in only one of the four populations. When the initial neutral divergence time T is null, the population tree is star-like, and the assumption of the F model is valid. As the initial neutral divergence time T increases, the departure from the F model increases. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration.

First, we present results using the factor model with $K = 3$ factors that is optimal because there are four populations in the divergence model (Patterson et al. 2006). We consider a long-enough divergence time $T = 2$ so that the first factor corresponds to the initial and neutral divergence, whereas the second and third factors correspond to the subsequent divergence events during which biological adaptation took place (fig. 1). The SNPs that have been truly involved in biological adaptation are usually associated with the correct factor because, among the 400 truly adaptive SNPs, 81% are associated with the second and third factor and this proportion raises to 98% (respectively 92%) when considering the 195 (respectively 305) adaptive SNPs with Bayes factors larger than 10 (respectively 1) (fig. 2).

Then, we compare the FDRs of three different approaches including the proposed factor model, *BayeScan* (version 2.1), and genome scans based on the F_{ST} statistic. For both *BayeScan* and the proposed factor model, we use Bayes factors for ranking SNPs, whereas we use F_{ST} values for the last method. More precisely, we use the q values for ranking SNPs with *BayeScan*, but, by definition of the q value, it provides the same ranking as the Bayes factors. To determine a threshold above which SNPs are considered as outliers, we enlarge the lists of top-ranked SNPs, provided by each method, until each

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

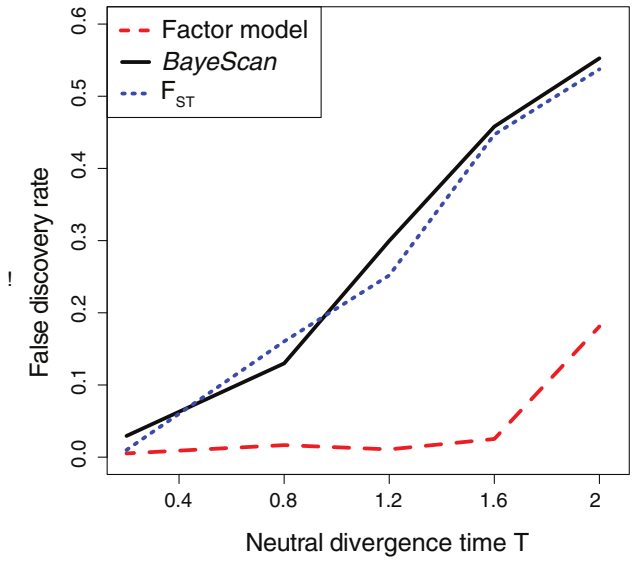


Fig. 3. False discovery rate as a function of the initial divergence time T in the population divergence model of figure 1. For the proposed factor model, Bayes factors are used for ranking SNPs, whereas we use q values with *BayeScan* and F_{ST} values for the genome scan based on F_{ST} values. To determine a threshold above which SNPs are considered as outliers, we constrain the lists of SNPs provided by each method to contain 50% of the 400 SNPs truly involved in local adaptation. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration, which is of 100 generations. For the simulations, the effective population size is $N_e = 1{,}000$ diploid individuals in each population and 50 individuals are sampled in each population. We assume that 400 SNPs, among a total of 10,000 SNPs, can confer selective advantage. The 400 adaptive SNPs are split into four sets of 100 SNPs, each set corresponding to adaptation in one of the four evolutionary lineages. We consider a selection coefficient $s = 0.1$ for homozygotes carrying two adaptive alleles and $s = 0.05$ for heterozygotes.

of them contains 50% of the 400 truly adaptive SNPs. This procedure amounts to setting the sensitivity to 50% (the sensitivity is also called recall rate in machine learning). Figure 3 shows that for all methods, the FDR is below 5% when the population tree is almost star-like ($T = 0.04$) but it increases with the initial neutral divergence time T. Although the FDR always increases with T, the FDRs of the factor model are always smaller than FDRs obtained with F_{ST} and *BayeScan*. For instance, when $T = 1$, the FDRs obtained with F_{ST} and *BayeScan* are between 20% and 30%, whereas it is smaller than 5% with the factor model. Instead of using a threshold of 50%, we also constrain the lists of SNPs to contain 25% or 75% of the truly adaptive SNP (i.e., setting the sensitivity to 25% or 75%). As for the 50% threshold, all methods have small FDR for sufficiently small initial divergence time T, and, as T increases, FDR increases at a slower rate for the factor model (supplementary fig. S1, Supplementary Material online). In summary, the FDR increases as the model of divergence deviates from a star-like phylogeny, but compared with other methods, the factor model reduces the proportion of false discoveries by a factor of 2 or more when there is a strong-enough deviation from the star-like assumption ($T > 0.8$).

The results presented so far were obtained using the factor model with $K = 3$ factors. By increasing the values of K from 1 to 6, we find that, compared with $K = 3$, the FDR drastically increases for underspecification of K ($K < 3$) but is almost insensitive to overspecification of K ($K > 3$, supplementary fig. S2, Supplementary Material online). We also compute the mean squared error (MSE) of equation (2) for different values of K to determine if the MSE can be a guide for choosing K. The MSE decreases from $K = 1$ to $K = 3$ before staying almost constant as K continues to grow (supplementary fig. S3, Supplementary Material online). In this example of population divergence, the MSE suggests choosing $K = 3$, but choosing a more complex model ($K > 3$) would provide comparable FDRs.

Isolation-By-Distance Model

The second simulation study provides an example of how to search for biological adaptation when there is isolation-by-distance. Approaches based on F statistics would require to group individuals into populations, and we want to avoid that. On a two-dimensional 10×10 grid, we simulate a stepping-stone model with selection acting on individuals located in the lower-right corner of the grid. We sample 10 diploid individuals at each of the 100 demes. A total of 50 out of 2,050 SNPs confer selective advantage in this region and the selection coefficient decreases gradually when moving away from the point where selection is maximal.

With the factor model, the selection gradient is reflected in a different factor depending on the value of the selection coefficients (results not shown). Here, we choose the intensity of selection so that the selection gradient becomes visible in the third factor (fig. 4). The other factors have spatial patterns that are typical for isolation-by-distance models (Novembre and Stephens 2008; Engelhardt and Stephens 2010). We choose $K = 4$ because the MSE decreases from $K = 1$ to $K = 4$ before being almost constant (supplementary fig. S3, Supplementary Material online). In terms of FDR, this choice of K is not optimal because $K = 3$ would provide smaller FDR (supplementary fig. S4, Supplementary Material online). However, as in the first example, smaller values of K, compared with the optimal value ($K < 3$), increase FDR drastically, whereas too large values of K ($K > 3$) increase the optimal FDR more moderately (supplementary fig. S4, Supplementary Material online). With $K = 4$, the FDR is equal to 0% when considering the top 25 SNPs, which corresponds to a sensitivity of 50%. However, when setting the sensitivity to 75%, the FDR increases to 30%, which corresponds to 38 true positive SNPs among a list of 54 SNPs. The 50 truly adaptive SNPs are all correctly associated with the factor corresponding to biological adaptation, which is the third factor here (fig. 2). When decreasing the number of sampled individuals from 10 to 1, the FDR, obtained with a sensitivity of 50%, increases considerably from 0% to 91% (supplementary fig. S5, Supplementary Material online).

We also investigate to what extent scaling the data matrix Y such that all columns have unit variance affects the results. For both models of population divergence and isolation-by-distance, we compare the FDR obtained when fitting the

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

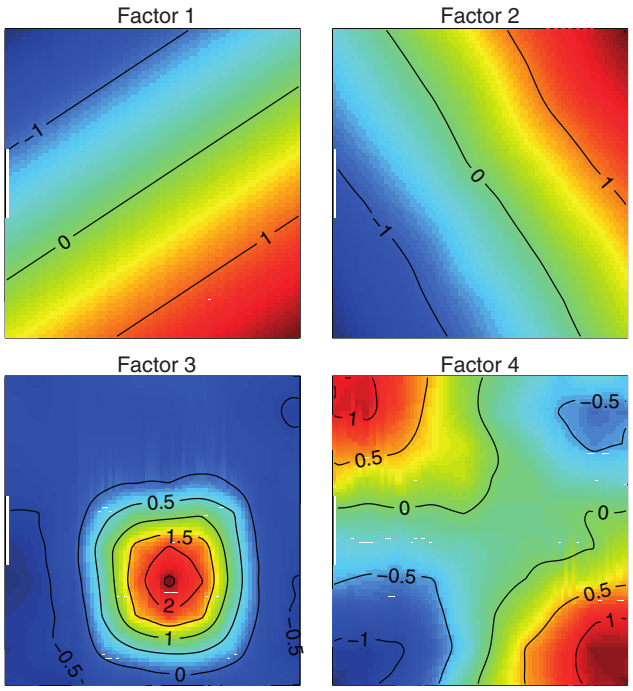


FIG. 4. Values of the $K = 4$ factors for the two-dimensional isolation-by-distance model. The different panels show the values of the first four factors when projected onto the two-dimensional space. As candidates for local adaptation, the factor model with $K = 4$ looks for SNPs whose variation is atypically well explained by one of the four factors. Spatial interpolation of the factors is obtained using the *Krig* function that is available from the *fields* R package. Each of the 100 populations in the grid has an effective population size of $N_e = 1{,}000$ diploid individuals with 10 sampled individuals in each population. After simulating an equilibrium stepping-stone model with a migration rate of $4N_em = 8$ per generation, we impose a selection gradient where 50 SNPs confer selective advantage during 100 generations. The selection coefficient $s = 0.1$ is maximal in population 64, which is located in the lower-right quarter of the grid. It is divided by 2 in the eight neighboring populations and again divided by 2 for the second layer of neighbors.

factor model with the scaled or unscaled data matrix Y. We do not find univocal evidence in favor or against the scaling. For the model of population divergence, no scaling is preferable, whereas the results are more complex for the isolation-by-distance model: with a sensitivity of 25%, no scaling is again preferable but scaling the data decreases the FDR when the sensitivity is larger than 50% (supplementary fig. S6, Supplementary Material online). For the model of isolation-by-distance, we compute the minimum allele frequency of the SNPs that are part of the list with largest Bayes factors (setting the sensitivity at 25%). We find that scaling the data matrix increases the proposition of low-frequency variants among the list of top SNPs, and this phenomenon is much more pronounced when looking at the false positives only in the top list (supplementary fig. S7, Supplementary Material online).

Analysis of Human SNP Data

The HGDP data set contains 644,199 SNPs, after removal of the SNPs on the sex chromosomes and on the mitochondrion, which have been typed for 1,043 individuals coming from 53 different populations (Li et al. 2008). First, we fit the factor model to the unscaled matrix Y of SNPs using different values of K. In contrast with the two previous examples, there is no value of K at which the MSE stops to decrease (supplementary fig. S3, Supplementary Material online). By looking at the different factors (fig. 5 and supplementary fig. S8, Supplementary Material online), we decide to consider a model that captures genetic differentiation between, but not within, continents. When considering $K = 8$, we find that factors 5–8 capture population structure within continents (supplementary fig. S8, Supplementary Material online) and we choose not to consider evolutionary processes acting at this scale. Using the fact that we are interested in genetic differentiation between continents, we consider a factor model with $K = 4$. The first factor mostly contrasts African

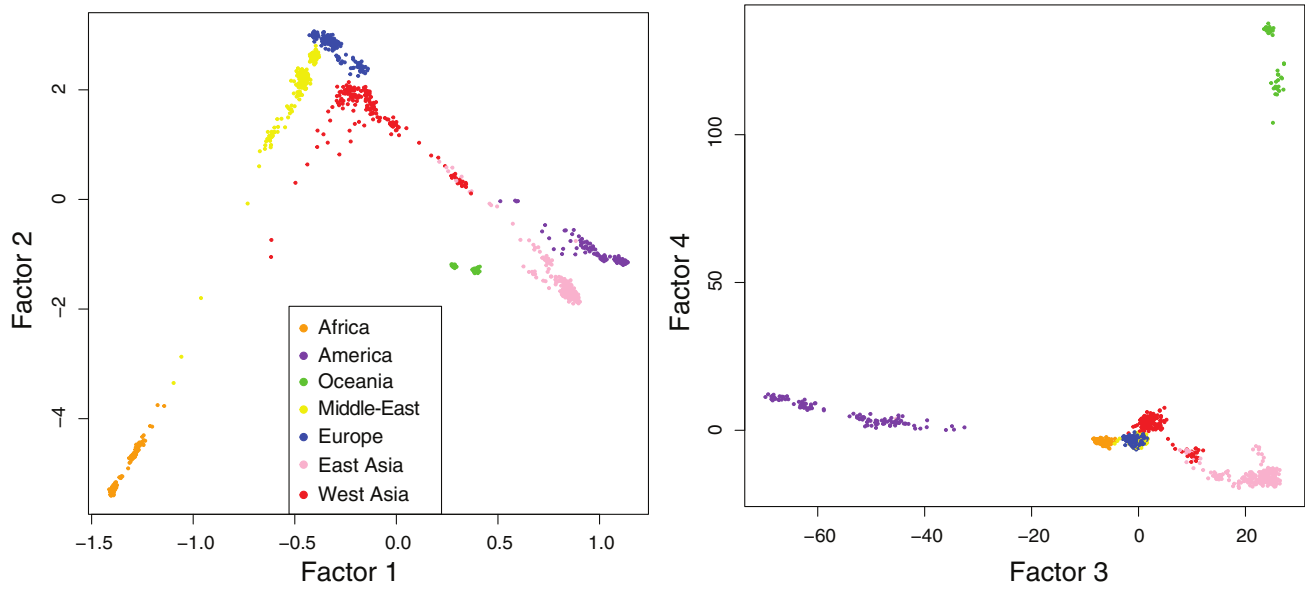


FIG. 5. Values of the $K = 4$ factors for the HGDP data set.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

from Asiatic and Native American individuals and the second factor mainly discriminates African from Middle-Eastern and Western Eurasian individuals. The third factor distinguishes Native American individuals—coming from Central and South America—from the rest of the sample, whereas the last factor separates individuals from Oceania from the rest of the sample (fig. 5). We additionally check that genomic regions of strong linkage disequilibrium do not distort the latent factors. Using window sizes of 100 SNPs and a step size of 10 SNPs, we prune with *PLINK* the SNPs so that there are no more pairs of SNPs with a r^2 larger than 0.2 resulting in a total of 160,257 SNPs (Purcell et al. 2007). Comparing the four latent factors with and without pruning, we find r^2 values larger than 0.93 indicating that LD does not distort the ascertainment of population structure.

We choose to restrict our analysis to the 5,000 top-hit SNPs (supplementary table S1, Supplementary Material online). Their values of the Bayes factors range from 1.03 to 5.05 on a $\log_{10}$ scale. The two SNPs with the largest Bayes factors (rs1834640 and rs2250072) are correlated with the second factor. They are located on chromosome 15 and the closest gene is *SLC24A5*, which is located at 20–30 kb from the SNPs. Among the 5,000 SNPs with largest Bayes factors, 851 are related to factor 1, 844 with factor 2, 1,982 with factor 3, and 1,323 with factor 4. For each of the four sublists, we further provide information for the ten SNPs with the largest Bayes factors (table 1):

- For the first factor, although we consider ten different SNPs, only two genomic regions are found. One of the two genomic regions is located on chromosome 10 and downstream of the oncogene *CYP26A1* whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011; Mallick et al. 2013). The other SNPs were found in the *SM6* gene which is implicated in the structural maintenance of chromosome protein 6 and which has already been picked as a candidate for selection in another scan with the HGDP sample (Hao et al. 2013). For all the ten SNPs, we investigate the worldwide repartition of allele frequencies with the ALFRED database (Rajeevan et al. 2012). East Asiatic and Native American populations have allele frequencies that are different from the rest of the sample (supplementary table S2, Supplementary Material online) as can be predicted when looking at the geographic repartition of the first factor (fig. 5).
- For the outlier SNPs associated with the second factor, the allelic frequencies were mostly different when comparing Western Eurasian individuals with the rest of the sample (supplementary table S2, Supplementary Material online). In addition to the SNPs close to the *SLC24A5* gene that is associated with light skin in Western Eurasia (Canfield et al. 2013), we also find four other regions located close to the following genes: *EDAR* in chromosome 2 which has been associated with various traits including hair thickness and sweating (Kamberov et al. 2013), *SLC35F3* in chromosome 1, *KIF3A* in chromosome 5, *RABGAP1* and *STRBP* in chromosome 9 with the latter being involved in spermatogenesis, and *MYO5C* and *DUT* in chromosome 15.
- For the third factor, eight of the ten SNPs with the largest Bayes factor are found in a 1-Mb region of chromosome 22, which encompasses many different genes (table 1). For the SNPs in this large region of chromosome 22, the allelic frequencies mostly differ between Native Americans and the rest of the sample. For sub-Saharan African populations, allelic frequencies of these SNPs are intermediate with Pygmies populations having frequencies that are often the most similar to the Native Americans (supplementary table S2, Supplementary Material online).
- The allele frequencies of the SNPs that are the most associated with the fourth factor mostly differ between individuals from Oceania (Papuan and Melanesian) and the rest of the sample with Native Americans and Pygmies population having, for some SNPs, allele frequencies that are the most similar to the Oceanians (supplementary table S2, Supplementary Material online). Among the ten outlier SNPs, four SNPs are located in chromosome 8 and four SNPs are located in chromosome 17. Among the 1,323 SNPs associated with the fourth factor, there is an excess of outlier SNPs in chromosome 8 (supplementary fig. S9, Supplementary Material online) pointing to a prominent role of its genes in adaptation to the local conditions of Oceania. There are different genomic regions with large Bayes factors in chromosome 8 and one of these genomic regions encompasses *RP1L1*, a gene often found in selection scans (Barreiro et al. 2008) and related to eye diseases (Davidson et al. 2013).

We also perform a Gene Ontology (GO) enrichment analysis on human genes using the 5,000 SNPs with the largest Bayes factors. We find significant enrichment of biological processes for each of the four factors (supplementary table S3, Supplementary Material online). Some interesting instances of the enriched gene ontologies include three different GO terms related to regulation of hormone secretion for the first factor, enrichment of homophilic cell adhesion for the third factor, and aging for the fourth factor. Finally, we look at a catalog of published GWAS (Welter et al. 2014) to search for enrichment of outlier SNPs related to a particular phenotype (supplementary table S4, Supplementary Material online). The traits that are the most associated with the outlier SNPs are height (6 SNPs), obesity and weight (5 SNPs), and Crohn's disease (5 SNPs).

Discussion

Based on a Bayesian factor model, we provide a new approach to detect loci subject to local adaptation. The hierarchical factor model considers the SNPs that are atypically related to population structure as outliers and candidates for local adaptation. Population structure is captured by a set of K latent factors. Simulations showed that latent factors can adequately describe clustering of individuals into populations (fig. 1), isolation-by-distance patterns, and gradients of selection (fig. 4). Similarly to approaches based on measures of genetic differentiation, the hierarchical factor model is a statistical outlier approach where local adaptation is not

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

Table 1. List of the Ten SNPs with Largest Bayes Factors for Each of the Four Factors Obtained with the HGDP Data Set.

Chromosome	rs Identifier	Physical Position[a]	Closest Gene	Dist.[b]	$Log_{10}(BF)$[c]	Factor No.
chr 10	4918664	94911055	*CYP26A1*	83	2.8	1
chr 10	10882168	94919424	*CYP26A1*	92	2.7	1
chr 10	7091054	95008434	*MYOF*	48	2.7	1
chr 10	11187300	94910281	*CYP26A1*	83	2.6	1
chr 2	7556886	17771611	*SMC6*	0	2.6	1
chr 10	12220128	94965001	*MYOF*	91	2.5	1
chr 10	6583859	94883463	*CYP26A1*	56	2.5	1
chr 10	4918924	94966946	*MYOF*	89	2.4	1
chr 2	1834619	17764966	*SMC6*	0	2.4	1
chr 2	4578856	17716869	*SMC6*	0	2.4	1
chr 15	1834640	46179457	*SLC24A5*	21	5.1	2
chr 15	2250072	46172199	*SLC24A5*	29	4.2	2
chr 2	260714	108928927	*EDAR*	0	3.2	2
chr 1	7531501	232404926	*SLC35F3*	0	2.9	2
chr 15	11637235	46420445	*DUT*	0	2.9	2
chr 9	10760260	124753347	*RABGAP1*	0	2.9	2
chr 9	2416899	125054924	*STRBP*	0	2.8	2
chr 5	2406410	132093779	*KIF3A*	0	2.8	2
chr 15	3751631	50321636	*MYO5C*	0	2.8	2
chr 9	618746	124777344	*RABGAP1*	0	2.8	2
chr 22	139553	40517145	*MEI1*	0	4.1	3
chr 22	5996039	40311903	*PMM1*	0	4.0	3
chr 22	8139993	40325281	*DESI1*	0	4.0	3
chr 22	126092	40508387	*MEI1*	0	4.0	3
chr 22	1005402	39621676	*XPNPEP3*	0	3.6	3
chr 22	8137373	40059162	*ZC3H7B*	0	3.6	3
chr 22	133074	39408419	*MCHR1*	0	3.6	3
chr 22	9611613	40291777	*CSDC2*	0	3.5	3
chr 20	2424641	24665867	*SYNDIG1*	71	3.5	3
chr 14	2600814	46054248	*LINC00871*	13	3.5	3
chr 8	16892216	120271074	*MAL2*	19	2.8	4
chr 8	6990312	110671493	*SYBU*	0	2.8	4
chr 17	9908046	50918781	*MMD*	64	2.7	4
chr 17	575873	39055489	*MEOX1*	18	2.6	4
chr 4	4691075	164468935	*NPY1R*	0	2.6	4
chr 17	4471745	50923883	*MMD*	70	2.5	4
chr 14	12891534	80069114	*CEP128*	0	2.5	4
chr 8	6988341	110653602	*SYBU*	2	2.5	4
chr 8	12216712	9933221	*MSRA*	16	2.5	4
chr 17	11869714	45942306	*MYCBPAP*	0	2.5	4

[a]The positions are given for the NCBI36 assembly.

[b]Dist. is the distance from the closest gene and is measured in kilobase pairs.

[c]$Log_{10}(BF)$ is the logarithm (in base 10) of the Bayes factor.

modeled explicitly. In contrast, another recent statistical approach for selection scans provides a generative model for local adaptation based on a diffusion approximation (Vitalis et al. 2014).

However, there are also important differences with approaches based on measures of genetic differentiation. Compared with the software *BayeScan* or genome scans based on F_{ST}, the factor model does not assume a particular model of population structure. In a model of population divergence, we show that removing the assumptions of the *F* model considerably reduces the FDR. To explain why the factor model generates fewer false discoveries, we introduce the notions of mechanistic and phenomenological models (Hilborn and Mangel 1997). Mechanistic models aim to mimic the biological processes that are thought to have given rise to the data, whereas phenomenological models seek only to best describe the data using a statistical model. In the spectrum between mechanistic and phenomenological models, the *F* model would stand close to mechanistic models, whereas factor models would be closer to the phenomenological ones. Mechanistic models are appealing because they provide quantitative measures that can be related

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

to biologically meaningful parameters. For instance, F statistics measure genetic drift which can be related to migration rates, divergence times, or population sizes. In contrast, phenomenological models work with mathematical abstractions such as latent factors that can be difficult to interpret biologically. The downside of mechanistic models is that violation of the modeling assumption can invalidate the proposed framework and generate many false discoveries in the context of selection scans. The F model assumes a particular covariance matrix between populations which is found with star-like population trees for instance (Bonhomme et al. 2010). However, more complex models of population structure can arise for various reasons including non-instantaneous divergence and isolation-by-distance, and they will violate the mechanistic assumptions and make phenomenological models preferable.

Although PCA or the related factor model is generally used to investigate population structure, there have already been several attempts at performing selection scans based on these statistical approaches. The first idea is to compute F_{ST} values between pairs of populations that contain the top and bottom individuals for each principal component (Abdellaoui et al. 2013). This approach provides a list of outliers that are specific to each principal component in the same way as the hierarchical factor model of equations (2)–(4) provides outliers that are related to one of the K factors. A second proposition involves new interpretations of PCA related to F statistics, which provide statistical measures to detect local adaptation (Laloë and Gautier 2011). A last and recent proposition called "logistic factor analysis" adds a logistic link function to the factor model (eq. 2) in order to guarantee that the predicted values can be interpreted as frequencies because they lie between 0 and 1 (Hao et al. 2013). Loci involved in biological adaptation were scanned using a deviance statistic (Hao et al. 2013). These related approaches are built on the success of PCA and factor models to capture population structure with a small number of variables.

Choosing the dimension K of the statistical model that ascertains population structure is a recurrent problem. One possibility is to use an *objective* approach based on a quantitative criterion. Examples of such objective criteria include the Δ_K measure to detect the number of clusters using the software STRUCTURE (Evanno et al. 2005) or the Tracy–Widom statistic to choose the dimension K in PCA (Patterson et al. 2006). Another possibility is to adopt a *subjective* approach and to choose a value of K such that increasing K would provide results that are considered to be of too little interest. With the proposed Bayesian factor model, we implemented both approaches. For the simulations, choosing K based on the MSE of equation (2) works well because the MSE stops to decrease when K increases beyond a certain value. However, for the human data, the choice is more complex because the MSE decreases regularly as K increases. We chose $K = 4$ because we were only interested in biological adaptation that is related to genetic differentiation between continents, but we acknowledge that major adaptive processes also occur within continents (Jarvis et al. 2012). The subjective choice of K is related to the choice of pairwise comparisons when using F_{ST} between pairs of populations for genome scans (Nosil et al. 2008). Choosing which populations should be compared or which latent factors should be kept pertains to the biological questions addressed by the genome scan. To provide recommendations for choosing K, we suggest to fit the hierarchical factor model with different values of K in order to investigate if there is a value of K, at which the MSE stops or almost stops to decrease. If not, the choice of K can be based on subjective arguments where the latent factors of too little interest can be discarded.

Although the proposed factor model provides fewer false discoveries than approaches based on F statistics, there remain caveats and possible improvements. First, one of our objectives was to propose a method for selection scans that avoids the computational burden of some Bayesian approaches, which can become a serious obstacle when analyzing large-scale SNP data. However, this objective is only partly fulfilled. The downside of our approach is that it relies on a Markov chain Monte Carlo (MCMC) algorithm that quickly grinds to a halt under the sheer mass of SNP data (Lange et al. 2014). Fortunately, the MCMC algorithm is based on a Gibbs sampler that alternates the computation of least square solutions, which are fast to compute. For the HGDP data set (644,199 SNPs), the run-time ranges from 13 to 16 h using a single computer processor (2.4 GHz, 64-bit Intel Xeon) when K increases from 1 to 8. Second, we considered a particular outlier model (eq. 4) for the sake of interpretability which assumes that outliers should be atypically explained by one latent factor. However, other outlier models would be possible, for instance, by imposing a sparsity prior that forces most of the factor loadings to be null (Carvalho et al. 2008) or by assuming that outliers should be atypically related to the K latent factors and not to only one of them. For handling the two mentioned caveats, we are currently working on the development of a faster version of our software where various outlier models are evaluated. Third, we investigate the effect of scaling the data matrix Y and find that it can increase or decrease the FDR because of two opposing effects. The advantage of the scaling is that it makes the regression coefficients (the factor loadings) more comparable because all centered allele frequencies are at the same scale. The downside of scaling is that it gives more importance to the low-frequency variants (supplementary fig. S7, Supplementary Material online) for which the statistical estimates of population differentiation are highly variable (Meirmans and Hedrick 2011). The fourth caveat occurs if local adaptation has acted in directions that do not align with the underlying population structure. For example, in investigating worldwide convergent evolution of high-altitude adaptation in humans, it is unlikely that a latent factor will separate high-altitude populations (Tibet, Andes, Ethiopia) from lowland populations. An approach incorporating an environmental variable, that is, a fixed factor, corresponding to altitude would be more appropriate (Foll et al. 2014). The fifth caveat concerns the choice of the threshold for the Bayes factor. When analyzing the HGDP data, we use a threshold of 10 which corresponds to strong evidence for the outlier model according to

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

the Jeffrey's scale of evidence for Bayes factors, but other choices would have also been possible (Kass and Raftery 1995). The last caveat concerns the inability to pinpoint the geographic location at which adaptation took place. Although outlier SNPs are related to latent factors, the factor model does not provide the population or group of individuals concerned by adaptation, whereas alternative statistical models can estimate selection coefficients for each population separately (Vitalis et al. 2014).

Although the purpose of our article is to present a novel method and not to provide an in-depth analysis of local adaptation in the HGDP, fitting the factor model to the HGDP data provides interesting results. The first two latent factors mainly measure differentiation between Africa, Western Eurasia, and East Asia and some outliers related with these two factors are involved in morphological traits, which are often reported to be enriched with genes having signatures of positive selection (Barreiro et al. 2008). In the list of outliers, we found the genes *SLC24A5* and *EDAR*, which are often reported as top hits in selection scans (Pickrell et al. 2009; Hao et al. 2013) and are related to skin pigmentation and hair thickness, respectively (Kamberov et al. 2013; Mallick et al. 2013). We also found that SNPs close to the oncogene *CYP26A1*, whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011), are part of the top list for outliers. The third and fourth factors correspond to genetic differentiation between Native Americans, individuals from Oceania, and the rest of the sample. There are many regions in chromosome 8 enriched with outlier SNPs. One of these regions encompasses the gene *RP1L1* that is associated with retinal diseases and which has already been reported to have one of the strongest signatures of positive selection along with other genes related to sensory functions (Barreiro et al. 2008). Many outlier SNPs strongly related to the third and fourth factors have allele frequencies that are similar between Southern Native Americans and Pygmies (third factor) or between individuals from Oceania, Southern Native Americans, and Pygmies (fourth factor). Because these individuals all live in tropical rain forests and have similar diet consisting of roots and tubers, our findings support the importance of diet, climate, and potentially pathogen load to explain human adaptation (Hancock et al. 2010; Fumagalli et al. 2011). The SNPs with similar allele frequencies in different geographic regions are good candidates for convergent evolution and would deserve further analysis.

Factor models are enriching the toolbox of population genetic methods. The main principle is to model population structure via latent variables called factors. Factors models have already been proposed to ascertain population structure (Engelhardt and Stephens 2010) and to account for population structure when testing for gene-environment association (Frichot et al. 2013). We showed that factor models also provide a convenient individual-based framework to find loci that have atypical patterns of genetic differentiation. A major argument supporting the proposed hierarchical factor model is that it produces fewer false discoveries compared with genome scans based on F_{ST}.

Materials and Methods

Hierarchical Bayesian Modeling

We provide the prior distributions for the latent variables of the hierarchical factor model defined by equations (2)–(4). To account for linkage disequilibrium in the genome and encourage outlier loci to be clustered along the genome, we consider a Potts model with an external field (Winkler 2003)

$$p(z_1,\ldots,z_p) \propto (1-\pi)^{p_0}\pi^{(p-p_0)}e^{\beta\sum_{i\sim j}1_{z_i=z_j}}, \qquad (5)$$

where the sum in the exponential ranges over all pairs of neighboring loci. We consider that each locus has two neighbors except at the beginning and at the end of the chromosome where a locus has only one neighbor. In equation (5), the variable p_0 is the number of loci such that $z_i = 0$, 1 is the indicator function, β is the parameter of the Potts model and is set to $\beta = 1$, and π is the prior proportion of outliers. To model the proportion of outliers, we consider a uniform prior on the $\log_{10}$ scale reflecting that we are interested in the order of magnitude of the proportion of outlier loci (Guan and Stephens 2011). In the following, we consider -4 and -1 for the lower and upper bound of the uniform prior, respectively. The proportion of loci under selection is a priori expected to be between 10^{-1} and 10^{-4} reflecting the working hypothesis that most loci are neutral and that loci under selection are rare. For the variance parameters σ_k^2, $k = 1,\ldots,K$, that are specific to each factor (eqs. 3 and 4), we consider the parameterization $\sigma_k^2 = \sigma^2\rho_k^2$, where σ^2 is the residual variance in equation (2) (Oba et al. 2003). We consider the non-informative prior for the variance parameters $p(\sigma^2) \propto 1/\sigma^2$ and $p(\rho_k^2) \propto 1/\rho_k^2$, $k = 1,\ldots,K$. For the variance–inflation parameters c_k^2 of equation (4), we consider uniform priors with 1 and 10 for the lower and upper bounds, and we find that a slight variation of the upper bound does not change the ranking of the SNPs (results not shown).

With the factor model of equation (2) as with many models with latent structure (Allman et al. 2009), there is a well-known issue of identifiability because identical likelihood values can be obtained from a solution ($\mathbf{U}$,$\mathbf{V}$) after using orthogonal rotations (West 2003). To add constraints to the model, we consider a prior with unit variance for each of the factors

$$\mathbf{U}_k \sim \mathcal{N}(0,\mathbf{I_n}),$$

where $\mathbf{I_n}$ is the squared $n \times n$ identity matrix (Oba et al. 2003). To further prevent the MCMC algorithm to produce alternative rotations of the factors (Engelhardt and Stephens 2010), we consider the solution of the singular value decomposition as starting values for the factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$ in the MCMC algorithm.

To evaluate the strength of evidence for outlyingness at each locus, we compute the Bayes factor on a $\log_{10}$ scale. The Bayes factor is defined as the ratio of the posterior odds $p(z_\ell > 0 \mid \mathbf{Y})/p(z_\ell = 0 \mid \mathbf{Y})$ and the prior odds

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

$p(z_\ell > 0)/p(z_\ell = 0) = \pi/(1-\pi)$. The description of the MCMC algorithm and the computation of the Bayes factor is given in the Supplementary Material online.

Simulation of the Four-Population Divergence Model

The first simulation scenario is a divergence model with four populations. These populations have constant effective population sizes of $N_e = 1{,}000$ diploid individuals, with 50 individuals sampled in each population. The genotypes consist of 10,000 independent SNPs. The simulations are performed in two steps. In the first step, we use the software *ms* to simulate a neutral divergence model (Hudson 2002). When looking backward in time, we instantly merge population A_1 with A_2 and population B_1 with B_2, then after waiting a number $T = 20$, 80, 120, 160, 200 of generations, we merge the two remaining populations A and B. We keep only variants with a minor allele frequency larger than 5% at the end of this first step. The second step is performed with the software *SimuPOP* (Peng and Kimmel 2005). To run *SimuPOP*, we provide the allele frequencies in each of the four populations that have been generated with *ms*. Looking forward in time, we simulate 100 generations after the 2 concomitant divergence events. We assume no migration between populations. In each evolutionary lineage, we assume that 100 SNPs confer selective advantage using a selection coefficient of $s = 0.1$ for homozygotes carrying two adaptive alleles. In both simulation schemes, we assume an additive model for selection.

