

Thèse

Pour obtenir le grade de

**DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRE-
NOBLE ALPES**

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

Arrêté ministériel : 25 mai 2016

Présentée par

Keurcien LUU

Thèse dirigée par **Michael BLUM**

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

**Application de l'Analyse en
Composantes Principales pour étudier
l'adaptation biologique en génomique
des populations.**

Thèse soutenue publiquement le 21 décembre 2017,
devant le jury composé de :

Michael BLUM

Directeur de Recherche, CNRS, Directeur de thèse

Stéphane DRAY

Directeur de Recherche, CNRS, Rapporteur

Yves VIGOUROUX

Directeur de Recherche, IRD, Rapporteur

Hélène BADOUIN

Maître de Conférences, CNRS, Examinateur

Olivier FRANÇOIS

Professeur des Universités, Grenoble INP, Examinateur



Table des matières

Chapitre 1 : Introduction	1
1.1 La génétique des populations	1
1.1.1 L'évolution comme point de départ	1
1.2 À l'origine de la variabilité génétique	2
1.2.1 La théorie de l'évolution	2
1.2.2 L'évolution d'une théorie	3
1.2.3 Forces évolutives	4
1.2.4 Adaptation locale	6
1.3 Données de polymorphismes génétiques	8
1.3.1 Des données en grande dimension	9
1.3.2 Les marqueurs génétiques	10
1.3.3 Encodage des données génétiques	12
1.4 Les motivations de la thèse	14
1.4.1 Résultats principaux et organisation du manuscript	16
Chapitre 2 : Adaptation locale	19
2.1 L'état de l'art pour les scans génomiques	19
2.1.1 Modèles démographiques	19
2.1.2 L'indice de fixation	21
2.1.3 Test de Lewontin-Krakauer	21
2.1.4 Le modèle F	23
2.2 L'Analyse en Composantes Principales en génétique des populations	24
2.2.1 Principe de l'Analyse en Composantes Principales	24
2.2.2 Apparentement génétique interindividuel	25
2.2.3 Applications en génétique des populations	26
2.3 Statistiques de test basées sur l'Analyse en Composantes Principales	30
2.3.1 La communalité	31
2.3.2 La distance robuste de Mahalanobis	35
2.4 Article 1	40
Abstract	40
Significance Statement	41
Introduction	41
New Method	42
Results	44

Discussion	54
Materials and Methods	60
Acknowledgments	62
2.5 Article 2	63
Abstract	63
Introduction	63
Statistical and computational approach	65
Materials and methods	68
Results	70
Discussion	78
Acknowledgements	79
Data accessibility	79
Chapitre 3 : Introgression adaptative	81
3.1 Introgression par métissage	82
3.1.1 Coefficients de métissage locaux	82
3.1.2 Statistique de test	83
3.2 Introgression par flux de gènes	84
3.2.1 Diversité nucléotidique par paires de séquences	84
3.2.2 Le modèle ABBA-BABA	85
3.3 Une nouvelle statistique pour les scans d'introgression	86
3.3.1 Analyse en Composantes Principales locale	86
3.3.2 Résultats principaux	88
3.4 Article 3	93
Scanning genomes for adaptive introgression using principal component analysis	93
Introduction	93
A PCA-based approach to detect local introgression	95
Materials and Methods	96
Results	99
Chapitre 4 : Aspect computationnel	103
4.1 Du langage C au langage R	103
4.2 Du calcul de la matrice de covariance à l'algorithme IRAM	103
4.3 ACP et valeurs manquantes	105
Algorithme IRAM et données manquantes	106
Précision de la SVD en présence de données manquantes	107
4.4 Du format .pcadapt au format .bed	108
4.5 Interface Shiny	109
Chapitre 5 : Perspectives et conclusions	111
5.1 Substitution de l'Analyse en Composantes Principales	111
5.2 Utilisation du déséquilibre de liaison pour améliorer l'inférence de la structure	112
5.3 Utilisation de variables environnementales	112

5.4	Scans pour l'introgression et données manquantes	113
5.5	L'approche IRAM pour les données manquantes	114
5.6	Conclusion	115
Annexe A : Détails	117
Rapport entre la communalité et l'indice de fixation	117	
Une généralisation de la statistique T_{F-LK}	118	
Annexe B : Informations supplémentaires	119
Article 1	119	
Article 2	133	
Article 3	139	
Annexe C : R & Python	147
C.1	Simulations et modèles démographiques	147
C.1.1	Modèle en îles	147
C.1.2	Modèle de divergence	149
Bibliographie	153

Résumé

Titre : Application de l'Analyse en Composantes Principales pour étudier l'adaptation biologique en génomique des populations.

Résumé : L'identification de gènes ayant permis à des populations de s'adapter à leur environnement local constitue une des problématiques majeures du domaine de la génétique des populations. Les méthodes statistiques actuelles répondant à cette problématique ne sont plus adaptées aux données de séquençage nouvelle génération (NGS). Nous proposons dans cette thèse de nouvelles statistiques adaptées à ces nouveaux volumes de données, destinées à la détection de gènes sous sélection. Nos méthodes reposent exclusivement sur l'Analyse en Composantes Principales, dont nous justifierons l'utilisation en génétique des populations. Nous expliquerons également les raisons pour lesquelles nos approches généralisent les méthodes statistiques existantes et démontrons l'intérêt d'utiliser une approche basée sur l'Analyse en Composantes Principales en comparant nos méthodes à celles de l'état de l'art. Notre travail a notamment abouti au développement de padapt, une librairie R permettant l'utilisation de nos statistiques de détection sur des données génétiques variées.

Mots-clés : génétique des populations, analyse en composantes principales, adaptation locale, introgression, bio-informatique.

Title : Application of Principal Component Analysis to study biological adaptation in population genomics.

Abstract : Identifying genes involved in local adaptation is of major interest in population genetics. Current statistical methods for genome scans are no longer suited to the analysis of Next Generation Sequencing (NGS) data. We propose new statistical methods to perform genome scans on massive datasets based on Principal Component Analysis. We explain the reason why our PCA-based approaches can be seen as extensions of existing methods based on the F_{ST} measure of genetic differentiation. We additionally show that another PCA-based approach we have developed can be used to detect regions of the genome that have become adaptive because of introgression. For the local adaptation and adaptive introgression approaches, we report comparisons with state-of-the-art methods. Our statistical routines are implemented in the R package called padapt, which is designed for outlier detection in population genomics.

Keywords : population genetics, principal component analysis, local adaptation, introgression, bioinformatics.

Chapitre 1

Introduction

1.1 La génétique des populations

“Il est bon de rappeler que ce qui nous rend semblables est plus important que ce qui nous rend différents. Les milliards d’êtres humains éparpillés sur toute la planète se différencient par la couleur de la peau et par la forme du corps, par la langue et par la culture. Et cette variété, qui témoigne de notre capacité à changer, à nous adapter à des milieux différents et à y développer des modes de vie originaux, est la meilleure garantie pour l’avenir de l’espèce humaine. Les connaissances que nous avons acquises sur nous-mêmes montrent cependant que toute cette diversité, comme la surface changeante des océans ou de la voûte du ciel, est bien peu de chose par rapport à cet immense patrimoine que nous avons en commun, nous, les humains.”

— L. Cavalli-Sforza ([1994](#))

1.1.1 L’évolution comme point de départ

« La génétique est la science de l’hérédité. Elle est la clé de toute la biologie, parce qu’elle explique les mécanismes qui sont responsables de la reproduction des êtres vivants, du fonctionnement et de la transmission du matériel héréditaire, des différences entre les individus, de l’évolution biologique. »

Cette définition, donnée par Cavalli-Sforza et traduite ici de l’italien par Françoise Brun (L. Cavalli-Sforza, [1994](#)), restitue également les motivations à l’origine de l’émergence du domaine de la génétique des populations, à savoir l’étude de la variabilité interindividuelle d’un point de vue évolutionniste. Pour John H. Gillespie, il s’agit de la « discipline qui fait le lien entre la génétique et l’évolution » (Gillespie, [2010](#)) : « La génétique des populations s’intéresse à l’évolution d’un point de vue génétique. Elle diffère de la biologie en ce que ses idées les plus importantes ne sont pas expérimentales ou observationnelles mais davantage théoriques. Il pourrait difficilement en être autrement. Les objets d’étude sont principalement la fréquence et la valeur sélective des génotypes dans les populations naturelles. »

Malgré cette caractérisation, les fondements de la génétique des populations trouvent en réalité leurs origines bien avant la formalisation en 1909 par Wilhelm Johannsen du concept même de gène (Roll-Hansen, 2014), en témoignent les travaux de Charles Darwin (1809-1882) et de Gregor Mendel (1822-1884). *L'Origine des espèces*, publié en 1859 et considéré encore à ce jour comme le texte fondateur de la théorie de l'évolution (Darwin, 1980), énonce les premiers principes de la sélection naturelle. Les travaux de Mendel, figurent quant à eux parmi les premiers à se pencher sur les mécanismes de l'hérédité d'un point de vue statistique, notamment via l'étude de phénotypes en termes de proportions et de fréquences.

1.2 À l'origine de la variabilité génétique

1.2.1 La théorie de l'évolution

En 1859, Darwin soutenait l'idée selon laquelle la principale force évolutive est la sélection naturelle (Darwin, 1980). « Je me propose de passer brièvement en revue les progrès de l'opinion relativement à l'origine des espèces. Jusque tout récemment, la plupart des naturalistes croyaient que les espèces sont des productions immuables créées séparément. De nombreux savants ont habilement soutenu cette hypothèse. Quelques autres, au contraire, ont admis que les espèces éprouvent des modifications et que les formes actuelles descendent de formes préexistantes par voie de génération régulière. »

C'est de cette manière qu'en 1920, Edmond Barbier, dans sa notice relative à la traduction française de *L'Origine des espèces* (Darwin, 1980), décide de présenter le contexte dans lequel il a été amené à effectuer ce travail de traduction. Bien que la théorie de Darwin fut globalement bien accueillie par la communauté scientifique, elle fut tout de même en proie à de nombreuses critiques. L'une des principales critiques émises à son encontre fut relative à la croyance de Darwin selon laquelle l'hérédité *par mélange* serait le principal mode de transmission des caractères héréditaires (Gayon, 1992). Or, si sélection naturelle il y a, la conservation et la transmission des caractères sélectionnés sont essentielles. Si bien qu'une hérédité *par mélange* n'est pas envisageable pour soutenir la thèse de la sélection naturelle, puisque tout caractère transmis de cette façon se verrait altéré (ou dilué si l'on souhaite conserver l'idée de mélange) à chaque génération et donc éliminé après un certain temps. Cependant, sa théorie bénéficiera par la suite des travaux de Mendel qui, lors de leur redécouverte en 1902 (Bateson & Mendel, 1913), apporteront l'élément fondamental manquant à la théorie darwinienne : le principe d'hérédité *mendélienne*. Cette théorie de l'évolution néo-darwinienne, née de la conciliation de la théorie darwinienne et du principe d'hérédité de Mendel, constitue le paradigme évolutionniste tel que nous le connaissons aujourd'hui et porte le nom de *théorie synthétique de l'évolution*.

1.2.2 L'évolution d'une théorie

À la théorie néo-darwinienne est souvent opposée la théorie neutraliste développée par Motoo Kimura dans son ouvrage *The neutral theory of molecular evolution* (Kimura, 1983), bien que ces deux théories ne soient pas incompatibles. La première suggère que les mutations apparaissent à la faveur de la sélection naturelle. La seconde affirme quant à elle que l'évolution ne serait que le résultat de mutations qui surviennent de façon tout à fait aléatoire, tout en étant sélectionnées selon le même mécanisme de sélection naturelle proposé par Darwin (Figure 1.1).

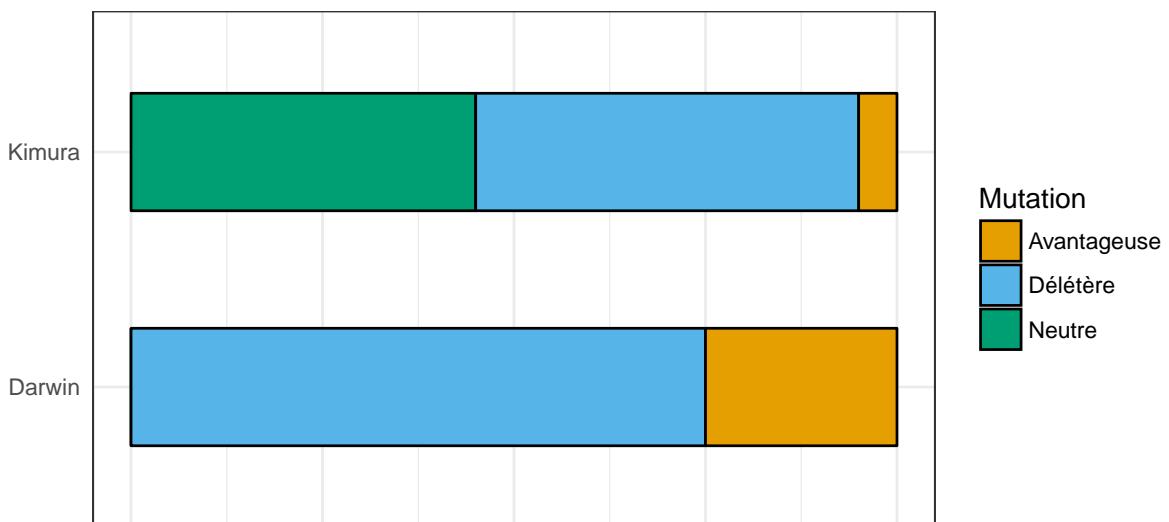


FIGURE 1.1 – Représentation schématique des probabilités d'occurrence pour chaque type de mutation pour la théorie sélectionniste de Darwin et pour la théorie neutraliste de Kimura (Bromham & Penny, 2003). Selon la théorie de Darwin, la plupart des mutations sont délétères et le reste des mutations confère un avantage sélectif. Selon la théorie de Kimura, une partie des mutations qui apparaissent n'ont pas d'effet sur la valeur sélective. Ces mutations sont dites neutres.

Une des composantes principales de cette nouvelle théorie consiste à affirmer que les fluctuations aléatoires dans les fréquences d'allèle, n'affectant que très peu ou pas du tout la valeur sélective, constituent la principale source de variabilité de l'ADN (B. Charlesworth & Charlesworth, 2009). Une grande partie de la variation génétique observée est fonctionnellement neutre et n'occasionne pas de changement de valeur sélective.

L'approbation de cette théorie, bien que conceptuellement intéressante, aura un retentissement beaucoup plus important d'un point de vue de la méthodologie statistique. La formulation d'une hypothèse permettant de décrire un processus évolutif en l'absence de sélection, portant généralement le nom d'*hypothèse nulle* ou encore de *modèle neutre*, est souvent de première nécessité dans toute démarche visant à caractériser un mécanisme de sélection. La donnée d'observations mettant en défaut le modèle neutre aura pour conséquences de créditer davantage une hypothèse invoquant

un processus de sélection. Historiquement, la statistique D de Tajima fut l'une des premières statistiques développées à partir d'une hypothèse nulle bâtie pour les mutations neutres (Tajima, 1989).

1.2.3 Forces évolutives

Ce changement de paradigme nous invite de ce fait à observer la sélection naturelle à travers le prisme de la théorie neutraliste, et donc à identifier les mutations sélectives comme des mutations dont les origines ne peuvent être uniquement expliquées par des processus biologiquement neutres. La génétique des populations distingue trois types de processus neutres : la dérive génétique, les mutations aléatoires et le flux de gènes. Ces processus, tout comme la sélection naturelle, constituent les principales *forces évolutives*.

“La mutation propose, la sélection dispose.”

— L. Cavalli-Sforza (1994)

La dérive génétique

La dérive génétique correspond à tout ce qu'il y a d'aléatoire dans l'évolution d'une population. C'est le *hasard des choses*. Le nombre de descendants ou le choix de partenaire sexuel sont des exemples de phénomènes aléatoires participant à la dérive génétique. Le principe de dérive génétique est illustré en figure 1.2, à l'aide du modèle de Wright-Fisher tel qu'il est présenté dans l'ouvrage *Population Genetics* (Gillespie, 2010). Ce modèle simpliste suppose que la taille de la population est constante et que chaque individu de la $n + 1^{\text{ème}}$ génération est issu du brassage génétique de deux individus tirés aléatoirement de la $n^{\text{ème}}$ génération.

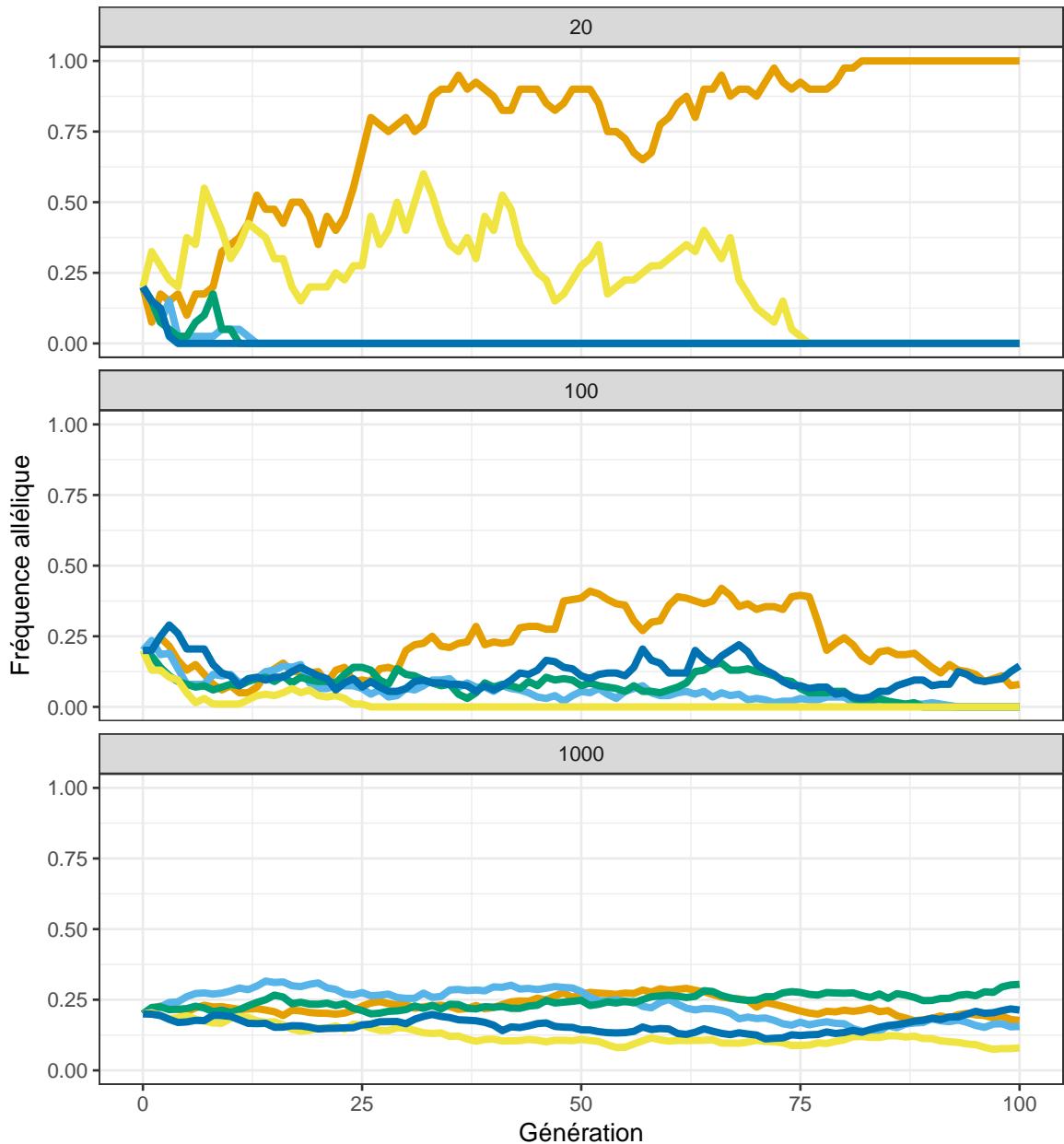


FIGURE 1.2 – Simulation numérique de la dérive génétique à l'aide du modèle de Wright-Fisher. La fréquence de l'allèle étudié est simulée pour 5 populations constituées chacune de 20, 100 ou 1000 individus sur une période de 100 générations. Dans chaque population, la fréquence de l'allèle est initialement de 0.20 (Gillespie, 2010).

En particulier, la figure 1.2 met en évidence deux caractéristiques de la dérive génétique :

- Les fréquences alléliques évoluent de façon indépendante d'une population à une autre.
- Pour un nombre de générations fixé, la dérive génétique entraîne une perte de diversité allélique plus rapidement au sein des populations de plus petite

TABLE 1.1 – Exemples de taux de mutation par paire de base et par réPLICATION chez différentes espèces (J. W. Drake et al., 1998).

Espèce	Taux de mutation
<i>E. coli</i>	5.4e-10
<i>C. elegans</i>	2.3e-10
Drosophile	3.4e-10
Souris	1.8e-10
Homme	5.0e-11

taille. Dans le modèle de Wright-Fisher, les fréquences alléliques finissent éventuellement par atteindre les états dits absorbants que sont 0 et 1.

Les mutations aléatoires

Si la dérive génétique entraîne une perte de diversité allélique, les mutations favorisent quant à elles le maintien des variations génétiques entre les populations (Gillespie, 2010). Les mutations apparaissent principalement lors de la phase de réPLICATION de l'ADN. Une mutation peut survenir à un locus donné avec une probabilité spécifique à chaque espèce (J. W. Drake, Charlesworth, Charlesworth, & Crow, 1998), appelée *taux de mutation* (Table 1.1).

Le flux de gènes

Le flux de gènes est généralement le résultat d'événements migratoires initiés par des individus appartenant à une population donnée, vers une seconde population dont les fréquences d'allèles diffèrent éventuellement de la population d'origine. Le flux de gènes influe sur la diversité génétique initialement présente si par exemple, parmi les allèles *migrants* figurent des allèles qui n'existaient pas dans la population receveuse.

La sélection naturelle

En biologie évolutive, la sélection naturelle est la force qui tend à préserver les allèles conférant des avantages quant à la viabilité ou la fertilité d'un individu. Elle agit sur les traits qui sont héritables. La grande majorité de ces traits sont transmis aux descendants en suivant le mécanisme d'hérédité génétique.

1.2.4 Adaptation locale

Principe

La diversité climatique et la diversité géologique terrestre ont naturellement façonné des environnements aux constitutions physiques et chimiques variées. À celles-ci viennent s'ajouter des caractéristiques écologiques résultant des interactions entre

TABLE 1.2 – Exemples de phénotypes liés à l'adaptation à la haute altitude distinguant les populations andines des populations tibétaines (C. M. Beall, 2007; Jeong & Di Rienzo, 2014).