Simulation of the Stepping-Stone Model

The second simulation scenario is a two-dimensional stepping-stone model with a 10×10 grid. Each of the 100 populations has an effective population size of $N_e = 1{,}000$ diploid individuals. We sample ten individuals in each population and there are 2,050 independent SNPs. We also consider a two-step procedure for the simulations. First, we simulate an equilibrium stepping-stone model with the software *ms*. Neighboring populations exchange migrants with a rate of $4N_em = 8$ per generation. Then we superimpose a selection gradient using *SimuPOP*. During 100 generations, we consider that 50 SNPs confer selective advantage. The selection coefficient $s = 0.1$ is maximal in population 64, which is located in the lower-right quarter of the grid. In the eight neighboring populations, the selection coefficient is $s = 0.05$, and in the second layer of neighbors, the selection coefficient is $s = 0.025$. The selection coefficient is equal to 0 for the rest of the grid.

False Discovery Rate and Sensitivity

To compare the performances of the different methods for selection scans, we compute the FDR for fixed values of the sensitivity. Denoting the number of false positives by FP, the number of false negatives by FN, and the number of true positives by TP, the FDR is defined as FP/(FP + TP) and the sensitivity (also called recall rate) is defined as TP/(TP + FN).

Gene Ontology Analysis

We perform a GO enrichment analysis with the software *Gowinda* (Kofler and Schlötterer 2012). The list of genes is built by considering all genes which contain outlier SNPs with a tolerance of 5,000 bp upstream and downstream. We use a threshold of 0.05 for the FDR, and we remove GO terms that are shared by less than 10 genes or more than 1,000 genes. We consider the *–snp* flag in *Gowinda* that assumes independence of SNPs within a gene. If we rather use the *–gene* flag, which assumes complete dependence of SNPs within a gene, there is no GO discovery with an FDR smaller than 5%. For each factor, we consider the ten GO terms with the smallest FDRs, and we only report GO terms that are related to biological processes.

Software Availability

The computer program PCAdapt for fitting the factor model is available from the authors' websites (http://membres-timc.imag.fr/Michael.Blum/, http://membres-timc.imag.fr/Nicolas.Duforet-Frebourg/, last accessed June 12, 2014).

Supplementary Material

Supplementary material, figures S1–S9, and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

This work has been supported by the French National Research Agency (DATGEN project, ANR-2010-JCJC-1607-01) and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

Abdellaoui A, Hottenga J-J, de Knijff P, Nivard MG, Xiao X, Scheet P, Brooks A, Ehli EA, Hu Y, Davies GE, et al. 2013. Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet.* 21:1277–1285.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.

Allman ES, Matias C, Rhodes JA. 2009. Identifiability of parameters in latent structure models with many observed variables. *Ann Stat.* 37: 3099–3132.

Antao T, Lopes A, Lopes R, Beja-Pereira A, Luikart G. 2008. Lositan: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinformatics* 9:323.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Bazin E, Dawson KJ, Beaumont MA. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* 185:587–602.

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol.* 13:969–980.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London B Biol Sci.* 263: 1619–1626.

Bierne N, Roze D, Welch JJ. 2013. Pervasive selection or is it. . .? Why are FST outliers sometimes so frequent? *Mol Ecol.* 22 2061–2064.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M. 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186:241–262.

Bonin A, Taberlet P, Miaud C, Pompanon F. 2006. Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol.* 23: 773–783.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

Box GE, Tiao GC. 1968. A Bayesian approach to some outlier problems. *Biometrika* 55:119–129.
Canfield VA, Berg A, Peckins S, Wentzel SM, Ang KC, Oppenheimer S, Cheng KC. 2013. Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3 (Bethesda)* 3:2059–2067.
Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. 2008. High-dimensional sparse factor modeling: applications in gene expression genomics. *J Am Stat Assoc.* 103:1438–1456.
Davidson AE, Sergouniotis PI, Mackay DS, Wright GA, Waseem NH, Michaelides M, Holder GE, Robson AG, Moore AT, Plagnol V, et al. 2013. RP1L1 variants are associated with a spectrum of inherited retinal diseases including retinitis pigmentosa and occult macular dystrophy. *Hum Mutat.* 34:506–514.
Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.
Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* 6:e1001117.
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.
Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193:929–941.
Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977–993.
Foll M, Gaggiotti OE, Daub JT, Excoffier L. 2014. Hierarchical Bayesian model of population structure reveals convergent adaptation to high altitude in human populations. arXiv preprint arXiv:1402.4348.
Fourcade Y, Chaput-Bardy A, Secondi J, Fleurant C, Lemaire C. 2013. Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. *Mol Ecol.* 8: 2065–2073.
Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol.* 30:1687–1699.
Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.
Gompert Z, Buerkle CA. 2011. A hierarchical Bayesian model for next-generation population genomics. *Genetics* 187:903–917.
Guan Y, Stephens M. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat.* 5:1780–1815.
Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195:205–220.
Guo F, Dey DK, Holsinger KE. 2009. A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *J Am Stat Assoc.* 104:142–154.
Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, et al. 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A.* 107:8924–8930.
Hao W, Song M, Storey JD. 2013. Probabilistic models of genetic variation in structured populations applied to global human studies. arXiv preprint arXiv:1312.2041.
Hilborn R, Mangel M. 1997. The ecological detective: confronting models with data, Vol. 28. Princeton (NJ): Princeton University Press.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, Froment A, Bodo J-M, Beggs W, Hoffman G, et al. 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet.* 8:e1002641.
Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected *EDAR* variant. *Cell* 152:691–702.
Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc.* 90:773–795.
Kofler R, Schlötterer C. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* 28: 2084–2085.
Laloë D, Gautier M. 2011. On the genetic interpretation of between-group PCA on SNP data. HAL hal-00661214.
Lange K, Papp JC, Sinsheimer JS, Sobel EM. 2014. Next-generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Ann Rev Stat Appl.* 1:279–300.
Lewontin R, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175–195.
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet.* 4:981–994.
Mallick CB, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Romero IG, Crivellaro F, et al. 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet.* 9:e1003912.
Manel S, Holderegger R. 2013. Ten years of landscape genetics. *Trends Ecol Evol.* 28:614–621.
Meirmans PG, Hedrick PW. 2011. Assessing population structure: Fst and related measures. *Mol Ecol Resour.* 11:5–18.
Nicholson G, Smith AV, Jónsson F, Gústafsson.Ó, Stefánsson K, Donnelly P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc Series B Stat Methodol.* 64:695–715.
Nosil P, Buerkle A. 2010. Population genomics. *Nat Educ Knowl.* 3:8.
Nosil P, Egan SP, Funk DJ. 2008. Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. *Evolution* 62:316–336.
Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.* 40:646–649.
Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S. 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19:2088–2096.
Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci.* 365: 185–205.
Osanai M, Lee G-H. 2011. Enhanced expression of retinoic acid-metabolizing enzyme *CYP26A1* in sunlight-damaged human skin. *Med Mol Morphol.* 44:200–206.
Patterson N, Price AL, Reich D. 2006. Population structure and eigen analysis. *PLoS Genet.* 2:e190.
Peng B, Kimmel M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687.
Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559–575.
Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK. 2012. ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res.* 40:D1010–D1015.
Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178:1817–1829.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the drosophila genome? *PLoS Genet.* 5:e1000495.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

Vitalis R, Dawson K, Boursot P. 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158: 1811–1823.

Vitalis R, Dawson K, Boursot P, Belkhir K. 2003. DetSel 1.0: a computer program to detect markers responding to selection. *J Hered.* 94: 429–431.

Vitalis R, Gautier M, Dawson KJ, Beaumont MA. 2014. Detecting and measuring selection from gene frequency data. *Genetics* 196: 799–817.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42:D1001–D1006.

West M. 2003. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.* 7:723–732.

Winkler G. 2003. Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction. Vol. 27. Berlin: Springer.

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

Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, Yang L, Pan X, Wang J, Shen Y, et al. 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol.* 28:1003–1011.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen JE, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329: 75–78.

Downloaded from http://mbe.oxfordjournals.org/ by guest on July 1, 2014

Supplementary Material of "Genome scans for detecting footprints of local adaptation using a Bayesian factor model."

- Description of the MCMC algorithm and of the computation of the Bayes factors.
- Figure S1: Population divergence model: false discovery rate as a function of the initial divergence time when the sensitivity (recall) has been set to 25% and 75%.
- Figure S2: Population divergence model: false discovery rate obtained with the factor model using different values of the number of factors K.
- Figure S3: Two example models and HGDP: average mean square error as a function of the number of factors K.
- Figure S4: IBD model: false discovery rate as a function of the number of factors K.
- Figure S5: IBD model: false discovery rate as a function of the number of sampled individuals per deme.
- Figure S6: Population divergence and IBD model: false discovery rate with or without scaling the data matrix Y.
- Figure S7: IBD model: minimum allele frequency of the top SNPs with or without scaling the data matrix Y.
- Figure S8: HGDP dataset: Plot of factors 5-8.
- Figure S9: HGDP dataset: Number of outlier SNPs per chromosome.
- Table S1: HGDP dataset: List of the 5,000 SNPs with largest Bayes factors (csv file).
- Table S2: HGDP dataset: List of the 4×10 SNPs with largest Bayes factors and links to the *HGDP BROWSER* database to display allele frequencies.
- Table S3: Results of the Gene Ontology based on the set of genes found in Table S1.
- Table S4: List of GWAS SNPs that are found for the outlier SNPs (csv file).

MCMC algorithm

We model the variance parameter of the regression coefficients using the parametrization $\sigma_k^2 = \sigma^2 \rho_k^2$. In the following, we consider $\Sigma_V = \sigma^2 \times \text{Diag}(\rho_1^2, \cdots, \rho_K^2)$ where $\text{Diag}(\rho_1^2, \cdots, \rho_K^2)$ is a diagonal matrix with $(\rho_1^2, \cdots, \rho_K^2)$ in the diagonal. $\Sigma_{V,k0}$ is the same as Σ_V except that the k_0^{th} element in the diagonal is $\sigma^2 c_{k_0}^2 \rho_{k_0}^2$ instead of $\sigma^2 \rho_{k_0}^2$. For a given value of K, we describe below a single step of the MCMC algorithm where we use Gibbs updating steps except when updating $c_1^2, \ldots, c_K^2$ and π.

- $$\mathbf{V}_\ell \leftarrow \mathcal{N}(\mathbf{m}_\ell, \Sigma_\ell),\ \ell = 1, \ldots, p,$$
  with $\mathbf{m}_\ell = (\sigma^2 \Sigma_{V,k_0}^{-1} + \mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}\mathbf{Y}_\ell)$ and $\Sigma_\ell = (\Sigma_{V,k_0}^{-1} + \frac{1}{\sigma^2}\mathbf{U}^T\mathbf{U})^{-1}$.

- $$\rho_k^2 \leftarrow IG(\frac{p}{2}, \frac{1}{2\sigma^2}\sum_{\ell=1}^{p} \frac{V_{k\ell}^2}{(c_k^2)^{\mathbb{1}_{z_\ell = k}}}),$$
  where $\mathbb{1}$ is the indicator function and IG is the inverse-gamma distribution.

- $$\sigma^2 \leftarrow IG(\frac{(n+K)p}{2}, \frac{1}{2}\sum_{\ell=1}^{p}(\sum_{i=1}^{n} |Y_{i,\ell} - U_i V_\ell|^2 + \sum_{k=1}^{K} \frac{V_{k\ell}^2}{\rho_k^2 (c_k^2)^{\mathbb{1}_{z_\ell = k}}})),$$

- $$\mathbf{U}_i \leftarrow \mathcal{N}(\mathbf{m}_i, \Sigma),\ i = 1, \ldots, n,$$
  with $\mathbf{m}_i = (I_K + \frac{1}{\sigma^2}\mathbf{V}\mathbf{V}^T)^{-1}\frac{1}{\sigma^2}\mathbf{V}\mathbf{Y}_i$ and $\Sigma = (\frac{1}{\sigma^2}\mathbf{V}\mathbf{V}^T + I_K)^{-1}$. The $\mathbf{U}_i$'s are vectors of dimension K and are the row vectors of the matrix $\mathbf{U}$. Note that in the main text, we rather refer to the column vectors of the matrix $\mathbf{U}$ when using the notation $\mathbf{U}_i$.

- $$z_\ell = k \text{ with a probability } p(z_\ell = k | \pi, V_l, \Sigma_V, c_k^2),\ k = 1, \ldots, K$$
  with
  $$p(z_\ell = k | \pi, V_l, \Sigma_V, c_k^2) \propto \frac{\pi}{K} e^{\beta \sum_{i \sim \ell} \mathbb{1}_{z_i, z_\ell}} \frac{1}{c} e^{\frac{V_{k\ell}^2}{2\sigma^2\rho_k^2}(1 - 1/c_k^2)}, \quad (1)$$
  and $p(z_\ell = 0 | \pi, V_l, \Sigma_V, c_k^2) \propto (1 - \pi) e^{\beta \sum_{i \sim \ell} \mathbb{1}_{z_i, z_\ell}}$ for $k = 0$.

- Update $(c_1^2, \ldots, c_K^2)$ using a Metropolis-Hasting step where the proposal for each element is a Gaussian random walk of variance equal to 1.

- Update $\log_{10}(\pi)$ using a Metropolis-Hasting step where the proposal is a Gaussian random walk of variance equal to 0.5.

We find that MCMC runs of 400 iterations (burn-in of 200 iterations) are long-enough to provide convergence.

Bayes factors

For each SNP, the Bayes factor is defined as the ratio between the posterior odds (outlier vs non-outlier) and the prior odds $\pi/(1-\pi)$. For numerical convenience, we rather compute the logarithm (in base 10) of the Bayes factor using a Monte-Carlo approximation. The $\log_{10}$ of the posterior odds (LPO) can be computed along the MCMC run by averaging

$$LPO = \log_{10}\{\sum_{k=1}^{K} p(z_\ell = k|\cdot)\} - \log_{10}\{p(z_\ell = 0|\cdot)\},\ \ell = 1, \ldots, p. \tag{2}$$

where $p(z_\ell|\cdot)$ is the conditional probability of z_ℓ given the other parameter values. Because the maximum of the probabilities $p(z_\ell = 1|\cdot), \cdots, p(z_\ell = K|\cdot)$ is on a much larger scale compared to the other probabilities, we approximate equation (2) using

$$LPO \sim \log_{10}\{\max_{k=1,\ldots,K} p(z_\ell = k|\cdot)\} - \log_{10}\{p(z_\ell = 0|\cdot)\},$$

and the conditional probabilities $p(z_\ell|\cdot)$ are given in the equation (1) of the supplementary material.

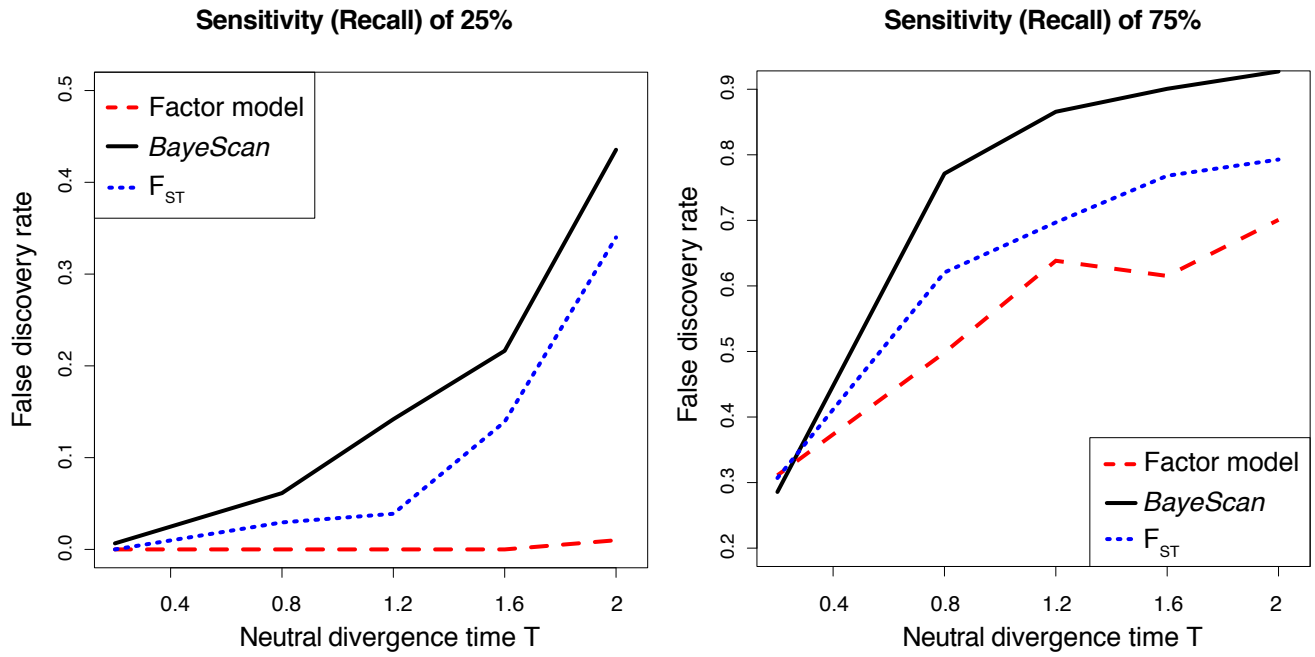


Figure S1: False discovery rate as function of the initial divergence time in the population divergence model depicted in the Figure 1 of the main text. For both *BayeScan* and the proposed factor model, Bayes factors are used for ranking SNPs whereas we use F_{ST} values for genome scans based on F_{ST} values. To determine a threshold above which SNPs are considered as outlier, we constrain the lists of top-ranked SNPs provided by each method to contain 25%(left panel) or 75% (right panel) of the 400 SNPs truly involved in local adaptation. Time is counted in units of time during which there is selection.

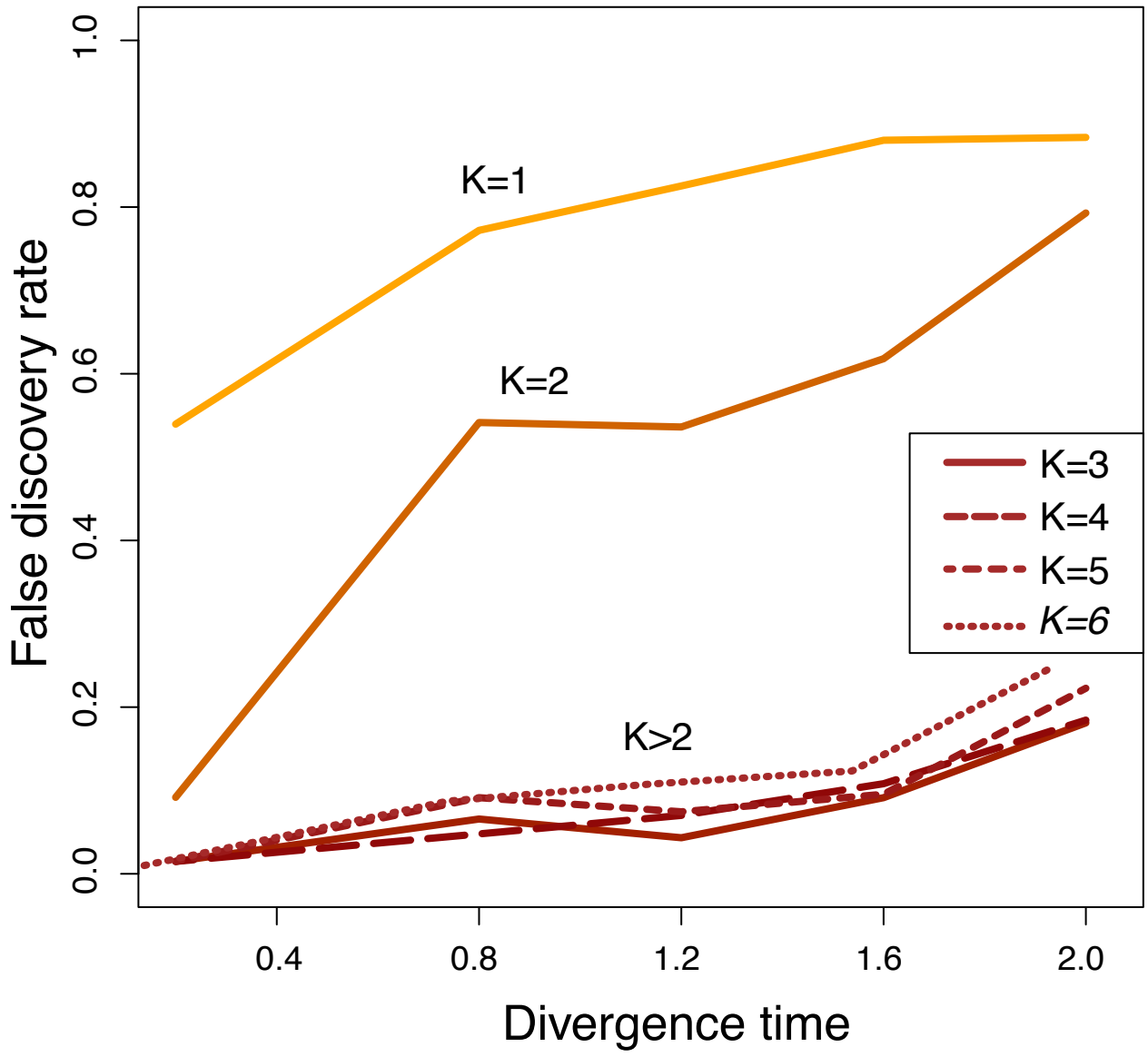


Figure S2: False discovery rate obtained with the factor model using different values of the number of factors K. Data are simulated with the population divergence model of the Figure 1 of the main text. To determine a threshold above which SNPs are considered as outlier, we constrain the lists of top-ranked SNPs provided by each method to contain 50% of the 400 SNPs truly involved in local adaptation.

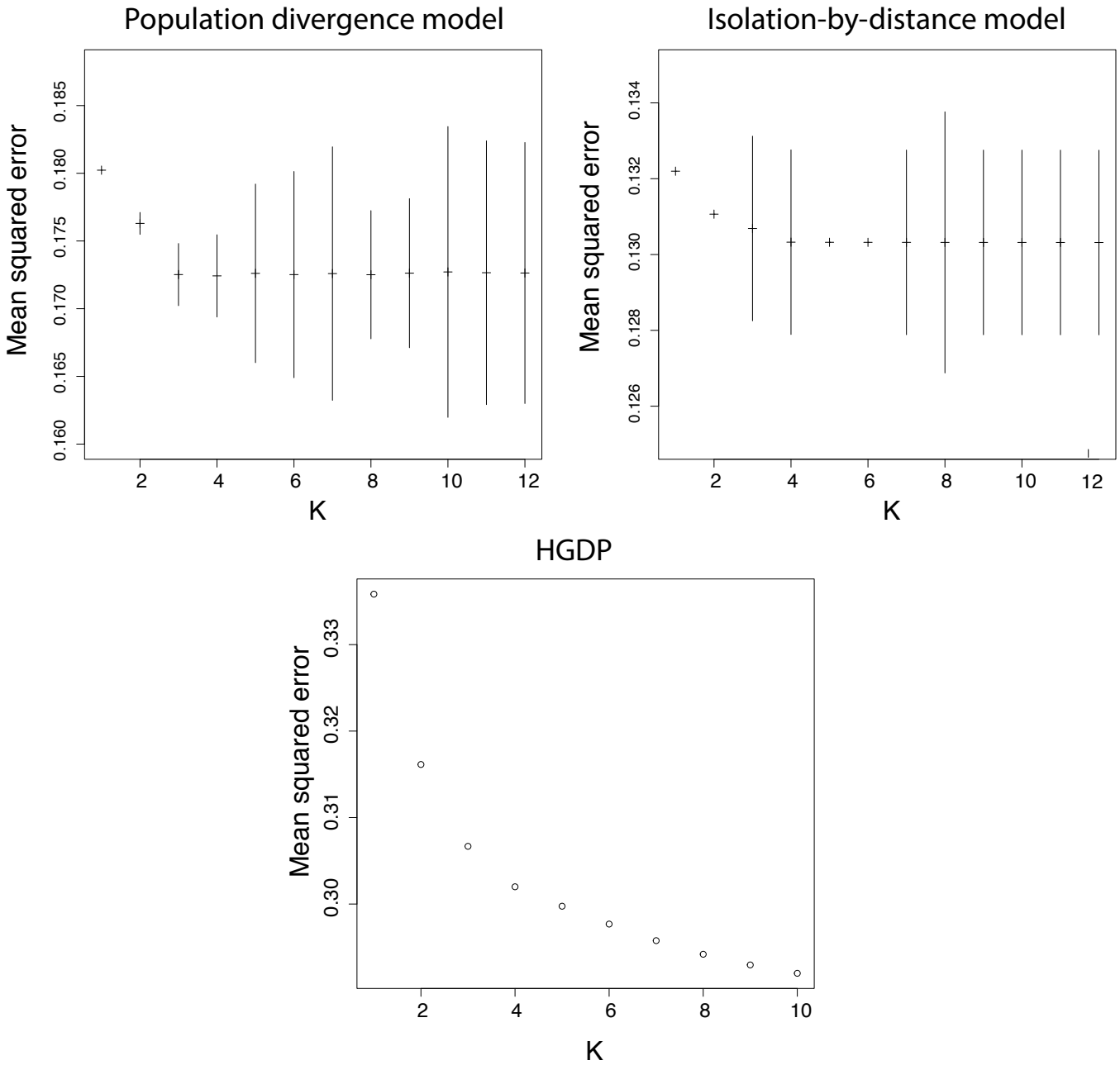


Figure S3: Average mean square error as a function of the number of factors K. Error bars correspond to 2 standard deviations and are estimated using a total of 10 different MCMC runs.

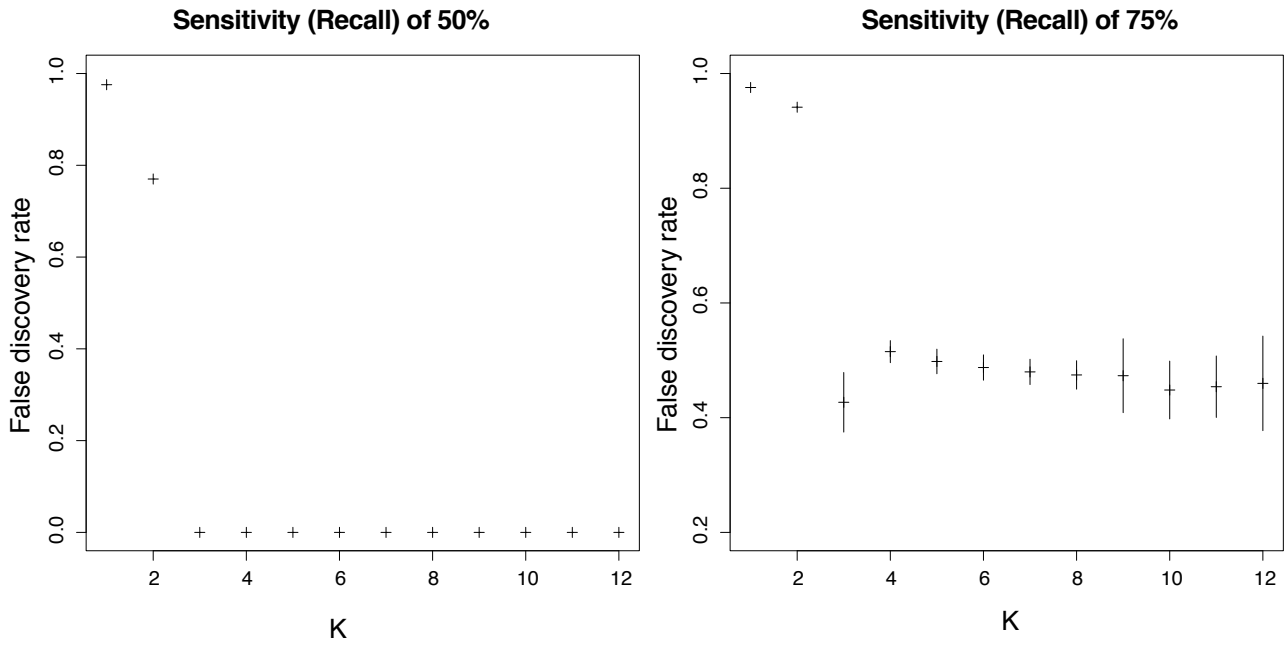


Figure S4: False discovery rate obtained with the factor model using different values of the number of factors K in the IBD model. Error bars correspond to 2 standard deviations and are estimated using a total of 10 different MCMC runs.

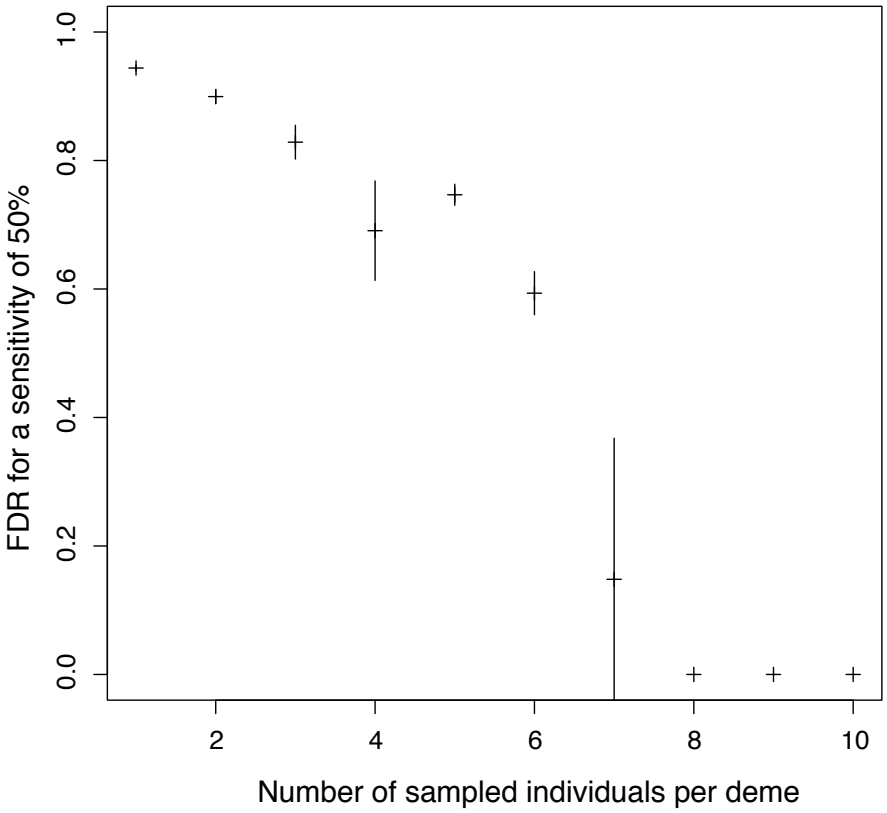


Figure S5: False discovery rate as a function of the number of sampled individuals per deme for the IBD model.

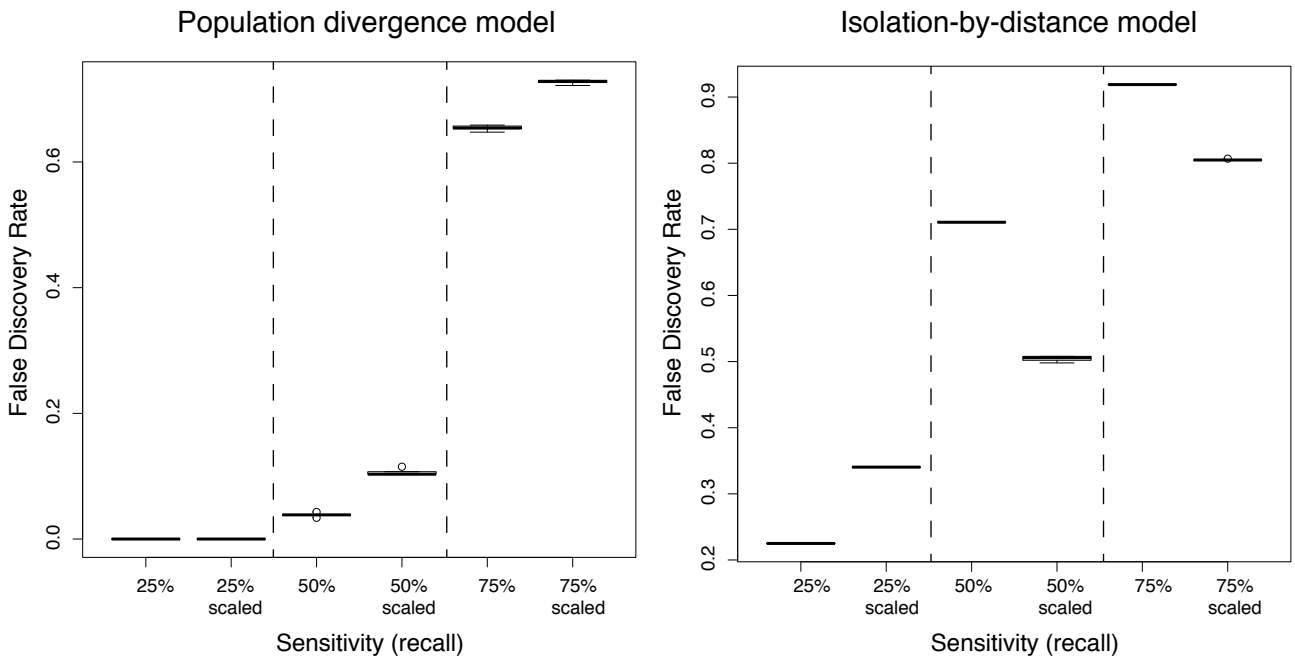


Figure S6: False discovery rate with or without scaling the data matrix Y. Each bar plot corresponds to the FDR obtained after running 5 different runs of MCMC. In the IBD model, the number of SNPs was increased compared to the simulations of the main text because we consider a total of 10,000 SNPs with 250 SNPs under selection.

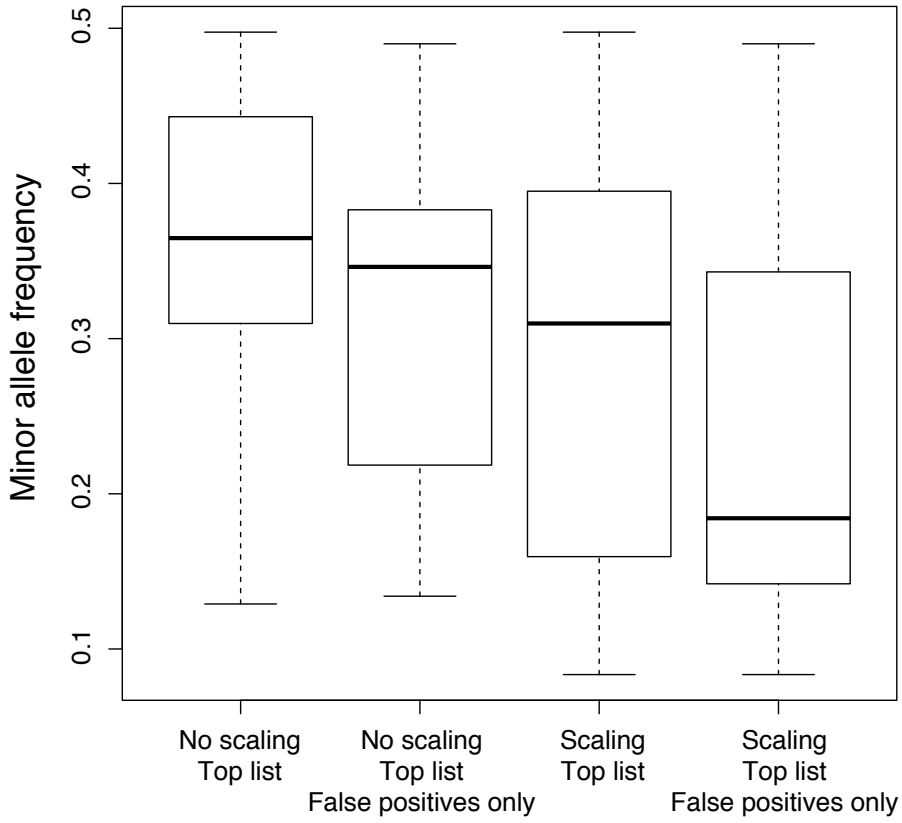


Figure S7: Minimum allele frequency of the top SNPs with or without scaling the data matrix Y. The simulations were performed under the IBD model; the number of SNPs was increased compared to the simulations of the main text because we consider a total of 10,000 SNPs with 250 SNPs under selection. The sensitivity (recall) is set at 50%.

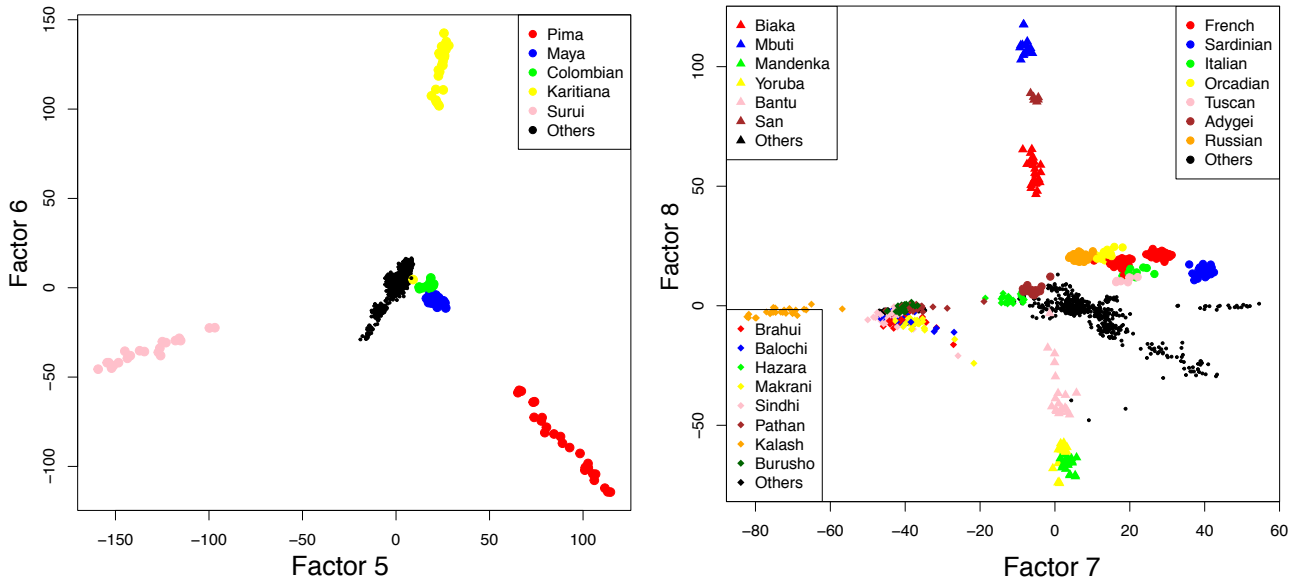


Figure S8: Values of factors $5-8$ for the HGDP dataset. Factors 5 and 6 correspond to differentiation within the Americas, factor 7 corresponds to differentiation within Eurasia and factor 8 corresponds to differentiation within Africa.

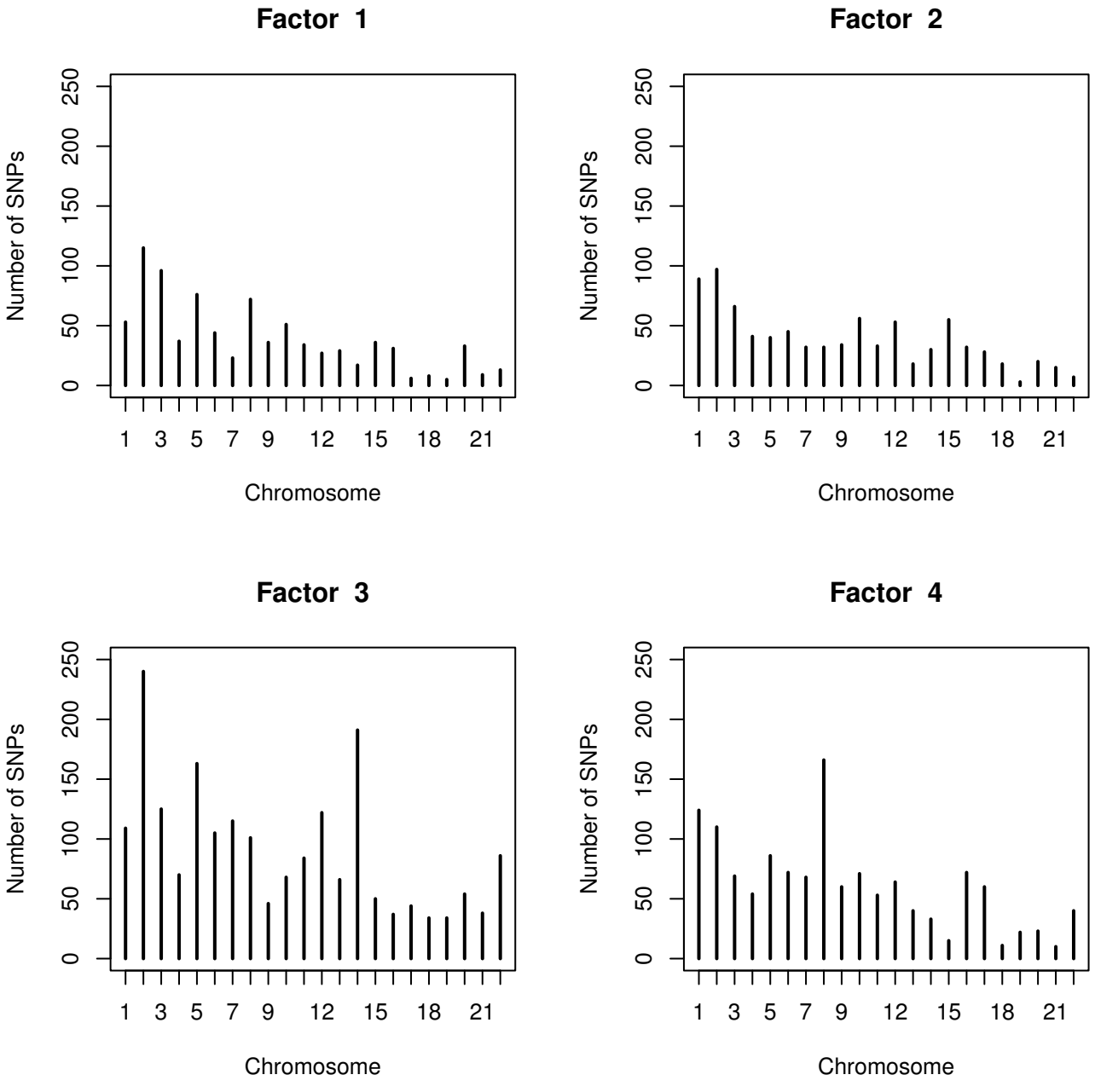


Figure S9: Number of outlier SNPs per chromosome for each of the four different factors.

rs identifier	Factor	Freq. Map
4918664	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94911055&chr=chr10&rs=rs4918664&imp=false
10882168	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94919424&chr=chr10&rs=rs10882168&imp=false
7091054	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=95008434&chr=chr10&rs=rs7091054&imp=false
11187300	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94910281&chr=chr10&rs=rs11187300&imp=false
7556886	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=17771611&chr=chr2&rs=rs7556886&imp=false
12220128	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94965001&chr=chr10&rs=rs12220128&imp=false
6583859	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94883463&chr=chr10&rs=rs6583859&imp=false
4918924	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94966946&chr=chr10&rs=rs4918924&imp=false
1834619	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=17764966&chr=chr2&rs=rs1834619&imp=false
4578856	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=17716869&chr=chr2&rs=rs4578856&imp=false
1834640	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46179457&chr=chr15&rs=rs1834640&imp=false
2250072	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46172199&chr=chr15&rs=rs2250072&imp=false
260714	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=108928927&chr=chr2&rs=rs260714&imp=false
7531501	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=232404926&chr=chr1&rs=rs7531501&imp=false
11637235	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46420445&chr=chr15&rs=rs11637235&imp=false
10760260	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=124753347&chr=chr9&rs=rs10760260&imp=false
2416899	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=125054924&chr=chr9&rs=rs2416899&imp=false
2406410	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=132093779&chr=chr5&rs=rs2406410&imp=false
3751631	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=50321636&chr=chr15&rs=rs3751631&imp=false
618746	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=124777344&chr=chr9&rs=rs618746&imp=false
139553	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40517145&chr=chr22&rs=rs139553&imp=false
5996039	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40311903&chr=chr22&rs=rs5996039&imp=false
8139993	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40325281&chr=chr22&rs=rs8139993&imp=false
126092	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40507762&chr=chr22&rs=rs6002481&imp=false
1005402	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=39621676&chr=chr22&rs=rs1005402&imp=false
8137373	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40059162&chr=chr22&rs=rs8137373&imp=false
133074	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=39408419&chr=chr22&rs=rs133074&imp=false
9611613	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=40291777&chr=chr22&rs=rs9611613&imp=false
2424641	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=24665867&chr=chr20&rs=rs2424641&imp=false
2600814	3	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46054248&chr=chr14&rs=rs2600814&imp=false
16892216	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=120271074&chr=chr8&rs=rs16892216&imp=false
6990312	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=110671493&chr=chr8&rs=rs6990312&imp=false
9908046	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=50918781&chr=chr17&rs=rs9908046&imp=false
575873	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=39055489&chr=chr17&rs=rs575873&imp=false
4691075	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=164468935&chr=chr4&rs=rs4691075&imp=false
4471745	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=50923883&chr=chr17&rs=rs4471745&imp=false
12891534	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=80069114&chr=chr14&rs=rs12891534&imp=false
6988341	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=110653602&chr=chr8&rs=rs6988341&imp=false
12216712	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=9933221&chr=chr8&rs=rs12216712&imp=false
11869714	4	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=45942306&chr=chr17&rs=rs11869714&imp=false

Table S2: List of the 10 SNPs with largest Bayes factors for each of the four factors obtained with the HGDP dataset. For each SNP, we provide the link to the *HGDP BROWSER* database to display pie charts of allele frequencies.

Gene Ontology	no. SNP*	no. genes+	on chip−	GO description	Factor
GO:0001522	6	2	15	pseudouridine synthesis	1
GO:0030324	19	5	79	lung development	1
GO:0032024	9	3	28	positive regulation of insulin secretion	1
GO:0090277	9	3	35	positive regulation of peptide hormone secretion	1
GO:0046887	9	3	51	positive regulation of hormone secretion	1
GO:0008045	11	1	16	motor axon guidance	1
GO:0051896	9	4	56	regulation of protein kinase B signaling cascade	1
GO:0042993	6	1	18	positive regulation of transcription factor import into nucleus	2
GO:0042346	6	1	14	positive regulation of NF-kappaB import into nucleus	2
GO:0007156	61	12	115	homophilic cell adhesion	3
GO:0009896	30	4	97	positive regulation of catabolic process	3
GO:0031331	29	3	76	positive regulation of cellular catabolic process	3
GO:0000096	20	1	29	sulfur amino acid metabolic process	4
GO:0042982	10	1	13	amyloid precursor protein metabolic process	4
GO:0006516	10	1	15	glycoprotein catabolic process	4
GO:0006555	20	1	17	methionine metabolic process	4
GO:0007568	26	5	175	aging	4

* number of outlier SNPs
+ number of outlier genes
− number of genes on SNP array

Table S3: Results of the Gene Ontology based on the set of genes found in Table S1. We use a threshold of 0.05 for the false discovery rate, we remove gene ontology terms that are shared by less than 10 genes or more than 1,000 genes. For each factor, we consider the ten GOs with the smallest FDRs and we only report the gene ontology terms that are related to biological processes.

doi:10.1111/evo.12342

NONSTATIONARY PATTERNS OF ISOLATION-BY-DISTANCE: INFERRING MEASURES OF LOCAL GENETIC DIFFERENTIATION WITH BAYESIAN KRIGING

Nicolas Duforet-Frebourg[1] and Michael G.B. Blum[1,2]

[1]*Laboratoire TIMC-IMAG, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France*

[2]*E-mail: michael.blum@imag.fr*

Received May 22, 2013
Accepted December 13, 2013

Patterns of isolation-by-distance (IBD) arise when population differentiation increases with increasing geographic distances. Patterns of IBD are usually caused by local spatial dispersal, which explains why differences of allele frequencies between populations accumulate with distance. However, spatial variations of demographic parameters such as migration rate or population density can generate nonstationary patterns of IBD where the rate at which genetic differentiation accumulates varies across space. To characterize nonstationary patterns of IBD, we infer local genetic differentiation based on Bayesian kriging. Local genetic differentiation for a sampled population is defined as the average genetic differentiation between the sampled population and fictive neighboring populations. To avoid defining populations in advance, the method can also be applied at the scale of individuals making it relevant for landscape genetics. Inference of local genetic differentiation relies on a matrix of pairwise similarity or dissimilarity between populations or individuals such as matrices of F_{ST} between pairs of populations. Simulation studies show that maps of local genetic differentiation can reveal barriers to gene flow but also other patterns such as continuous variations of gene flow across habitat. The potential of the method is illustrated with two datasets: single nucleotide polymorphisms from human Swedish populations and dominant markers for alpine plant species.

KEY WORDS: **Gene flow, genetic barrier, isolation-by-distance, landscape genetics, nonstationary kriging.**

Characterizing patterns of genetic differentiation within a species is a recurring task in population genetics. Wright (1943) introduced the model of isolation-by-distance (IBD), which assumes that differences of allele frequencies between populations accumulate under the assumption of local spatial dispersal. Because of local dispersal, IBD models predict the *pattern of IBD* where population differentiation increases with increasing geographic distances (Slatkin 1993; Rousset 1997). This pattern is observed in many model and nonmodel organisms as well as in humans suggesting that local dispersal is a leading evolutionary force (Sharbel et al. 2000; Ramachandran et al. 2005; Hardy et al. 2006; Hellberg 2009).

However, the pattern of IBD can mask complex variations of demographic parameters resulting in differential increases of genetic differentiation in different regions of the habitat. Variations of demographic parameters can arise when population densities or migration rates vary across space (Slatkin 1985). With the advent of landscape genetics (Manel et al. 2003; Manel and Holderegger 2013), the spatial variation of demographic parameters is an important topic because spatial heterogeneity (or landscape characteristics) is now recognized to be a key factor to explain population differentiation and gene flow (McRae and Beier 2007). Examples of spatial heterogeneity influencing population differentiation include varying local subpopulation size (Serrouya et al. 2012) as well as fragmented landscapes in urban and agricultural area where there are "corridors" for gene flow (Arnaud 2003; Munshi-South 2012). Barriers to gene flow, which can be caused by anthropogenic or geographic factors, are also emblematic examples of spatial heterogeneity influencing population structure (e.g., Castella et al. 2000; Epps et al. 2005; Riley et al. 2006;

© 2013 The Author(s). *Evolution* published by Wiley Periodicals, Inc. on behalf of The Society for the Study of Evolution.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

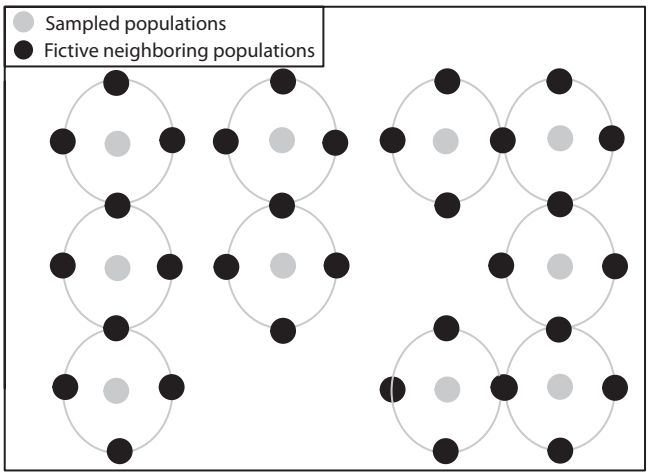


Figure 1. A two-dimensional range with putative sampled locations (in gray) and neighboring locations (in black). Local differentiation at each sampled location corresponds to the average (over neighbors) pairwise measure of differentiation between the population or individual at the sampled location and its neighbors.

Gauffre et al. 2008; Zalewski et al. 2009). Because the identification of barriers to gene flow has attracted considerable attention (Storfer et al. 2010), there is a large variety of statistical methods to detect them (Barbujani et al. 1989; Bocquet-Appel and Bacro 1994; Dupanloup et al. 2002; Manni et al. 2004; Cercueil et al. 2007; Crida and Manel 2007; Manel et al. 2007; Safner et al. 2011). Here, we propose a more general method that characterizes *non-stationary* patterns of IBD. A nonstationary pattern of IBD occurs when the rate at which differentiation between individuals or populations accumulates with distance depends on space. Nonstationary patterns of IBD arise for instance when there is a barrier to gene flow because genetic differentiation accumulates more rapidly with distance around the barrier but they can also occur on different situations such as continuous variations of gene flow across the species range.

To characterize nonstationary patterns of IBD, our approach provides a measure of local differentiation at each location where genetic data are available. The principle of the method is to estimate for each sampled location z_i, $i = 1, \ldots, n$, a local pairwise measure of population differentiation or of dissimilarity between the population located at the sampled location and fictive neighboring populations located at a fixed distance d of z_i (see Fig. 1). Considering for instance F_{ST} as a pairwise measure of genetic differentiation, the method provides estimates of F_{ST} for pairs of populations separated by a distance d and located in the vicinity of the sampling sites. The distance d has to be set in advance and should be small compared to the dimension of the region under study. Fictive neighboring populations are introduced as a mean to provide measures of local genetic differentiation—F_{ST} between populations separated by a distance d here—that are comparable between sampling sites. Compared to common tests for IBD (Hardy and Vekemans 1999), the method is more informative because it quantifies how local genetic differentiation varies across space; the rate at which genetic differentiation increases with distance may vary across space and the proposed approach provides a quantitative assessment of this variation. To determine if variation of local differentiation is sufficiently large to reject stationary IBD, we additionally provide an hypothesis-testing procedure based on simulations. The method is not restricted to pairwise population measurements and can also accommodate individual pairwise measures. Working at the scale of individuals is a desirable feature because using individuals as the operational unit avoids potential bias in identifying populations in advance and offers the opportunity to conduct studies at a finer scale (Manel et al. 2003, 2007). Using a detailed simulation study, we demonstrate that the method can correctly infer local variation of genetic differentiation and we present applications to human single nucleotide polymorphism (SNP) data (Humphreys et al. 2011) and amplified fragment length polymorphism (AFLP) from alpine plants (Gugerli et al. 2008).

Methods

For the sake of the presentation, we assume that the data consist of allele frequencies in each population and that the method relies on the empirical correlation matrix between populations. In the Results section, we show that the proposed approach is also appropriate with other pairwise matrices such as F_{ST} matrices between populations or correlation matrices between individuals.

To assess local genetic differentiation around a given sampled site, we estimate the correlation of allele frequencies between the sampled population and fictive populations located in the neighborhood of the sampling sites. Neighboring populations are located at a fixed and short distance from the sampled populations, and we measure the expected local correlation (averaged over neighbors) of allele frequencies between the sampled population and the neighboring fictive populations (Fig. 1). Because we aim at providing local genetic differentiation values that should be larger in regions of abrupt genetic changes, we consider one minus the local correlation as a measure of local differentiation.

We estimate local correlation using a Gaussian process approach (Bishop 2006), which is known as *kriging* in geostatistics (Cressie 1993). Kriging refers to a set of interpolation methods where a variable of interest is estimated at unsampled locations based on values measured at the sampling sites. Interpolation relies on a weighted average of the values measured at the sampling sites and the weights depend on a parametric function C, which describes how the correlation or the covariance decreases with distance (Cressie 1993). A direct application of kriging would

consist of interpolating the allele frequencies at the neighboring sites based on the allele frequencies estimated at the sampled sites. However, the proposed approach is nonstandard and requires methodological developments because we rather aim at estimating the correlation matrix between sampled and unsampled neighboring sites based on the correlation matrix between sampled sites. There is a vast literature of kriging procedures with nonstationary covariance when the function C describing the decay of correlation with distance varies in space (Nott and Dunsmuir 2002; Schmidt and O'Hagan 2003; Paciorek and Schervish 2006). The covariance between sampled and unsampled sites is usually estimated using a parametric model (Paciorek and Schervish 2006) or at least using a given functional model for the covariance function (Schmidt and O'Hagan 2003). However, compared to geostatics where only one or a few variables are observed at the sampling sites, we are in a favorable situation in population genetics to estimate how the covariance or the correlation varies across space. Because each locus is a statistical replicate, there is enough information to estimate the correlation between the sampled sites using the empirical correlation matrix for instance. Estimating local correlation amounts at interpolating the correlation between sampled and neighboring sites from the correlation matrix between sampled sites. We explain below how we perform the interpolation step.

THE KRIGING/GAUSSIAN PROCESS APPROACH

In the following, we denote by $\mathbf{X}$ and $\mathbf{Y}$ the vectors of allele frequencies at sampled and unsampled sites. We assume independence between loci and the vectors $\mathbf{X}$ and $\mathbf{Y}$ contain allele frequencies for an arbitrary locus. The objective of the kriging approach is to interpolate the covariance (or correlation) matrix between $\mathbf{X}$ and $\mathbf{Y}$ based on the empirical covariance matrix between sampled sites. The covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ is denoted $\mathbf{E}[(\mathbf{Y}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T]$, where $\mathbf{m}$ is a constant mean over the range. The main principle is to use weighted means of covariance values between sampled sites to estimated covariance between sampled and unsampled sites. As usual in kriging, the weights depend on a parametric function C that gives the decay of correlation with distance. We explain below how we compute these weights.

The Gaussian process viewpoint of kriging is to model the joint values of the variable at sampled and unsampled sites as a multivariate Gaussian variable (Bishop 2006)

$$(\mathbf{X}, \mathbf{Y}) \rightsquigarrow \mathcal{N}(\mathbf{m}, \Psi), \tag{1}$$

where

$$\Psi = \begin{pmatrix} \Psi_{\mathbf{xx}} & \Psi_{\mathbf{xy}} \\ {\Psi_{\mathbf{xy}}}^T & \Psi_{\mathbf{yy}} \end{pmatrix},$$

where $\Psi_{\mathbf{xx}}$ (respectively, $\Psi_{\mathbf{yy}}$) denote the covariance matrix between the sampled sites (respectively, unsampled sites) and $\Psi_{\mathbf{xy}}$ contain the covariances between the sampled and unsampled sites. The interpolation of the variable of unknown allele frequencies Y is obtained using the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, which can be written in the following regression form

$$\mathbf{Y} - \mathbf{m} = \boldsymbol{\tau}_{\Psi}(\mathbf{X} - \mathbf{m}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\tau}_{\Psi} = {\Psi_{\mathbf{xy}}}^T {\Psi_{\mathbf{xx}}}^{-1}$ and $\boldsymbol{\epsilon}$ is a residual independent of $\mathbf{X}$ (Brown et al. 1994). A naive computation of local covariance would consist of simulating with equation (2) the vector $\mathbf{Y}$ containing the allele frequencies at the neighboring sites and then to evaluate numerically the empirical covariance between allele frequencies at sampled and at neighboring sites. Although it is a valid approach, we can actually derive what is the expected covariance between sampled and neighboring sites using equation (2) and we obtain

$$\mathbf{E}[(\mathbf{Y}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T] = \boldsymbol{\tau}_{\Psi}\mathbf{E}[(\mathbf{X}-\mathbf{m})(\mathbf{X}-\mathbf{m})^T]. \tag{3}$$

In the computations, we replace $\mathbf{m}$ by the empirical mean so that we estimate the covariance matrix with $\boldsymbol{\tau}_{\Psi}\mathbf{Var}(\mathbf{X})$, where $\mathbf{Var}(\mathbf{X})$ denotes the empirical covariance matrix of $\mathbf{X}$. The matrix $\boldsymbol{\tau}_{\Psi}$ provides the weights of the weighted means, which are used to interpolate the covariance values between sampled and unsampled sites based on the covariance values between sampled sites.

More generally, we can estimate local similarities by multiplying the weight matrix $\boldsymbol{\tau}_{\Psi}$ with a similarity matrix between sampled sites. In the Results section, we consider similarity matrices that are not correlation or covariance matrices, and we use the pairwise matrix of $(1 - F_{ST})$ values for instance. When using individuals as operational units, numerical problems can arise if they are multiple individuals by site because the matrix Ψ_{xx} can be difficult to invert. Potential solutions are to consider a population—with one or more individuals—at each sampling site or to add a small perturbation to the geographical coordinates of the individuals.

Providing the correlation instead of the covariance between sampled and unsampled sites requires the standardization of the covariance equation (3) and the renormalization formula is provided in Appendix A. The final estimate for the covariance matrix is finally obtained by averaging equation (3) over posterior replicates of $\boldsymbol{\tau}_{\Psi}$. The parametric model for Ψ, which is needed to generate the posterior distribution of $\boldsymbol{\tau}_{\Psi}$, is given below.

A MODEL FOR THE CORRELOGRAM

To compute the weight matrix $\boldsymbol{\tau}_{\Psi}$, we consider the standard model of *stationary* kriging that assumes that the correlation between two points only depends on the distance between these two points. Using these assumptions, we should model how the correlation

decreases with increasing distance. We assume that this function C, called the correlogram, decays exponentially

$$C(d) = ((1-\alpha) + \alpha e^{-d/r} + \lambda 1_{d=0})/(1+\lambda), \qquad (4)$$

where d is the distance between two points, 1 denotes the indicator function, α determines the sill, which measures the limiting value of the correlation function, r is the *range* parameter, and λ is the *regularization* parameter. The parameter λ is introduced for numerical reasons because it ensures that the matrix Ψ_{xx} is invertible, which is required for the computation of the weight matrix $\boldsymbol{\tau}_\Psi$ (Bishop 2006). The range parameter r is inversely related to the rate at which correlation decays with distance. Denoting by d_{ij} the geographical distance between the ith and jth sites, then the entry of Ψ at the ith row and jth column is given by $C(d_{ij})$. We sample the triplet (α, λ, r) from the posterior distribution using an MCMC algorithm that contains both Gibbs and Metropolis–Hastings updating steps, and the details of the algorithm are provided in Appendix B (Handcock and Stein 1993).

HYPOTHESIS-TESTING PROCEDURE

We introduce two test statistics to test if the variation of local genetic differentiation is significant. The first test statistic is the coefficient of variation of local genetic differentiation values, that is the ratio between standard deviation and mean of local differentiation measures. The second statistic is the distance correlation statistic and it measures the dependence between local genetic differentiation and geographical coordinates. The distance correlation statistic extends Pearson correlation coefficient because it can measure nonlinear dependence (Székely et al. 2007). Because we use two test statistics, we consider the conservative Bonferroni correction and reject stationarity when one the two observed values of the test statistics is larger than the 97.5% quantile obtained for the null distribution of stationarity.

We consider two options for generating distribution of the test statistic under the null hypothesis. In the first option, we consider the parametric model of equation (4). We compute M pairwise covariance matrices $\mathbf{Var}(\mathbf{X})^i$, $i = 1, \ldots, M$, using the stationary correlogram of equation (4). The parameters $(\alpha_i, \lambda_i, r_i)$ of equation (4), which are used to compute the covariance matrices, are sampled according to the posterior distribution. The $2 \times M$ values of the tests statistics are then obtained after running the MCMC algorithm (Appendix B) M times for each of the simulated covariance matrix $\mathbf{Var}(\mathbf{X})^i$, $i = 1, \ldots, M$. When the sample size is too large, we have to limit the computational burden of the procedure, and we do not perform M MCMC runs. Instead, for the ith covariance matrix $\mathbf{Var}(\mathbf{X})^i$, we use the ith triplet $(\alpha_i, \lambda_i, r_i)$ to compute the weight matrix $\boldsymbol{\tau}_\Psi^i$ and to obtain values of local genetic differentiation. However, equation (4) is only an approximation of the correlation pattern found for IBD models. It is exact, for instance, in the one-dimensional stepping-stone model with infinite range (Kimura and Weiss 1964). To avoid the approximation of equation (4), we also consider explicit simulations of a stationary stepping-stone model using *ms* (Hudson 2002). We consider uniformly sampled migration rates such as $1 \leq 4N_0m \leq 20$ and we choose a sampling scheme that mimics the sampling of the data.

Results

SIMULATION STUDY

In the simulation study, we consider two different models for generating nonstationary patterns of IBD. First, we consider nonhomogeneous stepping stone models in one and two dimensions. We simulate with *ms* 2000 independent SNPs using spatially dependent effective migration rate $4N_0m$, where N_0 is the population size of each deme and m is the migration rate per generation between two neighboring demes. Because we assume independence between SNPs, each SNP is simulated with a coalescent simulation that is conditioned on having one segregating site. The second model is analytic and has been developed for performing nonstationary kriging when the correlogram function (equation (4)) is assumed to vary across space (Paciorek and Schervish 2006). The range parameter r of equation (4), which measures the rate at which correlation decays with distance, is assumed to be a function of space. For the second model, zones of abrupt changes such as genetic barriers correspond to regions with a smaller range parameter because correlation decays more rapidly with distance in these regions.

Barrier in a one-dimensional model

We investigate an example of a one-dimensional model with a genetic barrier. We simulate a stepping stone model with 100 populations of effective sizes $N_0 = 1000$ diploid individuals. Depending on the simulations, we sample either 20 equidistant populations or 20 uniformly sampled populations. We consider 20 chromosomes in each of the population. Migrations are constant between neighboring populations and we consider $4N_0m = 4$ and $4N_0m = 20$. The barrier is located between populations 50 and 51 and arose 8 units of time ago ($4N_0m = 0$) where time is counted in units of $4N_0$ generations. As similarity matrix, we consider the pairwise correlation of allele frequencies for 20 sampled populations. For each sampled deme, local genetic differentiation corresponds to one minus the expected correlation between the sampled deme and its two neighbors.

With equidistant sampling, we find that the parameters of the correlogram function (equation (4)) affect the estimated values of local genetic differentiation (Fig. S1). However, for all values of the correlogram parameters (α, λ, r) we consider, local differentiation is larger in the middle of the range, which is consistent with

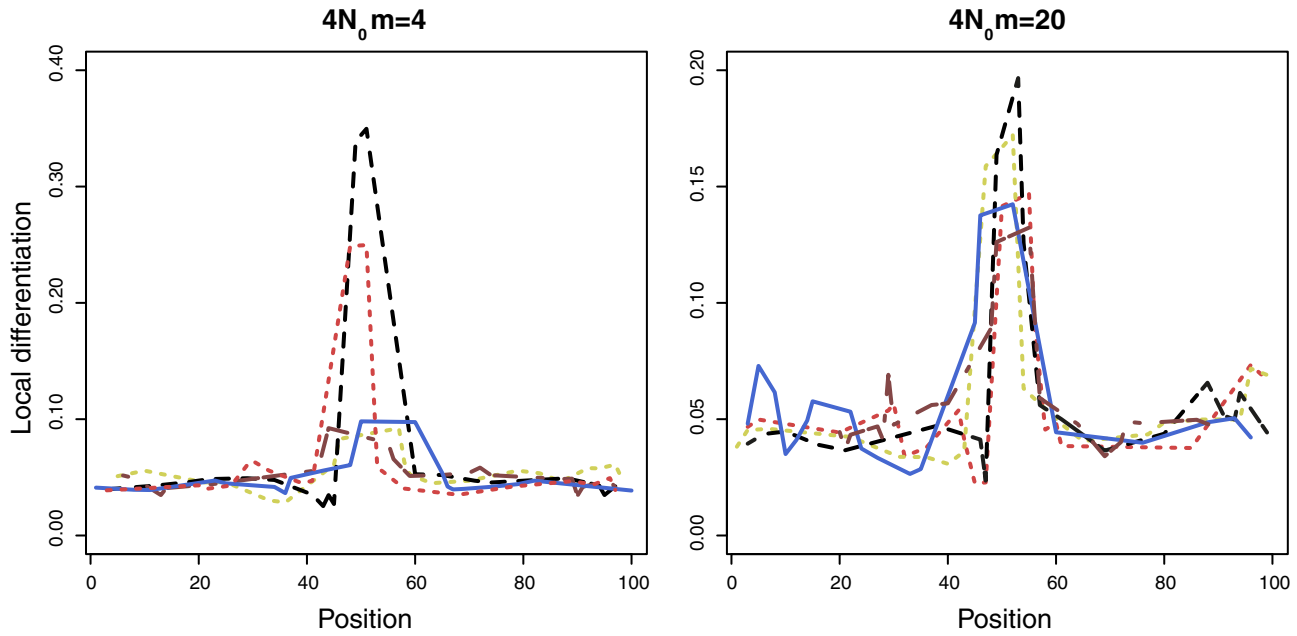


Figure 2. **Estimation of local genetic differentiation for five population sets where each set contains 20 randomly picked populations among 100 populations. Simulations are performed in a one-dimensional stepping-stone model with a barrier to gene flow in the middle of the range ($4N_0m = 20$ or $4N_0m = 4$ except between population 50 and population 51 where $4N_0m = 0$). Local genetic differentiation corresponds to one minus the expected correlation of allele frequencies between sampled demes and their two closest neighboring demes.**

the presence of a barrier to gene flow. Nonetheless the detailed trajectory of local differentiation as a function of space depends on the correlogram parameters and edge effects can be large for some parameter values (Fig. S1). To account for the uncertainty associated with the parameters of the correlogram function, we integrate the values of local genetic differentiation over the posterior distribution of (α, λ, r) (Fig. S1). To investigate if a barrier to gene flow is still detectable with irregular sampling, we also sampled randomly 20 populations among the 100 populations. For both intensities of barrier ($4N_0m = 20$ or $4N_0m = 4$ except at the barrier where $4N_0m = 0$) and for each replicate of population sampling, local differentiation is larger around the barrier to gene flow (Fig. 2). However, for the less stringent and more difficult to detect barrier, local differentiation increases less markedly around the barrier when sampling in the vicinity of the barrier is sparse (Fig. 2).