Trait	Andins	Tibétains
Augmentation du taux d'hémoglobine	oui	à partir de 4000m d'altitude
Augmentation de la pression artérielle pulmonaire	oui	non
Augmentation de la ventilation au repos	oui	non
Prévalence du mal chronique des montagnes	5%	1%

l'environnement et les organismes qui y évoluent *lato sensu*. La valeur sélective de ces organismes, désignant leur capacité de survie et de reproduction, peut être impactée par les caractéristiques environnementales auxquelles ils sont exposés. Si cette valeur sélective est associée à un trait phénotypique, on dit que ce trait confère un avantage adaptatif et on parle dans ce cas d'*adaptation locale*. Une population sera ainsi dite adaptée à son environnement si elle développe, par le biais de la sélection naturelle, un ou plusieurs allèles associés à un trait adaptatif augmentant la valeur sélective des individus la constituant. Ceci en réponse aux pressions environnementales auxquelles ces individus sont soumis. À titre d'exemple, nous pouvons citer l'adaptation des populations tibétaines et andines à la haute altitude (Table 1.2).

Chez l'Homme moderne, les habitudes alimentaires et les modes de vie constituent également des caractéristiques environnementales importantes. L'agriculture et le pastoralisme¹ ont notamment participé à la diversification des environnements humains (Jeong & Di Rienzo, 2014). En Europe, l'adaptation biologique à ces nouveaux modes de vie s'est par ailleurs manifesté par la sélection du phénotype *LP*, dit de *persistence de la lactase*, caractérisant l'aptitude à digérer le lactose à l'âge adulte (Itan, Powell, Beaumont, Burger, & Thomas, 2009).

L'hybridation et le flux de gènes comme sources d'adaptation

Un caractère (un variant allélique par exemple) peut être sélectionné au sein d'une population de différentes manières :

- le caractère y est déjà présent et devient adaptif suite à un changement d'environnement (*standing genetic variation*).
- le caractère y apparaît à la suite d'une mutation spontanée (*de novo mutations*).
- le caractère est hérité d'une autre population par flux de gènes (*adaptive introgression*).

1. Le pastoralisme décrit la relation interdépendante entre les éleveurs, leurs troupeaux et les milieux exploités.

Cette dernière possibilité est d'ailleurs très intéressante pour une espèce d'un point de vue de la diversité génétique. Pour enrichir le catalogue d'allèles, nous avons vu plus haut que le flux de gènes constituait un processus clé. La reproduction sexuée est connue pour jouer un rôle important dans la sélection naturelle, offrant aux individus la possibilité, par le biais du brassage allélique, d'hériter des "meilleurs allèles" que possèdent leurs parents. Augmenter la diversité génétique augmente ainsi les chances de voir se transmettre des allèles favorables.

Nous parlons d'hybridation ou de métissage lorsque deux individus, appartenant à deux populations différentes, se reproduisent entre eux. Pour ce qui est de la distinction de populations, nous nous en tiendrons à la suggestion faite par Harrison & others (1990), considérant que deux individus issus de populations différentes doivent chacun posséder des traits héritables qui les différencient. Ainsi, l'hybridation constitue un excellent moyen pour permettre à une espèce d'intégrer de nouveaux allèles (Stevison, 2008). Bien que l'hybridation conduise fréquemment à la naissance d'individus à la valeur sélective hautement diminuée voire d'individus non fertiles, il arrive que certains descendants aient tout de même la capacité de se reproduire et ainsi être à l'origine de l'apparition de nouveaux allèles. La sélection de tels allèles transmis par flux de gènes ou par hybridation correspond à ce que l'on appelle *l'introgression adaptative*. Des exemples d'introgression sont observés à la fois dans la nature (*introgression naturelle*) et à la fois chez des espèces domestiquées (*introgression délibérée*). Nous pouvons citer l'exemple de la souris domestique, *Mus musculus domesticus*, qui a hérité d'un gène de résistance à la coumadine (communément appelée « mort aux rats ») en s'hybridant avec une espèce sauvage, *Mus spretus* (Y. Song et al., 2011). Ce gène de résistance est localisé sur le chromosome 7 de la souris domestique et appartient à une région qui a été identifiée comme provenant effectivement de la souris sauvage.

L'identification des mutations génétiques responsables de l'adaptation est particulièrement cruciale pour la compréhension des processus évolutifs et plus spécifiquement ceux liés à la spéciation². L'introgression adaptative permet quant à elle de comprendre l'adaptation rapide de certaines espèces et en quoi l'hybridation constitue un vecteur important pour l'adaptation (Hufford et al., 2013). Bien que l'étude de l'adaptation locale puisse être menée de façon expérimentale par le biais de la *transplantation réciproque* (Kawecki & Ebert, 2004), nous proposons ici d'exploiter les génotypes de populations naturelles.

1.3 Données de polymorphismes génétiques

Avant d'aborder la partie sur les méthodes développées au cours de cette thèse, nous donnons une rapide présentation du type de données que nous traitons, à savoir des données de séquençage. Le séquençage de l'ADN consiste à déterminer l'ordre dans lequel sont agencées les paires de bases pour un fragment d'ADN donné. Son apparition a offert de nouvelles voies d'exploration en biologie évolutive. Les variations génétiques étaient jusqu'alors appréciées par le biais des différences phénotypiques. En

2. épisode de divergence d'une espèce ayant conduit à la formation d'au moins une nouvelle espèce.

phylogénie par exemple, le séquençage de l'ADN a donné naissance à la phylogénétique moléculaire, qui se distingue de la phylogénétique traditionnelle en ce qu'elle ne considère que les séquences de nucléotides pour évaluer la proximité entre deux espèces. Et c'est précisément l'accès à ces séquences qui va permettre à toute une population de méthodologies statistiques de voir le jour et de se développer.

1.3.1 Des données en grande dimension

Le séquençage nouvelle génération (appelé encore séquençage à haut débit ou NGS) a connu un essor considérable au cours des dernières décennies. Si bien que les prouesses techniques et les progrès technologiques réalisés dans ce domaine ont permis de réduire d'un facteur 100,000 les coûts de séquençage en l'espace de seulement 15 ans (Figure 1.3).

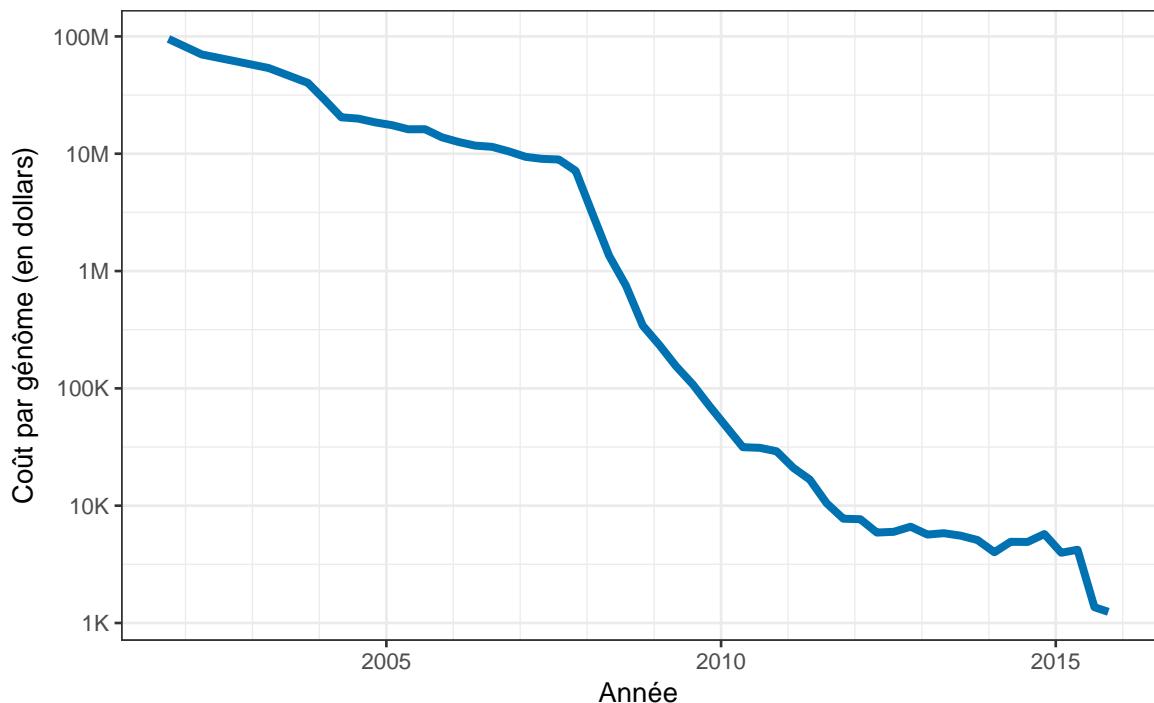


FIGURE 1.3 – Évolution des coûts de séquençage depuis 2001 (Wettersstrand, 2013).

Toutefois, compte tenu de la popularité croissante des technologies NGS (Muir et al., 2016) et des considérables volumes de données qu'elles génèrent, de nouvelles problématiques se posent quant à leur stockage et leur analyse, nécessitant l'utilisation de puissantes ressources de calcul ainsi que le développement d'algorithmes plus adaptés (Gogol-Döring & Chen, 2012).

1.3.2 Les marqueurs génétiques

Le séquençage de l'ADN a également permis de faire évoluer le concept de *marqueur génétique*. Un marqueur génétique correspondait autrefois à un gène *polymorphe*³ identifié sur la base d'observations phénotypiques. Grâce au séquençage de l'ADN, une nouvelle définition tenant compte de la position sur le chromosome a été adoptée pour caractériser un marqueur génétique. Différents types de marqueurs génétiques ont été identifiés, parmi lesquels figurent les microsatellites, les insertions, les délétions et les SNPs⁴. La structure spatiale de l'ADN n'étant pas prise en compte dans les travaux présentés ici, nous en garderons une représentation unidimensionnelle.

Microsatellite

Jusqu'à présent, les microsatellites ont connu un succès important, notamment grâce à la popularisation de techniques telles que la PCR (*Réaction en Chaîne par Polymérase*). Cependant, grâce aux nouvelles avancées technologiques que nous évoquerons un peu plus loin, ils sont progressivement délaissés au profit des SNPs. Un microsatellite est repérable par la répétition successive de petits motifs chacun composé de 1 à 4 nucléotides (Figure 1.4).

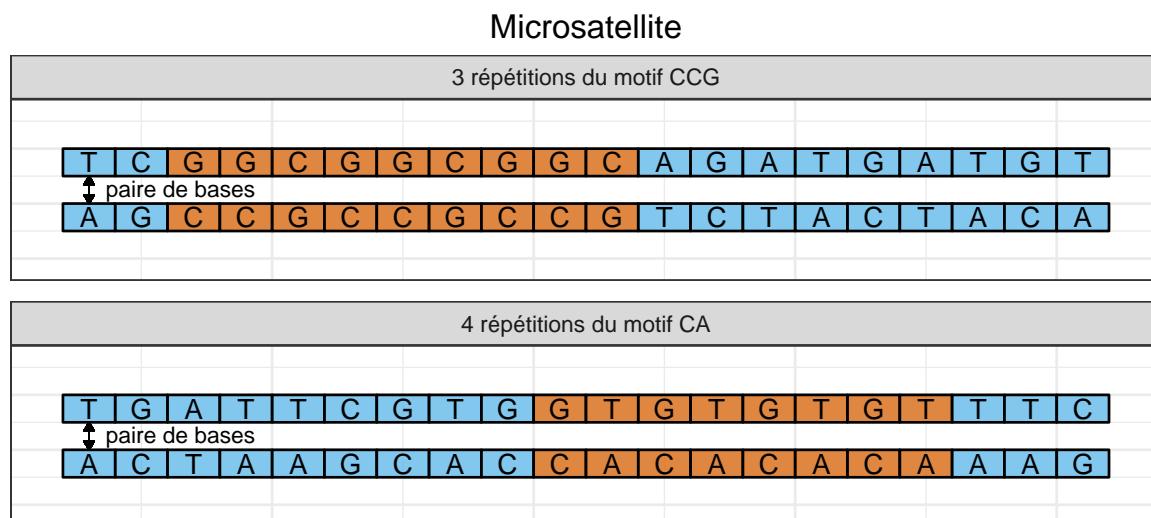


FIGURE 1.4 – Exemples de microsatellites. La première séquence comporte 3 répétitions du motif CCG, tandis que la seconde inclut 4 répétitions du motif CA.

Insertion/Délétion (Indel)

L'insertion ou la délétion d'une base constituent également des polymorphismes génétiques. Relativement à une séquence de nucléotides de référence, une insertion

3. présentant des variations à l'échelle de l'espèce.

4. SNP (*Single-Nucleotide Polymorphism*) : polymorphisme d'un seul nucléotide.

consiste en la présence d'une base supplémentaire tandis que la délétion consiste en l'absence d'une base (Figure 1.5).

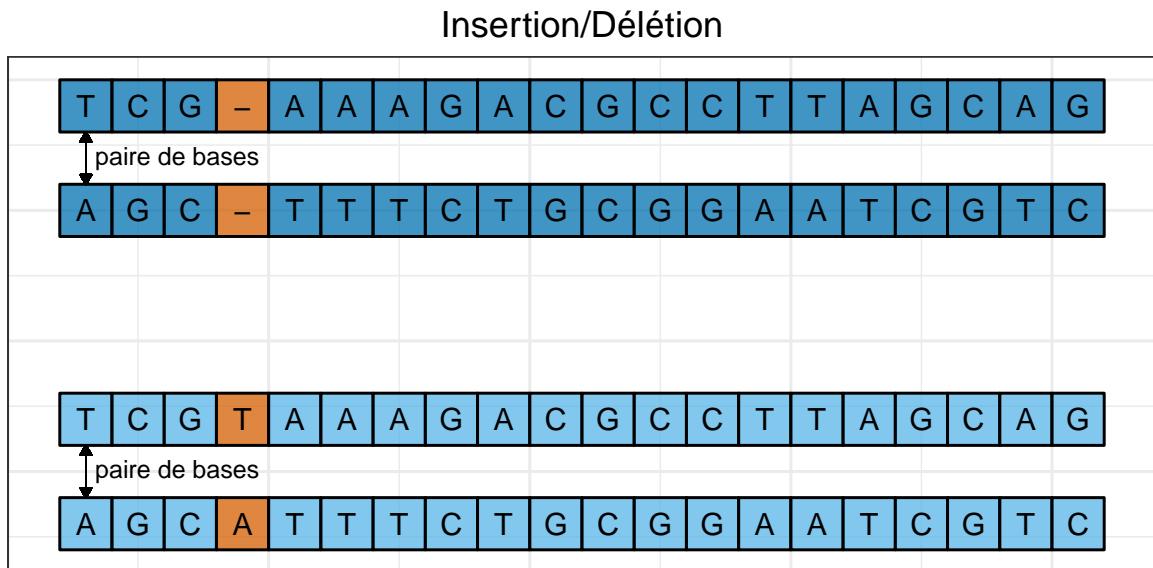


FIGURE 1.5 – Exemple d'Indel.

Polymorphisme d'un seul nucléotide (SNP)

Le polymorphisme d'un seul nucléotide correspond au polymorphisme génétique le plus simple, et correspond à l'emplacement d'un nucléotide présentant des variations appréciables à l'échelle d'une population (Figure 1.6).

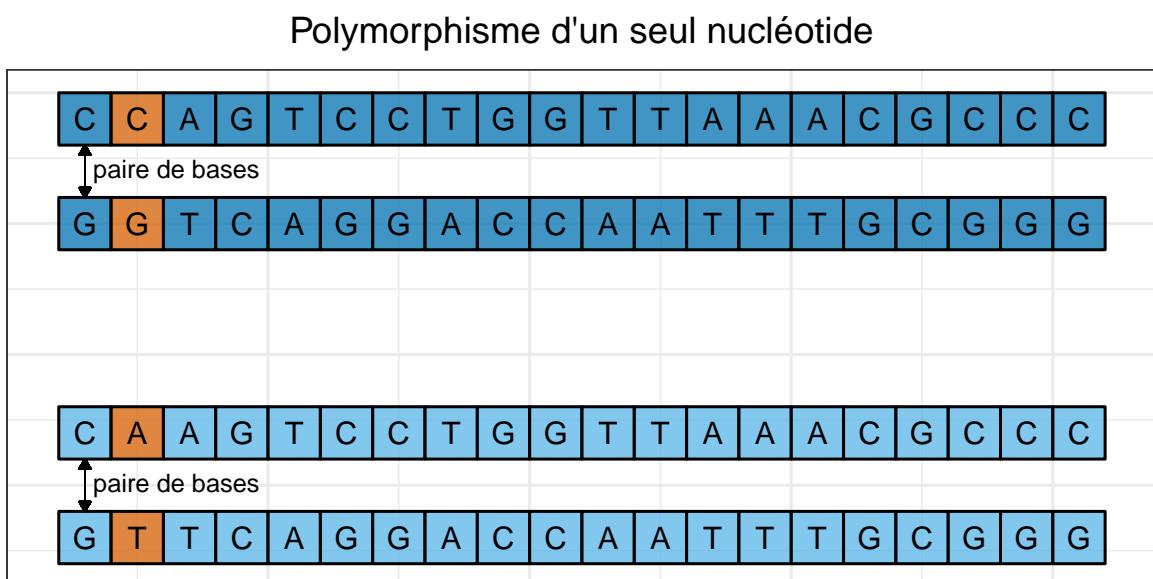


FIGURE 1.6 – Exemple de SNP.

1.3.3 Encodage des données génétiques

Dans ce paragraphe nous présentons le format des données sur lesquelles nous travaillons et définissons quelques notations qui seront utilisées par la suite. Conformément à l'usage qui en est fait par Gillespie (2010), nous emploierons le terme d'allèle pour désigner la version d'une base nucléique à un locus donné. Dans le cadre de nos travaux, nous considérerons que pour un locus donné, il n'existe que deux nucléotides possibles (SNP bi-allélique), dont l'un sera considéré comme *l'allèle de référence* et l'autre comme *l'allèle alternatif*. Pour un locus donné, l'encodage des données de SNPs consiste à attribuer à chaque individu (Figure 1.7) :

- la valeur 0 s'il est homozygote pour l'allèle de référence.
- la valeur 1 s'il est hétérozygote.
- la valeur 2 s'il est homozygote pour l'allèle alternatif.

Une base de données de génomes constituée de séquences de nucléotides pourra donc en pratique être encodée par une matrice, appelée *matrice de génotypes*, composée uniquement de 0, 1 et 2.

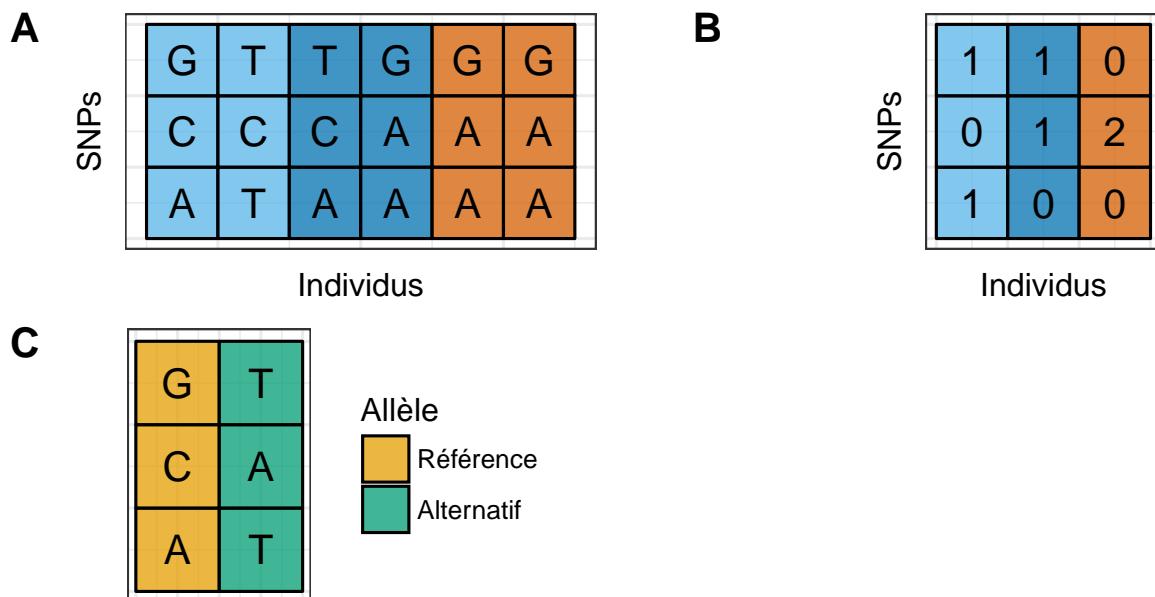


FIGURE 1.7 – Exemple d'encodage de données de SNPs. **A.** Chaque ligne de la matrice correspond à un SNP et chaque individu est représenté par une paire de colonnes (de la même couleur). **B.** Matrice de génotypes résultant de l'encodage. **C.** Chaque ligne de la matrice correspond à un SNP. La première colonne indique quel allèle est considéré comme allèle de référence. La seconde colonne indique quel allèle est considéré comme allèle alternatif.

Dans la mesure du possible, l'allèle de référence (resp. alternatif) est choisi de façon à correspondre à l'allèle ancestral (resp. dérivé). En pratique, le choix de l'allèle de référence et de l'allèle alternatif peut se faire de façon totalement arbitraire sans que cela n'influe sur les méthodes statistiques basées sur la variance des allèles, ce qui

sera généralement le cas pour celles qui sont présentées ici.

Les matrices de génotypes seront généralement notées G et seront considérées comme des éléments de $\mathcal{M}_{np}(\{0, 1, 2\})$, où p désigne le nombre de locus et n le nombre d'individus. Ainsi, $G_{i,:}$ désignera le vecteur de taille p composé du génotype de l'individu i . De même, $G_{:,j}$ désignera le vecteur de taille n contenant les comptages de l'allèle alternatif pour les différents individus au locus j . À la matrice G sera associée la matrice \tilde{G} , normalisée pour les lignes (c'est-à-dire pour les SNPs). Notant f_j la fréquence allélique de l'allèle alternatif (ou de l'allèle de référence), nous définissons $\tilde{G}_{:,j}$ par :

$$\tilde{G}_{:,j} = \frac{G_{:,j} - 2f_j}{\sqrt{2f_j(1-f_j)}} \quad (1.1)$$

1.4 Les motivations de la thèse

La production de données génétiques, volumineuses par la quantité d'information qu'elles renferment, laisse présager le meilleur pour les domaines de la médecine clinique et de la biologie évolutive. En génétique des populations, leur acquisition offre de nombreuses perspectives d'étude, notamment concernant la mise en évidence de gènes impliqués dans les processus évolutifs ou de gènes associés à certains phénotypes. Les méthodes répondant à cette problématique portent le nom de *scans génomiques*. Les scans génomiques pour la sélection, destinés à isoler des locus sous sélection, se sont largement répandus au cours des dernières années, principalement grâce au développement fulgurant qu'ont connu les technologies NGS. Pour illustrer le principe des scans génomiques, nous reprenons l'exemple de l'adaptation à l'altitude des populations tibétaines.

Pour détecter les gènes responsables de cette adaptation, Xu et al. (2010) ont réalisé un scan génomique en analysant la différenciation génétique entre une population constituée de tibétains et une autre constituée de Hans. Le résultat de ce scan génomique⁵ est présenté en figure 1.8.