To provide a comparison, we apply multidimensional scaling (MDS), which is a commonly used method to represent differentiation between populations. Based on the pairwise matrix of correlation between allele frequencies computed for each of the 20 equally spaced populations, we apply MDS. Figure 3 displays the scatter plots of the first two principal coordinates when $4N_0m = 4$, $4N_0m = 20$, and there are 20 equally spaced sampled populations. In the scenario where $4N_0m = 20$, the occurrence of a barrier of gene flow in the middle of the range is visible in the MDS plot whereas it is much less visible when $4N_0m = 4$ even with the perfectly regular sampling of populations. In the latter scenario, the MDS plot is above all influenced by the global IBD pattern that generates the inverted U-shaped pattern (Novembre and Stephens 2008). Patterns exactly similar to MDS were obtained with principal component analysis when using the population allele frequencies as raw data.

We additionally explore the running time of the algorithm. Once a pairwise matrix of F_{ST} or of other dissimilarity measures has been obtained, the most costly operations are the inversion of the matrix Ψ_{xx} required to compute the weight matrix $\boldsymbol{\tau}_\Psi$ as well as the computation of a determinant required to evaluate the likelihood in the MCMC algorithm (see Appendix B). The computing cost of both operations is proportional to the cube of the number of sampled sites. We check this theoretical prediction in this example by increasing the total number of demes and the number of sampled demes. We find that the cubic prediction is quite accurate although a bit pessimistic because the running time of the algorithm actually grows as the number of sampled sites at the power 2.5 (Fig. S2).

Barriers in a two-dimensional model

We consider two examples of a two-dimensional model with genetic barriers. In the first two-dimensional example, the data are simulated using a stepping-stone model in a 10×10 grid. In each population, there are $N_0 = 1000$ diploid individuals per population and we sample all populations considering 20 chromosomes

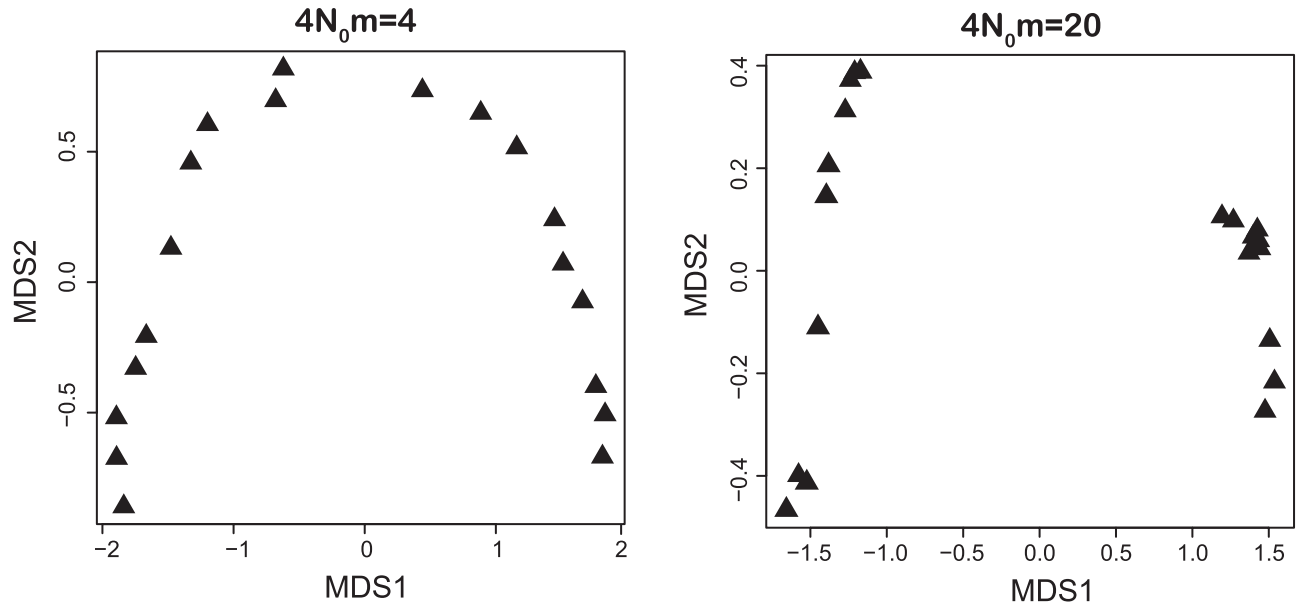


Figure 3. Multidimensional scaling plots in a one-dimensional range with a genetic barrier in the middle of the range.

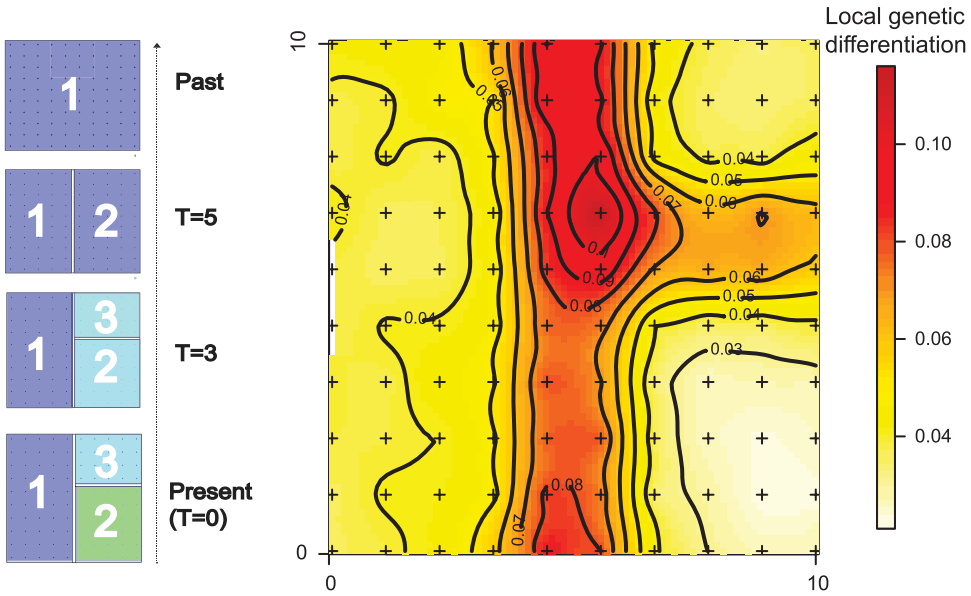


Figure 4. Map of local genetic differentiation in a two-dimensional stepping-stone model with two genetic barriers. On the left-hand side, the time line shows the time when the genetic barriers appeared. Local genetic differentiation was estimated using the pairwise correlation matrix, and it measures one minus the expected correlation between sampled populations and unsampled neighbors located at a distance of 0.1 from the sampled populations.

in each of them. The migration rate between neighboring populations is $4N_0m = 20$, where migration only occurs along horizontal and vertical lines but not along diagonals. We then assume that two barriers arose $T_1 = 5$ and $T_2 = 3$ units of time ago, where time is counted in units of $4N_0$ generations (see Fig. 4). In the second two-dimensional example, we specify explicitly the local decay of correlation using the nonstationary model of (Paciorek and Schervish 2006). We assume that there are three genetic barriers, which correspond to three different regions where the range parameter (r in equation (4)) is smaller (Fig. S3). In both examples, the input matrix of similarity is the correlation matrix between sampled sites, although the correlation is estimated based on simulated allele frequencies for the stepping stone example whereas it is obtained analytically using the convolution formula of Paciorek and Schervish (2006) for the second example.

Figures 4 and S3 show that estimated values of local genetic differentiation are larger around the genetic barriers as expected. For both examples, the relative importance of the barriers is

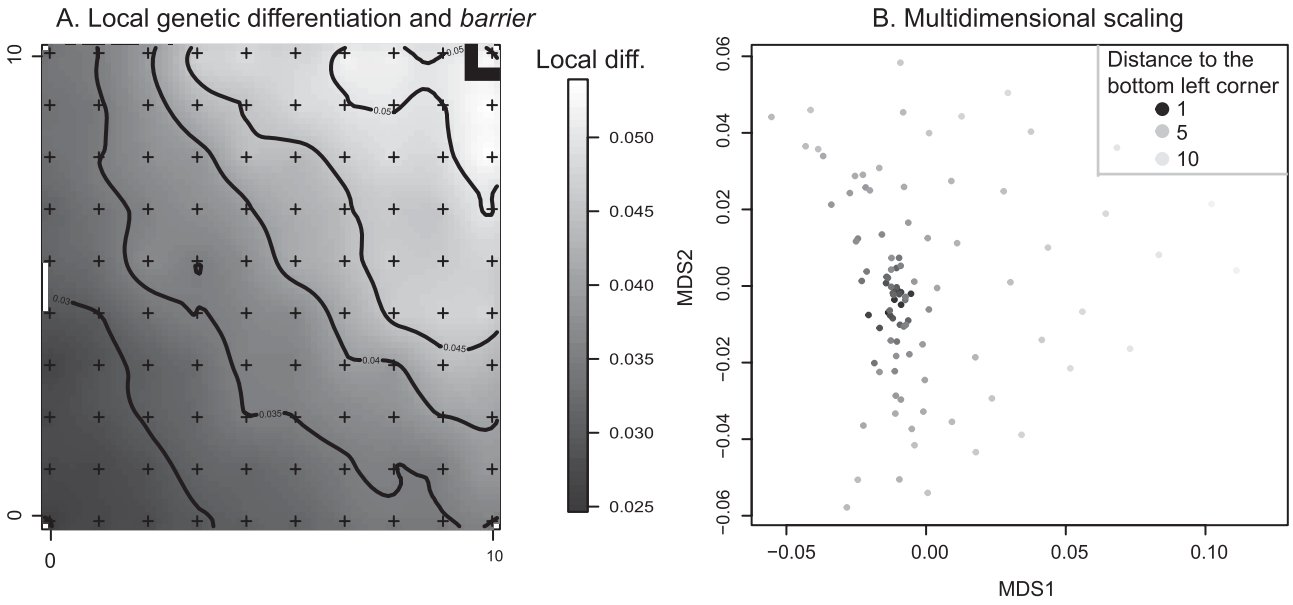


Figure 5. Investigating the pattern of population differentiation for a gradient of gene flow in a two-dimensional model. Gene flow is maximal at the lower left corner of the habitat and decreases quadratically to the distance from the lower left corner. (A) Estimation of local genetic differentiation values using the correlation with neighbors located at a distance of 0.1. The first genetic barrier that is found with the software *barrier* is shown with a thick black line. (B) Multidimensional scaling plot with a grayscale color scheme to represent the distance of each population to the lower left corner. Dark points correspond to populations that are close to the lower left corner whereas the lightest points are the farthest away from the lower left corner.

retrieved. The strongest barrier has the largest value of local genetic differentiation. For computing local genetic differentiation, we additionally consider F_{ST} pairwise values instead of correlation values in the stepping-stone model. In that case, the similarity matrix contains the pairwise $(1 - F_{ST})$ values that decay with increasing geographical distance as assumed by equation (4). The local genetic differentiation values now correspond to the expected F_{ST} between the sampled populations and their fictive neighbors. We find that the map of local genetic differentiation obtained with the F_{ST} measure is similar to the map obtained with the pairwise correlation matrix (Fig. S4). We also perform additional computations of local differentiation after incomplete sampling of the populations in the stepping-stone model. We sample, respectively, 50%, 33%, and 25% of the 100 populations present in the grid. We find that the oldest barrier is always recovered but not the most recent one, which is not detectable when sampling 33% or 25% of the populations (Fig. S5).

A gradient of gene flow

We also consider a different pattern of nonstationary IBD consisting of a two-dimensional stepping-stone model with a spatial gradient of gene flow. We assume that gene flow is maximum at the lower left corner of the habitat and decreases quadratically with distance from the lower left corner of the habitat (Fig. 5). As expected, we find that local genetic differentiation increases when moving away from the lower left corner of the habitat (Fig. 5). When sampling 50%, 33%, or 25% of the 100 populations present in the grid, we also find a gradient of local genetic differentiation (Fig. S6).

This example is an instance of nonstationarity, which cannot be described with barriers to gene flow. For the previous examples, the software *barrier* (Manni et al. 2004), which detects zones of abrupt genetic change, is able to find barriers in both the one- and two-dimensional models (Fig. S7). However, for the gradient of gene flow, *barrier* incorrectly finds a barrier in the upper right corner of the habitat, which is nonetheless consistent with the fact that gene flow is minimal here (Fig. 5). Additionally, MDS provides a meaningful representation for the examples of barriers in the one-dimensional model with $4N_0m = 20$ (Fig. 3) and in the two-dimensional model (Fig. S8) because the populations that live on the same side of the barriers cluster together in the MDS plot (but see $4N_0m = 4$ in Fig. 3 where the clustering is less evident). However, interpreting the pattern obtained with MDS is much more difficult for the example of a gradient of gene flow (Fig. 5). The observed pattern found with MDS is consistent with the gradient of gene flow because populations living in regions of high gene flow (dark points in Fig. 5) are located more closely on the MDS plot than populations living in regions of low gene flow (clear points in Fig. 5). Although consistent with a gradient of gene flow, the MDS plot is not as easily interpretable as the map of local genetic differentiation for this example.

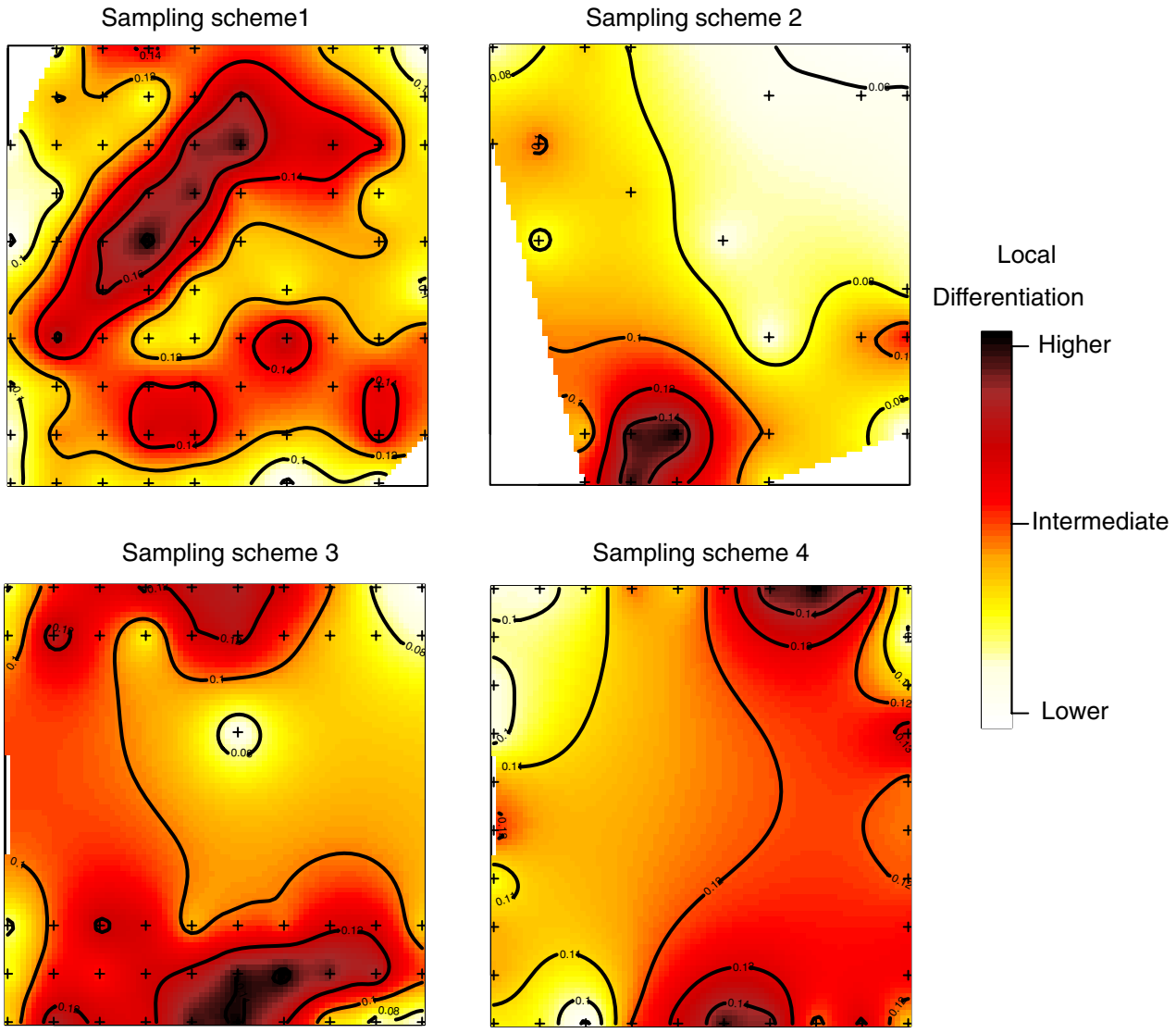


Figure 6. Effect of sampling scheme on local differentiation for data simulated with a stationary model of isolation-by-distance. Simulations are performed with *ms* under a stepping-stone model.

Testing nonstationary patterns of IBD

The estimates of local genetic differentiation may depend on the sampling scheme. Clustered or irregular sampling scheme in particular can be a matter of concern because they might generate false-positive patterns of nonstationarity. Here, we consider different sampling schemes in a two-dimensional stepping-stone model to study the risk of false positives. In the first and second sampling schemes, respectively, 25% and 75% of the total number of sites have been sampled (Fig. 6). The third sampling scheme consists of a clustered sampling scheme with two different geographic zones, which have been sampled, as well as an isolated sampled site between the two regions. In the last sampling scheme, only the perimeter of the two-dimensional square has been sampled. Figure 6 shows heat maps of local genetic differentiation for the stationary IBD simulations that have been generated with *ms*. To evaluate whether the observed variations of local genetic differentiation are sufficient evidence for nonstationarity, we perform 100 simulations of stationary patterns for each sampling scheme. When simulations of the null models are performed with *ms*, stationarity is rejected for 3–7% of the simulations, which is consistent with the nominal 5% type I error we use. However if equation (4) is used as a null model, stationarity is rejected for all the simulations performed with *ms*. The stepping-stone simulations of stationary IBD show that equation (4) should not be used for hypothesis testing and we should instead resort to explicit simulation of IBD models for generating distributions of the test statistics under the null hypothesis of stationarity. In addition, for all the simulations of nonstationary processes considered so far (barriers in one- and two-dimensional models and gradient of gene flow), we reject stationarity as expected.

APPLICATIONS

Nonstationary patterns of IBD among the swedish population

We first illustrate the kriging methodology using a human SNP dataset with a particularly dense geographic sampling. The data consist of genome-wide SNPs for 5174 Swedish individuals that cover all of the 21 Swedish counties Humphreys et al. (2011). To assign each individual to a county, Humphreys et al. (2011) used available geographic information with the following order of priority: city or village of birth, county of birth, municipality or city of residence, and county of residence if it is the only information available. They found strong differences between far northern counties and remaining counties, and also showed that northern counties are more clearly genetically differentiated from each other than southern counties are from each other.

Because our framework is an extension of IBD, we first check that population differentiation increases with increasing geographical distance. We confirm the prevalence of IBD in Sweden ($P < 10^{-7}$ for a Mantel test, see also Fig. S9). Then, we choose to quantify local genetic differentiation using the F_{ST} between a population living exactly in the barycentric center of the county and a putative neighboring population living 30 km away (see Fig. S10). We consider the pairwise $(1 - F_{ST})$ values between the counties as input matrix of pairwise similarities. Our hypothesis-testing procedure indicates significant nonstationary pattern of IBD in Sweden. We find that the northernmost counties (Nordbotten, Västerbotten, and Jämtlands) have the strongest values of local genetic differentiation (Fig. 7), whereas the smallest values are found in the regions around the Stockholm area (Östergötlands, Stockholms, Södermanlands, Västmanlands, and Jönköpings are the five counties with the lowest values of local differentiation). The fourth largest value is found for the Dalarna county (in north middle Sweden), which borders southern Norway, and counts more individuals with remote Finnish or Norwegian ancestry than other counties (Humphreys et al. 2011). As expected, the four counties with the largest local differentiation values are also the most differentiated from other counties even when controlling for geographic distance (Fig. S9).

In summary, we confirm the results of Humphreys et al. (2011) who found that there is more genetic differentiation within northern Sweden than within southern Sweden. The four counties with the largest values of local genetic differentiation are also the counties with the lowest population densities (Fig. S11) suggesting that low population density triggered population differentiation in northern Sweden.

Nonstationary patterns of IBD for alpine plant species

We consider a set of 20 alpine plant species that have been sampled across the Alps (Gugerli et al. 2008; Alvarez et al. 2009); Jay et al. 2012a). The sampling is particularly dense with one to three individuals per species collected for each cell of approximately 500 km^2. Individual genotypes consist of AFLPs. We compute allele frequencies at each sampling site—possibly using one individual only—and we consider the matrix of correlation between allele frequencies as input similarity matrix. Local genetic differentiation corresponds to one minus the expected correlation between sampled populations and neighboring populations located at 8 km.

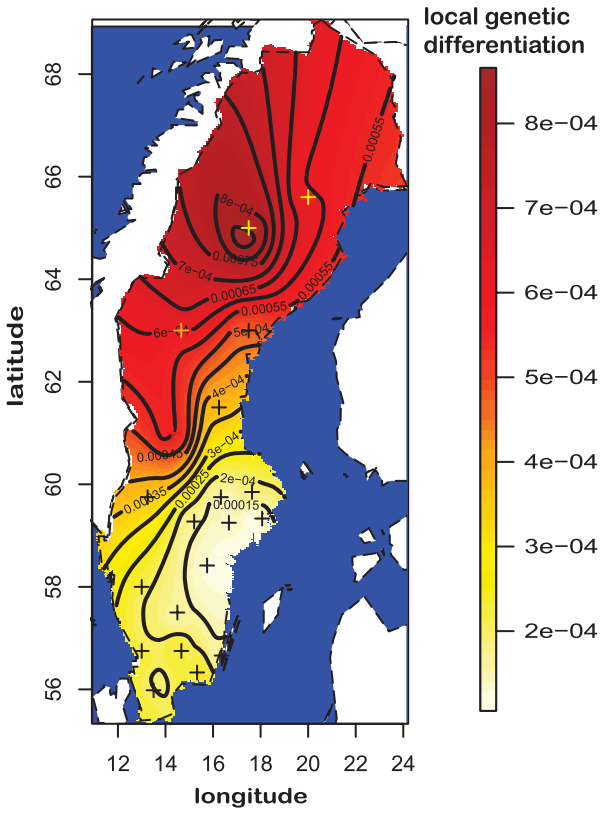


Figure 7. Map of local genetic differentiation computed for the 20 Swedish counties. Measures of local genetic differentiation correspond to the F_{ST} between the sampled counties and fictive neighboring populations, not shown, located at 30 km.

The test for nonstationarity is significant for seven species with a type I error rate of 5% and it increases to nine species when accepting a type I error rate of 10% (Table S1). However, for the species with nonstationary IBD, the detailed pattern of local genetic differentiation is idiosyncratic to each species (Fig. S12). For instance, the alpine species *Phyteuma hemisphaericum* exhibits larger values of local genetic differentiation in a large region ranging from the central Alps to the southwestern Alps, whereas there are disconnected regions of larger genetic differentiation for the species *Arabis Alpina* all located in the South of the Alps (Fig. 8). To integrate the results found for all species with nonstationary IBD, we compute, for each species, normalized rank values between 0 and 1, where 0 corresponds to the site of

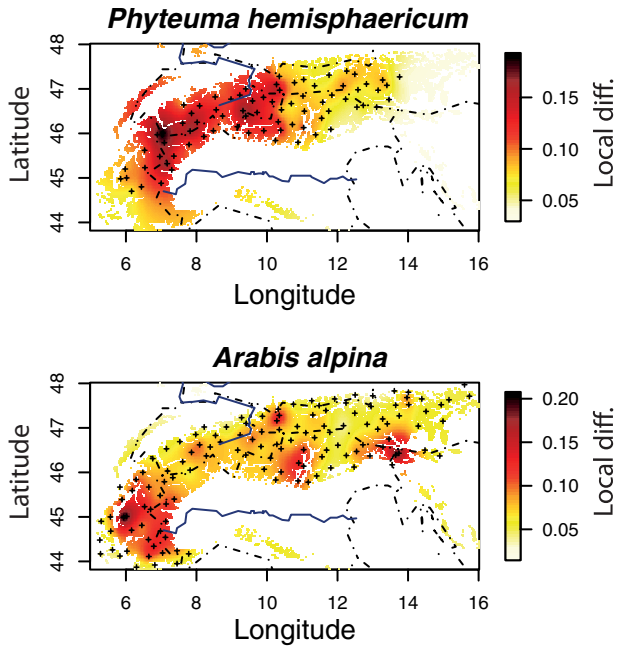


Figure 8. Heatmap of local genetic differentiation computed across the Alps for two alpine species: *Phyteuma hemisphaericum* and *Arabis alpina*. Local differentiation corresponds to one minus the correlation between sampled populations and fictive neighboring populations, not shown, located at 8 km around the sampling sites.

lowest local differentiation and 1 corresponds to the site of largest local differentiation. When averaging the normalized ranks across species, we found that a region of the western Alps encompassing the inner Alpine Aosta valley is the region with the larger values of local differentiation (Fig. S13). This region has already been found to be one of the two major break zones of allele distribution patterns for alpine plant species (Thiel-Egenter et al. 2011). Pleistocene glaciations are putative explanations for the occurrence of a break zone in this region: the populations of plants were initially fragmented into glacial refuges, then expanded via postglacial colonization routes, and a secondary contact zone finally arose where formerly allopatric populations admixed (Pawlowski 1970; Schönswetter et al. 2005; Thiel-Egenter et al. 2011).

Software

The software *LocalDiff* implementing the method is available at http://membres-timc.imag.fr/Michael.Blum/LocalDiff.html. It computes local genetic differentiation from a matrix of pairwise similarity score and can also handle raw genotype data that contain genotypes of individuals. In addition, it generates the *ms* command lines that are required for performing the stepping-stone simulations used in hypothesis testing.

Discussion

In this article, we present a new Bayesian method to characterize nonstationary patterns of IBD. Because the method aims at refining the description of IBD patterns, it should only be applied when IBD has already been detected. From *global* measures of pairwise similarity or dissimilarity, the method infers *local* measures of similarity or dissimilarity. Whatever is the exact measure of genetic (dis)similarity, we use the generic expression of *local genetic differentiation* when referring to the estimated local growth of genetic dissimilarity or differentiation. If considering for instance the F_{ST} pairwise matrix of genetic differentiation between populations, the inferred values correspond to the F_{ST} between the sampled populations and fictive neighboring populations located at a given distance. The method is not restricted to F_{ST} measures and can handle any type of measures of differentiation and is also valid at the individual scale. We consider for instance the correlation between the allelic types of individuals, but other measures would be valid such as identity by descent between individuals (Browning and Browning 2011) as well as coancestry measures (Lawson et al. 2012). Because the two latter measures are based on haplotypes instead of genotypes, they can provide information at a finer geographical scale (Gattepaille and Jakobsson 2012; Lawson et al. 2012).

GENETIC DIFFERENTIATION AND GENE FLOW

It is of course tempting to convert maps of local genetic differentiation into maps of gene flow or of dispersal distance. Assuming that differentiation occurs according to a stepping-stone model, such parameter estimates could be obtained using theoretical relationships between local F_{ST} and dispersal distance (Rousset 1997). However, relating F_{ST} or other measures of genetic differentiation to gene flow relies on many assumptions that may be unrealistic (Marko and Hart 2011). Although the estimation of gene flow with F_{ST}-based methods can be robust in some situations, such as temporal variation of gene flow (Leblois et al. 2004), there are other processes such as range expansion, local extinction, and recolonization that can modify drastically the pattern of genetic differentiation (Wade and McCauley 1988; Arenas et al. 2012). More generally, a map of local genetic differentiation is informative about the *pattern* of genetic differentiation but do not provide enough information to distinguish between the possible evolutionary *processes* that generated this pattern (for similar concerns about principal component analysis, see McVean 2009). To provide a concrete example, the same pattern, a zone of elevated local differentiation, can be interpreted as a barrier to gene flow in a equilibrium stepping-stone process (Figs. 2–4) or as a secondary contact zone following postglacial expansions in the case of the alpine plant species (Thiel-Egenter et al. 2011).

RELEVANCE TO LANDSCAPE GENETICS

Two key steps in landscape genetics are the detection of genetic discontinuities and the correlation of these discontinuities with landscape and environmental features such as barriers (Manel et al. 2003). Detection of genetic discontinuities is clearly provided by the proposed kriging method; for instance in the case of the alpine species *P. hemisphaericum*, we find genetic discontinuities, that is, larger local genetic differentiation, in a large region of the central and western Alps (Fig. 8). However, aiming at capture genetic discontinuities using barriers only might be too limited and the kriging method can reveal more complex patterns such as gradient of local differentiation across the species' range (Fig. 5). The second key step where the genetic discontinuities are correlated with landscape or environmental variable can also be obtained as a postprocessing step by correlating estimated local genetic differentiation with landscape variables. For instance, in the case of the human SNP Swedish data, we find that local genetic differentiation is correlated with population density. There are alternative and *integrative* approaches that account for both genetic data and landscape variables within the same statistical framework. Accounting for both sources of data can be performed either by a joint assessment of the pattern of population structure or differentiation and its correlation with landscape or environmental variable (Foll and Gaggiotti 2006; Jay et al. 2011), or by correlating genetic distances with distances based on landscape features (Cushman et al. 2006; McRae 2006). These integrative approaches are *hypothesis-driven* in the sense that each set of landscape features affecting population structure corresponds to one hypothesis that can be tested or compared to other ones. The proposed kriging approach is instead a technique of *exploratory data analysis*. It might be especially appropriate for large-scale conservation studies not focused on the underlying evolutionary processes but that should deal with reserve design and with the management of fragmented populations (Schwartz et al. 2007). To explore patterns of genetic differentiation, there are other statistical summaries of the data that can be computed. Population-specific F_{ST}'s based on the F-model can also provide local measures of differentiation by computing local values of F_{ST}'s (Gaggiotti and Foll 2010). However, compared to approaches based on the F-model, *LocalDiff* can also work with the individual-based sampling schemes often encountered in landscape genetics (Schwartz and McKelvey 2009). Moreover, F_{ST}'s based on the F-model relies on a parametric population-genetic model that may be sensitive to departures from the assumption of the F-model (Gaggiotti and Foll 2010). By contrast, the proposed approach relies on kriging (aka Gaussian process), which is a nonparametric approach that assumes a pattern of IBD only.

To compare or test the support of different evolutionary processes provided by the pattern of nonstationary IBD, we can rather resort to inference based on explicit simulations of evolutionary processes using for instance approximate Bayesian computation (Csilléry et al. 2010). Within this simulation framework, measures of local genetic differentiation can be included as statistical summaries of the data. When the number of sampling sites is large, the MCMC algorithm might be too slow to provide summary statistics for approximate Bayesian computation. To overcome this problem, we provide an option in *LocalDiff* where measures of local differentiation are computed by integrating the parameter triplet (α, λ, r) over the prior distribution instead of the posterior distribution. Although integration over the posterior with MCMC should be preferred when possible, integrating over the prior distribution provides relevant measures of local differentiation for the examples we investigated (see Fig. S14).

CAVEATS

Although, computing local differentiation is a descriptive technique of exploratory data analysis, we provide an hypothesis-testing procedure attached to it. This is desirable feature because it can prevent from overinterpreting maps of local differentiation (Fig. 6). The test of nonstationarity relies on two different test statistics and the null distribution is obtained from stepping-stone simulations performed with *ms*. To obtain the null distribution of the test statistic, we consider the same sampling as in the data, which is crucial because the sampling scheme can affect the inferred pattern of local differentiation (Fig. 6). For the simulations of the null model, we uniformly sample the parameter $4N_0m$ in a fixed range, but we acknowledge that using an estimated value of the effective migration parameter would also be a valid strategy. We also consider an alternative and approximate null model under which the correlation decays exponentially (equation (4)), but this approximation should not be used for hypothesis testing because it is much too liberal. However, even considering explicit stepping-stone simulations might have drawbacks because other processes such as range expansions can also generate IBD patterns, but they might induce different distributions of tests statistics (Edmonds et al. 2004). More generally, sampling scheme affects the different methods that characterize population differentiation and simulations can be informative about the effects of sampling scheme (McVean 2009; Schwartz and McKelvey 2009; Jay et al. 2012b).

Another concern about the kriging method concerns the choice of the distance between the sampled populations or individuals and the fictive neighboring populations or individuals. We consider various choices of distances for the Swedish dataset (10–150 km) and find that all these choices of distance provide similar patterns of local genetic differentiation although the patterns generated with the larger distances are smoother (Fig. S15). The last caveat to bear in mind concerns the choice of the pairwise dissimilarity or differentiation matrix. For instance, in a simulation study with five populations, Lawson and Falush (2012)

showed that the five populations were clearly distinguishable with some but not all pairwise dissimilarity matrices between individuals. Although a potential caveat, being able to choose the measure of dissimilarity also adds to the flexibility of the method and different measures can convey information about processes that occurred at different time periods.