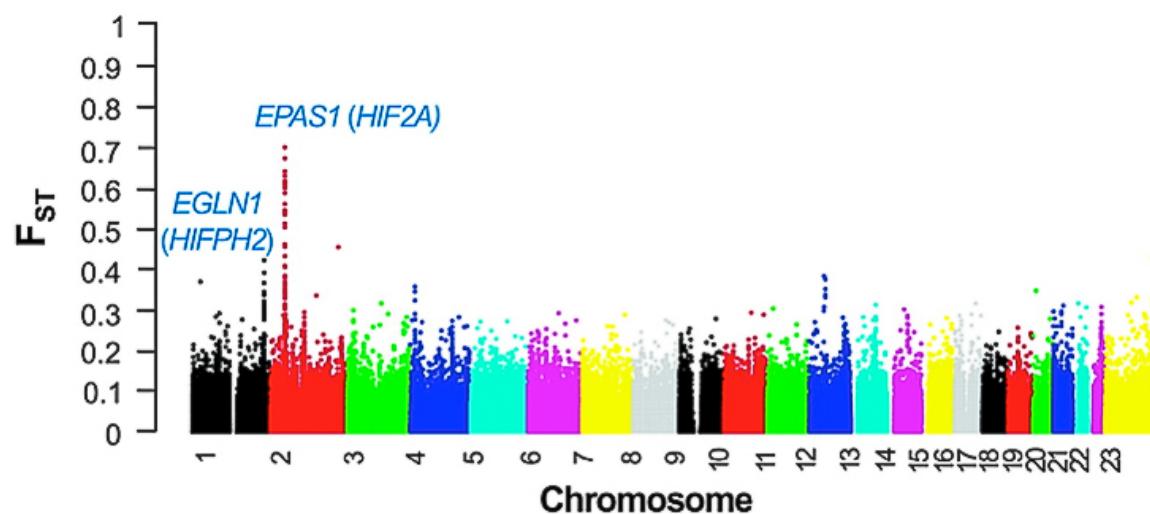


FIGURE 1.8 – Scan génomique réalisé sur des populations de tibétains et de Hans utilisant la F_{ST} (mesure de différenciation génétique de référence que nous introduirons dans le chapitre suivant). Deux régions génomiques présentant des valeurs de F_{ST} significativement élevées ont été identifiées sur les chromosomes 1 et 2 (Xu et al., 2010).

Les SNPs présentant la différenciation la plus élevée sont localisés sur le facteur de transcription EPAS1 connu pour être activé en condition d'hypoxie. À la suite de ces scans de différenciation, il a été montré à l'aide de la statistique D (Table 1.3)

5. se présente généralement sous la forme d'un graphique, appelé *Manhattan plot*, représentant la statistique de test utilisée en fonction de l'emplacement chromosomique.

TABLE 1.3 – Scan génomique pour l’introgression chez les tibétains réalisé par H. Hu et al. (2017) utilisant la statistique D (D^* correspond à la normalisation de la statistique D que nous introduirons dans le chapitre 3). La colonne de droite contient le nom des gènes potentiellement hérités de l’Homme de Denisova. Nous retrouvons dans cette liste le gène EPAS1 (H. Hu et al., 2017).

Chr	début	fin	D^*	Gènes
7	26800001	27000000	6.25	SKAP2
2	46400001	46600000	6.15	PRKCE, EPAS1
2	47600001	47800000	5.92	MIR559, MSH2, EPCAM, KCNK12
5	200001	400000	4.99	CCDC127, PDCD6, SDHA
12	8800001	9000000	4.63	MFAP5, RIMKLB, A2ML1
4	100400001	100600000	4.58	TRMT10A, C4orf17, MTTP

que les variants adaptatifs sur EPAS1 ont été transmis grâce à l’introgression issue de l’Homme de Denisova (H. Hu et al., 2017).

Principalement réalisés sur des données humaines, les scans génomiques se sont progressivement étendus à d'autres espèces, modèles et non-modèles (Haasl & Payseur, 2016), et les SNPs sont désormais les marqueurs les plus utilisés (Figure 1.9).

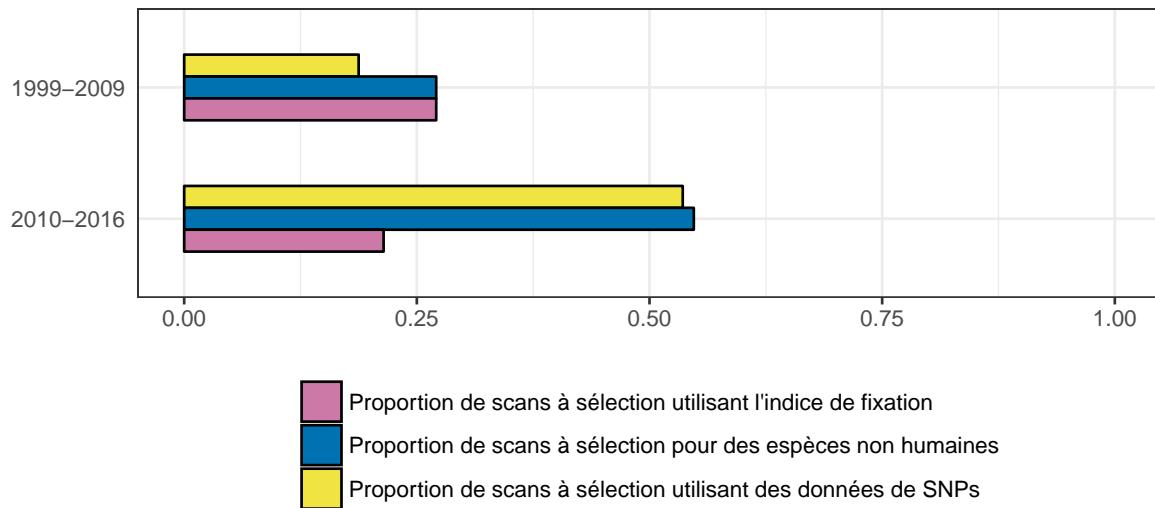


FIGURE 1.9 – Évolution de la proportion d’articles scientifiques s’intéressant à des espèces non humaines et de la proportion de ceux dont l’étude est réalisée sur des données de SNPs, dans le cadre des scans génomiques pour la sélection (Haasl & Payseur, 2016).

Toutefois, certains outils statistiques ayant été développés ne sont plus adaptés. La raison principale étant que les temps de calculs requis par ces outils pour l’analyse de données massives sont souvent prohibitifs. Les méthodes d’analyse exploratoire nécessitent souvent d’être expérimentées plusieurs fois. D’une part parce que le pré-traitement des données peut être amené à changer au cours de l’étude (filtration de marqueurs génétiques présentant une proportion trop élevée de données manquantes, retrait d’individus marginaux, etc.). D’autre part car les méthodes d’analyse proposent généralement un éventail de paramètres dont le choix ne s’impose pas spontanément. Expérimenter une même méthode avec différents critères de qualité et différents paramètres est souvent nécessaire pour minimiser les biais d’utilisation. Tout ceci justifie le recours à des méthodes rapides et adaptées à ces nouveaux volumes de données. Pour satisfaire à ces critères, des méthodes basées sur l’Analyse en Composantes Principales ont été proposées pour étudier l’adaptation locale (Duforet-Frebbourg, Luu, Laval, Bazin, & Blum, 2015 ; Galinsky et al., 2016).

1.4.1 Résultats principaux et organisation du manuscrit

Sur la base du travail de thèse de Nicolas Duforet-Frebbourg (Duforet-Frebbourg, 2014), nous avons cherché dans un premier temps à améliorer la méthodologie statistique implémentée dans le logiciel PCAdapt, ainsi que les performances computa-

tionnelles, afin de garantir la possibilité de l'utiliser sur des génomes plus denses et disposant d'une plus grande résolution. Afin de permettre la publication de l'article (Duforet-Frebourg et al., 2015), j'ai réalisé des simulations de modèles en îles et de divergence afin de valider l'approche basée sur la communalité pour réaliser des scans à sélection. J'ai ensuite développé la librairie R pcadapt et proposé une statistique de scan à sélection plus puissante que la communalité qui est la distance robuste de Mahalanobis. Ce travail correspond à la publication (Luu, Bazin, & Blum, 2017). La dernière contribution principale de ma thèse correspond au développement d'une méthode statistique basée sur l'ACP qui permet d'identifier les régions du génome impliquées dans les événements d'introgression adaptive. Ce travail correspond au manuscript *Scanning genomes for adaptive introgression using principal component analysis* et fera l'objet d'une soumission prochainement.

Dans le chapitre 2, nous commençons par décrire les différents modèles, historiques et contemporains, utilisés pour réaliser des scans génomiques pour la sélection. Ce premier chapitre rappelle ensuite quelle est l'utilisation de l'Analyse en Composantes Principales en génétique des populations. La fin du chapitre 2 rappelle les résultats obtenus avec la communalité et la distance de Mahalanobis, qui sont implémentées dans la librairie pcadapt, et qui correspondent aux publications (Duforet-Frebourg et al., 2015 ; Luu et al., 2017). Le chapitre 3 traite de l'introgression adaptive. Les chapitres 2 et 3 montrent comment l'ACP peut être exploitée de façon à détecter des signaux de sélection ainsi que des régions d'introgression adaptive. Enfin, nous terminerons avec les aspects computationnels et numériques qui nous ont préoccupé tout au long de la thèse, en présentant quelques comparatifs de performances réalisés avec divers algorithmes.

Chapitre 2

Adaptation locale

Cette première partie fera dans un premier temps un état de l'art des méthodes destinées à identifier des locus impliqués dans des processus d'adaptation locale. Nous présentons différentes méthodes classiques de scan génomique pour la sélection. Ensuite, nous présentons l'utilisation de l'Analyse en Composantes Principales en génétique des populations et nous montrons comment l'utiliser pour faire des scans à sélection. Par souci de clarté, nous ne considérons ici que des espèces diploïdes, bien qu'une grande partie des résultats présentés ici puisse être adaptée au cas d'espèces haploïdes. Les locus seront par ailleurs supposés bi-alléliques, c'est-à-dire que pour un locus donné, au plus deux allèles sont observés sur ce locus à l'échelle de la population étudiée.

2.1 L'état de l'art pour les scans génomiques

2.1.1 Modèles démographiques

Afin de mieux comprendre l'heuristique des méthodes de scan génomique présentées ici, nous donnons dans ce paragraphe une brève description des modèles démographiques fréquemment utilisés en génétique des populations. En effet l'idée de sélection dans une population est généralement relative à (au moins) une autre population, et l'histoire démographique de ces populations joue un rôle important sur la distribution théorique des fréquences alléliques.

Modèle en îles

Dans un modèle en îles, les différentes populations échangent entre elles des individus au cours du temps (Figure 2.1). La proportion d'individus échangés est appelée taux de migration. De forts taux de migration vont avoir tendance à homogénéiser les variations génétiques entre les populations. Des faibles taux de migration vont en revanche conduire à une différenciation plus forte. Les différences de taux de migration peuvent par exemple être expliquées par l'existence de barrières naturelles (Landguth, Cushman, Murphy, & Luikart, 2010).

Modèle *star-like*

Le modèle *star-like* suppose l'existence d'une population ancestrale de laquelle sont issues différentes populations (Figure 2.1, panels B et C). Contrairement au modèle en îles, les populations évoluent de façon indépendante sans s'échanger d'individus, et se différencient éventuellement sous l'effet de la dérive génétique et de mutations aléatoires. Il existe un modèle de divergence instantanée dans lequel la quantité de dérive est la même pour toutes les branches de l'arbre de populations (Figure 2.1, panel C).

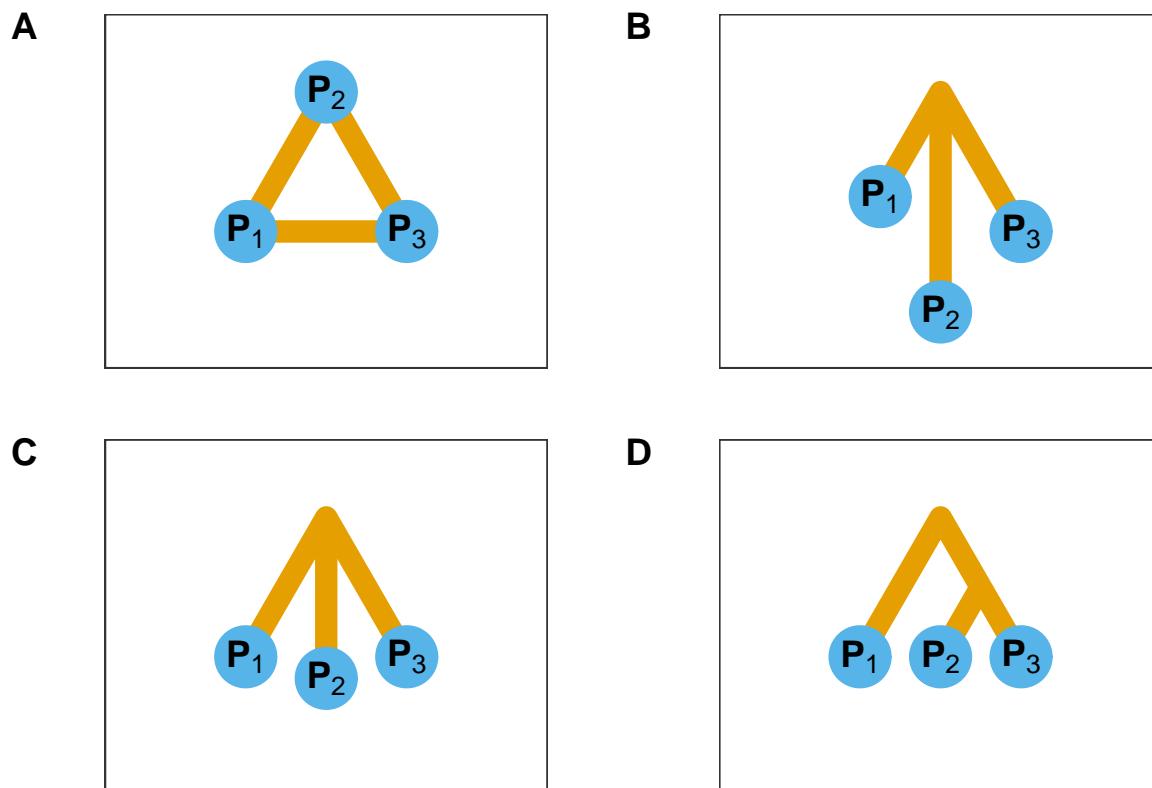


FIGURE 2.1 – Modèles démographiques. **A.** Représentation schématique d'un modèle en îles à trois populations. **B.** Représentation schématique d'un modèle *star-like* à trois populations. Les longueurs de branches correspondent à la dérive génétique depuis la divergence initiale et sont ici différentes les unes des autres. Dans un modèle de divergence, la quantité de dérive est proportionnelle au temps de divergence divisé par la taille efficace de la sous-population. **C.** Représentation schématique d'un modèle *star-like* à trois populations où les trois branches ont subi la même quantité de dérive génétique. **D.** Représentation schématique d'un modèle de divergence présentant une structure hiérarchique.

2.1.2 L'indice de fixation

Indépendamment de l'histoire démographique, un allèle sélectionné voit généralement sa prévalence augmenter au sein d'une population. Si bien que l'observation d'une fréquence allélique anormalement élevée dans une population relativement aux autres donne à suggérer que cet allèle a été favorablement sélectionné. Dans l'optique de détecter des signaux de sélection, il semble alors naturel de proposer une statistique testant si au moins deux populations présentent des fréquences alléliques significativement différentes l'une de l'autre (Holsinger & Weir, 2009). L'indice de fixation, ou encore F_{ST} en abrégé, est une statistique basée sur cette heuristique.

Définition 2.1 (Indice de fixation). Pour un locus donné, dénotant N le nombre de populations considérées et (p_1, p_2, \dots, p_N) les fréquences d'un des deux allèles existant, la F_{ST} est définie par la relation suivante (Wright, 1943) :

$$F_{ST} = \frac{\frac{1}{N-1} \sum_{i=1}^N (p_i - \bar{p})^2}{\bar{p}(1 - \bar{p})} \quad (2.1)$$

où $\bar{p} = \frac{1}{N-1} \sum_{i=1}^N p_i$.

Dans l'expression de la F_{ST} donnée en 2.1, le numérateur correspond à la variance génétique interpopulationnelle tandis que le dénominateur correspond à la variance génétique mesurée à l'échelle de la *métapopulation*¹. La F_{ST} peut être vue comme la réduction de variance génétique due à la structure de populations. Autrement dit, on regarde si le fait de grouper les individus dans des populations a une incidence ou non sur la variance génétique.

2.1.3 Test de Lewontin-Krakauer

Comme expliqué plus haut, la détection de locus sous sélection passe généralement par la caractérisation d'un modèle décrivant l'évolution de locus sous l'effet de processus neutres tels que la dérive génétique. Dans le cas des scans génomiques, il s'agit d'estimer la distribution neutre de la statistique de test (calculée en chaque locus), afin d'identifier les locus qui s'en écartent le plus. Suivant ce principe, Lewontin et Krakauer proposent pour la F_{ST} un test d'adéquation du χ^2 (Lewontin & Krakauer, 1973).

Définition 2.2 (Test de Lewontin-Krakauer). Notant N le nombre de populations, la statistique de test introduite par Lewontin & Krakauer (1973), dénotée T_{LK} , a pour expression :

$$T_{LK} = \frac{N-1}{\bar{F}_{ST}} F_{ST} \quad (2.2)$$

Dans un scénario de divergence instantanée (Figure 2.1), sous l'hypothèse que les fréquences alléliques sont distribuées selon une loi normale ou binomiale, T_{LK} suit une

1. ensemble de populations d'individus appartenant à la même espèce.

loi du χ^2 à $N - 1$ degrés de libertés (Lewontin & Krakauer, 1973). En effet, notant $p = (p_1, p_2, \dots, p_N)$, en utilisant la définition 2.1 de la F_{ST} et en remarquant que :

$$(N - 1)F_{ST} = \frac{1}{\bar{p}(1 - \bar{p})} \sum_{i=1}^N (p_i - \bar{p})^2 = \left(\frac{p - \bar{p}}{\sqrt{\bar{p}(1 - \bar{p})}} \right) \left(\frac{p - \bar{p}}{\sqrt{\bar{p}(1 - \bar{p})}} \right)^T, \quad (2.3)$$

il est possible de réécrire T_{LK} sous la forme d'une somme quadratique de lois normales. Cependant, les contraintes sur le modèle démographique sous-jacent sont extrêmement fortes et dans certaines situations elles ne seront pas vérifiées. Par exemple, dans les exemples de la figure 2.1, l'approximation χ^2 est correcte pour le modèle en îles (Figure 2.1, panel A) et pour le modèle *star-like* (Figure 2.1, panel C). En revanche, pour le panel C où la longueur des branches est différente, l'approximation χ^2 n'est plus correcte parce que les variances des quantités $p_i - \bar{p}$ ne sont pas identiques pour différentes valeurs de i . Des variantes de ce test ont donc été proposées pour s'adapter à des modèles de structure de populations plus flexibles (Bonhomme et al., 2010 ; Excoffier, Hofer, & Foll, 2009 ; Whitlock & Lotterhos, 2015).

Estimation du nombre de degrés de liberté effectif

Une manière de s'adapter à des modèles plus flexibles est de garder la statistique de test de l'équation (2.3) et d'améliorer l'approximation χ^2 . Pour améliorer l'approximation χ^2 , Whitlock & Lotterhos (2015) proposent d'approcher la statistique T_{LK} avec un χ^2 à df degrés de libertés au lieu de $N - 1$ degrés de libertés où df est un paramètre à estimer. Pour estimer df , Whitlock & Lotterhos (2015) dérivent un modèle de vraisemblance basé sur la distribution de la F_{ST} . Partant de la densité d'une variable aléatoire suivant un χ^2 à df degrés de libertés, la densité de la F_{ST} s'écrit :

$$f(F_{ST}) = \frac{df}{\bar{F}_{ST}} \times \frac{1}{2^{\frac{df}{2}} \Gamma\left(\frac{df}{2}\right)} \times \left(\frac{df}{\bar{F}_{ST}} F_{ST} \right)^{-1+\frac{df}{2}} \quad (2.4)$$

L'estimation de df se fait en maximisant la fonction de log-vraisemblance $\sum_{i=1}^p \log(f(F_{ST}^i))$ (où F_{ST}^i désigne la F_{ST} observée pour l'allèle i). La correction du test de Lewontin-Krakauer consiste alors à tester l'adéquation de $\frac{df}{\bar{F}_{ST}} F_{ST}$ à un χ^2 à df degrés de liberté.

Dérivation du test de Lewontin-Krakauer dans le cas de populations structurées

Dans le cas de structure hiérarchique (Figure 2.1, panel D), les quantités $p_i - \bar{p}$ ne sont plus indépendantes entre elles, ce qui remet en cause l'approximation χ^2 . Pour tenir compte de la structure hiérarchique, Bonhomme et al. (2010) proposent de corriger la statistique T_{LK} pour l'apparentement génétique des populations, modélisé par une matrice $\mathcal{F} = (f_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}_N(\mathbb{R})$, où f_{ij} mesure la corrélation entre p_i et p_j et peut être interprétée comme la probabilité qu'un individu de la population i et un individu de la population j aient hérité de cet allèle d'un même ancêtre commun.

La statistique de test T_{F-LK} , correspondant au test de Lewontin-Krakauer corrigé pour l'apparentement génétique \mathcal{F} , garde une forme analogue à celle développée en (2.3) :

$$T_{F-LK} = \left(\frac{p - \hat{p}_0}{\sqrt{\hat{p}_0(1 - \hat{p}_0)}} \right) \mathcal{F}^{-1} \left(\frac{p - \hat{p}_0}{\sqrt{\hat{p}_0(1 - \hat{p}_0)}} \right)^T \quad (2.5)$$

où \hat{p}_0 désigne un estimateur de la fréquence p_0 de l'allèle dans la population ancestrale. Dans le cas où les locus sont soumis uniquement à la dérive génétique, $T_{F-LK} \sim \chi^2_{N-1}$ (Bonhomme et al., 2010). En pratique, la méthode implémentée dans le logiciel hapflk cherche à estimer les paramètres p_0 et \mathcal{F} .