PERSPECTIVES

When carefully addressing the aforementioned caveats, measures of local genetic differentiation can provide interpretable patterns for describing nonstationary patterns of IBD. Here, the expression *non-stationary* refers to spatial variations but note that if temporal data are available, a similar kriging framework can be used to study how genetic drift evolves as a function of time. The present method relies on a matrix of pairwise (dis)similarity between individuals or populations located on georeferenced sampling sites. An approach based on dissimilarity matrices is an appropriate methodology for the new genomic era where we have to deal with massive data. Computing pairwise dissimilarity matrix can be computationally efficient (e.g., Browning and Browning 2011) and can even be parallelized to compute different parts of the matrix (Lawson and Falush 2012). Having a statistical method able to scale with the dimension of the genetic data should make it a valuable tool for investigating patterns of genetic differentiation in a wide range of studies.

ACKNOWLEDGMENTS

M. G. B. Blum is supported by the French National Research Agency (DATGEN project, ANR-2010-JCJC-1607-01).

REFERENCES

Alvarez, N., C. Thiel-Egenter, A. Tribsch, R. Holderegger, S. Manel, P. Schönswetter, P. Taberlet, S. Brodbeck, M. Gaudeul, L. Gielly, et al. 2009. History or ecology? Substrate type as a major driver of spatial genetic structure in Alpine plants. Ecol. Lett. 12:632–640.

Arenas, M., N. Ray, M. Currat, and L. Excoffier. 2012. Consequences of range contractions and range shifts on molecular diversity. Mol. Biol. Evol. 29:207–218.

Arnaud, J. 2003. Metapopulation genetic structure and migration pathways in the land snail *Helix aspersa*: influence of landscape heterogeneity. Landscape Ecol. 18:333–346.

Barbujani, G., N. Oden, and R. Sokal. 1989. Detecting regions of abrupt change in maps of biological variables. Syst. Biol. 38: 376–389.

Bishop, C. 2006. Pattern recognition and machine learning, Vol. 4. Springer, New York.

Bocquet-Appel, J., and J. Bacro. 1994. Generalized wombling. Syst. Biol. 43: 442–448.

Brown, P., N. Le, and J. Zidek. 1994. Multivariate spatial interpolation and exposure to air pollutants. Can. J. Stat. 22:489–509.

Browning, B., and S. Browning. 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88:173–182.

Castella, V., M. Ruedi, L. Excoffier, C. Ibanez, R. Arlettaz, and J. Hausser. 2000. Is the Gibraltar strait a barrier to gene flow for the bat *Myotis myotis* (Chiroptera: Vespertilionidae)? Mol. Ecol. 9:1761–1772.

Cercueil, A., O. François, and S. Manel. 2007. The genetical bandwidth mapping: a spatial and graphical representation of population genetic structure based on the wombling method. Theor. Popul. Biol. 71:332–341.

Cressie, N. A. C. 1993. Statistics for spatial data (Wiley series in probability and statistics). Wiley-Interscience, New York.

Crida, A., and S. Manel. 2007. Wombsoft: an R package that implements the wombling method to identify genetic boundary. Mol. Ecol. Notes 7: 588–591.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François. 2010. Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. 25: 410–418.

Cushman, S., K. McKelvey, J. Hayden, and M. Schwartz. 2006. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. Am. Nat. 168:486–499.

Dupanloup, I., S. Schneider, and L. Excoffier. 2002. A simulated annealing approach to define the genetic structure of populations. Mol. Ecol. 11: 2571–2581.

Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza. 2004. Mutations arising in the wave front of an expanding population. Proc. Natl. Acad. Sci. USA 101:975–979.

Epps, C., P. Palsbøll, J. Wehausen, G. Roderick, R. Ramey II, and D. McCullough. 2005. Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. Ecol. Lett. 8:1029–1038.

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

Gaggiotti, O. E., and M. Foll. 2010. Quantifying population structure using the F-model. Mol. Ecol. Res. 10:821–830.

Gattepaille, L. M., and M. Jakobsson. 2012. Combining markers into haplotypes can improve population structure inference. Genetics 190:159–174.

Gauffre, B., A. Estoup, V. Bretagnolle, and J.-F. Cosson. 2008. Spatial genetic structure of a small rodent in a heterogeneous landscape. Mol. Ecol. 17:4619–4629.

Gugerli, F., T. Englisch, H. Niklfeld, A. Tribsch, Z. Mirek, M. Ronikier, N. Zimmermann, R. Holderegger, and P. Taberlet. 2008. Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation—a project synopsis. Perspect. Plant Ecol. Evol. Syst. 10:259–281.

Handcock, M., and M. Stein. 1993. A Bayesian analysis of kriging. Technometrics 35:403–410.

Hardy, O., and X. Vekemans. 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. Heredity 83: 145–154.

Hardy, O. J., L. Maggia, E. Bandou, P. Breyne, H. Caron, M. H. Chevallier, A. Doligez, C. Dutech, A. Kremer, C. Latouche-hallé, et al. 2006. Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species. Mol. Ecol. 15:559–571.

Hellberg, M. 2009. Gene flow and isolation among populations of marine animals. Annu. Rev. Ecol. Syst. 40:291–310.

Hudson, R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Humphreys, K., A. Grankvist, M. Leu, P. Hall, J. Liu, S. Ripatti, K. Rehnström, L. Groop, L. Klareskog, B. Ding, et al. 2011. The genetic structure of the Swedish population. PLoS One 6:e22547.

Jay, F., O. François, and M. G. B. Blum. 2011. Predictions of native American population structure using linguistic covariates in a hidden regression framework. PLoS One 6:e16227.

Jay, F., S. Manel, N. Alvarez, E. Y. Durand, W. Thuiller, R. Holderegger, P. Taberlet, and O. François. 2012a. Forecasting changes in population genetic structure of alpine plants in response to global warming. Mol. Ecol. 21:2354–2368.

Jay, F., P. Sjödin, M. Jakobsson, and M. G. B. Blum. 2012b. Anisotropic isolation by distance: the main orientations of human genetic differentiation. Mol. Biol. Evol. 30:513–525.

Kimura, M., and G. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

Lawson, D. J., and D. Falush. 2012. Population identification using genetic data. Annu. Rev. Genomics Hum. Genet. 13:337–361.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8:e1002453.

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

Manel, S., and R. Holderegger. 2013. Ten years of landscape genetics. Trends Ecol. Evol. 28:614–621.

Manel, S., M. Schwartz, G. Luikart, and P. Taberlet. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18:189–197.

Manel, S., F. Berthoud, E. Bellemain, M. Gaudeul, G. Luikart, J. Swenson, L. Waits, P. Taberlet, et al. 2007. A new individual-based spatial approach for identifying genetic discontinuities in natural populations. Mol. Ecol. 16:2031–2043.

Manni, F., E. Guerard, and E. Heyer. 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum. Biol. 76:173–190.

Marko, P., and M. Hart. 2011. The complex analytical landscape of gene flow inference. Trends Ecol. Evol. 26:448–456.

McRae, B. 2006. Isolation by resistance. Evolution 60:1551–1561.

McRae, B., and P. Beier. 2007. Circuit theory predicts gene flow in plant and animal populations. Proc. Natl. Acad. Sci. USA 104:19885–19890.

McVean, G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5:e1000686.

Munshi-South, J. 2012. Urban landscape genetics: canopy cover predicts gene flow between white-footed mouse (*Peromyscus leucopus*) populations in New York City. Mol. Ecol. 21:1360–1378.

Nott, D. J., and W. T. Dunsmuir. 2002. Estimation of nonstationary spatial covariance structure. Biometrika 89:819–829.

Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40:646–649.

Paciorek, C. J., and M. J. Schervish. 2006. Spatial modelling using a new class of nonstationary covariance functions. Environmetrics 17:483–506.

Pawlowski, B. 1970. Remarques sur l'endémisme dans la flore des alpes et des carpates. Plant Ecol. 21:181–243.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 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.

Riley, S., J. Pollinger, R. Sauvajot, E. York, C. Bromley, T. Fuller, and R. Wayne. 2006. Fast-track: a southern California freeway is a physical and social barrier to gene flow in carnivores. Mol. Ecol. 15:1733–1741.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

Safner, T., M. Miller, B. McRae, M. Fortin, and S. Manel. 2011. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. Int. J. Mol. Sci. 12:865–889.

Schmidt, A., and A. O'Hagan. 2003. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. J. R. Stat. Soc. Ser. B 65:743–758.

Schönswetter, P., I. Stehlik, R. Holderegger, and A. Tribsch. 2005. Molecular evidence for glacial refugia of mountain plants in the European Alps. Mol. Ecol. 14:3547–3555.

Schwartz, M., and K. McKelvey. 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conserv. Genet. 10:441–452.

Schwartz, M., G. Luikart, and R. Waples. 2007. Genetic monitoring as a promising tool for conservation and management. Trends Ecol. Evol. 22:25–33.

Serrouya, R., D. Paetkau, B. N. Mclellan, S. Boutin, M. Campbell, and D. A. Jenkins. 2012. Population size and major valleys explain microsatellite variation better than taxonomic units for caribou in western Canada. Mol. Ecol. 21:2588–2601.

Sharbel, T., B. Haubold, and T. Mitchell-Olds. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol. Ecol. 9:2109–2118.

Slatkin, M. 1985. Gene flow in natural populations. Annu. Rev. Ecol. Syst. 16:393–430.

———. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264–279.

Storfer, A., M. Murphy, S. Spear, R. Holderegger, and L. Waits. 2010. Landscape genetics: where are we now? Mol. Ecol. 19:3496–3514.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. Annals Stat. 35:2769–2794.

Thiel-Egenter, C., N. Alvarez, R. Holderegger, A. Tribsch, T. Englisch, T. Wohlgemuth, L. Colli, M. Gaudeul, L. Gielly, N. Jogan, et al. 2011. Break zones in the distributions of alleles and species in Alpine plants. J. Biogeogr. 38:772–782.

Wade, M., and D. McCauley. 1988. Extinction and recolonization: their effects on the genetic differentiation of local populations. Evolution 42:995–1005.

Wright, S. 1943. Isolation by distance. Genetics 28:114–138.

Zalewski, A., S. Piertney, H. Zalewska, and X. Lambin. 2009. Landscape barriers reduce gene flow in an invasive carnivore: geographical and local genetic structure of American mink in Scotland. Mol. Ecol. 18:1601–1615.

Associate Editor: J. Fordyce

Appendix A: Normalizing the Covariance Matrix

To renormalize equation (3) and compute the correlation matrix, we have to compute the diagonal elements of the variance–covariance matrix $\mathbf{Var}(\mathbf{Y})$ for the unsampled sites. Using the fact that the residuals of the regression equation (2) are of variance $\Psi_{yy} - \Psi_{xy}\Psi_{xx}^{-1}\Psi_{xy}$ (Bishop 2006), we can show that the variance–covariance matrix $\mathbf{Var}(\mathbf{Y})$ is given by

$$\mathbf{Var}(\mathbf{Y}) = \tau_\Psi \mathbf{Var}(\mathbf{X})\tau_\Psi^T + \Psi_{yy} - \Psi_{xy}\Psi_{xx}^{-1}\Psi_{xy}.$$

Appendix B: Gibbs Sampler

For the parameters of the correlogram model, we choose the following prior distributions

$$\alpha \sim \mathcal{U}(\min(\mathbf{Var}(\mathbf{X})) - \delta, \min(\mathbf{Var}(\mathbf{X})) + \delta) \quad (5)$$

$$\log_{10} \lambda \sim \mathcal{U}(-4, -1) \quad (6)$$

$$r \sim \mathcal{U}(\min D_{ij}, \max D_{ij}), \quad (7)$$

where $\min(\mathbf{Var}(\mathbf{X}))$ is the smallest element of the variance–covariance matrix of $\mathbf{X}$ and D denotes the pairwise geographical distance between sampled sites. Here, we adopt an empirical Bayes approach, because we partly use the data when defining the prior distributions.

Using the Bayes formula we have

$$\log(p(\alpha, \lambda, r|\mathbf{Var}(\mathbf{X}))) = \log(p(\mathbf{Var}(\mathbf{X})|\alpha, \lambda, r)) + \log(p(\alpha, \lambda, r)) + C_1, \quad (8)$$

where C_1 is a constant and the likelihood is given by the Wishart distribution Schmidt and O'Hagan (2003)

$$\log(p(\mathbf{Var}(\mathbf{X})|\alpha, \lambda, r)) = -\frac{l-1}{2}\log(|\Psi_{\mathbf{xx}}|) - \frac{l}{2}\mathrm{trace}(\Psi_{\mathbf{xx}}{}^{-1}\mathbf{Var}(\mathbf{X})) + C_2, \quad (9)$$

where C_2 is another constant, and l is the number of loci. We then simulate a sample of the joint posterior probability using an hybrid algorithm with Gibbs and Metropolis–Hastings (MH) updating steps. To obtain replicates from the conditional distributions $p(\alpha|\lambda, r, \mathbf{Var}(\mathbf{X}))$ and $p(\lambda|r, \alpha, \mathbf{Var}(\mathbf{X}))$, we evaluate the likelihood function over a grid for the triplet (α, λ, r). The evaluations of the likelihood function are performed before running the MCMC algorithm and are saved to be subsequently used during the course of the algorithm. We evaluate the conditional densities for each point of the grid using equations (8) and (9) and then we use a sampling algorithm for simulating discrete random variables with known probability masses. The updates for the parameter r are performed with an MH algorithm using as proposal a discrete random walk over the grid. The resolution of the grid is more acute for the parameter r than for the other two parameters, which explains why we consider an MH step that was less time-consuming than evaluating the likelihood function for all the points of the grid.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Figure S1. The parameters of the correlogram function affect the estimated values of local genetic differentiation.
Figure S2. Running time of the algorithm as a function of the number of sampled sites.
Figure S3. True and estimated maps of local genetic differentiation in a twodimensional range of dimension 15×10 containing three barriers.
Figure S4. Map of local genetic differentiation in a two-dimensional habitat with two genetic barriers when using the pairwise FST as dissimilarity matrix.
Figure S5. Effect of incomplete sampling when estimating local differentiation in a two-dimensional model with genetic barriers.
Figure S6. Effect of incomplete sampling when estimating local differentiation in a two-dimensional model with a gradient of gene flow.
Figure S7. Output of the software barrier in a one-dimensional and two-dimensional habitats that contain barriers to gene flow.
Figure S8. Multidimensional scaling in a two-dimensional model that contains two barriers to gene flow.
Figure S9. Isolation-by-distance for the Swedish dataset.
Figure S10. Barycentric locations of the Swedish counties and putative neighboring populations located 30 km away.
Figure S11. Local genetic differentiation as a function of the population density in Sweden.
Figure S12. Local genetic differentiation on a log scale for the 20 alpine plant species.
Figure S13. Average values across species of the normalized ranks of local differentiation.
Figure S14. Local differentiation in two-dimensional models when averaging the correlogram parameters over the prior distribution.
Figure S15. Local genetic differentiation for the Swedish dataset for various values of the distance between sampled sites and unsampled neighboring sites.
Table S1. Test of nonstationarity for the 20 alpine species.

Supplementary Material of "Non-stationary patterns of isolation-by-distance: inferring local genetic differentiation with Bayesian Kriging."

- Figure S1: The parameters of the correlogram function affect the estimated values of local genetic differentiation.
- Figure S2: Running time of the algorithm as a function of the number of sampled sites.
- Figure S3: True and estimated maps of local genetic differentiation in a two dimensional range of dimension 15×10 containing three barriers.
- Figure S4: Map of local genetic differentiation in a two-dimensional habitat with two genetic barriers when using the pairwise F_{ST} as dissimilarity matrix.
- Figure S5: Effect of incomplete sampling when estimating local differentiation in a two-dimensional model with genetic barriers.
- Figure S6: Effect of incomplete sampling when estimating local differentiation in a two-dimensional model with a gradient of gene flow.
- Figure S7: Output of the software *barrier* in a one-dimensional and two-dimensional habitats that contain barriers to gene flow.
- Figure S8: Multidimensional scaling in a two-dimensional model that contains two barriers to gene flow.
- Figure S9: Isolation by distance for the Swedish data set.
- Figure S10: Barycentric locations of the Swedish counties and putative neighboring populations located 30 km away.
- Figure S11: Local genetic differentiation as a function of the population density in Sweden.
- Figure S12: Local genetic differentiation on a log scale for the 20 alpine plant species.
- Figure S13: Average values across species of the normalized ranks of local differentiation.
- Figure S14: Local differentiation in 2-dimensional models when averaging the correlogram parameters over the prior distribution.
- Figure S15: Local genetic differentiation for the Swedish data set for various values of the distance between sampled sites and unsampled neighboring sites.
- Table S1: Test of non-stationarity for the 20 alpine species.

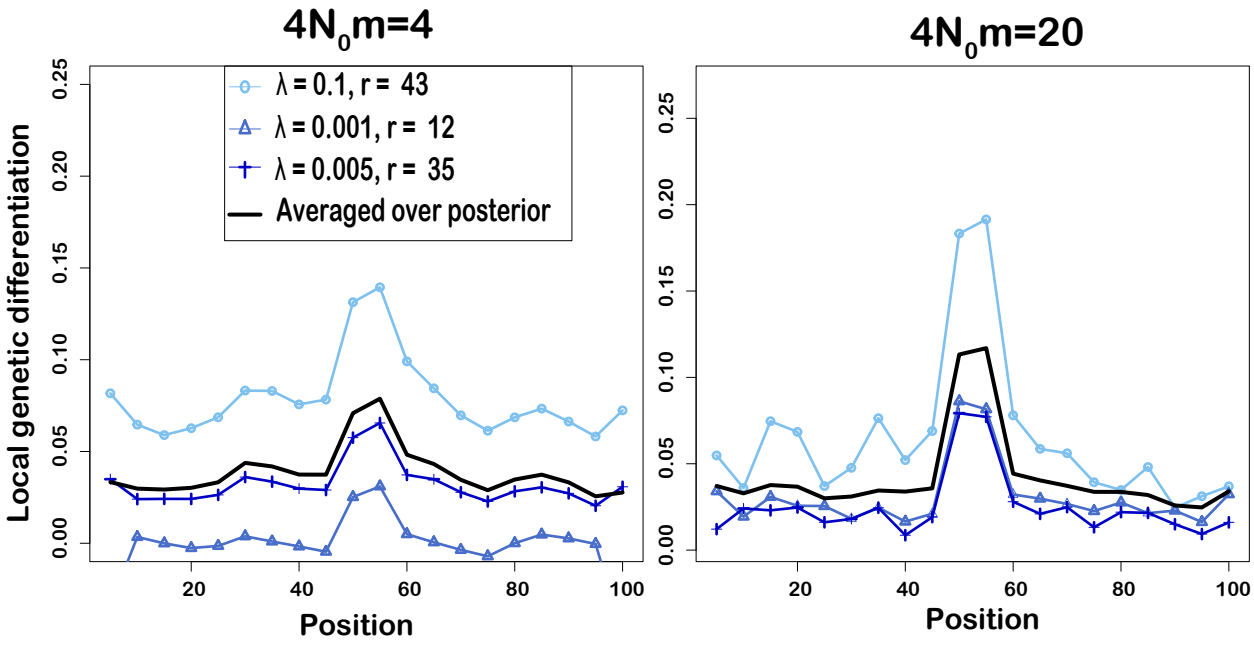


Figure S1: Estimation of local genetic differentiation in a one-dimensional range with a genetic barrier in the middle of the range. Local differentiation values are shown for different values of the correlogram parameters as well as when integrating over the posterior distribution of the corrrelogram parameters (α, λ, r). Local genetic differentiation corresponds to one minus the expected correlation of allele frequencies between sampled demes and their two closest neighboring demes. The genetic barrier is located between deme 50 and deme 51. Sampling was performed regularly with 20 equally-spaced populations picked among the total set of 100 populations.

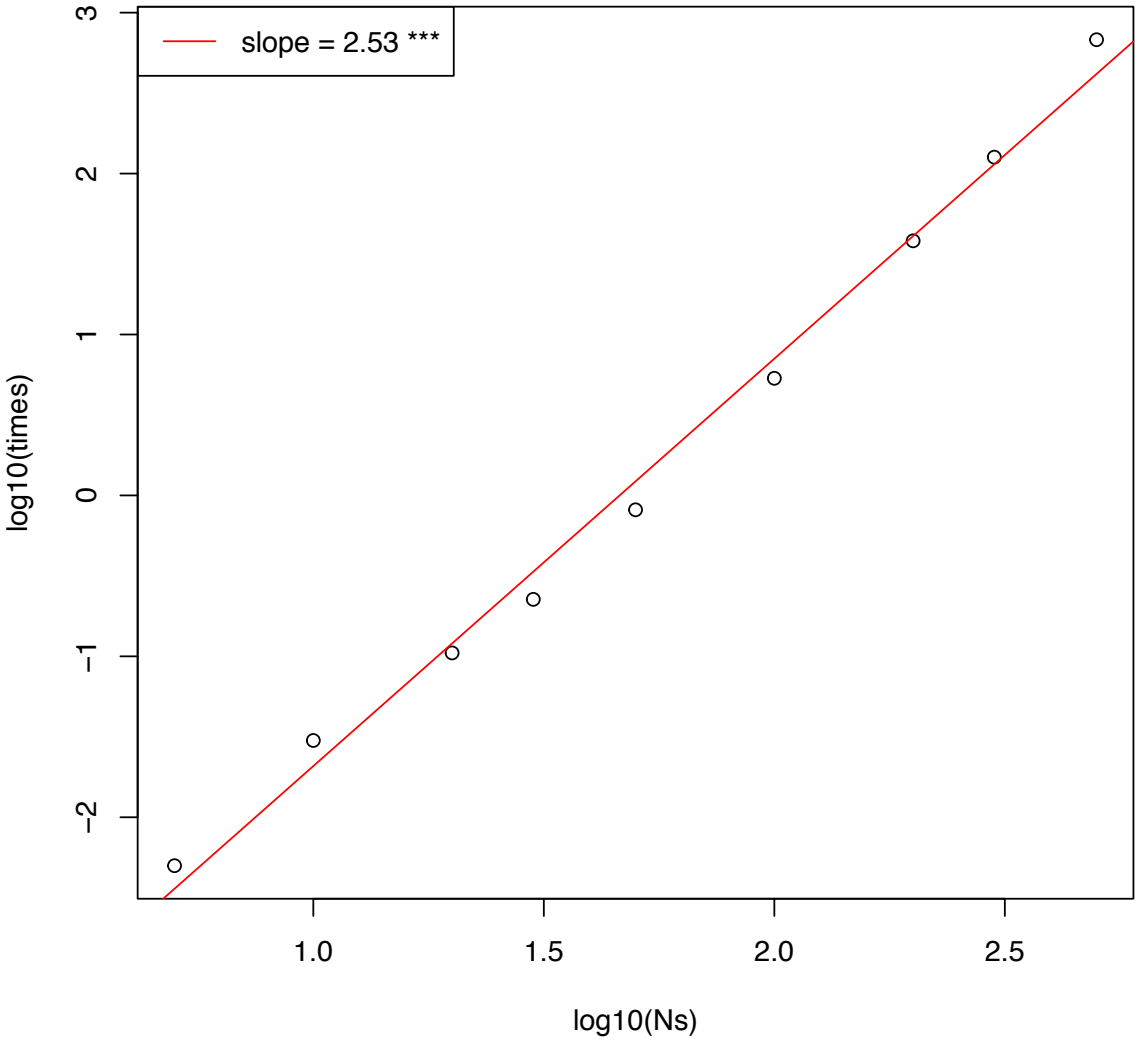


Figure S2: Running time of the algorithm (in seconds) as a function of the number (N_s) of sampled sites (log scale for both axes).

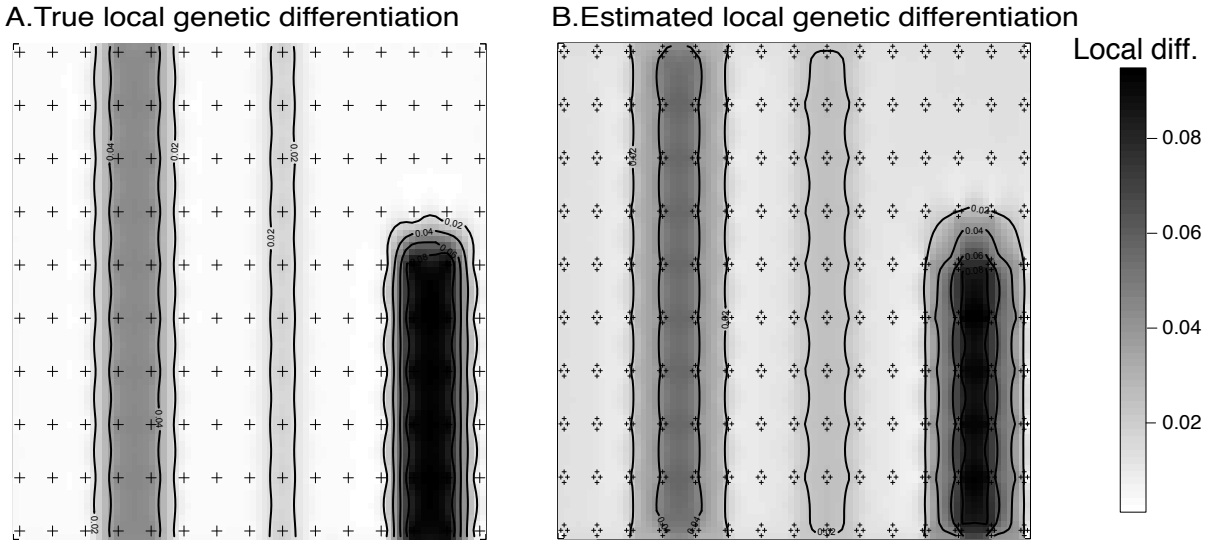


Figure S3: True and estimated maps of local genetic differentiation in a two dimensional range of dimension 15×10 containing three barriers. The local differentiation at sampled points corresponds to one minus the correlation between sampled points (large crosses in panel A) and fictive neighbors (small crosses in panel B) located at a distance of 0.1. The correlation matrix is specified with the non-stationary model of Paciorek and Schervish (2006) considering smaller range values (parameter r of equation (4) of the main text) for the zones of abrupt genetic change. The true local differentiation of panel A is obtained using the convolution formula of Paciorek and Schervish (2006) at distance 0.1 which provides pairwise correlation based on the landscape of the range parameter r.

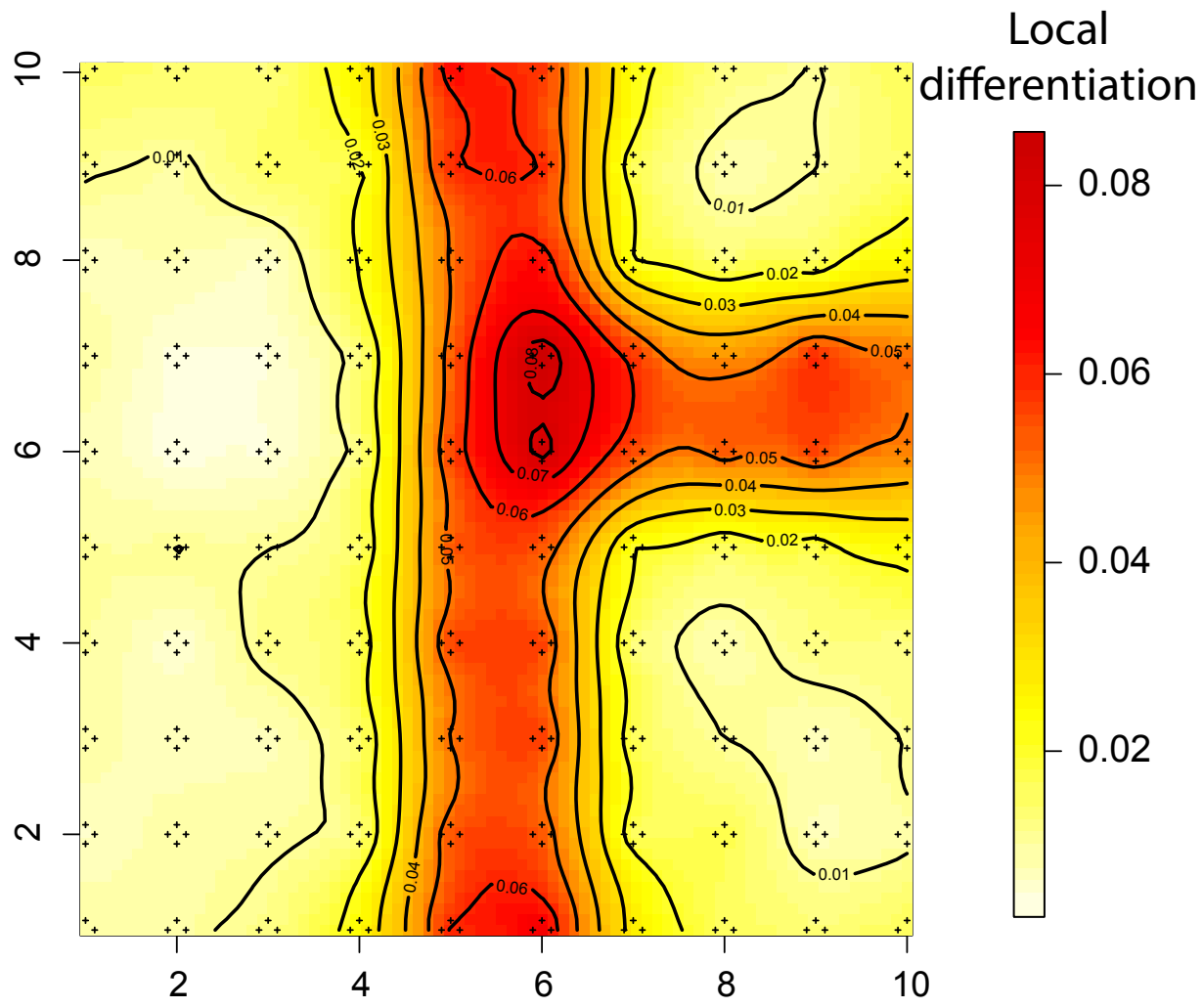


Figure S4: Map of local genetic differentiation in a two-dimensional model with two genetic barriers (see Figure 4 of the main text for the details of the generating process). Local genetic differentiation was estimated using the pairwise matrix of (1-F_{ST}). The measure of local differentiation corresponds to F_{ST} between sampled populations and unsampled neighbors located at a distance of 0.1 from the sampled populations.

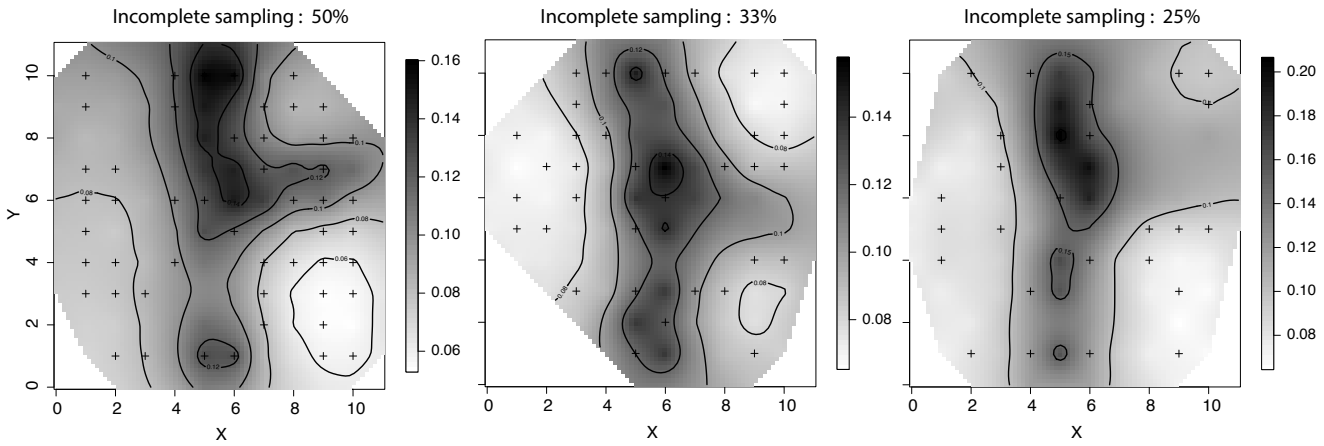


Figure S5: Map of local genetic differentiation in a two-dimensional model with two genetic barriers and incomplete sampling (see Figure 4 of the main text for the details of the generating process). Local genetic differentiation was estimated using the pairwise correlation matrix, and it measures one minus the expected correlation between sampled populations and unsampled neighbors located at a distance of 0.1 from the sampled populations.

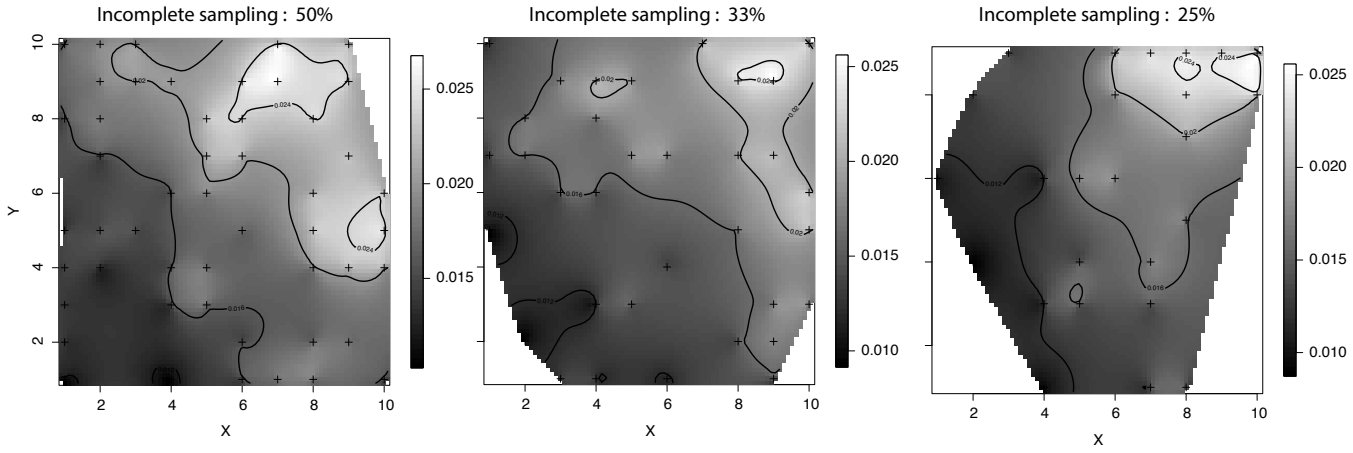
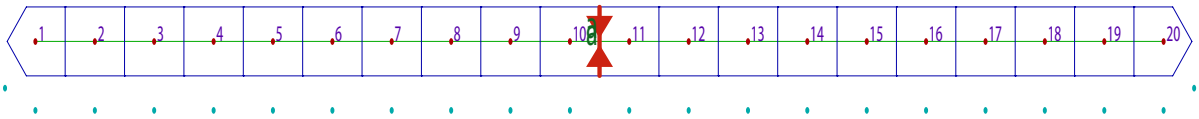


Figure S6: Map of local genetic differentiation in a two-dimensional model with a gradient of gene flow and incomplete sampling (see Figure 5 of the main text for the details of the generating process). Local genetic differentiation was estimated using the pairwise correlation matrix, and it measures one minus the expected correlation between sampled populations and unsampled neighbors located at a distance of 0.1 from the sampled populations.