2.1.4 Le modèle F

Une autre idée consiste à affirmer qu'en l'absence de sélection, dans un modèle en îles (Figure 2.1, panel A), la proportion d'allèles immigrants² doit être la même pour tous les locus (Beaumont & Balding, 2004). Cette proportion d'allèles immigrants mesure la dérive génétique subie par la population qui intègre ces allèles (Villemereuil & Gaggiotti, 2015). En effet, une population échangeant moins d'allèles avec les autres populations se retrouverait alors plus différenciée. Les locus susceptibles d'être sous sélection sont ceux présentant une proportion d'allèles migrants anormalement basse (Bazin, Dawson, & Beaumont, 2010 ; Petry, 1983). Beaumont & Balding (2004) proposent alors un modèle de régression logistique pour la F_{ST} (Balding, 2003 ; Balding, Bishop, & Cannings, 2008), en la modélisant par des effets α spécifiques au locus (taux de mutation, sélection) et des effets β spécifiques à la population (taille de population efficace, taux de migration), c'est le modèle F :

$$\log \left(\frac{F_{ST}}{1 - F_{ST}} \right) = \alpha + \beta \quad (2.6)$$

La décomposition de la F_{ST} , en une somme d'effets locus-spécifique et population-spécifique, permet d'identifier les locus à forte F_{ST} qui présentent un effet qui n'est pas partagé par les autres locus. L'un des défauts de ce modèle est qu'il ne prend pas en compte la structure hiérarchique (Foll, Gaggiotti, Daub, Vatsiou, & Excoffier, 2014). Le modèle F est implémenté dans le logiciel Bayescan.

La liste des méthodes présentées ici n'est pas exhaustive, mais elle met en avant plusieurs défauts qui sont communs à la plupart des méthodes de scan génomique.

Dans le cas de la F_{ST} , la nécessité de travailler avec des fréquences alléliques populationnelles impose de travailler à l'échelle des populations plutôt qu'à l'échelle des individus. Les conséquences directes d'une telle nécessité sont :

- l'assignation arbitraire de chaque individu à une population. La présence d'individus métissés peut s'avérer problématique.

2. venant d'une autre population.

— la supposition que la structure de populations n'est pas continue.
Par ailleurs, les méthodes de scan génomique basées sur des méthodes bayésiennes telles que Bayescan ou la première version de PCAdapt (Duforet-Frebourg, Bazin, & Blum, 2014; Foll & Gaggiotti, 2008) sont connues pour être computationnellement lourdes, et peuvent nécessiter plusieurs jours de calcul même dans le cas de jeux de données comportant seulement quelques milliers de SNPs.

2.2 L'Analyse en Composantes Principales en génétique des populations

L'Analyse en Composantes Principales est un outil incontournable pour l'analyse de données génétiques. Elle est notamment connue pour sa capacité à retrouver la structure génétique, et donne la possibilité de ne garder qu'un nombre réduit de variables tout en résumant l'essentiel de la variation génétique. Nous présentons ici l'Analyse en Composantes Principales ainsi que ses applications en génétique des populations.

2.2.1 Principe de l'Analyse en Composantes Principales

L'Analyse en Composantes Principales est une méthode statistique consistant à transformer un ensemble de variables possiblement corrélées en un nouvel ensemble de variables orthogonales et donc non corrélées appelées *composantes principales*. De plus, elle est définie de telle sorte que la première composante principale maximise la variance, ce qui signifie que parmi toutes les droites affines de l'espace de départ, la première composante correspond à celle où la projection orthogonale des observations présente la dispersion (ou la variance) la plus grande (Figure 2.2). De la même manière, la deuxième composante principale correspond à la droite affine maximisant la variance, sous la contrainte d'être orthogonale à la composante principale précédente (Figure 2.2). Les composantes principales suivantes se déduisent donc des précédentes en suivant ce schéma itératif. La contrainte d'orthogonalité impose que le nombre de composantes principales soit inférieur au nombre de variables et au nombre d'observations. En pratique, le calcul des composantes principales repose sur la diagonalisation de la matrice de covariance ou de la matrice de corrélation.

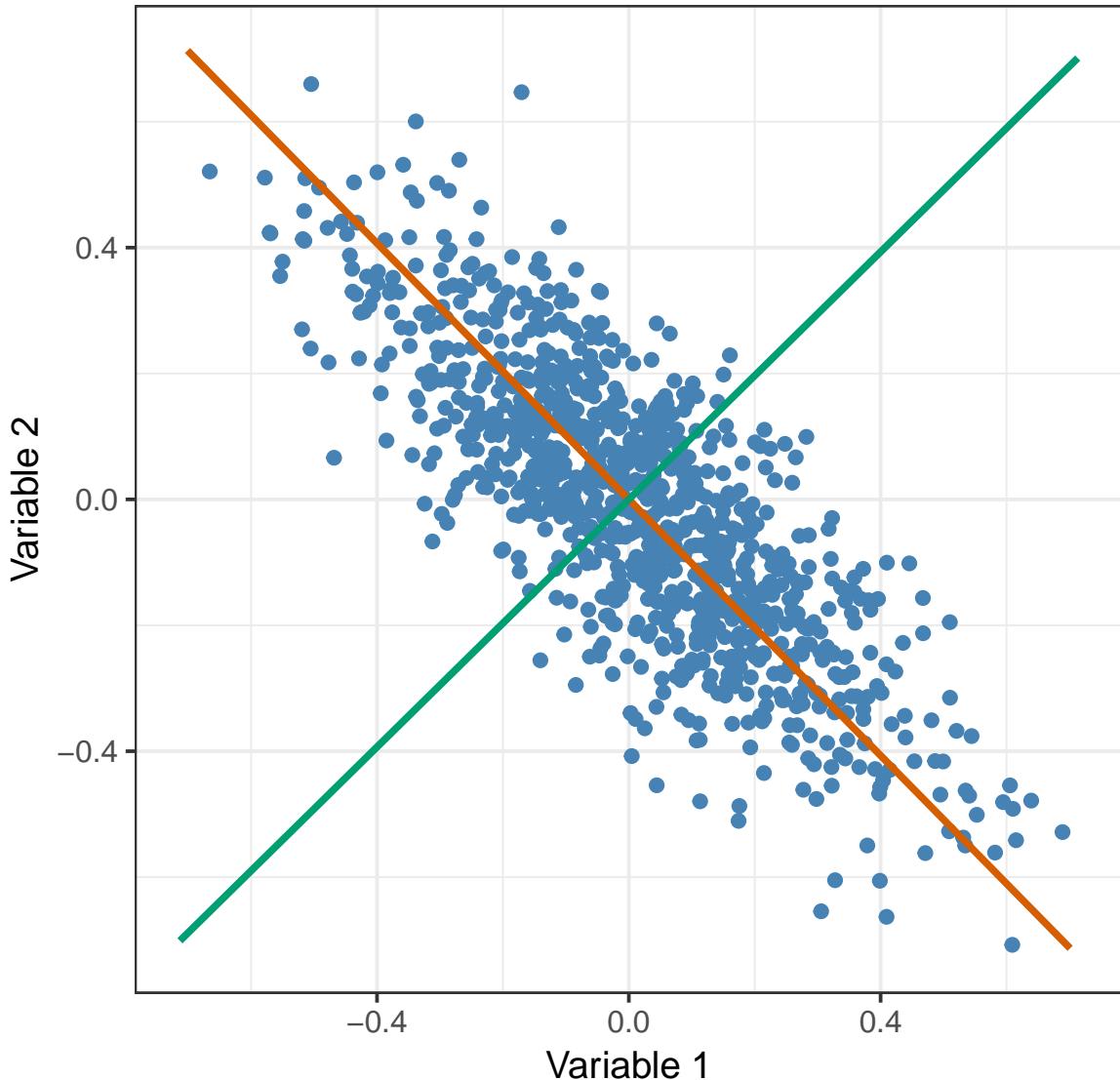


FIGURE 2.2 – Analyse en Composantes Principales de données distribuées selon une Gaussienne multivariée. La droite rouge correspond à l’axe de projection maximisant la variance, et donc par définition, à la première composante principale. La droite verte correspond à la deuxième composante principale et se déduit de la première grâce à la contrainte d’orthogonalité et au fait qu’il n’y a ici que deux variables.

2.2.2 Apparentement génétique interindividuel

L’utilisation de l’Analyse en Composantes Principales en génétique des populations a été popularisée par Cavalli-Sforza (P. Menozzi, Piazza, & Cavalli-Sforza, 1978). En génétique des populations, l’Analyse en Composantes Principales est réalisée à partir de la diagonalisation d’une matrice de covariance particulière, appelée matrice d’apparentement génétique (G. McVean, 2009). Nous avons vu un peu plus haut la notion d’apparentement génétique interpopulationnel ainsi que l’intérêt de corriger

la F_{ST} pour celui-ci. Depuis l'apparition des données génomiques, il est possible de définir des mesures de similarité génétique à l'échelle de l'individu à partir des données de génotype, contrairement à l'apparentement génétique interpopulationnel qui est défini à partir des fréquences alléliques. La figure 2.3 compare les deux matrices d'apparentement pour une même simulation d'un modèle démographique à trois populations, et met en évidence le fait que l'apparentement génétique peut s'apprécier à une échelle plus fine. Il existe cependant différentes définitions de l'apparentement génétique interindividuel (D. Speed & Balding, 2015). Nous utiliserons la définition basée sur la corrélation allélique, retenue par Galinsky et al. (2016) et G. Chen, Lee, Zhu, Benyamin, & Robinson (2016), auquel cas l'apparentement génétique entre l'individu i et l'individu j est donnée par la quantité suivante :

$$G_{RM,ij} = \frac{1}{p} \sum_{k=1}^p \frac{(G_{ki} - 2p_k) \times (G_{kj} - 2p_k)}{2p_k(1 - p_k)} \quad (2.7)$$

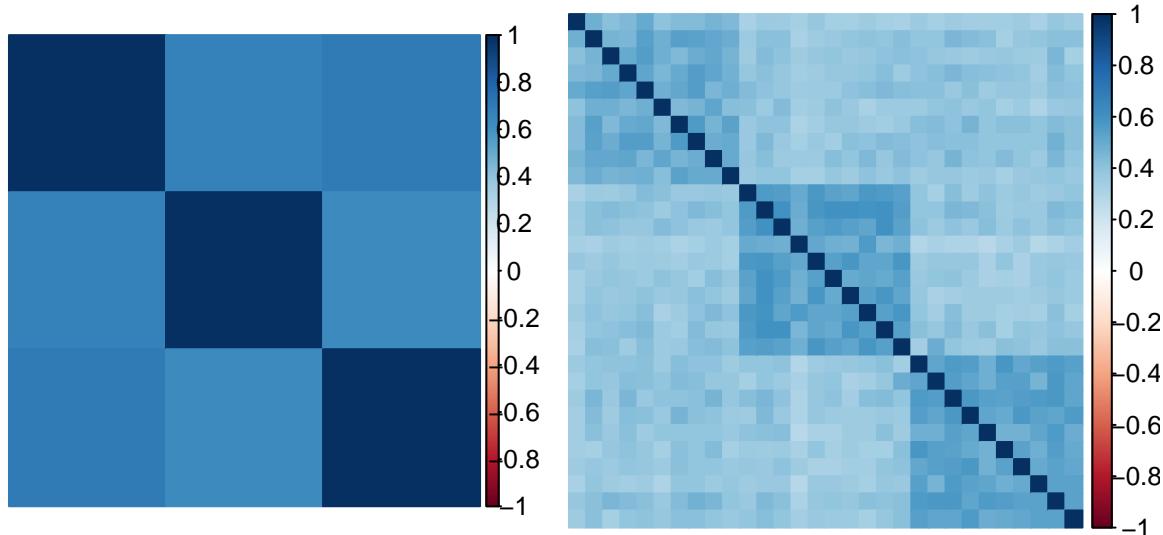


FIGURE 2.3 – À gauche une matrice d'apparentement génétique interpopulationnelle. À droite une matrice d'apparentement génétique interindividuelle. Ces matrices d'apparentement ont été estimées à partir d'une simulation d'un modèle en îles à trois populations.

2.2.3 Applications en génétique des populations

Visualisation

En génétique des populations, l'ACP est devenu un outil de visualisation extrêmement utilisé. Cela s'explique notamment par sa capacité à rendre compte de la structure de populations à l'aide d'un faible nombre d'axes principaux (Figure 2.4), appelés aussi *axes de variation génétique* (Price et al., 2006).

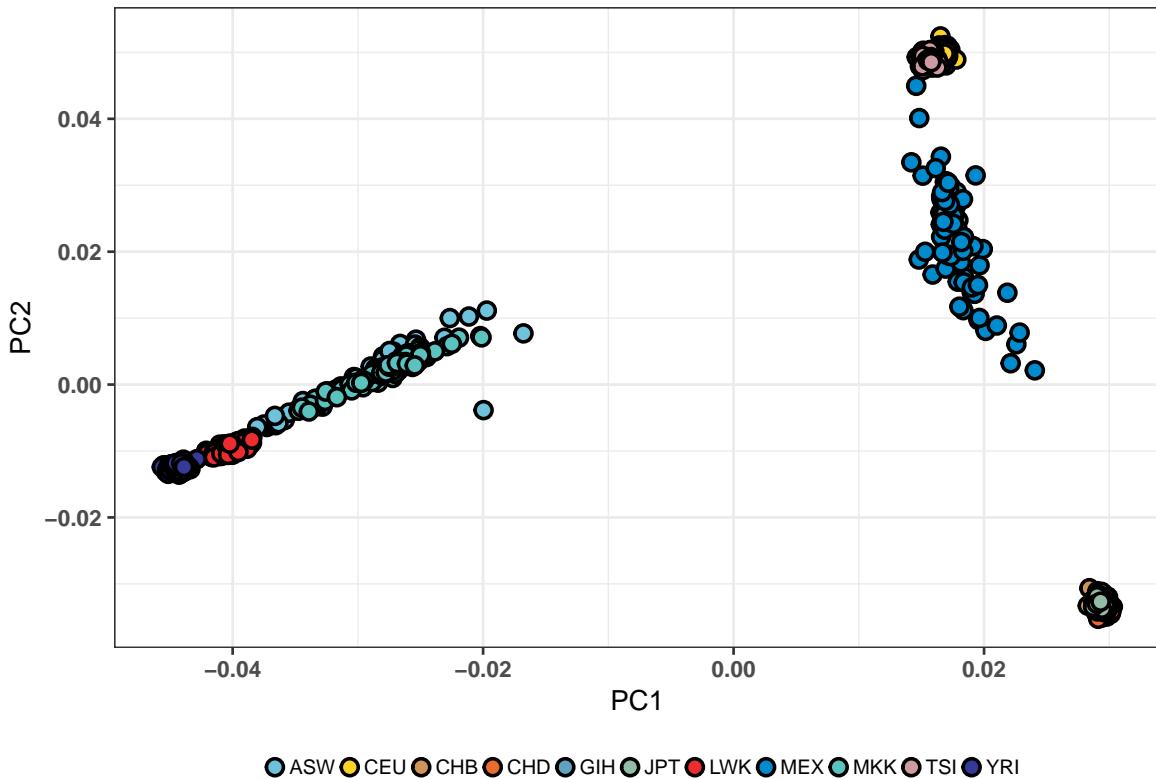


FIGURE 2.4 – ACP réalisée sur la phase 3 du jeu de données HapMap à l'aide de la librarie R pcadapt (Gibbs et al., 2003). Le jeu de données HapMap est un jeu de données humaines incluant une grande diversité de populations. Les deux premières composantes principales distinguent trois groupes génétiques correspondant aux populations africaines, asiatiques et européennes.

Correction pour la structure de populations

En génétique associative, où l'on cherche à détecter les gènes associés à un phénotype en comparant des individus porteurs et non-porteurs du phénotype, les composantes principales servent par exemple à corriger pour la structure de populations pour éviter les associations dues à de la différenciation génétique entre les individus porteurs et non-porteurs du phénotype (Price et al., 2006).

Structure géographique

Novembre et al. (2008) ont montré que ces axes de variation génétique pouvaient également être interprétés en terme d'axes géographiques. L'Analyse en Composantes Principales a été réalisée sur un échantillon d'individus européens³ issus du jeu de données POPRES (Nelson et al., 2008). En figure 2.5, nous observons en effet que la projection des individus sur les deux premiers axes de l'ACP reflète de façon

³. européens dont les grands-parents proviennent de la même région géographique.

particulièrement frappante la disposition géographique des différentes populations.

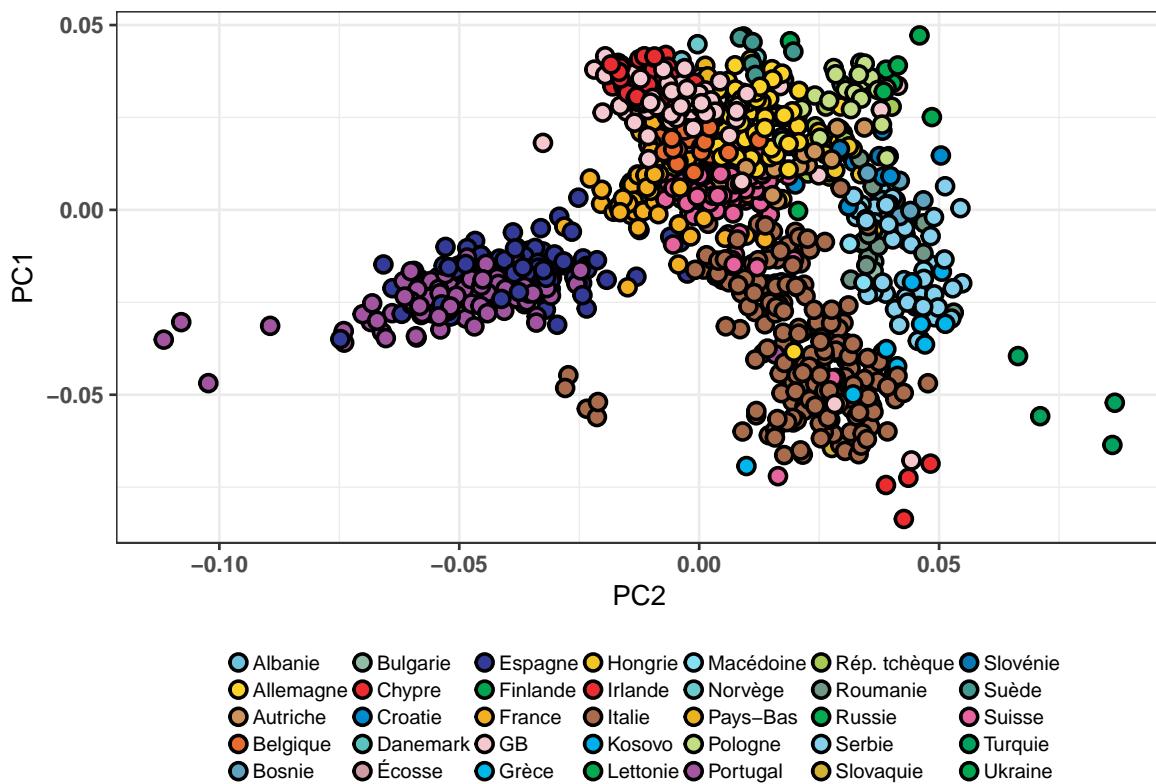


FIGURE 2.5 – ACP réalisée sur le jeu de données POPRES à l'aide de la librairie R pcadapt (Novembre et al., 2008).

Ascendance génétique

Une autre particularité de l'ACP réside dans la possibilité d'inférer les *coefficients de métissage* ou *coefficients d'ascendance* à partir des composantes principales (Ma & Amos, 2012 ; G. McVean, 2009). Un coefficient de métissage quantifie pour un individu donné la proportion de son génome provenant d'un groupe génétique spécifique (appelé aussi population source ou population ancestrale). L'un des premiers articles à établir un lien entre l'ACP et les coefficients de métissage global fut sur l'interprétation généalogique de l'ACP par G. McVean (2009) (Figure 2.6). Nous reviendrons sur cette notion dans le chapitre consacré à l'introgression.

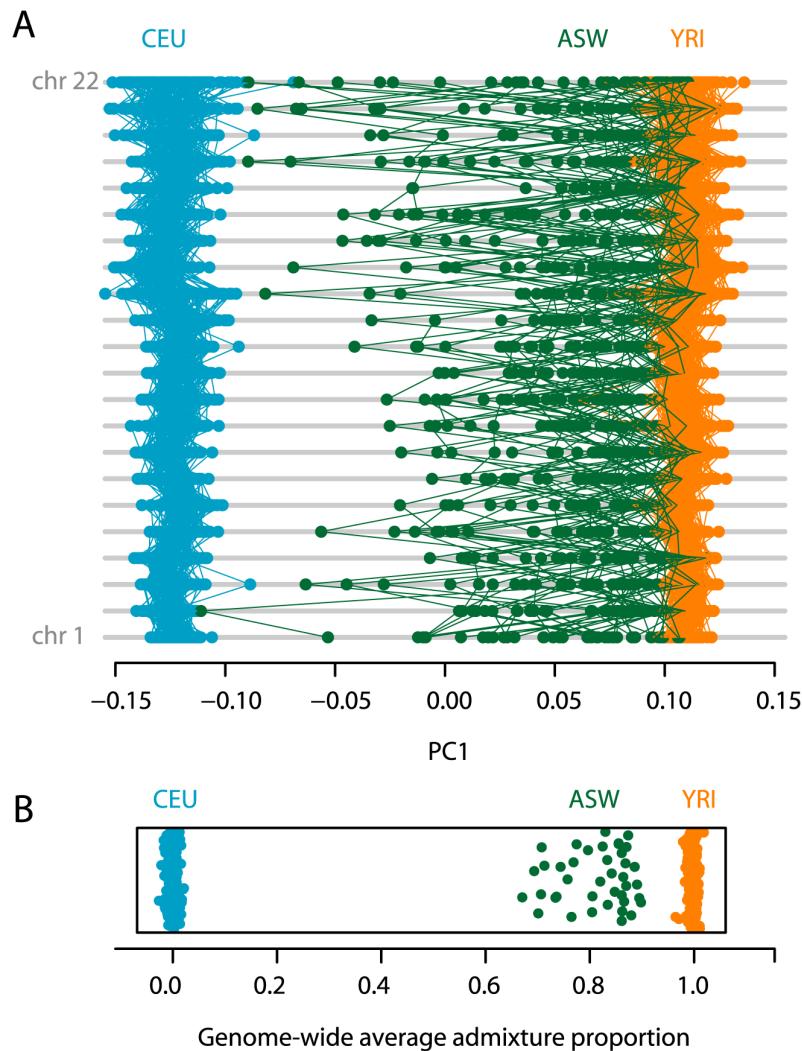


FIGURE 2.6 – Coefficients de métissage et ACP (G. McVean, 2009). **A.** Chaque ligne de la figure correspond à un chromosome. Chacune de ces lignes représente la projection des individus issus des populations CEU, ASW et YRI sur la première composante principale. **B.** Chaque point représente un individu correspondant à la moyenne des projections précédentes, ramenées à l'intervalle [0, 1] à l'aide d'une transformation affine.

Malgré l'importance et la popularité de l'Analyse Composantes Principales en génétique des populations, ce n'est que récemment que son utilisation a été étendue aux scans génomiques (Duforet-Frebbourg et al., 2015; Galinsky et al., 2016). Dans la partie qui suit, nous proposons de nouvelles statistiques basées sur l'Analyse en Composantes Principales et démontrons en quoi elles généralisent les tests classiques de différenciation.