A. Barrier outputs for a 1-dimensionnal stepping stone model with one barrier.



B. Barrier outputs for a 2-dimensionnal stepping stone model with two barriers.

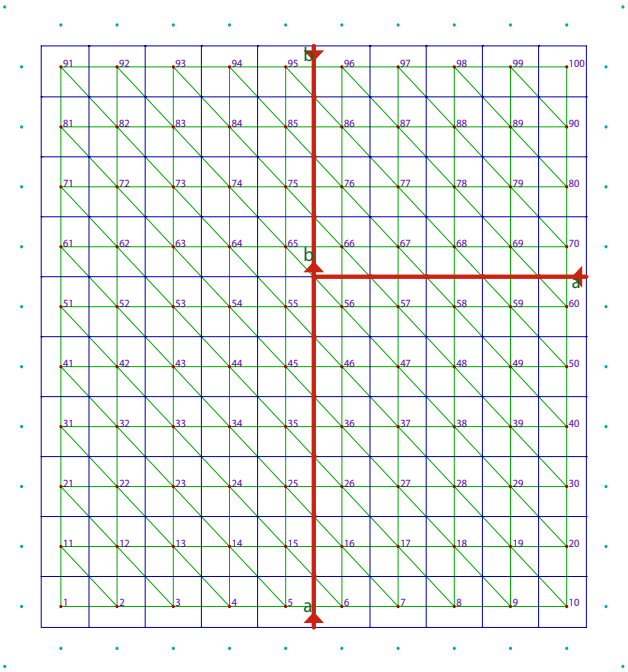


Figure S7: Outputs of the software *barrier*. Panel A) output for the 1-dimensional stepping stone example with 1 barrier. Panel B) output for the 2-dimensional stepping stone example with 2 barriers. In these examples, the barriers are correctly found by the software *barrier*.

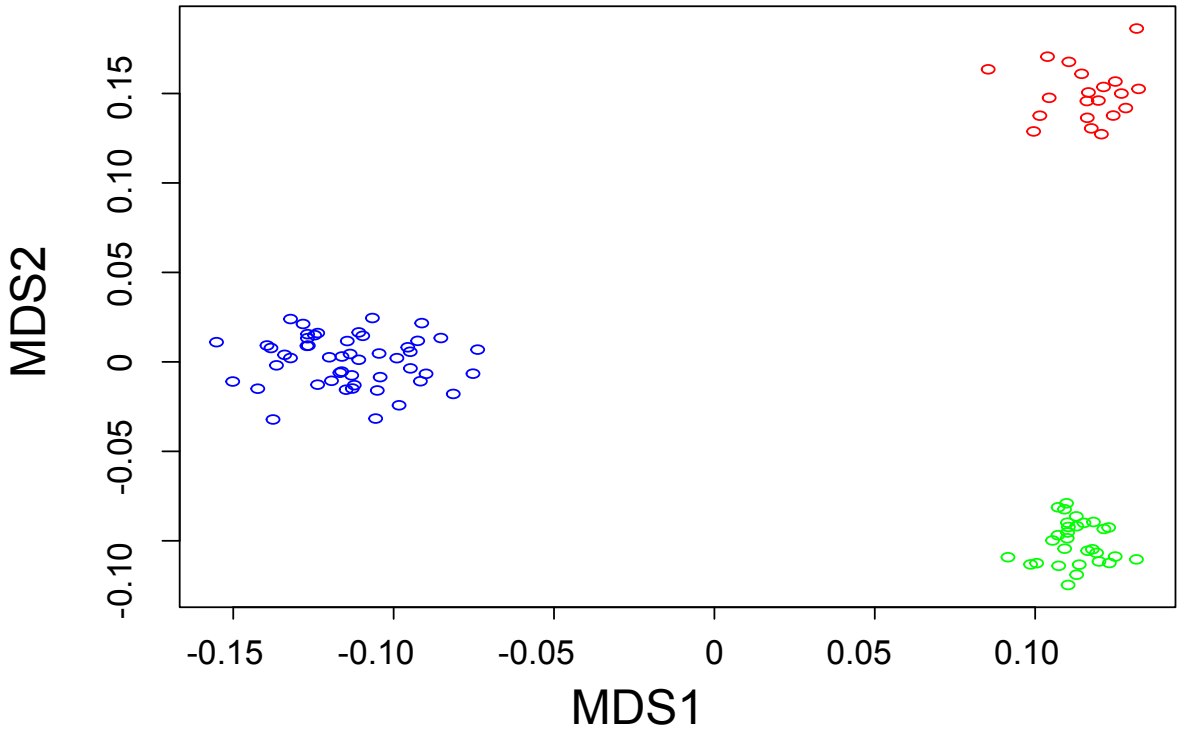


Figure S8: Multidimensional scaling in a 2-dimensional stepping stone model with 2 barriers to gene flow. Dots represent populations and are colored according to the location of the population with respect to the barriers.

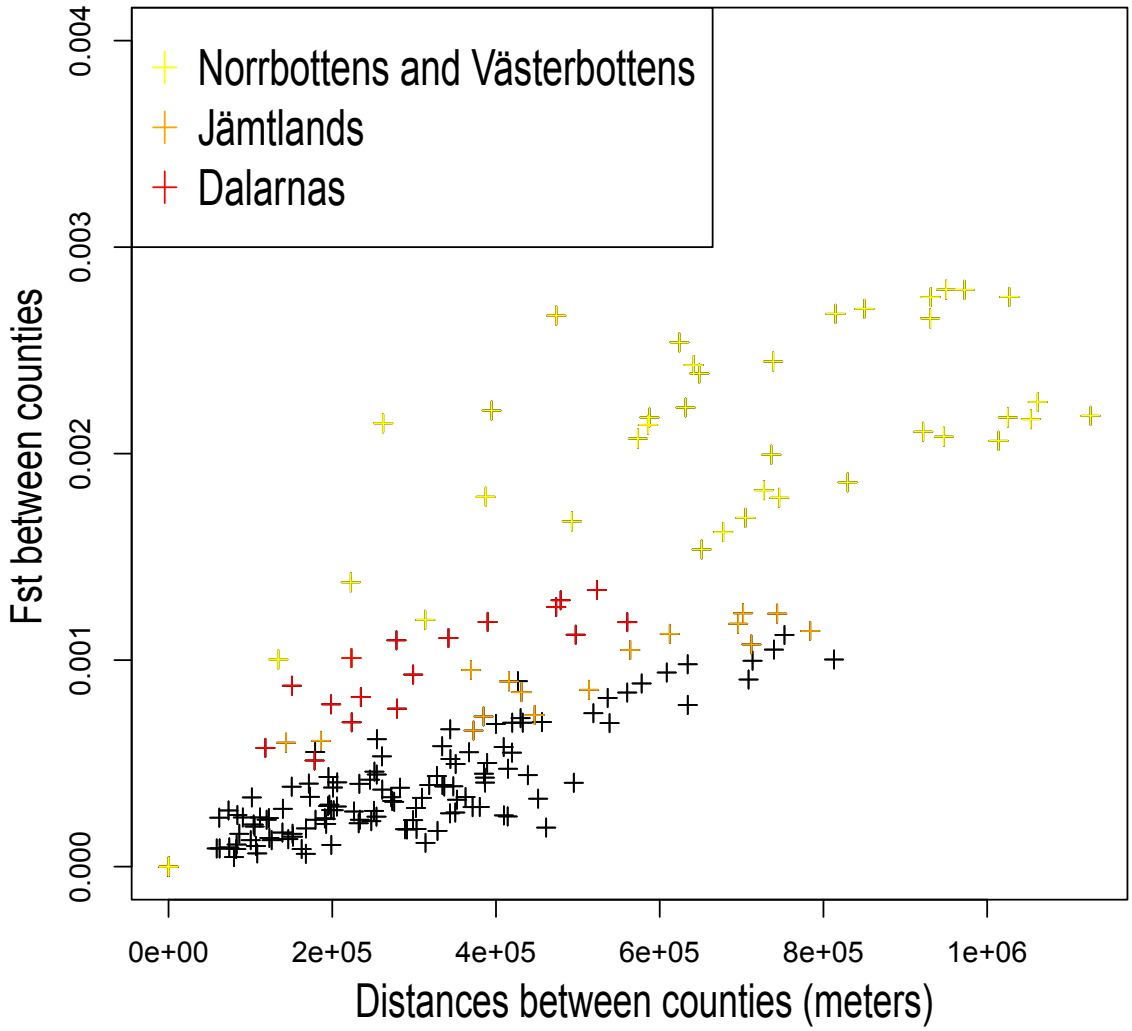


Figure S9: F_{st} explained by geographical distance for the Swedish counties. The pairwise comparisons involving the four counties with the largest values of local genetic differentiation are colored (yellow, red, orange).

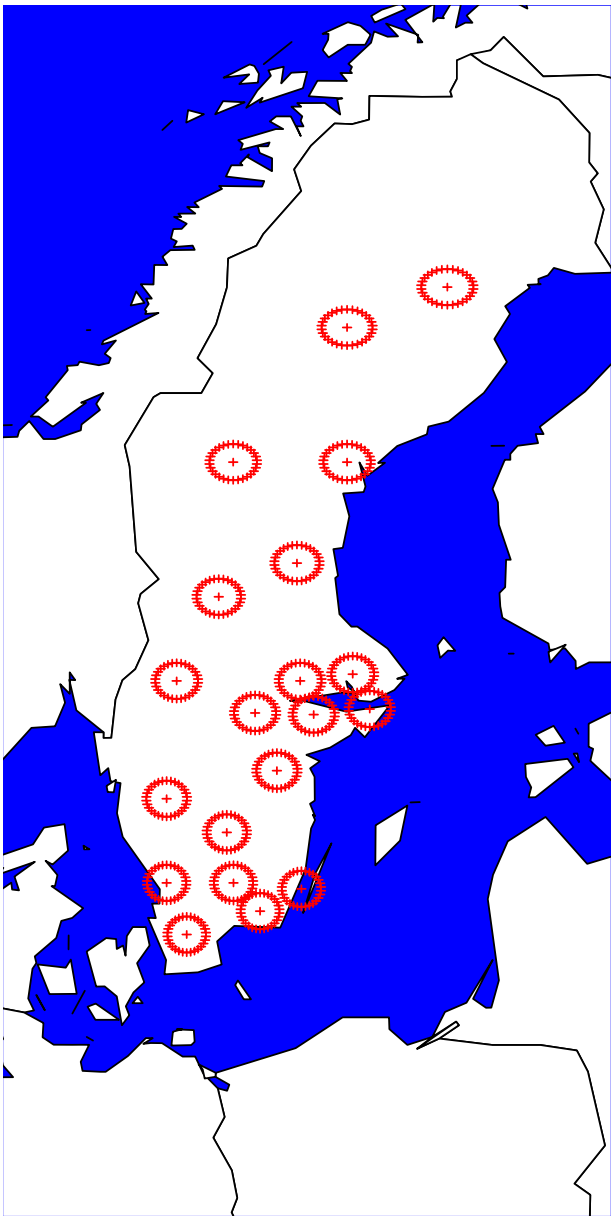


Figure S10: Barycentric locations of the Swedish counties and putative neighboring populations located 30 km away.

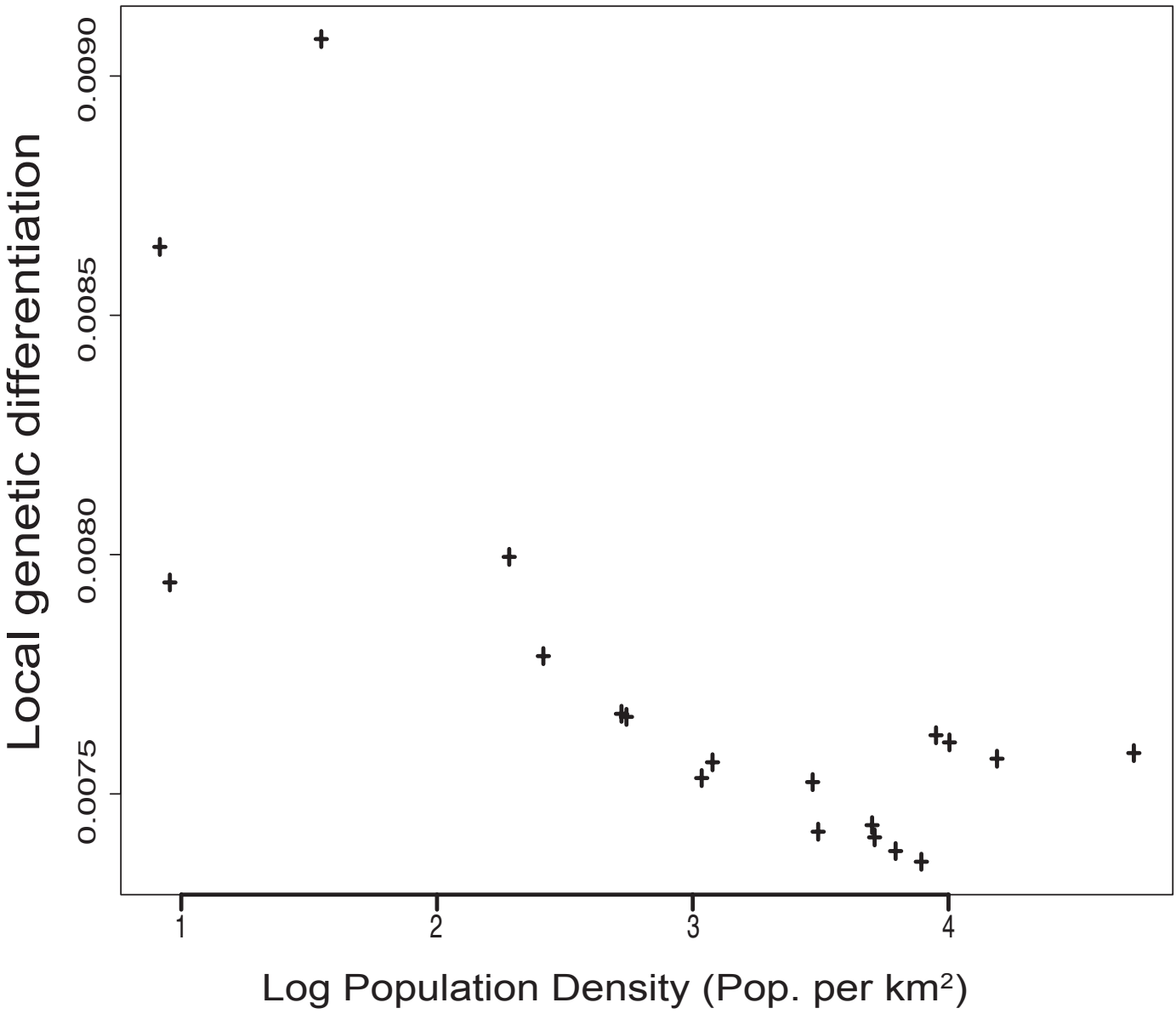


Figure S11: Local genetic differentiation as a function of population density in Sweden. The dot corresponding to the Stockholm county is not shown since its population density is out of range by several order of magnitude.

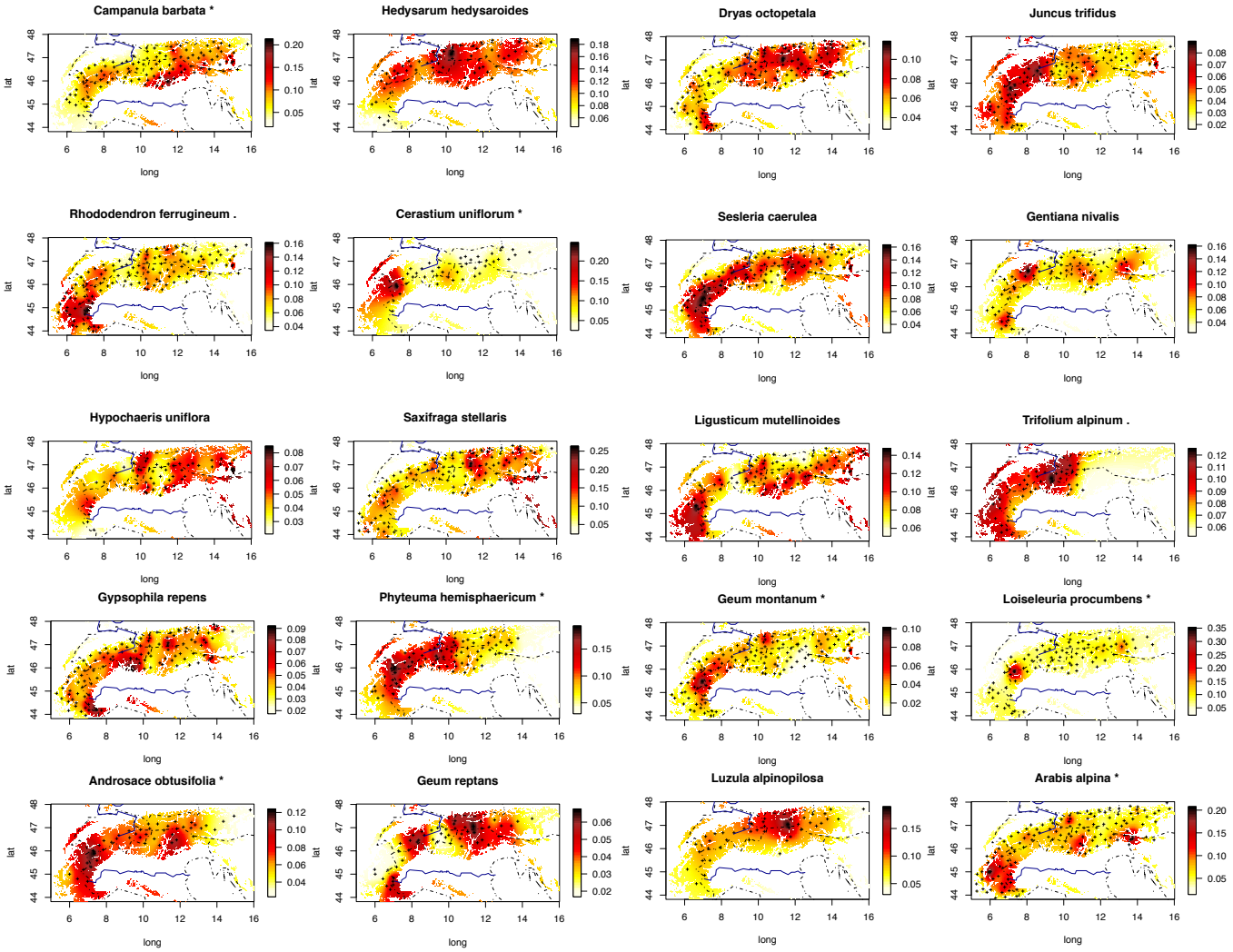


Figure S12: Local genetic differentiation on a log scale for the 20 alpine plant species. We compute allele frequencies at each sampling site–possibly using one individual only–and we consider the matrix of correlation between allele frequencies as input similarity matrix. Local genetic differentiation corresponds to one minus the expected correlation between sampled populations and neighboring populations located at 8 km. The symbol * (resp. ·) corresponds to a rejection of stationarity at $\alpha = 5\%$ (resp. $\alpha = 10\%$) accounting for Bonferroni correction.

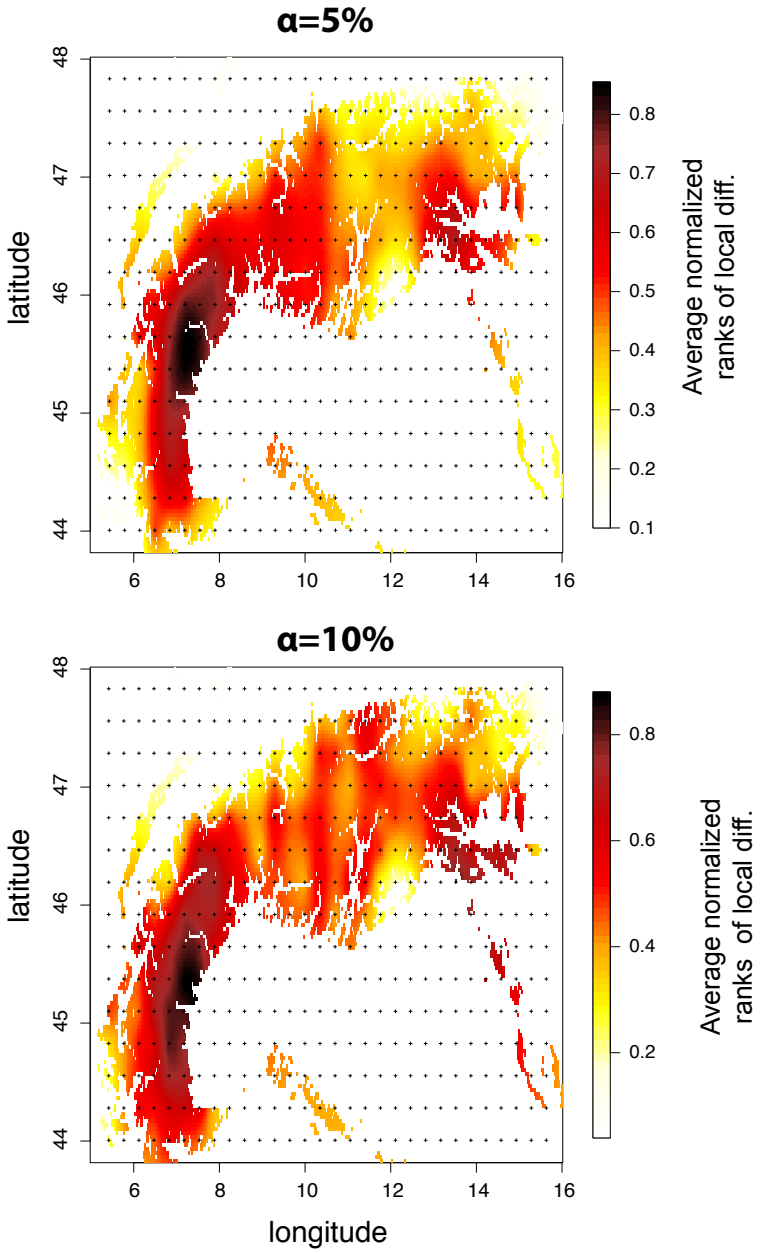


Figure S13: Normalized ranks of local differentiation (between 0 and 1) averaged across species. In the upper panel, the initial matrices for computing local differentiation are the correlation matrices whereas it is the F_{ST} matrix for the lower panel.

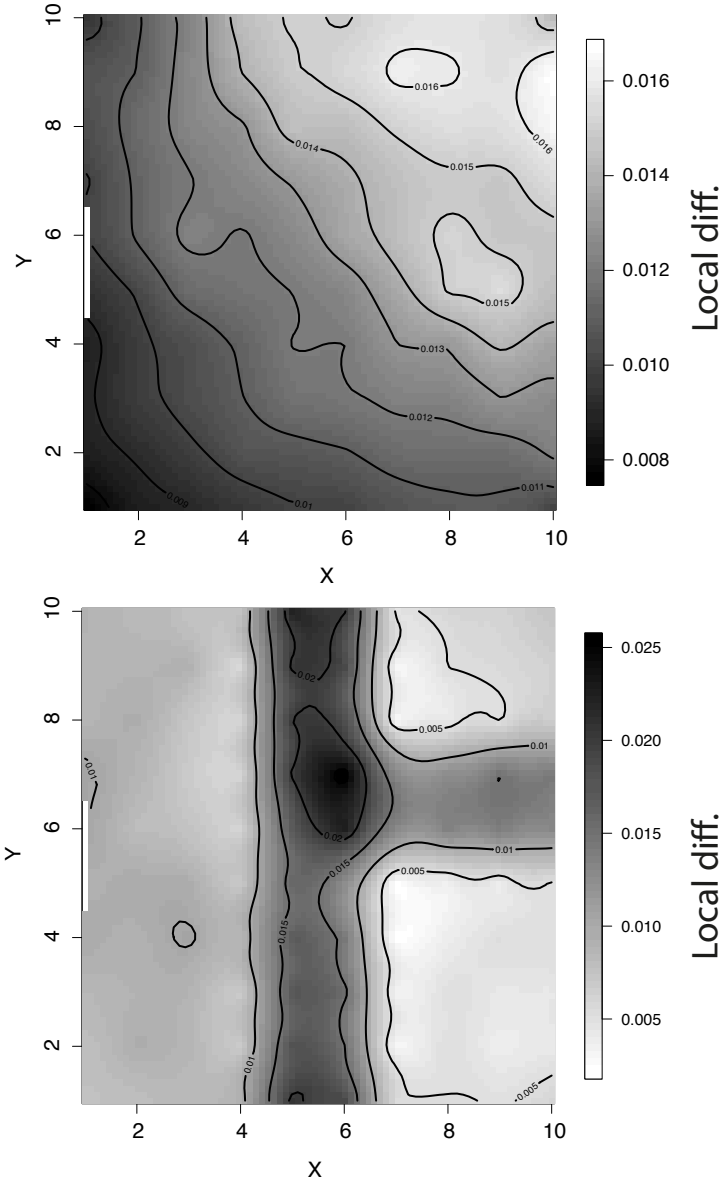


Figure S14: Maps of local genetic differentiation in a two-dimensional model with two genetic barriers (lower panel) and with a gradient of gene flow (upper panel) when integrating the correlogram parameters over the prior distribution instead of the posterior. Local genetic differentiation was estimated using the pairwise correlation matrix, and it measures one minus the expected correlation between sampled populations and unsampled neighbors located at a distance of 0.1 from the sampled populations.

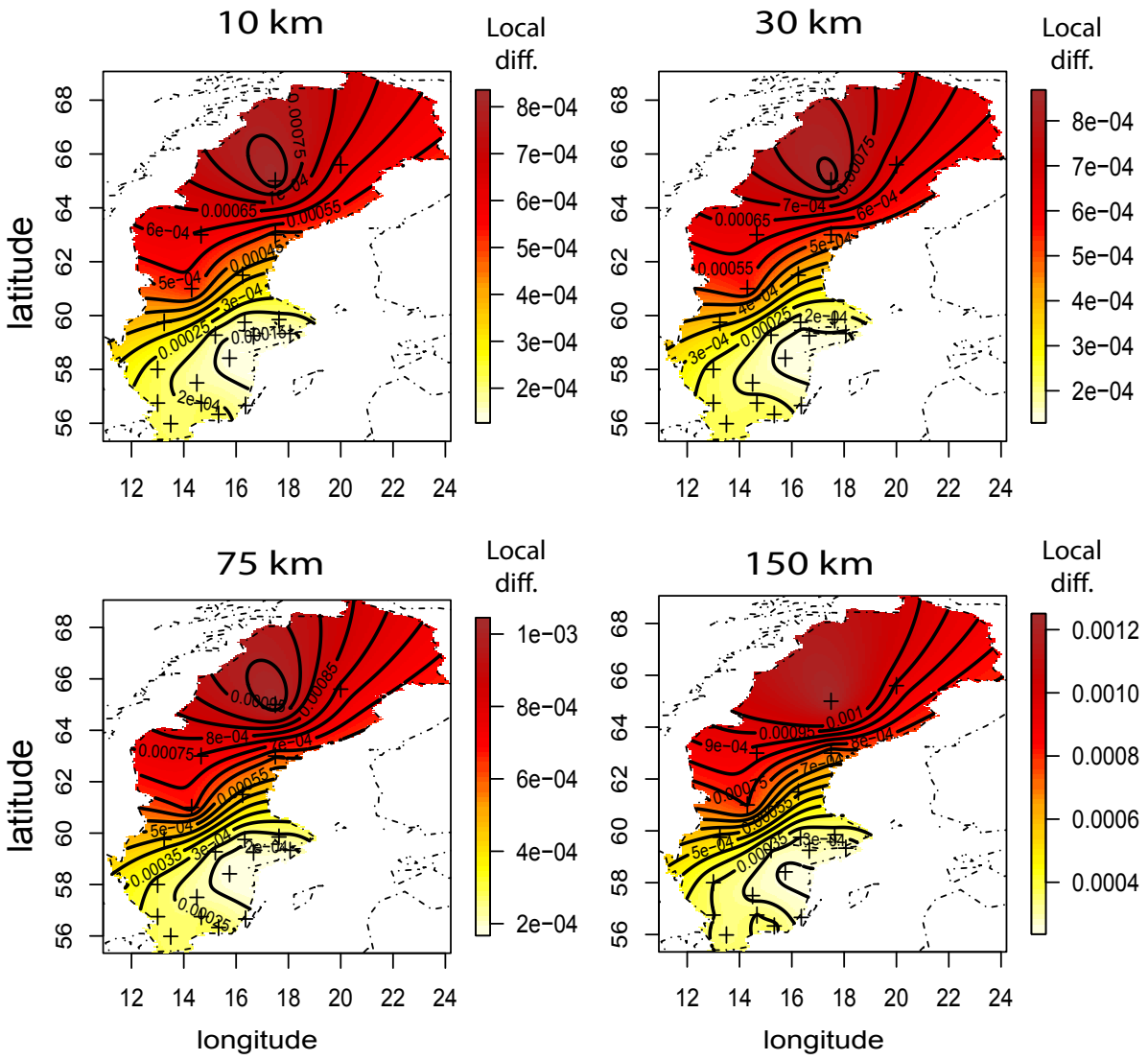


Figure S15: Local genetic differentiation for the Swedish data set for various values of the distance between sampled sites and unsampled neighboring sites. Distance to neighbors are of 10, 30, 75 or 150 km.

Names	P-value (CV)	P-value (Dcorr)	Significance
Androsace obtusifolia	0.54	0.01	*
Geum reptans	0.25	0.87	
Luzula alpinopilosa	0.31	0.49	
Arabis alpina	0.00	0.25	*
Gypsophila repens	0.15	0.65	
Phyteuma hemisphaericum	0.06	0.01	*
Campanula barbata	0.02	0.17	*
Hedysarum hedysaroides	0.99	0.66	
Rhododendron ferrugineum	0.43	0.05	·
Cerastium uniflorum	0.00	0.08	*
Hypochaeris uniflora	0.81	0.24	
Saxifraga stellaris	0.19	0.18	
Dryas octopetala	0.72	0.18	
Juncus trifidus	0.18	0.15	
Sesleria caerulea	0.47	0.11	
Gentiana nivalis	0.08	0.91	
Ligusticum mutellinoides	1.00	0.95	
Trifolium alpinum	0.05	0.95	·
Geum montanum	0.00	0.44	*
Loiseleuria procumbens	0.01	0.82	*

Table 1: Test of non-stationarity for the 20 alpine species. CV denotes the coefficient of variation and Dcorr corresponds to the distance correlation statistic. The symbol * (resp. ·) corresponds to a rejection of stationarity at $\alpha = 5\%$ (resp. $\alpha = 10\%$) after accounting for Bonferroni correction.

HaploPOP: a software that improves population assignment by combining markers into haplotypes

Nicolas Duforet-Frebourg*+[1] , Lucie M. Gattepaille+[2] , Michael G.B. Blum[1] and Mattias Jakobsson*[2,3]

[1]Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG, Grenoble, France
[2]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[3]Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Email: Nicolas Duforet-Frebourg*- nicolas.duforet@imag.fr; Lucie M. Gattepaille - lucie.gattepaille@gmail.com; Michael Blum - michael.blum@imag.fr; Mattias Jakobsson*- mattias.jakobsson@ebc.uu.se;

*Corresponding author

+ Both authors contributed equally

Abstract

In ecology and forensics, population assignment uses molecular markers to assign individuals into groups. However, assigning individuals to populations can be difficult when the level of genetic differentiation among populations is too small. To improve the accuracy of population assignment, we present a new algorithm, implemented in the *HaploPOP* software, that combines markers into haplotypes. The algorithm is based on a criterion called the Gain of Informativeness for Assignement that provides a measure to decide if a pair of markers should be combined into haplotypes or not. Because complete exploration of all possible solution for constructing haplotypes is computationally prohibitive, our new algorithm uses a greedy algorithm based on windows of fixed sizes. We evaluate the performance of population assignment using a split-validation approach. We consider both simulated SNP data and human genetic data of individuals from Spain and Portugal. Our results show that constructing haplotypes with *HaploPOP* can substantially reduce assignment error. The *HaploPOP* software is freely available as a command-line software at http://www.ebc.uu.se/Jakobsson/software/HaploPOP/.

Background

An application of molecular markers, useful in forensic science and ecology, concerns the assignment of individuals into predefined populations [1, 2]. With the advent of new sequencing technologies, the number of available molecular markers in different species is rapidly increasing. Due to recombination events, short-distant markers in the genome are in Linkage Disequilibrium (LD), and dense datasets tend to contain increasingly correlated markers. Such correlation is usually perceived as a nuisance factor in statistical analyses, since it violates the common asumption of independence between markers. This statistical nuisance can be overcome by pruning SNPs using for example the software *PLINK* [3]. However a lot of genetic information may be lost because of the pruning process. It can be beneficial in the context of population structure to combine correlated markers into haplotypes to provide information about population structure at a finer scale [4]. It is even more important to combine correlated markers when assignment methods are challenged because of a low level of genetic differentiation between populations [5].

Gattepaille and Jakobsson [6] introduced the Gain of Informativeness for Assignment (GIA), which is a statistic measuring the gain in information for population assignment by combining two markers into haplotypes. GIA is based upon an ancestry information criterion that measures to what extent a molecular marker is informative about population assignment [7]. GIA is defined as the difference between the ancestry information carried by the two markers independently and the ancestry information carried by the haplotypes resulting from the combination of the two markers. In the case of two markers in linkage equilibrium, the gain is non-positive indicating that no haplotypes should constructed. Using GIA, it has been shown that correct assignment can be substantially increased [6]. However, a major combinatorial challenge arises when using GIA because of the prohibitively large number of pairs of markers that can be combined into haplotypes.