2.3 Statistiques de test basées sur l'Analyse en Composantes Principales

Nous présentons dans cette partie des statistiques basées sur l'Analyse en Composantes Principales dans le cadre des scans génomiques. Les raisons pour lesquelles nous nous sommes intéressés à l'ACP sont multiples. Tout d'abord, l'ACP est particulièrement adaptée au traitement de données en grande dimension, justement parce qu'elle permet de les résumer à l'aide d'un nombre réduit de variables. De plus, comme expliqué précédemment, l'ACP permet de retrouver la structure génétique de façon non paramétrique et sans prior sur l'appartenance individuelle. Grâce à cette propriété, nous montrons la possibilité d'étendre la F_{ST} au cas de populations structurées sans information populationnelle a priori. L'idée de notre démarche repose sur l'hypothèse que les axes principaux reflètent la structure génétique et que les marqueurs génétiques les plus corrélés à ces axes sont des candidats crédibles pour l'adaptation locale. Cette partie sera dédiée à la présentation des méthodes statistiques développées à partir de cette hypothèse de travail. Leur présentation sera accompagnée de validations numériques conduites sur des simulations ainsi que de justifications théoriques quant aux similarités qu'elles présentent avec les méthodes classiques de scan génomique.

Dans l'article 1 (Duforet-Frebbourg et al., 2015), nous introduisons la communalité en tant que statistique de test pour la détection de signaux d'adaptation locale. La communalité est une notion empruntée à l'analyse factorielle et s'interprète comme la proportion de variance expliquée par le modèle à facteurs. Nous justifierons également d'un point de vue théorique les observations établissant la correspondance entre l'indice de fixation et la communalité. Pour ce faire, nous montrerons que la F_{ST} peut se réécrire sous la forme d'une statistique de communalité pour un modèle à facteurs discrets, nous invitant de ce fait à considérer la communalité comme une extension de la F_{ST} au cas continu. Cette généralisation est particulièrement intéressante lorsqu'il est difficile de définir des populations de façon claire comme cela peut être le cas en présence d'individus métissés. Cependant, de la même manière que la F_{ST} , la communalité n'est pas adaptée au cas de populations structurées.

L'article 2 présente une nouvelle statistique de test basée sur la distance robuste de Mahalanobis pour pallier au problème de la structure de populations (Luu et al., 2017). Enfin nous établissons le lien entre cette nouvelle statistique et le test de Lewontin-Krakauer corrigé pour l'apparentement génétique (2.5), ce qui permettra de conclure quant à la généralisation de la statistique de test T_{F-LK} par cette nouvelle statistique.

Ce travail a notamment abouti sur le développement d'une librairie R implémentant ces statistiques, appelée pcadapt (Duforet-Frebbourg et al., 2015; Luu et al., 2017). L'aspect computationnel de ces méthodes sera cependant traité dans le chapitre correspondant, et permettra notamment de discuter des problématiques liées à la présence de données manquantes ainsi que de la complexité algorithmique.

2.3.1 La communalité

Nous présentons dans ce paragraphe un résumé des travaux relatifs à l'article 1 (Duforet-Frebourg et al., 2015). Dans cet article, nous y abordons la possibilité de réaliser des scans génomiques pour la sélection en utilisant l'Analyse en Composantes Principales (ACP). Nous expliquons comment l'indice de différenciation génétique, communément appelé F_{ST} , peut être vu comme la proportion de variance expliquée par les composantes principales. La corrélation entre les variants génétiques et les composantes principales donne un cadre conceptuel permettant la détection de variants génétiques impliqués dans les processus d'adaptation locale sans qu'il n'y ait besoin de définir de populations *a priori*.

Définitions

La première approche présentée repose sur la corrélation des locus avec chaque axe principal :

Définition 2.3 (Corrélation à un axe principal). Soit $G \in \mathcal{M}_{np}(\{0, 1, 2\})$, où n désigne le nombre d'individus et p le nombre de locus. Soit $U\Sigma V^T$ la décomposition en valeurs singulières de rang K de \tilde{G} . Notant $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_K}$ les éléments diagonaux de Σ , la corrélation ρ_{jk} du locus $j \in [|1, p|]$ avec le k -ième axe principal est donnée par la formule ci-dessous (Cadima & Jolliffe, 1995) :

$$\rho_{jk} = \frac{\sqrt{\lambda_k} V_{jk}}{\sqrt{n-1}}. \quad (2.8)$$

La seconde approche statistique présentée est une approche multivariée et propose de considérer la somme quadratique de ces corrélations ainsi que la somme quadratique des loadings. L'intérêt de cette approche par rapport à la précédente est de considérer les locus comme des variables multidimensionnelles, généralisant ainsi l'approche composante par composante.

Définition 2.4 (Communalité). Reprenant les notations de la définition 2.3, la communalité h^2 est définie pour chaque locus j par la formule ci-dessous :

$$\begin{aligned} h_j^2 &= \sum_{k=1}^K \rho_{jk}^2 \\ &= \frac{1}{n-1} \sum_{k=1}^K \sqrt{\lambda_k} V_{jk} \\ &\simeq \|\tilde{G}_{:,j}^T U\|_2^2. \end{aligned} \quad (2.9)$$

Le choix de cette statistique est fondé sur l'interprétation usuelle de la communalité en tant que proportion de variance expliquée par les K premières composantes principales.

Résultats

La F_{ST} peut être interprétée comme une proportion de variance expliquée. Pour chacun des scénarios démographiques étudiés (modèle en île et modèle de divergence à 3 populations), nous avons montré numériquement que la F_{ST} et la communalité sont corrélées à plus de 90% (Duforet-Frebourg et al., 2015). Dans la suite de ce manuscript, nous explicitons un modèle à facteurs où la F_{ST} correspond à la proportion de variance expliquée, ce qui permet de réécrire la F_{ST} sous une forme analytique analogue à celle de la communalité h^2 .

Proposition 2.1. Soient N le nombre de populations, n_j le nombre d'individus appartenant à la j -ème population, n le nombre total d'individus et δ_{ji} le symbole de Kronecker tel que $\delta_{ji} = 1$ si l'individu i appartient à la j -ème population et 0 sinon. Notant $U_\delta = \left(\frac{\delta_{ji}}{\sqrt{2(N-1)n_j}} \right)_{1 \leq i \leq n, 1 \leq j \leq N} \in \mathcal{M}_{nN}(\mathbb{R})$, il existe une matrice $L \in \mathcal{M}_{Np}(\mathbb{R})$ telle que :

$$\tilde{G} = U_\delta L.$$

Démonstration. En constatant que $U_\delta^T U_\delta = \text{Diag}\left(\frac{1}{2(N-1)n_1}, \dots, \frac{1}{2(N-1)n_N}\right)$, nous pouvons définir $L = (U_\delta^T U_\delta)^{-1} U_\delta^T \tilde{G}$, si bien que $\tilde{G} = U_\delta L$. \square

En considérant cette factorisation matricielle $\tilde{G} = U_\delta L$ comme un modèle à facteurs et en reprenant la définition 2.4, la communalité pour ce modèle s'écrit $\|\tilde{G}_{\cdot,j}^T U_\delta\|_2^2$.

Proposition 2.2.

$$F_{ST} = \|\tilde{G}_{\cdot,j}^T U_\delta\|_2^2.$$

Démonstration. En reprenant les notations de la proposition 2.1 et en observant que $p_k = \frac{\sum_{i=1}^n \delta_{ki} G_{ij}}{\sum_{i=1}^n 2\delta_{ki}}$,

$$\begin{aligned} F_{ST} &= \frac{1}{N-1} \frac{\sum_{k=1}^N (p_k - \bar{p})^2}{\bar{p}(1-\bar{p})} \\ &= \frac{1}{(N-1)\bar{p}(1-\bar{p})} \sum_{k=1}^N \left(\frac{\sum_{i=1}^n \delta_{ki} G_{ij}}{\sum_{i=1}^n 2\delta_{ki}} - \bar{p} \right)^2 \\ &= \frac{1}{N-1} \sum_{k=1}^N \left(\frac{\sum_{i=1}^n \delta_{ki} (G_{ij} - 2\bar{p})}{2n_k \sqrt{\bar{p}(1-\bar{p})}} \right)^2 \end{aligned} \quad (2.10)$$

Or $\tilde{G}_{ij} = \frac{G_{ij} - 2\bar{p}}{\sqrt{2\bar{p}(1-\bar{p})}}$:

$$\begin{aligned} F_{ST} &= \sum_{k=1}^N \left(\sum_{i=1}^n \frac{\delta_{ki}}{\sqrt{2(N-1)n_k}} \tilde{G}_{ij} \right)^2 \\ &= \|\tilde{G}_{\cdot,j}^T U_\delta\|_2^2. \end{aligned} \quad (2.11)$$

\square

La proposition 2.2 permet d'identifier la F_{ST} comme une statistique de communalité dans le cas particulier où les facteurs U_δ sont discrets. La statistique h^2 utilise les scores continus de l'ACP en guise de facteurs, ce qui suggère l'utilisation de la communalité en tant que généralisation de la F_{ST} . En réalité, ceci ne nous permet pas tout à fait de conclure puisque les matrices U et U_δ peuvent en principe être bien différentes. Dans le cas de deux populations, nous montrons en annexe que la matrice U peut être assimilée à U_δ , ce qui permet d'établir le lien entre la F_{ST} et la communalité.

L'approche basée sur l'ACP permet de retrouver des signaux d'adaptation connus. Les statistiques de corrélation et de communalité ont été calculées sur deux jeux de données humaines présentant une structure de populations discrète (1000 Genomes) et une structure de populations continue (POPRES). Les approches basées sur l'ACP permettent de s'affranchir des considérations liées au caractère continu ou discret de la structure de populations, et par la même occasion de la difficulté de définir des populations. L'analyse de ces jeux de données a permis la détection de différentes régions du génome connues pour être impliquées dans des processus d'adaptation locale (Figure 2.7), achevant de valider l'utilisation de l'ACP en tant qu'outil statistique pour les scans à sélection.

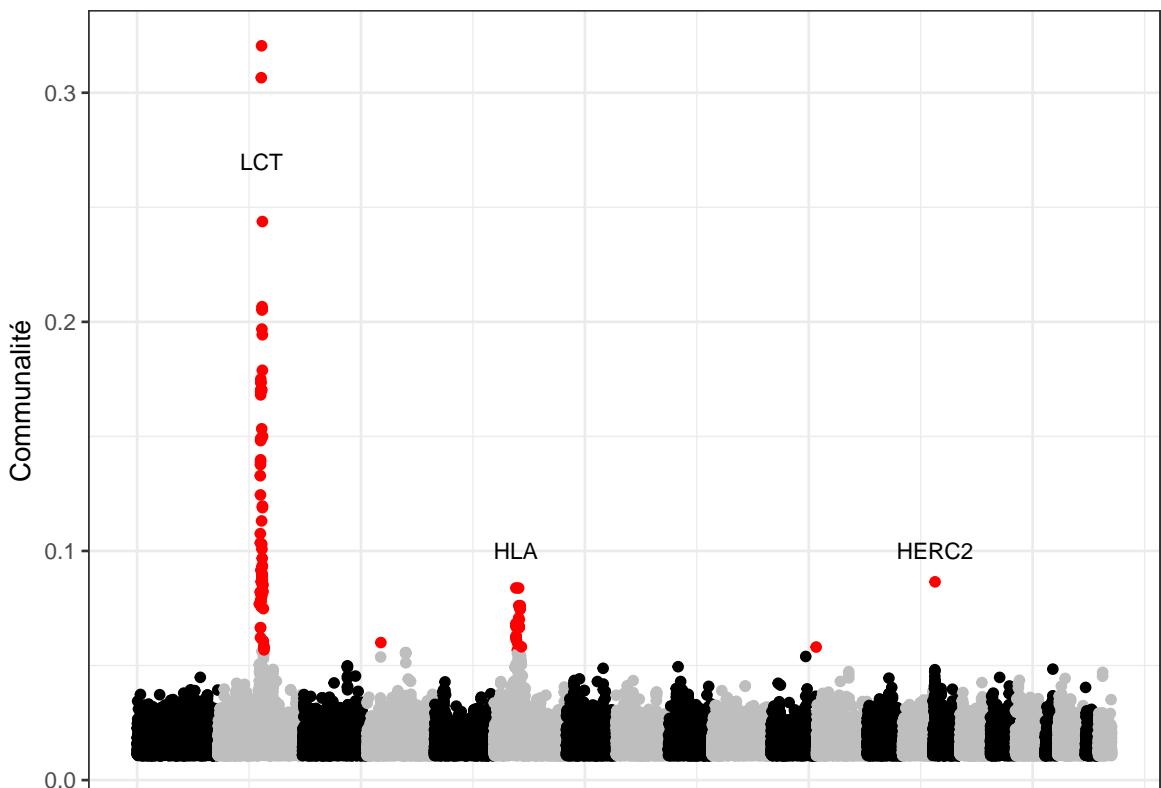


FIGURE 2.7 – Analyse du jeu données POPRES réalisée avec pcadapt en utilisant la statistique de la communalité. Les points rouges correspondent aux 100 locus ayant les valeurs de communalité les plus élevées.

Limites

De par sa définition, la communalité est dépendante de la proportion de variance expliquée par chaque composante principale retenue et donc de l'amplitude relative des valeurs singulières $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_K}$ de \tilde{G} . Dans des cas de figures où $\lambda_1 \gg \lambda_2$, h^2 est équivalente à ρ_1^2 , excluant la possibilité de détecter d'éventuels signaux d'adaptation sur les composantes suivantes. Or ceci s'explique très bien en exploitant le caractère géométrique des statistiques telles que la communalité qui sont distribuées selon un χ^2 . Puisqu'il s'agit de sommes quadratiques de lois normales, les courbes de niveau décrites par ces statistiques sont des ellipsoïdes. Les courbes de niveau peuvent être vues comme le pendant géométrique des seuils de p -valeur. Si l'on s'intéresse aux courbes de niveau de la communalité, nous nous apercevons en effet qu'elle aura tendance à favoriser la détection de SNPs corrélés avec la première composante, malgré la présence éventuelle de SNPs fortement corrélés avec les composantes suivantes. Pour l'illustrer, plaçons-nous dans le cas $K = 2$ (supposant que les deux premières composantes principales aient été retenues). Notant X (resp. Y) la variable aléatoire associée aux loadings de la première (resp. seconde) composante principale. La communalité s'écrit $h^2 = \lambda_1 X^2 + \lambda_2 Y^2 = \left(\frac{X}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{Y}{\sqrt{\lambda_2}}\right)^2$ avec $\lambda_1 > \lambda_2$. En dimension 2, les lignes de niveaux de h^2 décrivent des ellipses dont le petit axe est orienté selon X (et paramétré par $\frac{1}{\sqrt{\lambda_1}}$) (Figure 2.8), justifiant ainsi qu'il est plus facile de se trouver en-dehors de l'ellipse en ayant une forte valeur de X qu'une forte valeur de Y .

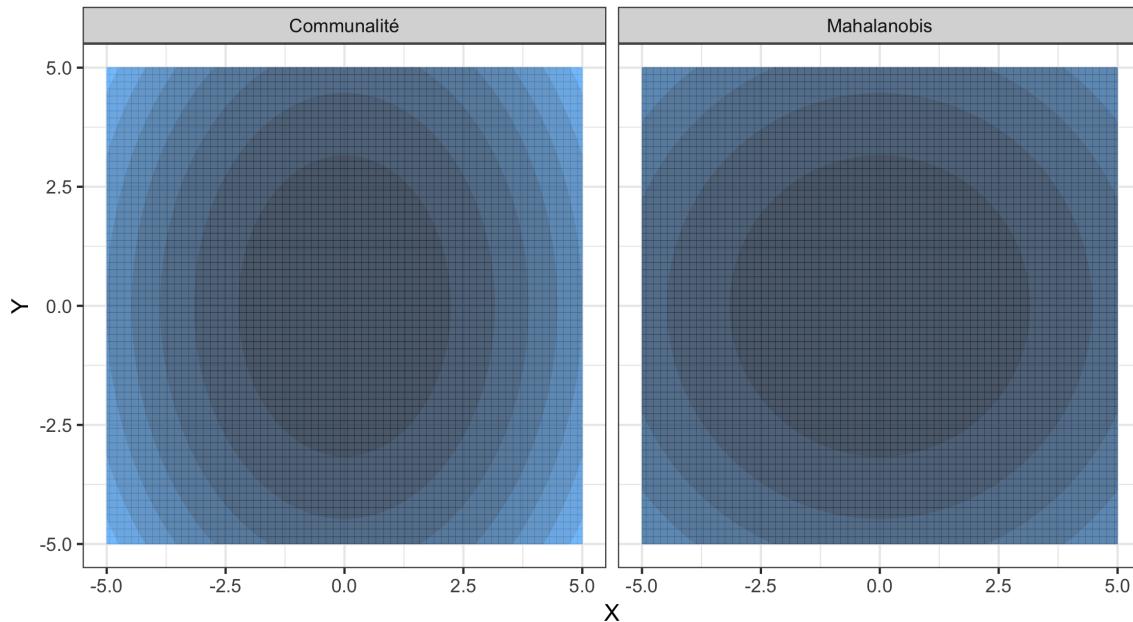


FIGURE 2.8 – À gauche les lignes de niveau de la communalité avec $\lambda_1 = 2\lambda_2$ où l'axe des abscisses et l'axe des ordonnées correspondent aux loadings de chaque composante principale. À droite les lignes de niveau de la distance de Mahalanobis.

L'interprétation de la F_{ST} en termes de proportion de variance expliquée fait de

la communalité une statistique de test intéressante d'un point de vue conceptuel, permettant de considérer la communalité comme une extension de la F_{ST} . Cependant, elle garde de la F_{ST} les défauts liés à la non-prise en compte de l'apparentement génétique interpopulationnel (cf. équation (2.3) (Bonhomme et al., 2010)). Afin de remédier à ce problème, nous proposons une statistique de test basée sur la distance robuste de Mahalanobis.

2.3.2 La distance robuste de Mahalanobis

Ce paragraphe résume de façon détaillée les résultats obtenus dans l'article 2, assortis de justifications qui ne figurent pas nécessairement dans l'article.

Définition

Nous choisissons d'utiliser les coefficients de régression standardisés plutôt que les loadings pour calculer la distance de Mahalanobis. Ce choix sera discuté dans la suite de ce paragraphe.

Définition 2.5 (Coefficients de régression standardisés). Soit $U\Sigma V^T$ la décomposition en valeurs singulières de rang K de \tilde{G} . Notant $\epsilon = \tilde{G} - U\Sigma V^T \in \mathcal{M}_{np}(\mathbb{R})$, les coefficients de régression standardisés z , appelés encore z -scores, sont définis pour chaque locus j par la relation ci-dessous (Saporta, 2006) :

$$z_j = \frac{U^T G_{\cdot,j}}{\sqrt{\frac{\|\epsilon_{\cdot,j}\|_2^2}{n-K-1}}}. \quad (2.12)$$

La distance de Mahalanobis est une statistique classique de détection de valeurs aberrantes pour des données multivariées, ce qui signifie qu'elle permet de déterminer si une observation x provient effectivement d'un ensemble d'observations X . Dans le cas de la communalité par exemple, nous cherchions à détecter les locus significativement isolés de l'ensemble des locus neutres, à partir de la distribution des corrélations. La distance de Mahalanobis apparaît donc comme une alternative naturelle à la communalité pour la détection de locus sous sélection.

Définition 2.6 (Distance de Mahalanobis). Soit $z \in \mathbb{R}^K$, et $Z = (z_1, \dots, z_p) \in \mathcal{M}_{pK}(\mathbb{R})$ une matrice représentant un ensemble de p z -scores K -dimensionnels. La distance de Mahalanobis de z relativement à cet ensemble est donnée par la formule ci-dessous :

$$D^2(z) = (z - \bar{z})^T \Sigma^{-1} (z - \bar{z}), \quad (2.13)$$

où \bar{z} (resp. Σ) désigne la moyenne des z -scores (resp. la matrice de covariance).

Si les z -scores z_i sont distribués selon une loi normale multidimensionnelle de moyenne \bar{z} et de matrice de covariance Σ définie positive, alors $D^2 \sim \chi_K^2$. En pratique, nous utilisons le facteur d'inflation génomique $\lambda = \text{median}(D^2)/\text{median}(\chi_K^2)$ et approchons D^2/λ par un χ^2 à K degrés de liberté (François, Martins, Caye, & Schoville, 2016).

Contrairement aux loadings, les coefficients de régression standardisés prennent en compte la variance résiduelle. Une comparaison de leur utilisation avec la distance de Mahalanobis est présentée dans la suite de ce paragraphe.

Estimation robuste de la matrice de covariance

La statistique de test T_{F-LK} de Bonhomme et al. (2010) et la F_{ST} généralisée proposée par Ochoa & Storey (2016) sont basées sur une estimation de la matrice d'apparentement génétique \mathcal{F} par une *méthode des moments*. Pour des données gaussiennes multivariées, cela consiste à estimer la moyenne et la matrice de covariance. Les estimateurs classiques de moyenne sont connus pour être sensibles à la présence de données aberrantes (Figure 2.9). Un estimateur est dit *robuste* s'il n'est pas ou s'il est peu affecté par la présence de données aberrantes. La distance de Mahalanobis peut être rendue robuste si l'estimation de la moyenne \bar{x} et de la matrice de covariance Σ se fait à l'aide d'estimateurs robustes. Nous montrons ici la nécessité d'utiliser un estimateur robuste pour la covariance et justifions notre choix d'estimateur. Nous proposons une comparaison géométrique (Figure 2.9) de deux estimateurs robustes de la matrice de covariance : MCD (Covariance de Déterminant Minimal (Rousseeuw, 1985)) et OGK (Gnanadesikan-Kettenring Orthogonalisé (Maronna & Zamar, 2002)). La procédure de comparaison est la suivante :

- les matrices de covariance sont estimées pour les méthodes OGK et MCD sur deux jeux de données simulées (96,7% de locus neutres pour le modèle de divergence contre 92,8% de locus neutres pour le modèle en îles). L'estimateur de covariance classique est également inclus dans la comparaison.
- les matrices de covariance ainsi estimées sont ensuite représentées par des ellipses relativement à un niveau de confiance fixé ici à 95%, signifiant que les ellipses de covariance (Figure 2.9) sont censées recouvrir 95% des observations neutres.
- nous regardons ensuite quelle ellipse contient le moins d'observations aberrantes, tout en couvrant au moins 95% des locus neutres.

La figure 2.9 montre que l'utilisation d'une méthode d'estimation robuste est nécessaire, même pour des jeux de données présentant moins de 10% de données aberrantes. Quant à la comparaison entre l'estimateur OGK et l'estimateur MCD, les deux méthodes semblent effectivement tenir compte de la présence de données aberrantes pour ajuster le calcul de la matrice de covariance. Nous notons néanmoins que l'estimateur OGK retient moins de données aberrantes tout en recouvrant une proportion de locus neutres visiblement supérieure à 95% au sein de son ellipse de confiance à 95%. Notre choix d'estimateur s'est donc porté sur l'estimateur OGK et a été réimplémenté dans la librairie pcadapt.