In this article, we present a new algorithm that efficiently uses GIA to build haplotypes for population assignment. The algorithm needs *labelled* individuals whose population of origin is known. Based on these *labelled* individuals, the algorithm uses GIA to construct haplotypes. To handle large number of molecular markers, we provide an heuristic where only markers located within the same genomic region can be combined to form haplotypes. Combining markers into haplotypes is a recursive process so that haplotypes can result from the combination of two or more molecular markers. Then we recode the raw genotype data to account for the fact that markers have been combined into haplotypes. The new data file containing both genotypic and haplotypic information can then be used to assign individuals to populations using Principal Component Analysis, or the software ADMIXTURE for instance [8, 9].

Because the construction of haplotypes uses predefined populations, there is a risk of overfitting. If the

evaluation of population assignment is performed with the same individuals that were used to construct the haplotypes, there is a risk of underestimating the assignment error. Additionally, the constructed haplotypes can generate artificial population structure although there is no true stratification between the predefined populations. Both problems arise because the construction of haplotypes can exagerate the differentiation between populations. To get a fair evaluation of population assignment and of the level of differentiation between population, we implement a split-validation approach [10] where we do not use the same individuals to construct the haplotypes and to evaluate assignment error. The haplotypes are built using a subset of the individuals, consisting of a training set. The quality of population assignment and the level of population stratification can then be assessed using the remaining individuals. The fact that these individuals cluster in accordance with the set of predefined populations provides an evidence for population structure, which may have been missed when using genotype makers only.

Our new algorithm for combining markers into haplotypes is implemented in the software *HaploPOP*. The software is a command-line software written in C. We give examples of how to use *haploPOP* to perform population assignment with SNP data simulated under a population divergence model. We also show that *HaploPOP* improves the assignment of individuals from Spain and Portugal using the POPRES dataset that contains 447,245 SNPs [11].

Implementation

Gain in Informativness for Assignment

The *Gain of Informativeness for Assignement* (*GIA*) is a one-dimensional statistic that provides a criterion to decide whether markers should be combined into haplotypes [6]. It is based on the *Informativeness for Assignment* (*IA*) statistic, which measures how informative a marker is for assigning individuals to different populations [7]. The more different the allele frequencies are in a set of predefined populations, the more informative the marker is to assign individuals of unknown origin to their source populations, and the larger is the *IA* statistic. Denoting by K the number of populations, by N the number of alleles of the marker under consideration, by $p_j^{(i)}$ the frequency of allele j in population i, and by $\overline{p}_j$ the average frequency of allele j across all populations, the *IA* statistic is computed as follows [7]

$$IA = \sum_{j=1}^{N} \left(-\overline{p}_j \log \overline{p}_j + \sum_{i=1}^{K} \frac{p_j^{(i)}}{K} \log p_j^{(i)} \right). \quad (1)$$

Given two multi-allelic markers M_1 and M_2, the question is whether combining M_1 with M_2 into a haplotype

marker H improves the assignment of individuals to the predefined populations. *GIA* computes the difference between the informativeness for assignment of H and the sum of the informativenesses of M_1 and M_2

$$GIA = IA(H) - (IA(M_1) + IA(M_2)). \tag{2}$$

If *GIA* is positive, it suggests that population assignment is improved by considering the haplotype H in contrast to using the two markers separately. However, if *GIA* is negative, there is no advantage of combining the two markers into a haplotype. In particular, theoretical result show that if the two markers are in linkage equilibrium, *GIA* is expected to be non-positive [6].

Maximizing information for assignment

We assume that genotype data are available for n individuals at l molecular markers $(M_1, ...M_l)$. We also assume that the dataset has been phased, where all individuals have been phased together in one go to avoid introducing any haplotype difference due to phasing (note that there may still be switch errors from the phasing, but these should affect all individuals similarly). The approach implemented in the software *HaploPOP* builds a set of haplotypes that increases the total informativeness for assignment contained in the raw genotype data. To find the optimal haplotype set Γ_0, we address the maximization problem

$$\begin{cases} \Gamma_0 = argMax_\Gamma \sum_{H \in \Gamma} IA(H) \\ \Gamma \in Part(M_1 ... M_l) \end{cases} \tag{3}$$

where $Part(M_1 ... M_l)$ is the set of all possible partitions of the l markers. Because the number of partitions in a dataset of l markers is too large (given by Bell's number, [12]), we can not evaluate the objective function for all possible partitions. A commonly used heuristic to approximate the maximum of such problem is to apply a greedy strategy, although it can perform arbitrarily good. In the case of increasing Informativeness for Assignment, the resulting haplotypes always provide genetic data with increased informativeness. Because the cost of this algorithm increases rapidly with the number of genetic markers, we limit potential combinations of markers to occur within fixed size windows.

Algorithm

In a first step, the algorithm constructs the haplotypes from the phased genotype file of individuals with known origin and returns a *haplotype coding* file that provides the correspondence between the haplotypes and the initial markers. This is the `LEARN` option of *HaploPOP*. The construction of haplotypes is constrained

by a window-type approach. The set $(M_1, \dots M_l)$ of markers is divided into subsets of contiguous markers corresponding to the genomic windows. Haplotypes should be combinations of markers being part of the same window. The window size is chosen by the user and can be defined based on the number of markers, on the number of base pairs, or based on genetic distance. In every window, the *GIA* statistic is computed for all pairs of markers, and the pair with the highest *GIA* statistic is merged to form haplotypes. Combinations proceed until there is no more information to gain in the window (i.e. no more pair for which $GIA > 0$, or a certain positive threshold). A particular haplotype-loci formed by a combination of markers is thereafter treated as any other marker of the particular window and can be combined with other markers in a recursive manner.

We denote by n the number of *labelled* individuals whose population of origin is known, and by l the total number of initial markers. The greedy algorithm proceeds as follows

- divide the $2n \times l$ data matrix in contiguous windows.
- for every window do
  - Calculate *GIA* for every pair of markers.
  - while for all markers M and M', $max_{M,M'}(GIA(M, M') > 0)$, do
    - $(M_0, M_0') = argmax_{M,M'} GIA(M, M') > 0$.
    - Combine the markers M_0 and M_0' to make a haplotypic marker H_0.
    - Remove the *GIA* statistics involving M_0 and M_0' and compute the new *GIA* statistics with pairs of markers that include H_0.

At every end of the inner loop, the algorithm partitions the markers into a set of haplotypes that increases the score of the objective function (3). It stops when no pairwise combination of haplotype-markers improves the total score of the partition. A warning is raised when the number of haplotype-alleles reaches the number of chromosomes $2n$ making haplotypes useless, since they become private to every individual and do not provide any useful information for assignment.

In a second step, *HaploPOP* combines the initial genotype file into a haplotype file according to the combinations of markers constructed at the previous step. The genotype file can contain individuals of unknown origin whose population assignment is the user's goal, as well as the population samples used to construct the haplotypes, or any additional individuals of interest. This corresponds to the `APPLY` function

of *HaploPOP*. When two markers are combined, the resulting haplotype alleles are coded in a range from 0 to the number of haplotypic alleles minus one, in order of appearance of the individuals.

Window size

A key parameter of the method is the window size. This parameter is important for both the speed of the algorithm, and the level of informativeness of the haplotypes. Window size governs the number of operation performed by the algorithm. In the case of constant window sizes of S markers, the number of windows is $n_{\text{window}} = l/S$, and the cost C of the algorithm in number of operations is

$$C(n_{window}, S, n, K) = O(n_{window} \times (2nS^2K + S^3)) \tag{4}$$

where K is the number of populations in the data. The algorithm scales very well for genome wide data sets, since for a given window size S, the cost of the algorithm is proportional to the number of markers l in the data.

Combinations with a large number of haplotype-alleles, fitting too closely genotypes of individuals from the training set, would perform poorly for another set of individuals from the same reference populations. We refer to such phenomenon as overfitting and restricting the computation of combinations to windows of limited size mitigates the effects of overfitting. By increasing the window size, we can consider more and more complex combinations of haplotypes which increases the number of degrees of freedom of the algorithm. Models or algorithms with too many degrees of freedom with respect to the data are usually prone to overfitting. Combining closely linked markers into haplotype may integrate extra information about ancestry, potentially from recombination events that should in principle harbor information about ancestry. Thus, information gained between physically close markers is more likely to be useful for population assignment than marker-combinations for markers that are far apart, which may be prone to overfitting. Limiting the size of the window is a way to avoid construction of haplotypes that artificially caracterises a subgroup of individuals in the traning set without link to past recombination effect. Such haplotypes are undesirable because they would carry no information at the population level. As we demonstrate in the Results section, the window size can have a strong impact on the performance of the created combinations of markers, and an optimal value generally exists. The optimal window size depends on multiple factors, such as the effective population sizes of the populations and the extent of Linkage Disequilibrium. In practice, an approximation of a well-performing window size can be determined using the decay of Linkage Disequilibrium in the genome. In the human example below, optimal window level is between 10 and 20 kb, which corresponds to the

distance at which LD decay vanishes [13].

Split-validation

To validate the gain in assignment provided by the constructed haplotypes, we implement a split-validation technique [10]. For each population, we randomly split the individuals into two subsets consisting of the training subset used to learn the haplotypes and of the validation subset used to compute assignment errors. To assign individuals to populations, we use Principal Component Analysis (PCA) as implemented in the software EIGENSOFT [8]. The number of principal components we consider equals the number of populations used for constructing the haplotypes minus one [8]. For each individual of the validation set, we compute Euclidean distances on the PC space between this individual and the barycentric coordinates of each population. We assign individuals to the population that has the closest barycenter. Because the origin of individuals in the validation set was known, we can measure the number of incorrectly assigned individuals from the validation set. Note that the assessment depends on the assignment procedure itself and that different assignment procedures might lead to different assignment errors (see [6] for a comparison of different assignment strategies). However, since we are primarily interested in the comparison between assignment using the raw genotype data and assignment using combined markers found with *HaploPOP*, we focus on a single assignment approach based on PCA. Since each haplotype-allele is treated as a unique allele (without any relationship with other alleles), we use a presence/absence coding for each haplotype-allele. We determine the PC axes using both the individuals from the training and validation sets and assign individuals as described above.

Results and Discussion

Application on simulated data

We evaluate the *HaploPOP* software on simulated data generated by the software *ms* [14]. We simulate 200kb sequences from a 3-population divergence model. We set the effective population sizes of all populations to $N_e = 1,000$, the mutation rate to $\theta = 48$, and we consider a sample of 100 individuals in each population. We generate four datasets with effective recombination rates of $\rho = 30, 60, 120, 240$ and replicate this procedure 10 times for each value of the recombination rate. The population divergence between population 2 and population 3 was set to occur $T_1 = 0.025$ coalescent time units and the population divergence between population 2 and population 1, was set to occur $T_2 = .05$ coalescent time units.

To validate the gain in assignment provided by the constructed haplotypes, we used a split-validation

technique. We consider 50 randomly picked individuals in each population of the training set and the validation set contains the 50 remaining individuals from each populations. Figure 1 shows that assignment is improved when constructing haplotypes rather than when using the raw genotype data (window size=1 for the raw genotype data in Figure 1). The error decreases when using larger window sizes for constructing haplotypes. However, most of the gain for assignment is found when moving from genotypes to haplotypes spanning at most 50 kb (Figure 1). Compared to the error of assignment obtained with genotype data, the error is decreased by $20-70\%$ depending on the recombination rate. The improved assignment if constructing haplotypes spanning 200 kb instead of 50 kb is less pronounced and the error is reduced by at most 8%. If we compare the results obtained with the different recombination rates, we find that the greater recombination rates increase information for assigning individuals to populations (Figure 1).

Application on human data

We investigate to what extent constructing haplotypes with *HaploPOP* improves population assignment of Spanish and Portuguese individuals from the POPRES phased dataset, which contains 447,245 SNPs [11]. Considering the first two PCs using PCA based on all SNP-genotype data, we find that the two populations can not be distinguished (Figure 2). PCA for all European individuals of the POPRES collection also could not distinguish between Spanish and Portuguese individuals [15]. Using *HaploPOP*, we construct the haplotypes that are informative to discriminate between Spanish and Portuguese individuals. We randomly select half of the Spanish and half of the Portuguese for constructing the training set. We then perform PCA based on the haplotype markers generated by *HaploPOP* and compute the PC scores for the individuals from the training set and the validation set (Figure 3A). On PC1, the Spanish and Portuguese samples from the training set are clearly separated. By contrast, the Spanish and Portuguese individuals from the validation set overlap but a large majority of these individuals are pulled in the direction of their candidate population of origin.

To show that the population labels, Portugal and Spain, correspond to true population differentiation, we generate a control dataset. We assign arbitrarily each individual to a label A or B. By construction, individuals labelled by A (or B) contain both individuals from Spain and Portugal. Using the training set, we learn the haplotypes that are discriminant between A and B. Using the validation set, we find that the individuals labelled by A and B cannot be separated by PCA. This show that the separation between Spanish and Portuguese individuals corresponds to true population differentiation.

This example emphasizes two important features of *HaploPOP*. First, the haplotypes constructed based on

the training set are very efficient in separating individuals of the training set, regardless of any stratification between candidate populations, and this should not be interpreted as an indication of actual population structure. Second, evidence for population structure comes from the ability of the constructed haplotypes to distinguish between individuals that were not used in the training process. Comparing only the individuals from the training set on a PC plot can either lead to the wrong impression that two populations can be distinguished (Figure 3B) or at least exaggerate the ability to distinguish between the two populations (Figure 3A) because of overfitting. However, if a validation set of individuals can be assigned to the candidate populations, it is a good indication of fine-scale level of stratification between the candidate populations.

We show that two populations that were not distinguishable with raw genotype data can be separated when constructing haplotypes. This highlights how *HaploPOP* can be used to study samples where prior belief suggests that there is population structure but SNP-genotype data fails to detect it. When computing the error for assignment of the validation set, we find that there is an optimal value in terms of window size (Figure 4) at which the assignment error can be reduced by 45%. Intuitively, combining SNPs into haplotypes can improve the power of separation between groups only up to a certain extent: for too large window sizes, we run into overfitting issues where trained haplotypes are well suited to separate the particular individuals of the training set but not of the validation set. The optimal value of the window size depends on many factors, such as the extent of linkage disequilibrium in the groups, or the degree of genetic differentiation between the groups. The strategy we advocate for choosing the window size is to try different window sizes and to find the minimal assignment error as estimated with a split-validation approach.

Conclusions

In the article, we present a new algorithm to use the GIA statistic to construct haplotypes using a window-based approach. The algorithm is implemented in the command-line software *HaploPOP*. The software allows users to apply a 2-step procedure. First *HaploPOP* learns combination of markers, refered as haplotypes, that are informative about population assignment. Second, *HaploPOP* recodes the genotype dataset of individuals to account for the discriminant haplotypes. This new data file can then be used to investigate population structure with statistical approaches such as PCA. We have shown that constructing haplotypes with *HaploPOP* can substantially reduce mis-assignment of individuals to candidate populations. It can also reveal fine-scale stratification that may have been missed when using the raw SNP data. For SNP data simulated in a 3-population divergence model, the assignment-error was reduced by 20% to 70%. Using the 447,245 SNPS of POPRES, the assignment-error was reduced by 45% when trying to distinguish Portuguese

from Spanish individuals.

Constructing Haplotypes with *HaploPOP* is a promising approach to assign individuals into population in forensic science and ecology. It can also confirm prior belief about fine scale population structure which is a main confounding factor for association studies with rare variants [16]. Additionnally, we anticipate that constructing haplotypes is particularly useful for conservation studies where it might be difficult to obtain a large number of molecular markers.

Availability and requirements

Linux and windows versions of the software are available at:
http://www.ebc.uu.se/Jakobsson/software/HaploPOP/.

Author's contributions

LG, and MJ designed the project and set the requirements of the software. NDF wrote the source code. NDF and LG ran simulations and analysed data. All authors wrote, read and approved the final manuscript.

Acknowledgements

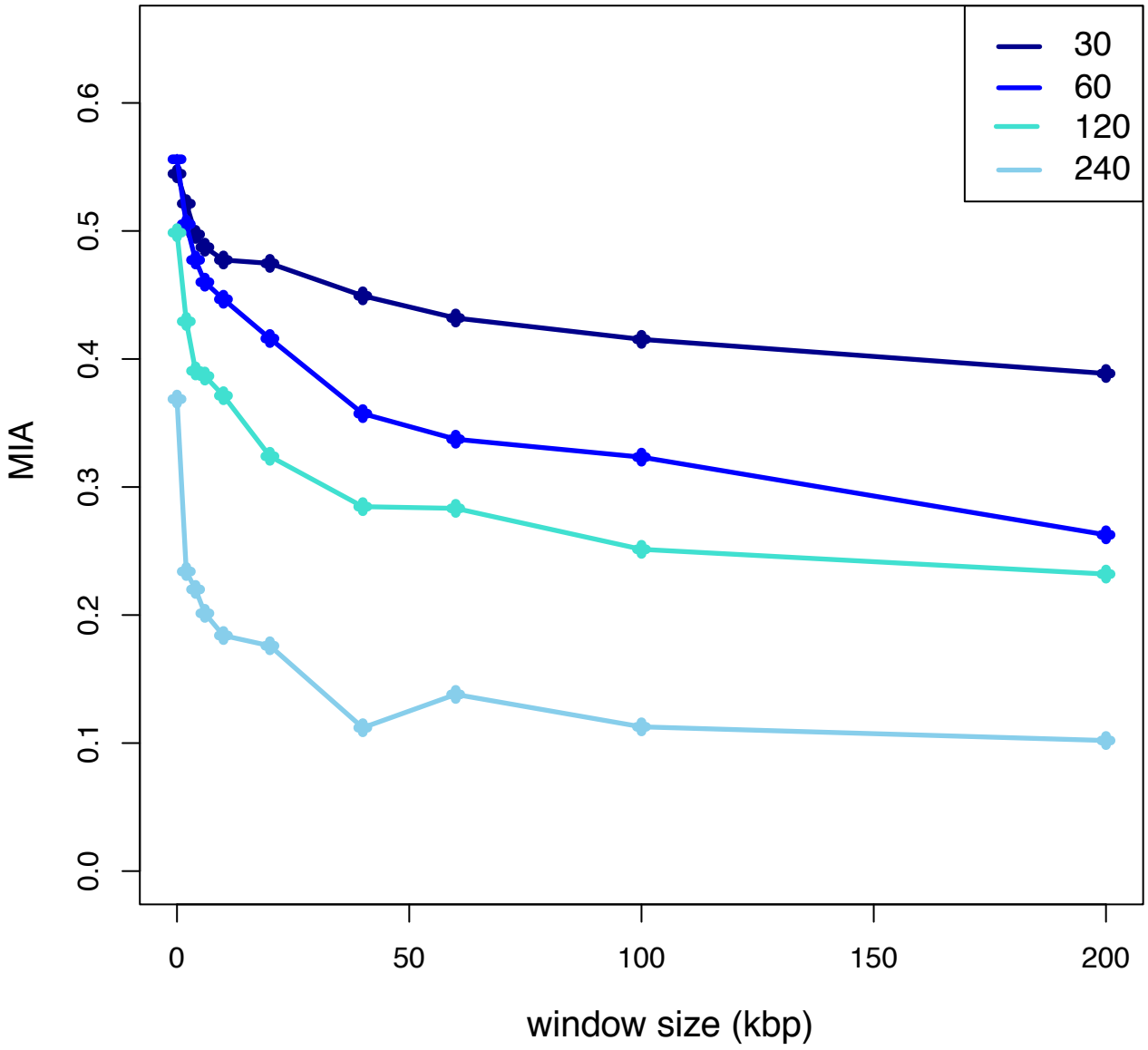
This work was supported by a grant from the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) awarded to Mattias Jakobsson and Michael Blum. A grant from the French national research agency provided support to Michael Blum and Nicolas Duforet-Frebourg (DATGEN project, ANR-2010-JCJC-1607-01) and a grant from the Swedish research council provided support to Mattias Jakobsson. The POPRES data were obtained from dbGaP (accession number phs000145.v1.p1).

References

1. Manel S, Gaggiotti OE, Waples RS: **Assignment methods: matching biological questions with appropriate techniques**. *Trends in Ecology & Evolution* 2005, **20**(3):136–142.
2. Kayser M, de Knijff P: **Improving human forensics through advances in genetics, genomics and molecular biology**. *Nature Reviews Genetics* 2011, **12**(3):179–192.
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al.: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *The American Journal of Human Genetics* 2007, **81**(3):559–575.
4. Lawson DJ, Hellenthal G, Myers S, Falush D: **Inference of population structure using dense haplotype data**. *PLoS genetics* 2012, **8**:e1002453.
5. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes Jr OE: **Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation**. *Conservation Genetics* 2006, **7**(2):295–302.
6. Gattepaille LM, Jakobsson M: **Combining markers into haplotypes can improve population structure inference**. *Genetics* 2012, **190**:159–174.
7. Rosenberg NA, Li LM, Ward R, Pritchard JK: **Informativeness of genetic markers for inference of ancestry**. *The American Journal of Human Genetics* 2003, **73**(6):1402–1422.
8. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *PLoS genetics* 2006, **2**(12):e190.
9. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals**. *Genome Research* 2009, **19**(9):1655–1664.
10. Hastie T, Friedman J, Tibshirani R: *The elements of statistical learning*, *Volume 2*. Springer 2009.
11. Nelson M, Bryc K, King K, Indap A: **The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research**. *American journal of human genetics* 2008, **83**:347–358, [http://www.sciencedirect.com/science/article/pii/S0002929708004436].
12. Rota GC: **The number of partitions of a set**. *The American Mathematical Monthly* 1964, **71**(5):498–504.
13. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: **A worldwide survey of haplotype variation and linkage disequilibrium in the human genome**. *Nature genetics* 2006, **38**(11):1251–1260.
14. Hudson RR: **Generating samples under a Wright–Fisher neutral model of genetic variation**. *Bioinformatics* 2002, **18**(2):337–338.
15. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD: **Genes mirror geography within Europe.** *Nature* 2008, **456**(7218):98–101, [http://www.ncbi.nlm.nih.gov/pubmed/18758442].
16. Mathieson I, McVean G: **Differential confounding of rare and common variants in spatially structured populations**. *Nature genetics* 2012, **44**(3):243–246.

Figures

Figure 1 - Mean percentage of Incorrect Assignment (MIA) for the simulations of the divergence model with 3 populations.



The mean incorrect assignment of individuals is evaluated with individuals from the validation set that were not used to construct the haplotypes.

Figure 2 - Principal Component Analysis on 447,245 SNPs for Spanish and Portuguese samples from POPRES.

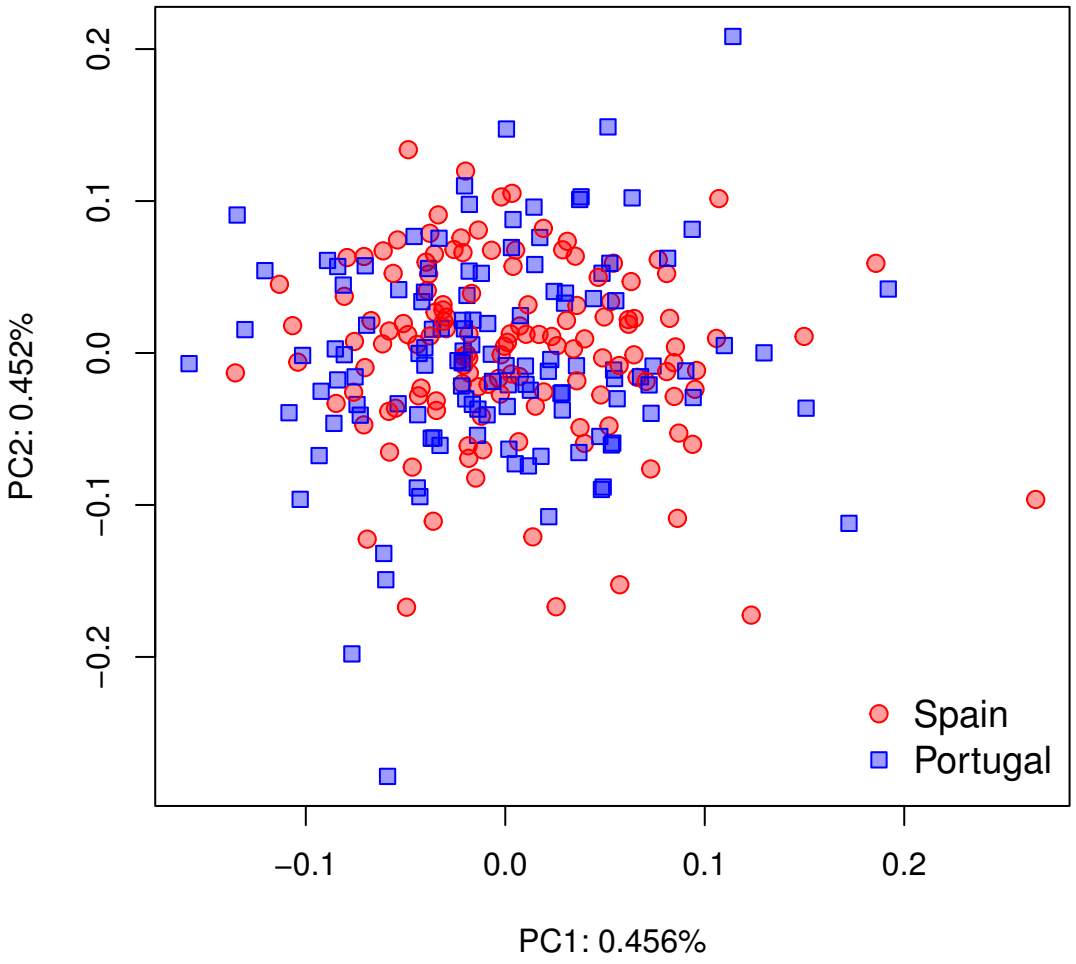
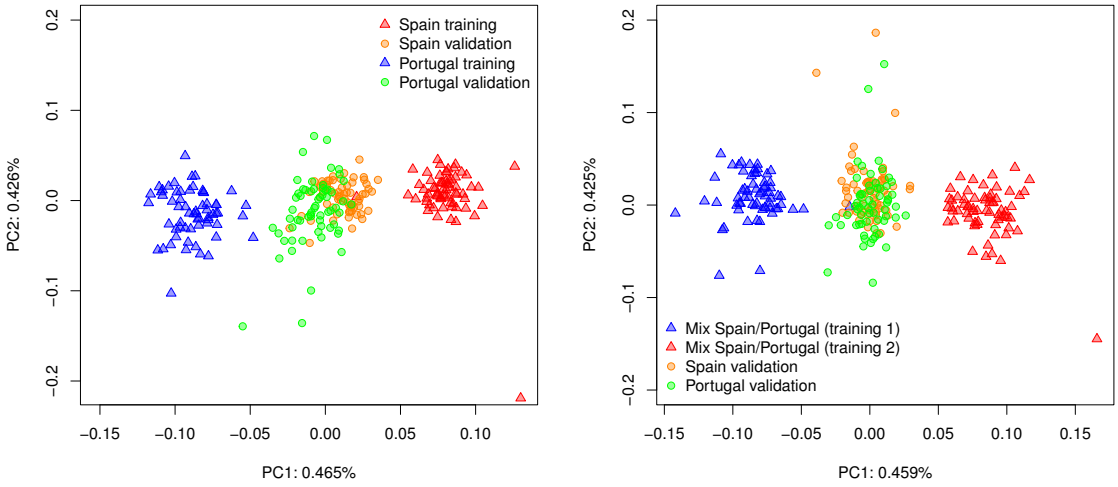
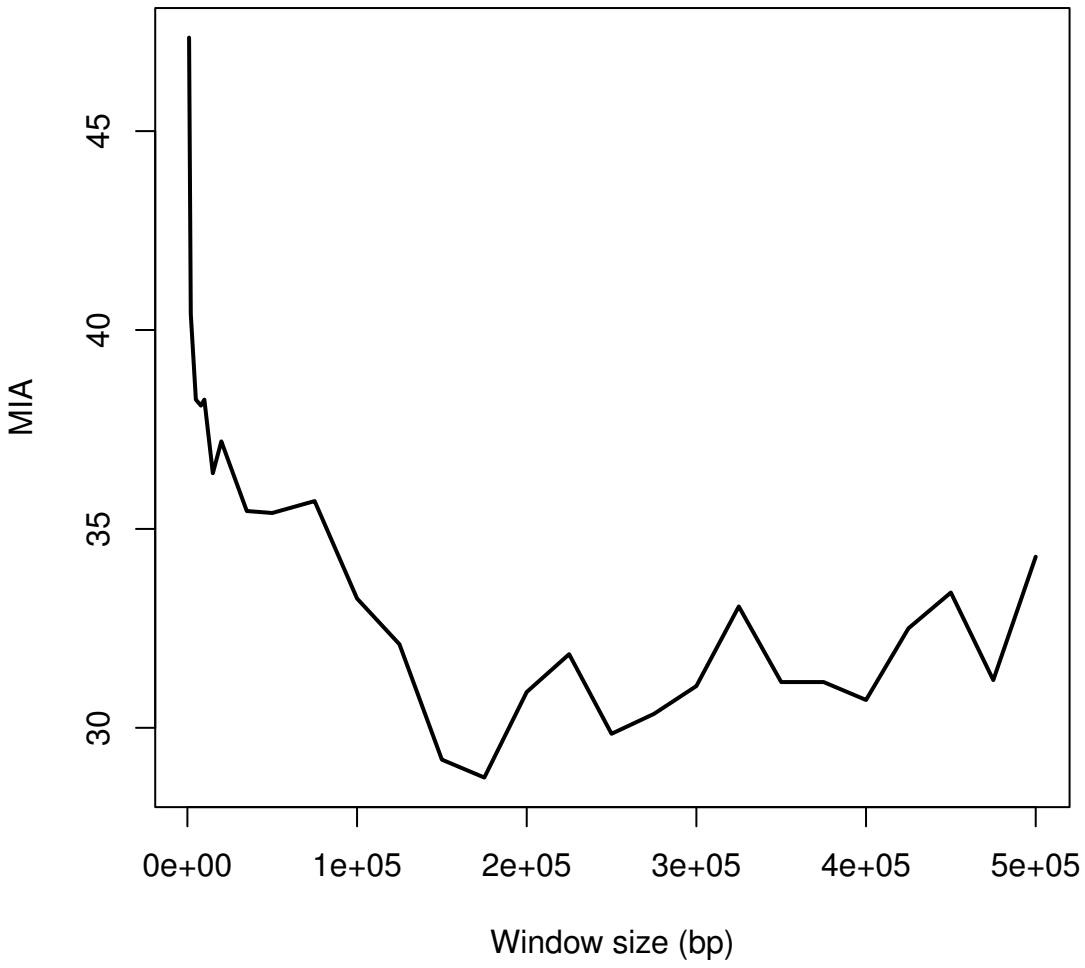


Figure 3 - Principal Component Analysis of the Spanish and Portuguese samples from POPRES using the haplotypes found with *HaploPOP*.



The haplotypes were built from 447,245 SNPs using a window size of 20 000 bp. For constructing haplotypes, the populations consist of the Portuguese and Spanish individuals (Panel A) or a mix of Portuguese and Spanish individuals in the two populations to be separated.

Figure 4 - Mean percentage of Incorrect Assignment (MIA) when distinguishing the Spanish and Portuguese samples from POPRES.



The error is evaluated with a split-validation approach.

European Journal of Human Genetics (2014), 1–6
© 2014 Macmillan Publishers Limited All rights reserved 1018-4813/14
www.nature.com/ejhg

ARTICLE

Fine-scale human genetic structure in Western France

Matilde Karakachoff[1,2,3,4,15], Nicolas Duforet-Frebourg[5,15], Floriane Simonet[1,3,4], Solena Le Scouarnec[1,3,4], Nadine Pellen[6], Simon Lecointe[1,3,4], Eric Charpentier[1,3,4], Françoise Gros[1,3,4], Stéphane Cauchi[7,8,9], Philippe Froguel[7,8,9,10], Nane Copin[11], the D.E.S.I.R. Study Group[11,16], Thierry Le Tourneau[1,3,4,12], Vincent Probst[1,3,4,12], Hervé Le Marec[1,3,4,12], Sabrina Molinaro[2], Beverley Balkau[13,14], Richard Redon[1,3,4,12], Jean-Jacques Schott[1,3,4,12], Michael GB Blum[5] and Christian Dina*[,1,3,4,12]

The difficulties arising from association analysis with rare variants underline the importance of suitable reference population cohorts, which integrate detailed spatial information. We analyzed a sample of 1684 individuals from Western France, who were genotyped at genome-wide level, from two cohorts D.E.S.I.R and CavsGen. We found that fine-scale population structure occurs at the scale of Western France, with distinct admixture proportions for individuals originating from the Brittany Region and the Vendée Department. Genetic differentiation increases with distance at a high rate in these two parts of Northwestern France and linkage disequilibrium is higher in Brittany suggesting a lower effective population size. When looking for genomic regions informative about Breton origin, we found two prominent associated regions that include the lactase region and the HLA complex. For both the lactase and the HLA regions, there is a low differentiation between Bretons and Irish, and this is also found at the genome-wide level. At a more refined scale, and within the Pays de la Loire Region, we also found evidence of fine-scale population structure, although principal component analysis showed that individuals from different departments cannot be confidently discriminated. Because of the evidence for fine-scale genetic structure in Western France, we anticipate that rare and geographically localized variants will be identified in future full-sequence analyses.
European Journal of Human Genetics advance online publication, 3 September 2014; doi:10.1038/ejhg.2014.175

INTRODUCTION

Advances in genotyping technologies have facilitated genome-wide association studies and have contributed to the identification of thousands of SNPs associated with disease[1]. However, most of the associated SNPs explain a modest proportion of heritability for most common diseases,[2] and they do not provide variants directly applicable for diagnosis, prevention and treatment.[3] With the advent of whole-genome sequencing, there is now the possibility to evaluate the role of genetic variants that were too rare to be picked up by GWAS.[4] Because rare and potentially deleterious variants arose more recently than common variants, they were not yet eliminated by selection and they might provide most of the medically actionable alleles.[3,5] These alleles, because of their recent origin, tend to cluster within a lineage and therefore, many disease-related traits have been found in population isolates.[6] Likewise, non-isolated populations displaying fine-scale genetic structure, within a short geographical distance, may be appropriate for gene mapping because they likely present high frequency of rare alleles.[7]

Moreover, there are specific difficulties arising from analyses based on rare variants, as they exhibit stronger stratification than common variants, which is not accounted for by existing statistical methods.[8] The difficulties arising in rare-variant analyses underscore the importance of suitable reference population cohorts, which integrate detailed spatial information.[8] In our study, we contribute to this effort by investigating fine-scale genetic structure in Western France based on 1684 genotyped individuals, for whom we have detailed geographic information. We provide a descriptive analysis of the pattern of population structure and genetic variation in Western France with the objective of building a reference panel useful for future association studies.