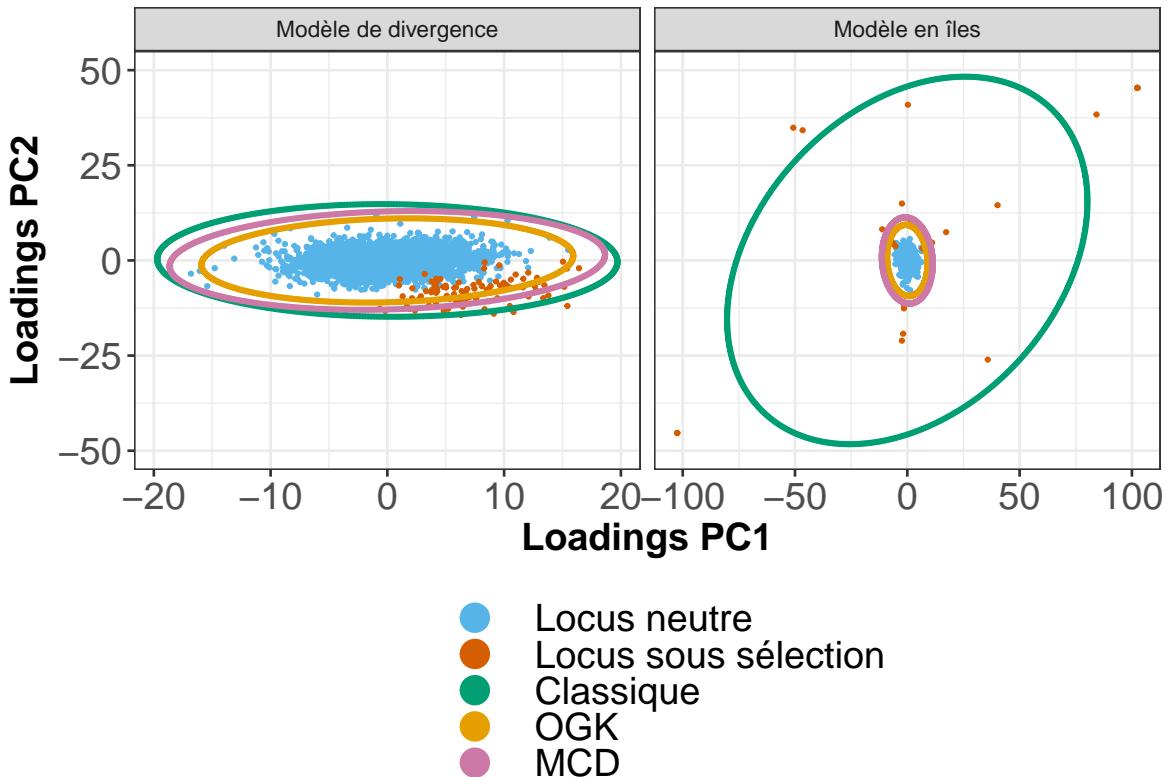


FIGURE 2.9 – Comparaison des estimations de la matrice de covariance. Les matrices de covariance peuvent être interprétées géométriquement en termes d’ellipses. Une ellipse de confiance au seuil α est censée contenir une proportion α de l’ensemble des observations effectivement issues de la distribution (correspondant aux observations neutres). Nous constatons que l’estimateur classique de covariance surestime les valeurs propres de la matrice de covariance même lorsque la proportion de données aberrantes est faible.

Loadings et coefficients de régression standardisés

Nous justifions ici de façon empirique l’utilisation des coefficients de régression standardisés (z -scores) au détriment des loadings pour calculer la distance de Mahalanobis. La procédure de comparaison est la suivante :

- la distance de Mahalanobis et les p -valeurs associées sont calculées à partir des loadings et des coefficients de régression standardisés.
- pour chacun des seuils de significativité (ici 5%, 10% et 20%), nous déterminons l’ensemble des SNPs présentant une p -valeur ajustée⁴ inférieure à ce seuil et calculons le taux de fausses découvertes et la puissance relativement à cet ensemble.

Cette procédure de comparaison nous permet par exemple d’évaluer si le taux de fausses découvertes est bien contrôlé ou encore si une méthode est trop conservative

4. nous utilisons en réalité la q -valeur associée.

ou non. En figure 2.10, nous constatons que les deux statistiques sont bien contrôlées, c'est-à-dire que pour chacun des seuils de significativité, aucune des deux statistiques ne présentent un taux de fausses découvertes supérieur à ce seuil. La figure 2.10 met en évidence le côté relativement conservatif de la distance de Mahalanobis calculée à partir des loadings, ce qui a pour effet de diminuer la puissance du test.

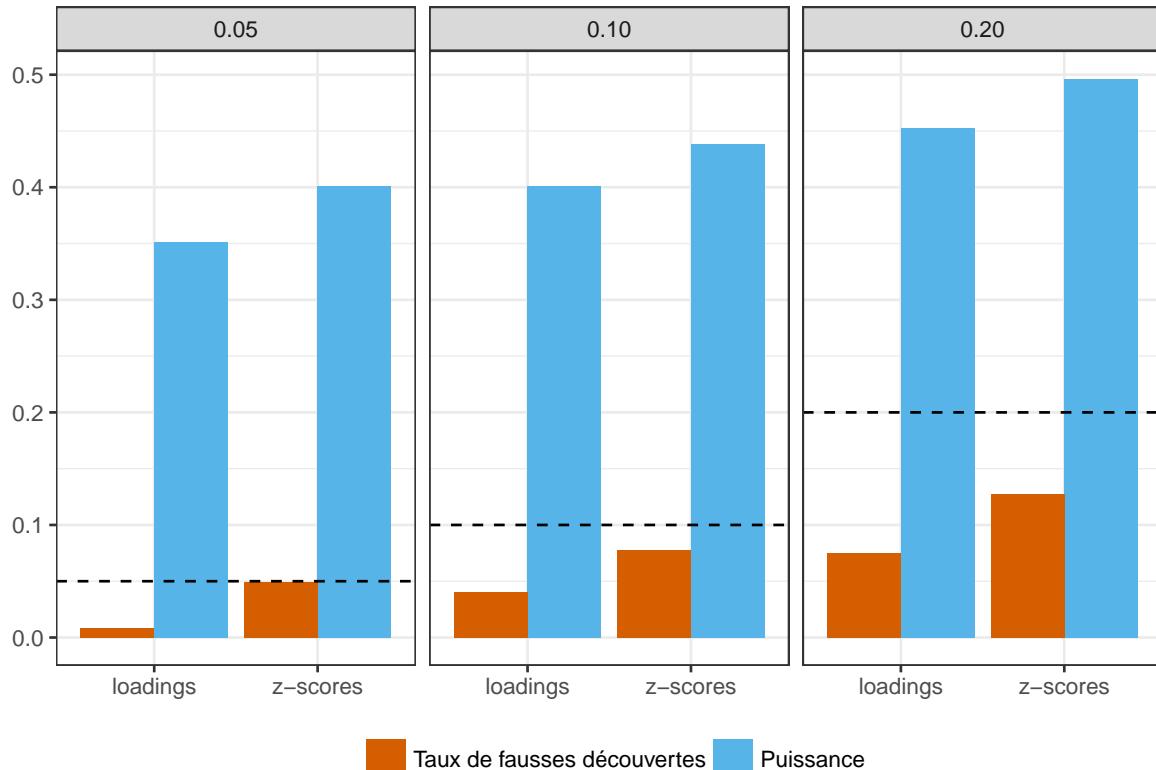


FIGURE 2.10 – Comparaison des distances de Mahalanobis calculées à partir des loadings et des z-scores. Les taux de fausses de découvertes et les puissances sont calculées puis moyennées sur l'ensemble des simulations de modèles en îles utilisées dans l'article 2. La ligne en pointillé représente le taux de fausses découvertes attendu pour chaque seuil de significativité.

Rappel des résultats principaux de l'article 2

Nous avons développé une méthode de scan à sélection valable pour différentes structures de populations, aussi bien adaptée au cas de populations discrètes qu'au cas de populations continues. À l'aide de simulations reproduisant différents scénarios démographiques, exhibant des structures de populations discrètes et continues, nous montrons que l'utilisation de la distance robuste de Mahalanobis permet de pallier aux défauts de la communalité. En termes de sensibilité statistique et de contrôle du taux de fausses découvertes, notre méthode affiche de meilleurs résultats en comparaison à d'autres méthodes de scan à sélection (OutFLANK, Bayescan, FLK), quelque soit le modèle démographique sous-jacent, à

l’exception du modèle en îles où les performances sont équivalentes. Nous étudions également l’impact du caractère discret ou continu de la structure de populations sur les méthodes de scan à sélection et montrons que notre méthode est moins sensible à la présence d’individus métissés, contrairement aux méthodes basées sur la F_{ST} .

Une généralisation de T_{F-LK} . La similarité des résultats numériques produits par pcadapt et FLK suggèrent l’existence d’un lien entre les deux méthodes et nous justifions ce résultat en annexe. La distance robuste de Mahalanobis calculée à partir de l’ACP généralise la statistique de test T_{F-LK} car elle peut être utilisée dans le cas de populations continues.

2.4 Article 1

Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data

Nicolas Duforet-Frebbourg,^{1,2,3} Keurcien Luu,^{1,2} Guillaume Laval,^{4,5} Eric Bazin,⁶ and Michael G.B. Blum^{*1,2}

¹TIMC-IMAG UMR 5525, Univ. Grenoble Alpes, Grenoble, France

²CNRS, TIMC-IMAG, Grenoble, France

³Department of Integrative Biology, University of California, Berkeley

⁴Department of Genomes and Genetics, Institut Pasteur, Human Evolutionary Genetics, Paris, France

⁵Centre National De La Recherche Scientifique, URA3012, Paris, France

⁶CNRS, Laboratoire D'écologie Alpine UMR 5553, Univ. Grenoble Alpes, Grenoble, France

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

Associate editor: John Novembre

Abstract

To characterize natural selection, various analytical methods for detecting candidate genomic regions have been developed. We propose to perform genome-wide scans of natural selection using principal component analysis (PCA). We show that the common F_{ST} index of genetic differentiation between populations can be viewed as the proportion of variance explained by the principal components. Considering the correlations between genetic variants and each principal component provides a conceptual framework to detect genetic variants involved in local adaptation without any prior definition of populations. To validate the PCA-based approach, we consider the 1000 Genomes data (phase 1) considering 850 individuals coming from Africa, Asia, and Europe. The number of genetic variants is of the order of 36 millions obtained with a low-coverage sequencing depth (3x). The correlations between genetic variation and each principal component provide well-known targets for positive selection (EDAR, SLC24A5, SLC45A2, DARC), and also new candidate genes (APPBP2, TP1A1, RTTN, KCNMA, MYO5C) and noncoding RNAs. In addition to identifying genes involved in biological adaptation, we identify two biological pathways involved in polygenic adaptation that are related to the innate immune system (beta defensins) and to lipid metabolism (fatty acid omega oxidation). An additional analysis of European data shows that a genome scan based on PCA retrieves classical examples of local adaptation even when there are no well-defined populations. PCA-based statistics, implemented in the *PCAdapt* R package and the *PCAdapt fast* open-source software, retrieve well-known signals of human adaptation, which is encouraging for future whole-genome sequencing project, especially when defining populations is difficult.

Significance Statement

Positive natural selection or local adaptation is the driving force behind the adaption of individuals to their environment. To identify genomic regions responsible for local adaptation, we propose to consider the genetic markers that are the most related with population structure. To uncover genetic structure, we consider principal component analysis that identifies the primary axes of variation in the data. Our approach generalizes common approaches for genome scan based on measures of population differentiation. To validate our approach, we consider the human 1000 Genomes data and find well-known targets for positive selection as well as new candidate regions. We also find evidence of polygenic adaptation for two biological pathways related to the innate immune system and to lipid metabolism.

Introduction

Because of the flood of genomic data, the ability to understand the genetic architecture of natural selection has dramatically increased. Of particular interest is the study of local positive selection which explains why individuals are adapted to their local environment. In humans, the availability of genomic data fostered the identification of loci involved in positive selection (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008; Grossman et al., 2013; Pickrell et al., 2009; Sabeti et al., 2007). Local positive selection tends to increase genetic differentiation, which can be measured by difference of allele frequencies between populations (Colonna et al., 2014; Nielsen, 2005; Sabeti et al., 2006). For instance, a mutation in the DARC gene that confers resistance to malaria is fixed in sub-Saharan African populations whereas it is absent elsewhere (Hamblin, Thompson, & Di Rienzo, 2002). In addition to the variants that confer resistance to pathogens, genome scans also identify other genetic variants, and many of these are involved in human metabolic phenotypes and morphological traits (Barreiro et al., 2008; Hancock et al., 2010).

In order to provide a list of variants potentially involved in natural selection, genome scans compute measures of genetic differentiation between populations and consider that extreme values correspond to candidate regions (G. Luikart, England, Tallmon, Jordan, & Taberlet, 2003). The most widely used index of genetic differentiation is the F_{ST} index which measures the amount of genetic variation that is explained by variation between populations (Excoffier, Smouse, & Quattro, 1992). However the F_{ST} statistic requires to group individuals into populations which can be problematic when ascertainment of population structure does not show well-separated clusters of individuals (e.g., Novembre et al. (2008)). Other statistics related to F_{ST} have been derived to reduce the false discovery rate (FDR) obtained with F_{ST} but they also work at the scale of populations (Bonhomme et al., 2010; Fariello, Boitard, Naya, SanCristobal, & Servin, 2013; Günther & Coop, 2013). Grouping individuals into populations can be subjective, and important signals of selection may be missed with an inadequate choice of populations (W.-Y. Yang, Novembre, Eskin, & Halperin, 2012). We have previously developed an individual-based approach for selection scan based

on a Bayesian factor model but the Markov chain Monte Carlo (MCMC) algorithm required for model fitting does not scale well to large data sets containing a million of variants or more (Duforet-Frebourg et al., 2014).

We propose to detect candidates for natural selection using principal component analysis (PCA). PCA is a technique of multivariate analysis used to ascertain population structure (N. Patterson, Price, & Reich, 2006). PCA decomposes the total genetic variation into K axes of genetic variation called principal components. In population genomics, the principal components can correspond to evolutionary processes such as evolutionary divergence between populations (G. McVean, 2009). Using simulations of an island model and of a model of population fission followed by isolation, we show that the common F_{ST} statistic corresponds to the proportion of variation explained by the first K principal components when K has been properly chosen. With this point of view, the F_{ST} of a given variant is obtained by summing the squared correlations of the first K principal components opening the door to new statistics for genome scans. At a genome-wide level, it is known that there is a relationship between F_{ST} and PCA (G. McVean, 2009), and our simulations show that the relationship also applies at the level of a single variant.

The advantages of performing a genome scan based on PCA are multiple : it does not require to group individuals into populations, the computational burden is considerably reduced compared with genome scan approaches based on MCMC algorithms (Foll & Gaggiotti, 2008) ; Riebler, Held, & Stephan (2008) ; Günther & Coop (2013) ; Duforet-Frebourg et al. (2014)], and candidate single nucleotide polymorphisms (SNPs) can be related to different evolutionary events that correspond to the different principal components. Using simulations and the 1000 Genomes data, we show that PCA can provide useful insights for genome scans. Looking at the correlations between SNPs and principal components provides a novel conceptual framework to detect genomic regions that are candidates for local adaptation.

New Method

New Statistics for Genome Scan

We denote by Y the $(n \times p)$ centered and scaled genotype matrix where n is the number of individuals and p is the number of loci. The new statistics for genome scan are based on PCA. The objective of PCA is to find a new set of orthogonal variables called the principal components, which are linear combinations of (centered and standardized) allele counts, such that the projections of the data onto these axes lead to an optimal summary of the data. To present the method, we introduce the truncated singular value decomposition (SVD) that approximates the data matrix Y by a matrix of smaller rank

$$Y \approx U\Sigma V^T, \quad (2.14)$$

where U is a $(n \times K)$ orthonormal matrix, V is a $(p \times K)$ orthonormal matrix, Σ is a diagonal $(K \times K)$ matrix and K corresponds to the rank of the approximation. The solution of PCA with K components can be obtained using the truncated SVD :

the K columns of V contain the coefficients of the new orthogonal variables, the K columns of U contain the projections (called “scores”) of the original variables onto the principal components and capture population structure (supplementary fig. B.1), and the squares of the elements of Σ are proportional to the proportion of variance explained by each principal component (Jolliffe, 1986). We denote the diagonal elements of Σ by $\sqrt{\lambda_k}$, $k = 1, \dots, K$ where the λ_k ’s are the ranked eigenvalues of the matrix YY^T . Denoting by V_{jk} , the entry of V at the j^{th} line and k^{th} column, then the correlation ρ_{jk} between the j^{th} SNP and the k^{th} principal component is given by $\rho_{jk} = \sqrt{\lambda_k} V_{jk} / \sqrt{n-1}$ (Cadima & Jolliffe, 1995). In the following, the statistics ρ_{jk} are referred to as “loadings” and will be used for detecting selection. The second statistic we consider for genome scan corresponds to the proportion of variance of a SNP that is explained by the first K PCs. It is called the communality in exploratory factor analysis because it is the variance of observed variables accounted for by the common factors, which correspond to the first K PCs. Because the principal components are orthogonal to each other, the proportion of variance explained by the first K principal components is equal to the sum of the squared correlations with the first K principal components. Denoting by h_j^2 the communality of the j^{th} SNP, we have

$$h_j^2 = \sum_{k=1}^K \rho_{jk}^2. \quad (2.15)$$

The last statistic we consider for genome scans sums the squared of normalized loadings. It is defined as $h'_j^2 = \sum_{k=1}^K V_{jk}^2$. Compared to the communality h_j^2 , the statistic h'_j^2 should theoretically give the same importance to each PC because the normalized loadings are on the same scale as we have $\sum_{j=1}^K V_{jk}^2 = 1$, for $k = 1, \dots, K$.

Numerical Computations

The method of selection scan should be able to handle a large number p of genetic variants. In order to compute truncated SVD with large values of p , we compute the $n \times n$ covariance matrix $\Omega = YY^T/(p-1)$. The covariance matrix Ω is typically of much smaller dimension than the $p \times p$ covariance matrix. Considering the $n \times n$ covariance matrix Ω speeds up matrix operations. Computation of the covariance matrix is the most costly operation and it requires a number of arithmetic operations proportional to pn^2 . After computing the covariance matrix Ω , we compute its first K eigenvalues and eigenvectors to find $\Sigma^2/(p-1)$ and U . Eigenanalysis is performed with the *dsyevr* routine of the linear algebra package LAPACK (Anderson et al., 1999). The matrix V , which captures the relationship between each SNPs and population structure, is obtained by the matrix operation $V^T = \Sigma^{-1} U^T Y$. The software *PCAdapt fast*, process data as a stream and never store in order to have a very low memory access whatever the size of the data.

Results

Island Model

To investigate the relationship between communality h^2 and F_{ST} , we consider an island model with three islands. We use $K = 2$ when performing PCA because there are three islands. We choose a value of the migration rate that generates a mean F_{ST} value (across the 1,400 neutral SNPs) of 4%. We consider five different simulations with varying strengths of selection for the 100 adaptive SNPs. In all simulations, the R^2 correlation coefficient between h^2 and F_{ST} is larger than 98%. Considering as candidate SNPs the 1% of the SNPs with largest values of F_{ST} or of h^2 , we find that the overlap coefficient between the two sets of SNPs is comprised between 88% and 99%. When varying the strength of selection for adaptive SNPs, we find that the relative difference of FDRs obtained with F_{ST} (top 1%) and with h^2 (top 1%) is smaller than 5%. The similar values of FDR obtained with h^2 and with F_{ST} decrease for increasing strength of selection (supplementary fig. B.2).

Divergence Model

To compare the performance of different PCA-based summary statistics, we simulate genetic variation in models of population divergence. The divergence models assume that there are three populations, A , B_1 and B_2 with B_1 and B_2 being the most related populations (figs. 2.11 and 2.12). The first simulation scheme assumes that local adaptation took place in the lineages corresponding to the environments of populations A and B_1 (fig. 2.11). The SNPs, which are assumed to be independent, are divided into three groups : 9,500 SNPs evolve neutrally, 250 SNPs confer a selective advantage in the environment of A , and 250 other SNPs confer a selective advantage in the environment of B_1 . Genetic differentiation, measured by pairwise F_{ST} , is equal to 14% when comparing population A to the other ones and is equal to 5% when comparing populations B_1 and B_2 . Performing PCA with $K = 2$ shows that the first component separates population A from B_1 and B_2 whereas the second component separates B_1 from B_2 (supplementary fig. B.1). The choice of $K = 2$ is evident when looking at the scree plot because the eigenvalues, which are proportional to the proportion of variance explained by each PC, drop beyond $K = 2$ and stay almost constant as K further increases (supplementary fig. B.3).

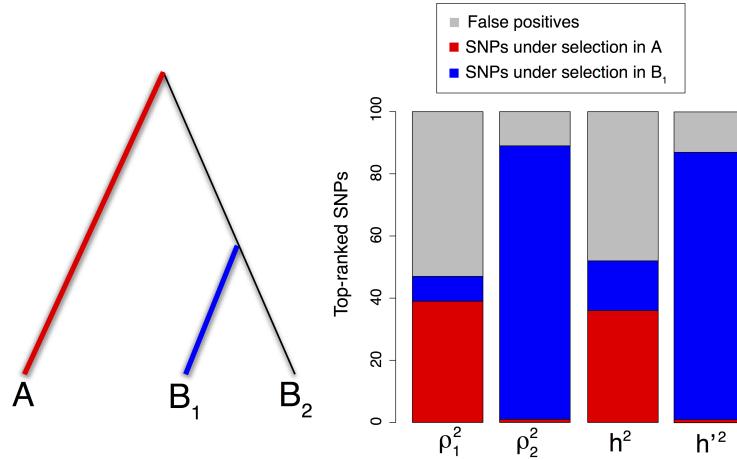


FIGURE 2.11 – Repartition of the 1% top-ranked SNPs for each PCA-based statistic under a divergence model with two types of adaptive constraints. Thicker and colored lineages correspond to lineages where adaptation took place. The squared loadings with PC1 ρ_{j1}^2 pick a large proportion of SNPs involved in selection in population A whereas the squared loadings with PC2 ρ_{j2}^2 pick SNPs involved in selection in population B_1 . This difference is reflected in the different repartition of the top-ranked SNPs for the communality h^2 and the statistic h'^2 .

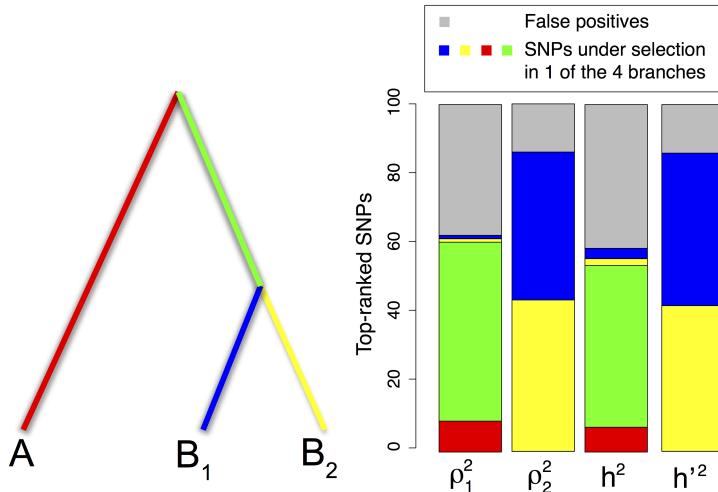


FIGURE 2.12 – Repartition of the 1% top-ranked SNPs of each PCA-based statistic under a divergence model with four types of adaptive constraints. Thicker and colored lineages correspond to lineages where adaptation occurred. The different types of SNPs picked by the squared loadings ρ_{j1}^2 and ρ_{j2}^2 are also found when comparing the communality h^2 and the statistic h'^2 .