Since the seminal work of Menozzi *et al*,[9] evidence has accumulated to show that allele frequencies vary within Europe, generating population structure at the continental scale.[10–12] Evidence of population structure accumulates at the scale of European countries[13–18] and even at a finer scale.[19,20] Fine-scale genetic structure can result from different processes including founder effect, endogamy, historical migration, cultural barriers to gene flow or selection processes.[11,21,22]

Our study focuses on genetic variation in Western France as defined by Historical Geography literature. Before the 19th century, Western France included the ancient regions of Brittany–Anjou–Poitou and of Maine–Basse–Normandie (actually consisting on the

[1]INSERM UMR 1087, Nantes, France; [2]Institute of Clinical Physiology, CNR, Pisa, Italy; [3]CNRS, UMR 6291 institut du thorax, Nantes, France; [4]Université de Nantes, Nantes, France; [5]Laboratoire TIMC-IMAG UMR 5525, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France; [6]Université de Bretagne Occidentale, Brest, France; [7]European Genomic Institute for Diabetes (EGID), Lille, France; [8]CNRS UMR8199, Pasteur Institute of Lille, Lille, France; [9]Lille Nord de France University, Lille, France; [10]Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK; [11]IRSA, Institut inter-Regional pour la Santé, La Riche, France; [12]CHU Nantes, l'institut du thorax, Service de Cardiologie, Nantes, France; [13]INSERM, CESP Center for Research in Epidemiology and Population Health, U1018, Villejuif, France; [14]Epidemiology of diabetes, obesity and chronic kidney disease over the lifecourse, University Paris Sud 11, UMRS 1018, Villejuif, France
*Correspondence: Dr C Dina, Inserm UMR 1087/CNRS UMR 6291 institut du thorax, 8 Quai Moncousu, BP, 70721 44007 Nantes cedex 1, France. Tel: +33 2 28 08 01 54; Fax: +33 2 28 08 01 30; E-mail: christian.dina@univ-nantes.fr
[15]These authors contributed equally to this work.
[16]Members of the D.E.S.I.R. Study Group are listed before References.
Received 8 October 2013; revised 21 July 2014; accepted 30 July 2014

departments of Calvados, Côtes-d'Armor, Finistère, Ille-et-Vilaine, Loire-Atlantique, Maine-et-Loire, Manche, Mayenne, Morbihan, Orne, Sarthe, Deux-Sèvres, Vendée, Vienne).[23,24] Western France is a large peninsula positioned in the northwest of France and delimited by the English Channel to the north, the Celtic and Atlantic Sea to the west and the Bay of Biscay to the south; the whole region covers ~90 000 km^2 extending ~470 km from west to east. This part of France is now subdivided into different administrative regions further subdivided into departments (Figure 1 and Supplementary Figure 1).

We ascertained fine-scale population structure in Western France. Using principal component analysis (PCA) and model-based clustering, we investigated the relationship between genetic structure and place of birth. To better characterize fine-scale population structure, we merged individuals from Western France with the European Population Reference Sample (POPRES),[25] and we computed measures of genetic proximity between Western France individuals and individuals from neighboring countries. We also better characterized fine-scale genetic variation by computing linkage disequilibrium (LD) and a local rate of genetic differentiation. Finally, we performed genome scans to map genomic regions correlated with population structure and to evaluate the genomic inflation of test statistics when not accounting for the place of birth.

MATERIALS AND METHODS

Samples

Analyses were performed with individuals from Western France coming from two different studies. The first study is the D.E.S.I.R. cohort, used to study the insulin resistance syndrome;[26,27] individuals from this cohort were born

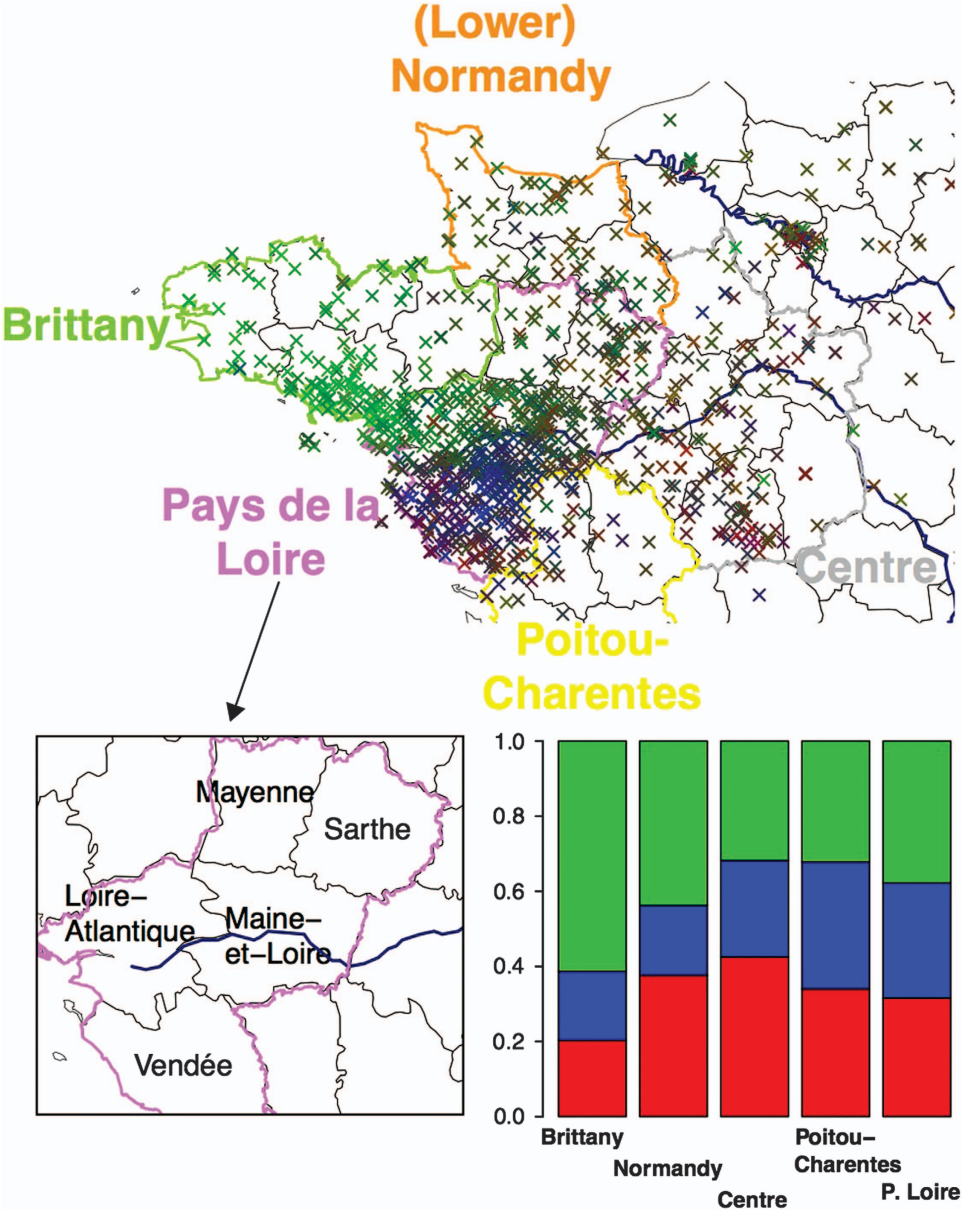


Figure 1 Ancestry coefficients displayed using the RGB color model. To visualize the admixture coefficient, each individual is colored using a RGB (red–green–blue) coding scheme. The RGB coefficients are proportional to the ancestry coefficients obtained with admixture. Departments of the Pays de la Loire region (bottom left). Admixture coefficients for the different regions of WF. Following the same RGB coding scheme, this figures shows the variation of admixture coefficients in WF (bottom right).

between 1930 and 1965 (median = 1950). The second group consists of patients admitted to Nantes, Angers and Rennes hospitals, between 2004 and 2010, and included in the cohort CavsGen, for the study of calcific aortic valve stenosis, a common pathology affecting mainly individuals over 65 years old. Individuals from CavsGen were born between 1916 and 1949 (median = 1931).

Individuals from both the D.E.S.I.R. and CavsGen cohorts were born in France and they came mainly from the Region Pays de la Loire (mainly departments of Loire-Atlantique, Maine-et-Loire and Vendée), but surrounding regions are also represented in the data (see Supplementary Table 1). The geographical locations of individuals were defined according to the place of birth, declared at the moment of enrollment in the cohorts. After excluding individuals with missing place of birth, the study population included 1684 unrelated individuals: 897 from D.E.S.I.R. and 787 from CavsGen. A third subgroup of 807 unrelated individuals, from the D.E.S.I.R. population, was used to replicate the relationship between genetic and geographical structure. This data set is called DESIR-REP in the following.

Genotyping

Samples were genotyped using the Axiom Genome-Wide CEU-1 Array (Affymetrix, Santa Clara, CA, USA). After quality control, we identified a set of 377 566 SNPs that were shared between D.E.S.I.R. and CavsGen. We pruned SNPs to account for LD ($r^2 < 0.2$), which provided a total of 78 261 SNPs used for all analyses. In the LD analysis and the association scan, all shared SNPs were used.

The additional sample DESIR-REP was genotyped using the HumanHap300 Genotyping BeadChip (Illumina, San Diego, CA, USA). The quality control procedure was similar and based on allele frequency and LD.

Model-based clustering and PCA

To investigate the population structure in Western France, we performed model-based clustering as implemented in admixture.[28] Admixture analysis was performed using $K = 2$ and $K = 3$ clusters. Before running admixture, we removed outlier individuals that would otherwise have formed a single cluster (four individuals in Maine-et-Loire and two individuals in Orne). Spatial kriging was used to interpolate ancestry coefficients. We also investigated population structure using PCA as implemented in the software *smartpca* (Cambridge, MA, USA) from the EIGENSOFT package.[29]

Identity-by-state statistics, LD and local differentiation

Pairwise identity-by-state (IBS) statistics were computed using Plink.[30] LD was estimated using the r^2 statistic. LD was computed based on the 377 566 SNPs in the data, after removal of the SNPs with minor allele frequencies under 5%. To compensate for unequal sampling in the different departments of interest, we used bootstrapping. In each department, 18 individuals were randomly selected (18 is the minimum sample size that is reached in the Manche Department) to compute r^2 values. The process was repeated five times to average over several outcomes, and to compensate for sampling effects. Finally, local genetic differentiation was computed based on the pairwise Fst matrix between departments. Estimated with a Bayesian Kriging method, local genetic differentiation provides an estimate of the Fst measure of differentiation between a population living in the center of a department and a putative neighboring population living 30 km away.[31]

Genomic regions informative about Breton origin

We analyzed the association between a binary variable indicating Breton origin, based on the place of birth and the genotype. We used Plink to compute a χ^2 statistic for each SNP, and we accounted for population stratification with the genomic inflation factor.[30] We then considered 50-SNP windows and for each window, we performed a test of enrichment of SNPs informative about Breton origin. The *P*-values corresponding to the test of enrichment were obtained with a binomial test by computing the proportion of *P*-values smaller than 10^{-4} among each window of 50 SNPs. Because the data included two different cohorts (D.E.S.I.R. and CavsGen), we also performed an additional analysis to account for a potential cohort effect. We performed two independent GWAS on each of the cohorts and we combined the resulting *P*-values using Stouffer's method[32] for meta-analysis (also called inverse normal method).

RESULTS

Population structure

Performing a clustering analysis with admixture at both $K = 2$ and $K = 3$, we found that the admixture coefficients were correlated with geography ($P < 10^{-16}$ when regressing admixture coefficients with latitude and longitude). With $K = 2$, admixture separated some of the Bretons from the rest of the sample (Supplementary Figure 2), and this cluster was again found with $K = 3$ (green cluster in Figure 1). When averaging the ancestry of individuals within each department, we found that the ancestry component corresponding to the Bretons (green cluster) was larger in the three departments at the western end of Brittany (Finistère, Côtes d'Armor, Morbihan, see Supplementary Figure 3). With $K = 3$, the ancestry of some individuals, many of them coming from the Vendée Department and from the south of the Maine-et-Loire department, differed from the rest of the sample (the blue cluster in Figure 1, see also Supplementary Figure 3).

The PCA confirmed the results obtained with admixture. The first principal component separated Bretons from the rest of the subjects (Figure 2 panel A) with individuals from the westernmost departments of Brittany having the lowest scores for PC1 (Supplementary Figure 4). The second component separated some individuals from the Vendée and Maine-et-Loire Departments from the rest of the sample (Figure 2 panel A and Supplementary Figure 4). The first two PC components were correlated with geography ($P < 10^{-16}$ when regressing the PC scores with latitude, longitude), also when accounting for genotyping plates and when considering the control DESIR-REP data set ($P < 10^{-9}$) (Supplementary Figure 5).

Then, we explored the PCA pattern for individuals born in the Pays de la Loire Region, because it was the most represented region in the data (926 out of 1684 sampled individuals). PCA showed genetic structure within the Pays de la Loire Region: some individuals from the western Loire-Atlantique department were located close to the Breton individuals, individuals from the northern Sarthe and Mayenne Departments were grouped with individuals from Normandy and some individuals from the southern Vendée Department and the eastern Maine-et-Loire department had the lowest scores for PC2 (Figure 2 panel B). However, PCA also showed considerable overlap between individuals from the different departments of the Region Pays de la Loire indicating that departmental origin could not be confidently assigned using the first two principal components.

Then, we merged the data with a collection of European individuals genotyped with the Affymetrix 500 K SNP panel (POPRES). From the first 10 PCs, we chose PC1 and PC3, because they were the most correlated with geography. Individuals from the DESIR-CavsGen data and from the control DESIR-REP data set were located between individuals from the UK, Ireland, Spain, the French-speaking part of Switzerland and the French POPRES population, which is consistent with their French origin (Supplementary Figures 6 and 7). Individuals from the DESIR-CavsGen data did not match exactly French individuals from POPRES, probably because of their westernmost origin.

When computing pairwise IBS, for individuals from this merged data set, we observed that the European countries showing the largest mean IBS with the DESIR-CavsGen individuals were France, UK, Ireland, followed by Germany and Belgium, followed by Spain and Italy (Figure 3 and Supplementary Figure 8). The IBS pattern differed according to French Regions, with the Brittany Region having a specific pattern. The countries that had the largest IBS with the Breton individuals were Ireland, UK and France in this order (Figure 3). The IBS between Irish and Bretons was significantly larger ($P < 10^{-12}$) than the IBS between Irish and individuals from the other regions of

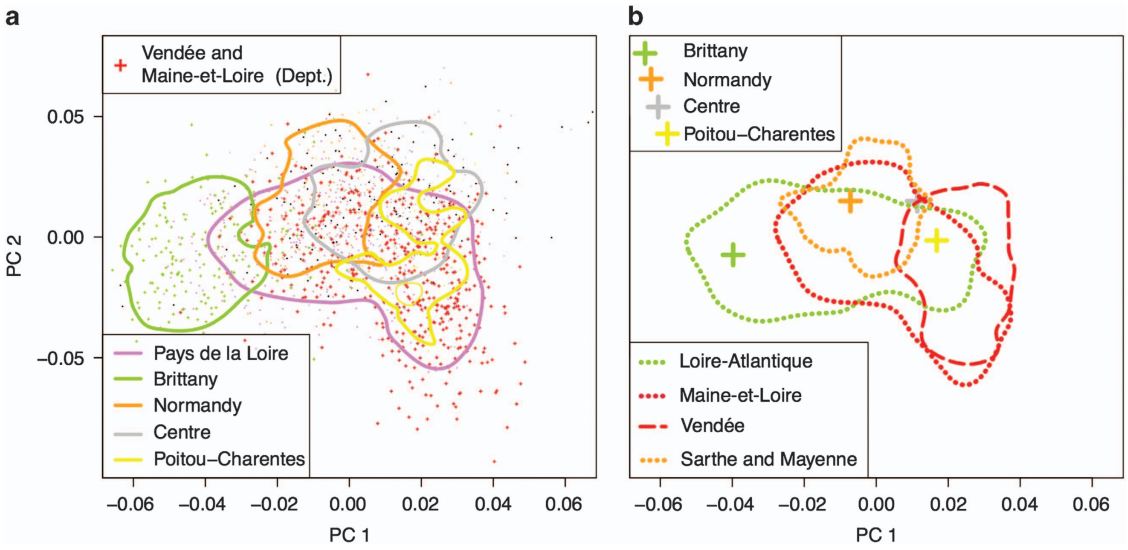


Figure 2 PCA of the CavsGen-DESIR data set. The envelopes correspond to the areas that contain 75% of the individuals of a given subdivision. (**a**) The envelopes correspond to the different regions of WF. Individuals from Brittany and from the Vendée and Maine-et-Loire Departments are displayed with larger crosses than the other individuals. (**b**) The envelopes correspond to the different departments of the Pays de la Loire Region, the Region with the majority of sampled individuals. Large crosses correspond to the barycentric coordinates of the individuals grouped by regions.

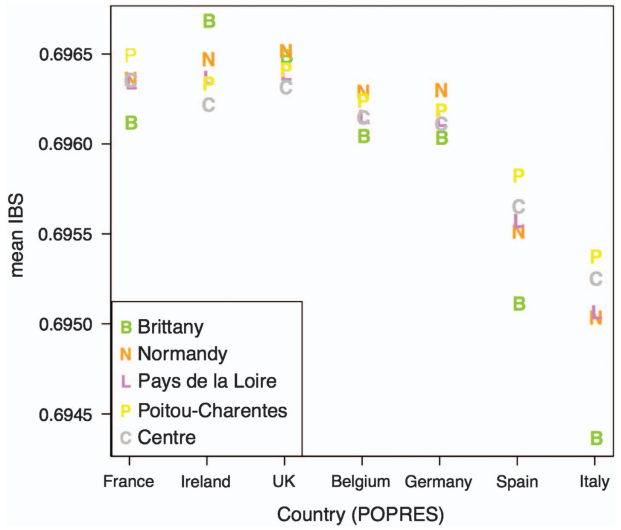


Figure 3 Mean IBS statistics between individuals from the French regions of the CavsGen-DESIR data set and individuals of POPRES from the neighboring countries of France.

Western France (Normandy, Pays de Loire, Poitou-Charente, Centre), and this pattern was also found with the control DESIR-REP data set (Supplementary Figure 9). Conversely, the IBS between Spanish (or Italians) and Bretons was significantly smaller ($P<10^{-12}$) than the IBS between Spanish (or Italian) and individuals from the other regions of Western France (Figure 3).

Spatial variation of LD and population genetic differentiation

To study how LD varies in Western France, we computed LD within each Department. Figure 4 shows that the extent of LD is greater at the western end of Brittany and decreases when moving eastward. We also computed a statistic that measures the amount of local genetic differentiation for each department. Genetic differentiation increases with distance (the isolation-by-distance pattern, Supplementary Figure 10) at a rate that varies across Western France. The departments with the largest values of local genetic differentiation—measured in Fst per 30 km—are the three departments at the western end of Brittany as well as the Vendée department. In summary, both LD and local differentiation suggest that local effective population size is smaller in the western part of Brittany and local differentiation points to a specific pattern in the Vendée Department.

Most informative genomic regions about breton origin

To look for genomic regions that are informative about Breton origin, we performed genome scans by correlating SNPs to self-declared Breton origin. We found a genomic inflation factor of $\lambda = 1.26$, confirming that there is a population structure in the data. When using Bonferonni correction, we found a SNP (rs6754311) in the lactase region significantly correlated with Breton origin. This SNP remained significantly correlated with Breton origin in a meta-analysis combining two independent association scans run for each cohort (CavsGen and D.E.S.I.R.). We also performed a window-based approach and searched for the 50-SNP regions enriched with ancestry-informative SNPs. The two genomic regions that were the most informative were around the lactase gene on chromosome 2, and in the HLA complex on chromosome 6 (Supplementary Figure 11). Around these two regions, we additionally computed for each 50-SNP window, the values of Fst between Ireland (POPRES) and the different French Regions. For both the HLA and the lactase regions, the Fst's between Ireland and Brittany was markedly smaller than the Fst's between Ireland and the other French Regions (Supplementary Figure 12).

For the lactase-persistence phenotype in Europe, the causative mutation is known and is located at SNP rs4988235. By computing the regional and departmental allele frequency of rs4988235, we found a west-to-east gradient of allele frequency with a frequency above 70% for 3 out of 4 departments in Brittany whereas the allele frequency decreased to between 45 and 55% in the easternmost departments of Western France (Supplementary Figure 13 and supplementary Table 2).

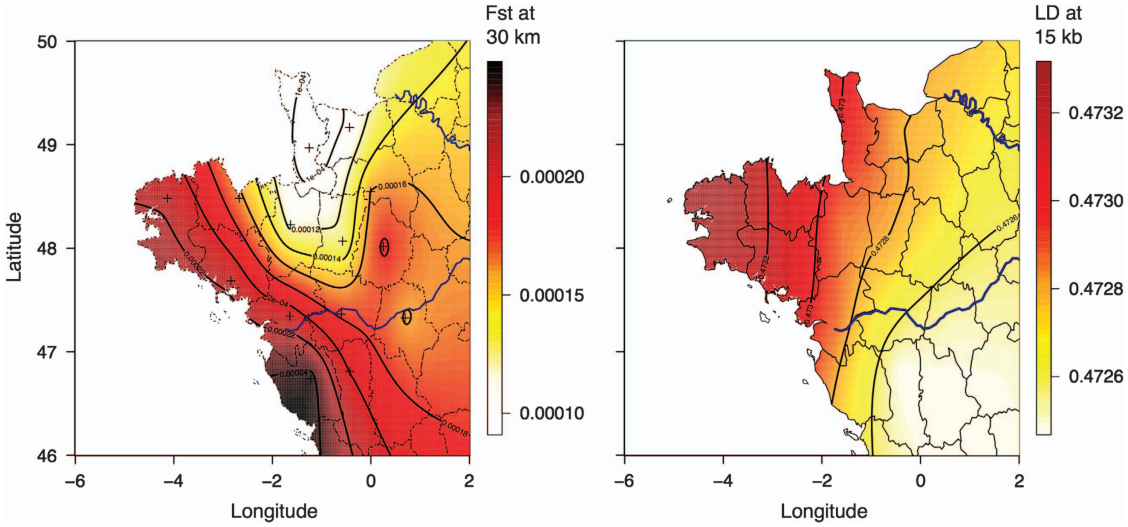


Figure 4 Spatial variation of local genetic differentiation (Fst at 30 km) and of LD (at 15 kb).

DISCUSSION

There is currently ongoing interest in fine-scale population structure because it may impact association studies based on rare variants[8] Our analysis shows that fine-scale population structure does occur at the scale of Western France. We found genetic ancestry related to Breton and Vendean origin. Uneven sampling can bias ascertainment of population structure[33] but cannot cause the particular ancestry of Breton individuals because they are not overrepresented in the data set. Our results support the importance of isolation by distance in Western France because genetic differentiation between departments is well explained by geographical distances. The finding of fine-scale population structure in historical Western France is compatible with the historical records. Inner rural areas lacked long distance migration that resulted in isolated and static populations.[34] Stability is also evident in the French rural population as a consequence of high endogamy and high percentages of marriages occurring within a short distance, at least until the 19th century.[23] In his studies about marriage, R. Leprohon[35] found that a high percentage of marriages took place between people living within a radius limited to 5–10 km. According to the analysis done by N. Pellen[22] in Kerlouan, a Western French county, about 90% of mating was from the same village during the 17th and 18th centuries.

Patterns of genetic variation are partly generated by genetic drift defined as the random fluctuations of allele frequencies. When two populations diverge, genetic differentiation increases because of genetic drift. The smaller the effective population size, the larger the effect of genetic drift. We argue that the larger values of local Fst and of LD that were found in Brittany are explained by the enhanced effect of genetic drift in populations of lower effective population size.[36] In the Vendée department, we posit that the presence of individuals with distinct ancestries (Figure 1) explains why only local Fst is increased. Because genetic differentiation increases more rapidly in these two regions, we expect that an excess of rare variants specific to these regions will be found with full-sequence data.

Because of population structure, association studies in Western France may generate false-positive associations. When using the Breton birthplace as an artificial phenotypic trait, two prominent associated genomic regions were found: the regions around the lactase gene and the HLA complex. By contrast to the lactase region, no single SNP within the HLA region was significantly associated with Breton origin. However, a correlation was found when looking for 50-SNP windows enriched with small enough *P*-values. The correlation between Breton origin and genotypes of the HLA region can be a true association, although there is strong LD within the HLA region,[37] and this is a matter of concern when using window-based approaches. For these two regions informative about ancestry, there was a marked genetic proximity between Irish and Bretons, in terms of allele frequencies. The genetic proximity between Bretons and Irish was also found with Y chromosome haplogroups and on a genome-wide scale, possibly reflecting some degree of common Celtic origin.[38–40] Both the lactase and the HLA regions are known to harbor prominent geographic variation in Europe and the increase in the lactase persistence allele in Brittany fits within the overall increase that is found when moving from southeastern to northwestern Europe.[41] In particular, the frequency of the lactase persistent allele is high in Ireland (95%).[41] The variation of lactase-persistence frequency is in line with the different admixture coefficients found for the Breton individuals pointing to something more than isolation-by-distance because of common ancestry between Bretons and individuals from the British isles. However, we acknowledge that admixture is not the only process that may have generated particular allele frequencies in Brittany; processes of positive selection that arose because of biological adaptation to agricultural practice can also explain spatial variation of gene frequencies.[42]

In summary, we emphasize the benefits of building reference panels with detailed information about geographic origin. In our study, we used geographic information to reveal fine-scale genetic structure in historical Western France. We found distinctive ancestries for Vendean and Breton individuals and the latter were found to have genetic proximities with Irish individuals. At a more subtle level, we also identified very fine-scale population structure within the Pays de la Loire Region. As for other European regions, such as the Netherlands, British Isles and Sardinia,[18,43,44] it is important to build a comprehensive reference panel of Western France genetic variation with fine-grained geographic information in view of performing rare-variant association studies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Emmanuelle Bourcereau who was involved in patient recruitment of the CavsGen cohort. We also thank Professors Le Guerrier and Baufreton from University Hospitals of Rennes and Angers. We are grateful to the French Regional Council of Pays-de-la-Loire (Regional grant Biliv, grant VACARME - www.vacarme-project.org), the French Ministry of Health (PHRC/RNI-PROG/09/61: Génétique du rétrécissement aortique) and the French Ministry of Research (CavsGen project, ANR-13-BSV6-0011-01). We wish to thank the Fondation Genavie for its financial support. The POPRES data were obtained from dbGaP (accession no. phs000145.v3.p2). FS was funded by a Eurostars E!6490 Cardiomarks grant. MGBB is supported by the French National Research Agency (DATGEN project, ANR-2010-JCJC-1607-01). The D.E.S.I.R. study has been supported by INSERM contracts with CNAMTS, Lilly, Novartis Pharma and Sanofi-Aventis; by INSERM (Réseaux en Santé Publique, Interactions entre les déterminants de la santé), Cohortes Santé TGIR, the Association Diabète Risque Vasculaire, the Fédération Française de Cardiologie, La Fondation de France, ALFEDIAM, Société francophone du diabète, ONIVINS, Abbott, Ardix Medical, Bayer Diagnostics, Becton Dickinson, Cardionics, Merck Santé, Novo Nordisk, Pierre Fabre, Roche, Topcon.

THE D.E.S.I.R. STUDY GROUP

INSERM U1018: B Balkau, P Ducimetière, E Eschwège; INSERM U367: F Alhenc-Gelas; CHU D'Angers: A Girault; Bichat Hospital: F Fumeron, M Marre, R Roussel; CHU de Rennes: F. Bonnet; CNRS UMR8090, Lille: S Cauchi, P Froguel; Centres d'Examens de Santé: Alençon, Angers, Blois, Caen, Chateauroux, Chartres, Cholet, Le Mans, Orléans, Tours; Institute de Recherche Médecine Générale: J Cogneau; General practitioners of the Region; Institute inter-Regional pour la Santé: C Born, E Caces, M Cailleau, O Lantieri, JG Moreau, F Rakotozafy, J Tichet, S Vol.

1 Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 2009; **106**: 9362–9367.
2 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
3 Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA: Clan genomics and the complex architecture of human disease. *Cell* 2011; **147**: 32–43.
4 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.
5 McClellan J, King M-C: Genetic heterogeneity in human disease. *Cell* 2010; **141**: 210–217.
6 Kristiansson K, Naukkarinen J, Peltonen L: Isolated populations and complex disease gene identification. *Genome Biol* 2008; **9**: 109.
7 Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD: Fine-scale population structure and the era of next-generation sequencing. *Hum Mol Genet* 2010; **19**: R221–R226.
8 Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; **44**: 243–246.
9 Menozzi P, Piazza A, Cavalli-Sforza L: Synthetic maps of human gene frequencies in Europeans. *Science* 1978; **201**: 786–792.
10 Heath SC, Gut IG, Brennan P *et al*: Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008; **16**: 1413–1429.
11 Novembre J, Johnson T, Bryc K *et al*: Genes mirror geography within Europe. *Nature* 2008; **456**: 98–101.
12 Jay F, Sjödin P, Jakobsson M, Blum MGB: Anisotropic isolation by distance: the main orientations of human genetic differentiation. *Mol Biol Evol* 2013; **30**: 513–525.
13 Humphreys K, Grankvist A, Leu M *et al*: The genetic structure of the Swedish population. *PLoS One* 2011; **6**: e22547.
14 O'Dushlaine CT, Morris D, Moskvina V *et al*: Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur J Hum Genet* 2010; **18**: 1248–1254.
15 Salmela E, Lappalainen T, Liu J *et al*: Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. *PLoS One* 2011; **6**: e16747.
16 Di Gaetano C, Voglino F, Guarrera S *et al*: An overview of the genetic structure within the Italian population from genome-wide data. *PLoS One* 2012; **7**: e43759.
17 Esko T, Mezzavilla M, Nelis M *et al*: Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* 2012; **21**: 659–665.
18 Abdellaoui A, Hottenga J-J, Knijff de P *et al*: Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet* 2013; **21**: 1277–1285.
19 O'Dushlaine C, McQuillan R, Weale ME *et al*: Genes predict village of origin in rural Europe. *Eur J Hum Genet* 2010; **18**: 1269–1270.
20 Colonna V, Nutile T, Ferrucci RR *et al*: Comparing population structure as inferred from genealogical versus genetic information. *Eur J Hum Genet* 2009; **17**: 1635–1641.
21 Biswas S, Scheinfeldt LB, Akey JM: Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am J Hum Genet* 2009; **84**: 641–650.
22 Pellen N: Hasard, coïncidence, prédestination... et s'il fallait plutôt regarder du côté de nos aïeux? Analyse démographique et historique des réseaux généalogiques et des structures familiales des patients atteints de mucoviscidose en Bretagne. Thèse de doctorat en démographie historique. Versailles, France: Université de Versailles-Saint-Quentin-en-Yvelines, 2012.
23 Blayo Y, Henry L: Données démographiques sur la Bretagne et l'Anjou de 1740 à 1829. *Ann Démographie Hist* 1967; **1967**: 91–171.
24 Dupaquier J: La dynamique démographique de la France de l'Ouest au XIXe s. *Norois* 1988; **140**: 409–414.
25 Nelson MR, Bryc K, King KS *et al*: The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 2008; **83**: 347–358.
26 Balkau B: [An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome]. *Rev Dépidémiologie Santé Publique* 1996; **44**: 373–375.
27 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
28 Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655–1664.
29 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
30 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
31 Duforet-Frebourg N, Blum MGB: Non-stationary patterns of isolation-by-distance: inferring measures of local genetic differentiation with Bayesian kriging. *Evolution* 2013; **68**: 1110–1123.
32 Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr: *The American Soldier: Adjustment During Army Life. (Studies in Social Psychology in World War II, Volume 1)*. Princeton University Press, NJ, USA, 1949.
33 McVean G: A genealogical interpretation of principal components analysis. *PLoS Genet* 2009; **5**: e1000686.
34 Hohenberg P: Migrations et fluctuations démographiques dans la France rurale, 1836-1901. *Ann Économies Sociétés Civilis* 1974; **29**: 461–497.
35 Leprohon R: *Vie et mort des Bretons sous Louis XIV.* Brasparts, France: Bibliophiles de Bretagne, 1984.
36 Pimenoff VN, Laval G, Comas D *et al*: Similarity in recombination rate and linkage disequilibrium at CYP2C and CYP2D cytochrome P450 gene regions among Europeans indicates signs of selection and no advantage of using tagSNPs in population isolates. *Pharmacogenet Genomics* 2012; **22**: 846–857.
37 Evseeva I, Nicodemus KK, Bonilla C, Tonks S, Bodmer WF: Linkage disequilibrium and age of HLA region SNPs in relation to classic HLA gene alleles within Europe. *Eur J Hum Genet* 2010; **18**: 924–932.
38 McEvoy B, Richards M, Forster P, Bradley DG: The Longue Durée of genetic ancestry: multiple genetic marker systems and Celtic origins on the Atlantic facade of Europe. *Am J Hum Genet* 2004; **75**: 693–702.
39 Dubut V, Chollet L, Murail P *et al*: mtDNA polymorphisms in five French groups: importance of regional sampling. *Eur J Hum Genet* 2004; **12**: 293–300.
40 Busby GB, Brisighelli F, Sánchez-Diz P *et al*: The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc Biol Sci* 2012; **279**: 884–892.
41 Itan Y, Jones BL, Ingram CJE, Swallow DM, Thomas MG: A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol Biol* 2010; **10**: 36.
42 Fix AG: Gene frequency clines in Europe: demic diffusion or natural selection? *J R Anthropol Inst* 1996; **2**: 625–643.
43 Winney B, Boumertit A, Day T *et al*: People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet* 2012; **20**: 203–210.
44 Chen W, Li B, Zeng Z *et al*: Genotype calling and haplotyping in parent-offspring trios. *Genome Res* 2013; **23**: 142–151.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)