We investigate the relationship between the communality statistic h^2 , which measures the proportion of variance explained by the first two PCs, and the F_{ST} statistic. We find a squared Pearson correlation coefficient between the two statistics larger than 98.8% in the simulations corresponding to figures 2.11 and 2.12 (supplementary fig. B.4). For these two simulations, we look at the SNPs in the top 1% (respectively, 5%) of the ranked lists based on h^2 and F_{ST} , and we find an overlap coefficient always larger than 93% for the lists provided by the two different statistics (respectively, 95%). Providing a ranking of the SNPs almost similar to the ranking provided by F_{ST} is therefore possible without considering that individuals originate from predefined populations.

We then compare the performance of the different statistics based on PCA by investigating if the top-ranked SNPs (top 1%) manage to pick SNPs involved in local adaptation (fig. 2.11). The squared loadings ρ_{j1}^2 with the first PC pick SNPs involved in selection in population A (39% of the top 1%), a few SNPs involved in selection in B_1 (9%), and many false positive SNPs (FDR of 53%). The squared loadings with the second PC ρ_{j2}^2 pick less false positives (FDR of 12%) and most SNPs are involved in selection in B_1 (88%) with just a few involved in selection in A (1%). When adaptation took place in two different evolutionary lineages of a divergence tree between populations, a genome scan based on PCA has the nice property that outlier loci correlated with PC1 or with PC2 correspond to adaptive constraints that occurred in different parts of the tree.

Because the communality h^2 gives more importance to the first PC, it picks preferentially the SNPs that are the most correlated with PC1. There is a large overlap of 72% between the 1% top-ranked lists provided by h^2 and ρ_{j1}^2 . Therefore, the communality statistic h^2 is more sensitive to ancient adaptation events that occurred in the environment of population A . In contrast, the alternative statistic h'^2 is more sensitive to recent adaptation events that occurred in the environment of population B_1 . When considering the top-ranked 1% of the SNPs, h'^2 captures only one SNP involved in selection in A (1% of the top 1%) and 88 SNPs related to adaptation in B_1 (88% of the top 1%). The overlap between the 1% top-ranked lists provided by h'^2 and by ρ_{j2}^2 is of 86%.

The h'^2 statistic is mostly influenced by the second principal component because the distribution of squared loadings corresponding to the second PC has a heavier tail, and this result holds for the two divergence models and for the 1000 Genomes data (supplementary fig. B.5). To summarize, the h^2 and h'^2 statistics give too much importance to PC1 and PC2, respectively, and they fail to capture in an equal manner both types of adaptive events occurring in the environment of populations A and B_1 .

We also investigate a more complex simulation in which adaptation occurs in the four branches of the divergence tree (fig. 2.12). Among the 10,000 simulated SNPs, we assume that there are four sets of 125 adaptive SNPs with each set being related to adaptation in one of the four branches of the divergence tree. Compared with the simulation of figure 2.11, we find the same pattern of population structure (supplementary fig. B.1). The squared loadings ρ_{j1}^2 with the first PC mostly pick SNPs involved in selection in the branch that predates the split between B_1 and B_2 (51%

of the top 1%), SNPs involved in selection in the environment of population A (9%), and false positive SNPs (FDR of 38%). Except for false positives (FDR of 14%), the squared loadings ρ_{j2}^2 with the second PC rather pick SNPs involved in selection in B_1 and B_2 (42% for B_1 and 44% for B_2). Once again, there is a large overlap between the SNPs picked by the communality h^2 and by ρ_1^2 (92% of overlap) and between the SNPs picked by h'^2 and ρ_2^2 (93% of overlap). Because the first PC discriminates population A from B_1 and B_2 (supplementary fig. B.1), the SNPs most correlated with PC1 correspond to SNPs related to adaptation in the (red and green) branches that separate A from populations B_1 and B_2 . In contrast, the SNPs that are most correlated to PC2 correspond to SNPs related to adaptation in the two (blue and yellow) branches that separate population B_1 from B_2 (fig. 2.12).

We additionally evaluate to what extent the results are robust with respect to some parameter settings. When considering the 5% of the SNPs with most extreme values of the statistics instead of the top 1%, we also find that the summary statistics pick SNPs related to different evolutionary events (supplementary fig. B.6). The main difference being that the FDR increases considerably when considering the top 5% instead of the top 1% (supplementary fig. B.6). We also consider variation of the selection coefficient ranging from $s = 1.01$ to $s = 1.1$ ($s = 1.025$ corresponds to the simulations of figs. 2.11 and 2.12). As expected, the FDR of the different statistics based on PCA is considerably reduced when the selection coefficient increases (supplementary fig. B.7).

In the divergence model of figure 2.11, we also compare the FDRs obtained with the statistics h^2 , h'^2 , and with a Bayesian factor model implemented in the software *PCAdapt* (Duforet-Frebourg et al., 2014). For the optimal choice of $K = 2$, the statistic h'^2 and the Bayesian factor model provide the smallest FDR (supplementary fig. B.8). However, when varying the value of K from $K = 1$ to $K = 6$, we find that the communality h^2 and the Bayesian approach are robust to overspecification of K ($K > 3$) whereas the FDR obtained with h'^2 increases importantly as K increases beyond $K = 2$ (supplementary fig. B.8).

We also consider a more general isolation-with-migration model. In the divergence model where adaptation occurs in two different lineages of the population tree (fig. 2.11), we add constant migration between all pairs of populations. We assume that migration occurred after the split between B_1 and B_2 . We consider different values of migration rates generating a mean F_{ST} of 7.5% for the smallest migration rate to a mean F_{ST} of 0% for the largest migration rate. We find that the R^2 correlation between F_{ST} and h^2 decreases as a function of the migration rate (supplementary fig. B.9). For F_{ST} values larger than 0.5%, R^2 is larger than 97%. The squared correlation R^2 decreases to 47% for the largest migration rate. Beyond a certain level of migration rate, population structure, as ascertained by principal components, is no more described by well-separated clusters of individuals (supplementary fig. B.10) but by a more clinal or continuous pattern (supplementary fig. B.10) explaining the difference between F_{ST} and h^2 . However, the FDRs obtained with the different statistics based on PCA and with F_{ST} evolve similarly as a function of the migration rate. For both types of approaches, the FDR increases for larger migration with almost no true discovery (only one true discovery in the top 1% lists) when considering the largest migration rate.

The main results obtained under the divergence models can be described as follows. The principal components correspond to different evolutionary lineages of the divergence tree. The communality statistic h^2 provides similar list of candidate SNPs than F_{ST} and it is mostly influenced by the first principal component which can be problematic if other PCs also convey adaptive events. To counteract this limitation, which can potentially lead to the loss of important signals of selection, we show that looking at the squared loadings with each of the principal components provide adaptive SNPs that are related to different evolutionary events. When adding migration rates between lineages, we find that the main results are unchanged up to a certain level of migration rate. Above this level of migration rate, the relationship between F_{ST} and h^2 does not hold anymore and genome scans based on either PCA or F_{ST} produce a majority of false positives.

1000 Genomes Data

Since we are interested in selective pressures that occurred during the human diaspora out of Africa, we decide to exclude individuals whose genetic makeup is the result of recent admixture events (African Americans, Columbians, Puerto Ricans, and Mexicans). The first three principal components capture population structure whereas the following components separate individuals within populations (fig. 2.13 and supplementary fig. B.11). The first and second PCs ascertain population structure between Africa, Asia, and Europe (fig. 2.13) and the third principal component separates the Yoruba from the Luhya population (supplementary fig. B.11). The decay of eigenvalues suggests to use $K = 2$ because the eigenvalues drop between $K = 2$ and $K = 3$ where a plateau of eigenvalues is reached (supplementary fig. B.3).

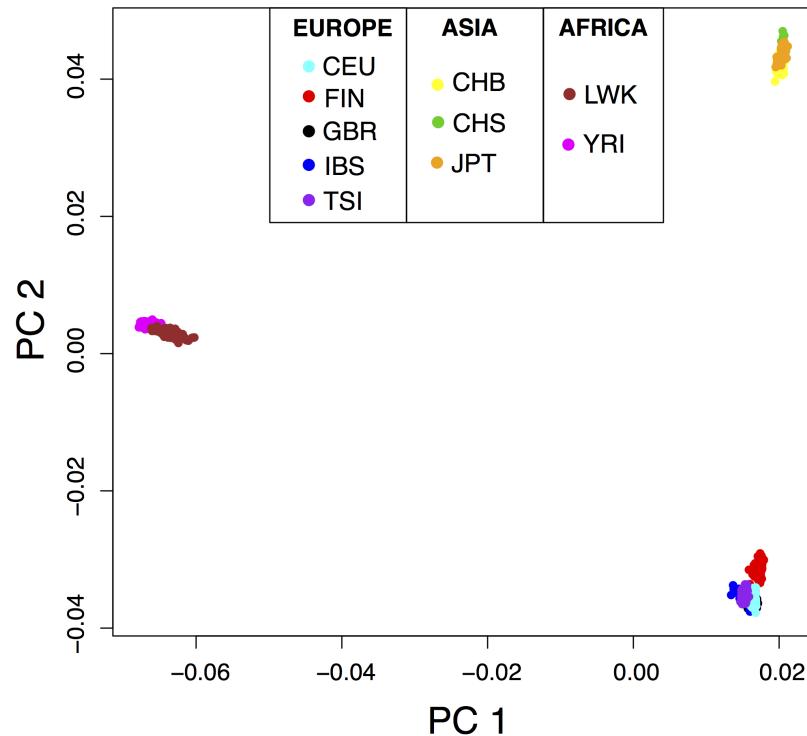


FIGURE 2.13 – PCA with $K = 2$ applied to the 1000 Genomes data. The sampled populations are the following : British in England and Scotland (GBR), Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), Iberian populations in Spain (IBS), Toscani in Italy (TSI), Han Chinese in Bejing (CHB), Southern Han Chinese (CHS), Japanese in Tokyo (JPT), Luhya in Kenya (LWK), Yoruba in Nigeria (YRI).

When performing a genome scan with PCA, there are different choices of statistics. The first choice is the h^2 communality statistic. Using the three continents as labels,

there is a squared correlation between h^2 and F_{ST} of $R^2 = 0.989$. To investigate if h^2 is mostly influenced by the first PC, we determine if the outliers for the h^2 statistics are related with PC1 or with PC2. Among the top 0.1% of SNPs with the largest values of h^2 , we find that 74% are in the top 0.1% of the squared loadings ρ_{j1}^2 corresponding to PC1 and 20% are in the top 0.1% of the squared loadings ρ_{j2}^2 corresponding to PC2. The second possible choice of summary statistics is the h'^2 statistic. Investigating the repartition of the 0.1% outliers for h' , we find that 0.005% are in the top 0.1% of the squared loadings ρ_{j1}^2 corresponding to PC1 and 85% are in the top 0.1% of the squared loadings ρ_{j2}^2 corresponding to PC2. The h'^2 statistic is mostly influenced by the second PC because the distribution of the V_{j2}^2 (normalized squared loadings) has a longer tail than the corresponding distribution for PC1 (supplementary fig. B.5). Because the h^2 statistic is mostly influenced by PC1 and h'^2 is mostly influenced by PC2, confirming the results obtained under the divergence models, we rather decide to perform two separate genome scans based on the squared loadings ρ_{j1}^2 and ρ_{j2}^2 .

The two Manhattan plots based on the squared loadings for PC1 and PC2 are displayed in figures 2.14 and 2.15. Because of linkage disequilibrium (LD), Manhattan plots generally produce clustered outliers. To investigate if the top 0.1% outliers are clustered in the genome, we count—for various window sizes—the proportion of contiguous windows containing at least one outlier. We find that outlier SNPs correlated with PC1 or with PC2 are more clustered than expected if they would have been uniformly distributed among the 36,536,154 variants (supplementary fig. B.12). Additionally, the clustering is larger for the outliers related to the second PC as they cluster in fewer windows (supplementary fig. B.12). As the genome scan for PC2 captures more recent adaptive events, it reveals larger genomic windows that experienced fewer recombination events.

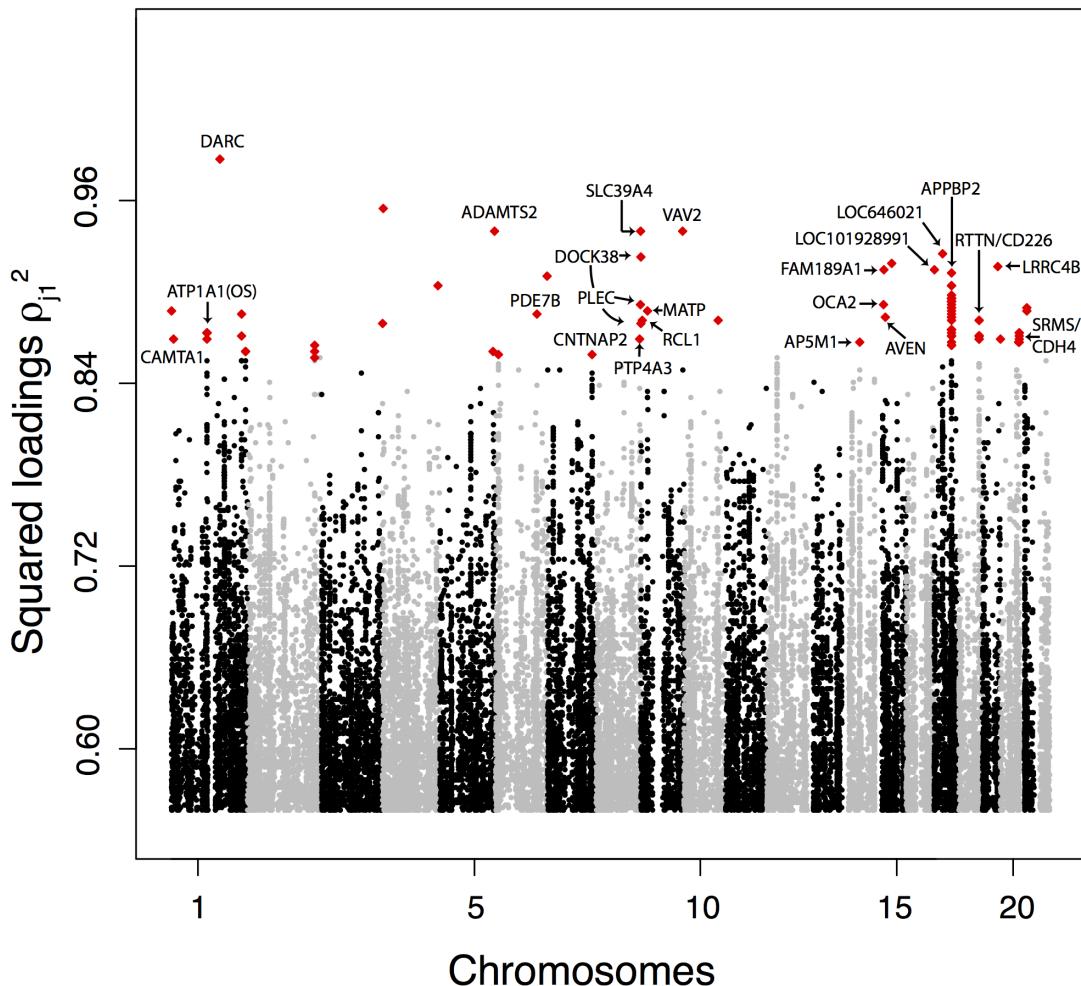


FIGURE 2.14 – Manhattan plot for the 1000 Genomes data of the squared loadings ρ_{j1}^2 with the first principal component. For sake of presentation, only the top-ranked SNPs (top 0.1%) are displayed and the 100 top-ranked SNPs are colored in red.

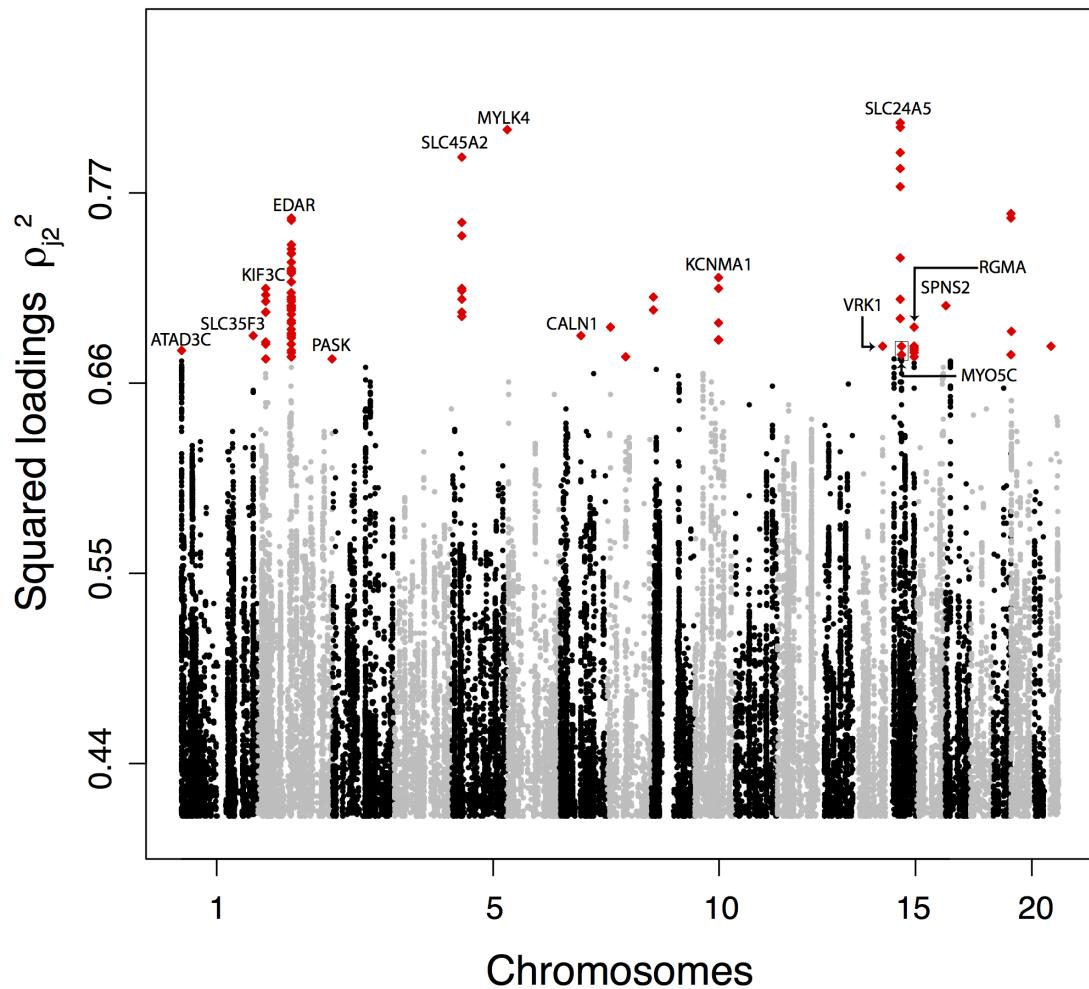


FIGURE 2.15 – Manhattan plot for the 1000 Genomes data of the squared loadings ρ_{j2}^2 with the second principal component. For sake of presentation, only the top-ranked SNPs (top 0.1%) are displayed and the 100 top-ranked SNPs are colored in red.

The 1000 Genome data contain many low-frequency SNPs; 82% of the SNPs have a minor allele frequency smaller than 5%. However, these low-frequency variants are not found among outlier SNPs. There are no SNP with a minor allele frequency smaller than 5% among the 0.1% of the SNPs most correlated with PC1 or with PC2.

The 100 SNPs that are the most correlated with the first PC are located in 24 genomic regions. Most of the regions contain just one or a few SNPs except a peak in the gene APPBP2 that contains 33 out of the 100 top SNPs, a peak encompassing the RTTN and CD226 genes containing 17 SNPs and a peak in the ATP1A1 gene containing seven SNPs (fig. 2.14). Confirming a larger clustering for PC2 outliers, the 100 SNPs that are the most correlated with PC2 cluster in fewer genomic regions. They are located in 14 genomic regions including a region overlapping with EDAR contains 44 top hits, two regions containing eight SNPs and located in the pigmentation genes SLC24A5 and SLC45A2, and two regions with seven top hit SNPs, one in the gene KCNMA1 and another one encompassing the RGLA/MYO5C genes (fig. 2.15).

We perform Gene Ontology (GO) enrichment analyses using Gowinda for the SNPs that are the most correlated with PC1 and PC2. For PC1, we find, among others, enrichment ($FDR \leq 5\%$) for ontologies related to the regulation of arterial blood pressure, the endocrine system and the immunity response (interleukin production, response to viruses). For PC2, we find enrichment ($FDR \leq 5\%$) related to olfactory receptors, keratinocyte and epidermal cell differentiation, and ethanol metabolism. We also search for polygenic adaptation by looking for biological pathways enriched with outlier genes (Daub et al., 2013). For PC1, we find one enriched ($FDR \leq 5\%$) pathway consisting of the beta defensin pathway. The beta defensin pathway contains mainly genes involved in the innate immune system consisting of 36 defensin genes and of two Toll-Like receptors (TLR1 and TLR2). There are additionally two chemokine receptors (CCR2 and CCR6) involved in the beta defensin pathway. For PC2, we also find one enriched pathway consisting of fatty acid omega oxidation ($FDR \leq 5\%$). This pathway consists of genes involved in alcohol oxidation (CYP, ALD, and ALDH genes). Performing a less stringent enrichment analysis which can find pathways containing overlapping genes, we find more enriched pathways : the beta defensin and the defensin pathways for PC1 and ethanol oxidation, glycolysis/gluconeogenesis and fatty acid omega oxidation for PC2.

To further validate the proposed list of candidate SNPs involved in local adaptation, we test for an enrichment of genic or nonsynonymous SNP among the SNPs that are the most correlated with the PC. We measure the enrichment among outliers by computing odds ratio (Fagny et al., 2014; Kudaravalli, Veyrieras, Stranger, Dermitzakis, & Pritchard, 2008). For PC1, we do not find significant enrichments (table 2.1) except when measuring the enrichment of genic regions compared with nongenic regions ($OR = 10.18$ for the 100 most correlated SNPs, $P \leq 5\%$ using a permutation procedure). For PC2, we find an enrichment of genic regions among outliers as well as an enrichment of nonsynonymous SNPs (table 2.1). By contrast with the enrichment of genic regions for SNPs extremely correlated with the first PC, the enrichment for the variants extremely correlated with PC2 outliers is significant when using different thresholds to define outliers (table 2.1).

TABLE 2.1 – Enrichment Measured with Odds Ratio (OR) of the Variants Most Correlated with the Principal Components Obtained from the 1000 Genomes Data. Enrichment significant at the 1% (resp. 5%) level are indicated with ** (resp. *).

	Top 0.1%	Top 0.01%	Top 0.005%	Top 100 SNPs
pc1-genic/noger	1.60*	1.24	1.09	1.93
pc1-nonsyn/all	1.70	1.18	2.42	10.07*
pc1-UTR/all	1.37	0.80	1.65	3.44
pc2-genic/noger	1.51*	2.27	4.73**	4.44*
pc2-nonsyn/all	1.72	4.66*	7.40	12.18*
pc2-UTR/all	1.68	4.01*	3.36	2.73

Discussion

The promise of a fine characterization of natural selection in humans fostered the development of new analytical methods for detecting candidate genomic regions (Vitti, Grossman, & Sabeti, 2013). Population-differentiation based methods such as genome scans based on F_{ST} look for marked differences in allele frequencies between population (Holsinger & Weir, 2009). Here, we show that the communality statistic h^2 , which measures the proportion of variance of a SNP that is explained by the first K principal components, provides a similar list of outliers than the F_{ST} statistic when there are $K + 1$ populations. In addition, the communality statistic h^2 based on PCA can be viewed as an extension of F_{ST} because it does not require to define populations in advance and can even be applied in the absence of well-defined populations.

To provide an example of genome scans based on PCA when there are no clusters of populations, we additionally consider the POPRES data consisting of 447,245 SNPs typed for 1,385 European individuals (Nelson et al., 2008). The scree plot indicates that there are $K = 2$ relevant clusters (supplementary fig. B.3). The first principal component corresponds to a Southeast–Northwest gradient and the second one discriminates individuals from Southern Europe along a East–West gradient (Jay, Sjödin, Jakobsson, & Blum, 2012; Novembre et al., 2008) (fig. 2.16). Considering the 100 SNPs most correlated with the first PC, we find that 75 SNPs are in the lactase region, 18 SNPs are in the HLA region, 5 SNPs are in the ADH1C gene, 1 SNP is in HERC2, and another is close to the LOC283177 gene (fig. 2.17). When considering the 100 SNPs most correlated with the second PC, we find less clustering than for PC1 with more peaks (supplementary fig. B.13). The regions that contain the largest number of SNPs in the top 100 SNPs are the HLA region (41 SNPs) and a region close to the NEK10 gene (10 SNPs), which is a gene potentially involved in breast cancer (Ahmed et al., 2009). The genome scan retrieves well-known signals of adaption in humans that are related to lactase persistence (LCT) (Bersaglieri et al., 2004), immunity (HLA), alcohol metabolism (ADH1C) (Han et al., 2007), and pigmentation (HERC2) (Wilde et al., 2014). The analysis of the POPRES data shows that genome scan based on PCA can be applied when there is a clinal or continuous pattern of

population structure without well-defined clusters of individuals.

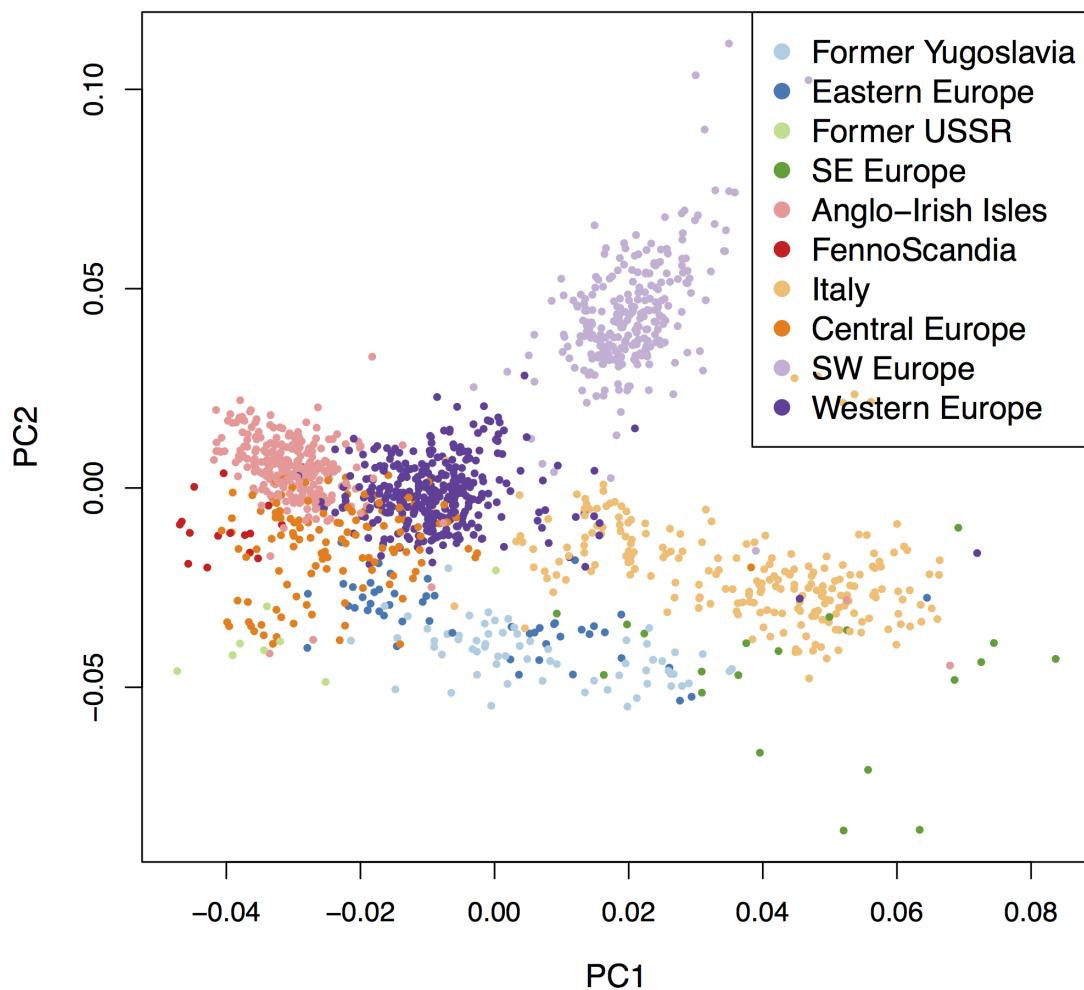


FIGURE 2.16 – PCA with $K = 2$ applied to the POPRES data.

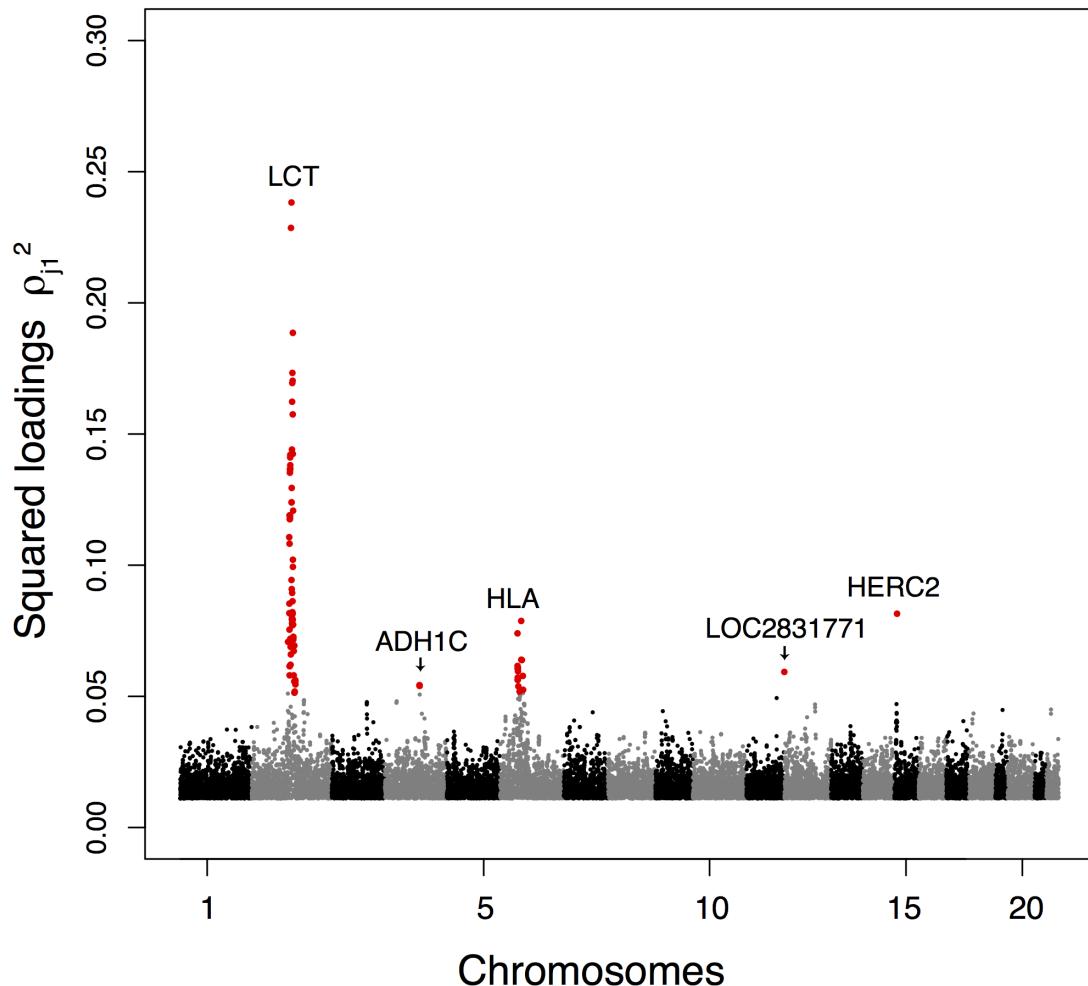


FIGURE 2.17 – Manhattan plot for the POPRES data of the squared loadings ρ_{j1}^2 with the first principal component. For sake of presentation, only the top-ranked SNPs (top 5%) are displayed and the 100 top-ranked SNPs are colored in red.

When there are clusters of populations, we have shown with simulations that genome scans based on F_{ST} can be reproduced with PCA. Genome scans based on PCA have the additional advantage that a particular axis of genetic variation, which is related to adaptation, can be pinpointed. Bearing some similarities with PCA, performing a spectral decomposition of the kinship matrix has been proposed to pinpoint populations where adaptation took place (Fariello et al., 2013). However, despite of some advantages, the statistical problems related to genome scans with F_{ST} remain. The drawbacks of F_{ST} arise when there is hierarchical population structure or range expansion because F_{ST} does not account for correlations of allele frequencies

among subpopulations (Bierne, Roze, & Welch, 2013; Lotterhos & Whitlock, 2014). An alternative presentation of the issues arising with F_{ST} is that it implicitly assumes either a model of instantaneous divergence between populations or an island-model (Bonhomme et al., 2010). Deviations from these models severely impact FDRs (Duforet-Frebourg et al., 2014). Viewing F_{ST} from the point of view of PCA provides a new explanation about why F_{ST} does not provide an optimal ranking of SNPs for detecting selection. The statistic F_{ST} or the proposed h^2 communality statistic are mostly influenced by the first principal component and the relative importance of the first PC increases with the difference between the first and second eigenvalues of the covariance matrix of the data. Because the first PC can represent ancient adaptive events, especially under population divergence models (G. McVean, 2009), it explains why F_{ST} and the communality h^2 are biased toward ancient evolutionary events. Following recent developments of F_{ST} -related statistics that account for hierarchical population structure (Bonhomme et al., 2010; Foll et al., 2014; Günther & Coop, 2013), we proposed an alternative statistic $h_j'^2$, which should give equal weights to the different PCs. However, analyzing simulations and the 1000 Genomes data show that $h_j'^2$ do not properly account for hierarchical population structure because outliers identified by $h_j'^2$ are almost always related to the last PC kept in the analysis. To avoid to bias data analysis in favor of one principal component, it is possible to perform a genome scan for each principal component.

In addition to ranking the SNPs when performing a genome scan, a threshold should be chosen to extract a list of outlier SNPs. We do not have addressed the question of how to choose the threshold and rather used empirical threshold such as the 99% quantile of the distribution of the test statistic (top 1%). If interested in controlling the FDR, we can assume that the loadings ρ_{kj} are Gaussian with zero mean (Galinsky et al., 2016). Because of the constraints imposed on the loadings when performing PCA, the variance of the ρ_{kj} 's is equal to the proportion of variance explained by the k^{th} PC, which is given by $\lambda_k/(p \times (n - 1))$ where λ_k is the k^{th} eigenvalue of the matrix YY^T . Assuming a Gaussian distribution for the loadings, the communality can then be approximated by a weighted sum of chi-square distribution. Approximating a weighted sum of chi-square distribution with a chi-square distribution, we have (Yuan & Bentler, 2010)

$$h^2 \times K/c \sim \chi_K^2, \quad (2.16)$$

where $c = \sum_{k=1}^K \lambda_k/(p \times (n - 1))$ is the proportion of variance explained by the first K PCs. The chi-square approximation of equation (2.16) bears similarity with the approximation of Lewontin & Krakauer (1973) that states that $F_{ST} \times (n_{\text{pops}} - 1)/\bar{F}_{ST}$ follows a chi-square approximation with $(n_{\text{pops}} - 1)$ degrees of freedom where \bar{F}_{ST} is the mean F_{ST} over loci and $(n_{\text{pops}} - 1)$ is the number of populations. In the simulations of an island model and of a divergence model, quantile-to-quantile plots indicate a good fit to the theoretical chi-square distribution of expression (2.16) (supplementary fig. B.14). When using the chi-square approximation to compute P -values, we evaluate if FDR can be controlled using Benjamini–Hochberg correction (Benjamini and Hochberg 1995). We find that the actual proportion of false discoveries corresponds

to the target FDR for the island model but the procedure is too conservative for the divergence model (supplementary fig. B.15). For instance, when controlling FDR at a level of 25%, the actual proportion of false discoveries is of 15%. A recent test based on F_{ST} and a chi-square approximation was also found to be conservative (Whitlock & Lotterhos, 2015). Analysing the phase 1 release of the 1000 Genomes data demonstrates the suitability of a genome scan based on PCA to detect signals of positive selection. We search for variants extremely correlated with the first PC, which corresponds to differentiation between Africa and Eurasia and with the second PC, which corresponds to differentiation between Europe and Asia. For variants most correlated with the second PC, there is a significant enrichment of genic and nonsynonymous SNPs whereas the enrichment is less detectable for variants related to the first PC. The enrichment analysis confirms that positive selection may favor local adaptation of human population by increasing differentiation in genic regions especially in nonsynonymous variants (Barreiro et al., 2008). Consistent with LD, we find that candidate variants are clustered along the genome with a larger clustering for variants correlated with the Europe–Asia axis of differentiation (PC2). The difference of clustering illustrates that statistical methods based on LD for detecting selection will perform differently depending on the time frame under which adaptation had the opportunity to occur (Sabeti et al., 2006). The fact that population divergence, and its concomitant adaptive events, between Europe and Asia is more recent than the out-of-Africa event is a putative explanation of the difference of clustering between PC1 and PC2 outliers. Explaining the difference of enrichment between PC1 and PC2 outliers is more difficult. The weaker enrichment for PC1 outliers can be attributed either to a larger number of false discoveries or to a larger importance of other forms of natural selection such as background selection (Hernandez et al., 2011).

When looking at the 100 SNPs most correlated with PC1 or PC2, we find genes for which selection in humans was already documented (9/24 for PC1 and 5/14 for PC2). Known targets for selection include genes involved in pigmentation (MATP, OCA2 for PC1 and SLC45A2, SLC24A5, and MYO5C for PC2), in the regulation of sweating (EDAR for PC2), and in adaptation to pathogens (DARC, SLC39A4, and VAV2 for PC1). A 100 kb region in the vicinity of the APPBPP2 gene contains one-third of the 100 SNPs most correlated with PC1. This APPBPP2 region is a known candidate for selection and has been identified by looking for miRNA binding sites with extreme population differentiation (J. Li et al., 2012). APPBPP2 is a nervous system gene that has been associated with Alzheimer disease, and it may have experienced a selective sweep (Williamson et al., 2007). For some SNPs in APPBPP2, the differences of allele frequencies between Eurasian populations and sub-Saharan populations from Africa are of the order of 90% (<http://popgen.uchicago.edu/ggv/>, last accessed December 2015) calling for a further functional analysis. Moreover, looking at the 100 SNPs most correlated with PC1 and PC2 confirms the importance of noncoding RNA (FAM230B, D21S2088E, LOC100133461, LINC00290, LINC01347, LINC00681), such as miRNA (MIR429), as a substrate for human adaptation (Grossman et al., 2013; J. Li et al., 2012). Among the other regions with a large number of candidate SNPs, we also found the RTTN/CD226 regions, which contain many SNPs correlated with PC1. In different selection scans, the RTTN genes have been detected (Barreiro et al., 2008 ; Carlson et al.,

2005), and it is involved in the development of the human skeletal system (Wu & Zhang, 2010). An other region with many SNPs correlated with PC1 contains the ATP1A1 gene involved in osmoregulation and associated with hypertension (Gurdasani et al., 2015). The regions containing the largest number of SNPs correlated with PC2 are well-documented instances of adaptation in humans and includes the EDAR, SLC24A5, and SLC45A2 genes. The KCNMA1 gene contains seven SNPs correlated with PC2 and is involved in breast cancer and obesity (Jiao et al., 2011; Oeggerli et al., 2012). As for KCNMA1, the MYO5C has already been reported in selection scans although no mechanism of biological adaption has been proposed yet (H. Chen, Patterson, & Reich, 2010; Fumagalli et al., 2010). To summarize, the list of most correlated SNPs with the PCs identifies well-known genes related to biological adaptation in humans (EDAR, SLC24A5, SLC45A2, DARC), but also provides candidate genes that deserve further studies such as the APPBPP2, TP1A1, RTTN, KCNMA1, and MYO5C genes, as well as the ncRNAs listed above.

We also show that a scan based on PCA can also be used to detect more subtle footprints of positive selection. We conduct an enrichment analysis that detects polygenic adaptation at the level of biological pathways (Daub et al., 2013). We find that genes in the beta-defensin pathway are enriched in SNPs correlated with PC1. The beta-defensin genes are key components of the innate immune system and have evolved through positive selection in the catarrhine primate lineages (Hollox & Armour, 2008). As for the HLA complex, some beta-defensin genes (DEFB1, DEFB127) show evidence of long-term balancing selection with major haplotypic clades coexisting since millions of years (Cagliani et al., 2008; Hollox & Armour, 2008). We also find that genes in the omega fatty acid oxidation pathways are enriched in SNPs correlated with PC2. This pathway was also found when investigating polygenic adaptation to altitude in humans (Foll et al., 2014). The proposed explanation was that omega oxidation becomes a more important metabolic pathway when beta oxidation is defective, which can occur in case of hypoxia (Foll et al., 2014). However, this explanation is not valid in the context of the 1000 Genomes data when there are no populations living in hypoxic environments. Proposing phenotypes on which selection operates is complicated by the fact that the omega fatty acid oxidation pathway strongly overlaps with two other pathways : ethanol oxidation and glycolysis. Evidence of selection on the alcohol dehydrogenase locus have already been provided (Han et al., 2007) with some authors proposing that a lower risk for alcoholism might have been beneficial after rice domestication in Asia (Y. Peng et al., 2010). This hypothesis is speculative and we lack a confirmed biological mechanism explaining the enrichment of the fatty acid oxidation pathway. More generally, the enrichment of the beta-defensin and of the omega fatty acid oxidation pathways confirms the importance of pathogenic pressure and of metabolism in human adaptation to different environments (Barreiro & Quintana-Murci, 2010; Daub et al., 2013; Fumagalli et al., 2011; Hancock et al., 2008).

In conclusion, we propose a new approach to scan genomes for local adaptation that works with individual genotype data. Because the method is efficiently implemented in the software PCAdapt fast, analyzing 36,536,154 SNPs took only 502 min using a single core of an Intel(R) Xeon(R) (E5-2650, 2.00GHz, 64 bits). Even with low-coverage sequence data (3x), PCA-based statistics retrieve well-known examples of biological

adaptation which is encouraging for future whole-genome sequencing project, especially for nonmodel species, aiming at sampling many individuals with limited cost.

Materials and Methods

Simulations of an Island Model

Simulations were performed with *ms* (Hudson, 2002). We assume that there are three islands with 100 sampled individuals in each of them. There is a total of 1,400 neutral SNPs, and 100 adaptive SNPs. SNPs are assumed to be unlinked. To mimic adaptation, we consider that adaptive SNP have a migration rate smaller than the migration rate of neutral SNPs ($4N_0m = 4$ for neutral SNPs) (Bazin et al. 2010). The strength of selection is equal to the ratio of the migration rates of neutral and adaptive SNPs. Adaptation is assumed to occur in one population only. The *ms* command lines for neutral and adaptive SNPs are given below (assuming an effective migration rate of $4N_0m = 0.1$ for adaptive SNPs).

The values of migrations rates we consider for adaptive SNPs are $4N_0m = 0.04, 0.1, 0.4, 1, 2$.

Simulations of Divergence Models

We assume that each population has a constant effective population size of $N_0 = 1000$ diploid individuals, with 50 individuals sampled in each population. The genotypes consist of 10,000 independent SNPs. The simulations were performed in two steps. In the first step, we used the software *ms* to simulate genetic diversity (Hudson, 2002) in the ancestral population. We kept only variants with a minor allele frequency larger than 5% at the end of the first step. The second step was performed with *simuPOP* (B. Peng & Kimmel, 2005) and simulations were started using the allele frequencies generated with *ms* in the ancestral population. Looking forward in time, we consider that there are 100 generations between the initial split and the following split between the two *B* subpopulations, and 200 generations following the split between the two *B* subpopulations. We assume no migration between populations. In the simulation of figure 2.11, we assume that 250 SNPs confer a selective advantage in the branch leading to population *A* and 250 other SNPs confer a selective advantage in the branch leading to population *B*. We consider an additive model for selection with a selection coefficient of $s = 1.025$ for heterozygotes. For the simulation of figure 2.12, we assume that there are four nonoverlapping sets of 125 adaptive SNPs with each set being related to adaptation in one of the four branches of the divergence tree. A SNP can confer a selective advantage in a single branch only.

When including migration, we consider that there are 200 generations between the initial split and the following split between the two *B* subpopulations, and 100 generations following the split between the two *B* subpopulations. We consider migration rates ranging from 0.2% to 5% per generation. Migration is assumed to occur only after the split between *B*₁ and *B*₂. The migration rate is the same for the

three pairs of populations. To estimate the F_{ST} statistic, we consider the estimator of Weir & Cockerham (1984).

1000 Genomes Data

We downloaded the 1000 Genomes data (phase 1 v3) (Consortium & others, 2012). We kept low-coverage genome data and excluded exomes and triome data to minimize variation in read depth. Filtering the data resulted in a total of 36,536,154 SNPs that have been typed on 1,092 individuals. Because the analysis focuses on biological adaptation that took place during the human diaspora out of Africa, we removed recently admixed populations (Mexican, Columbian, PortoRican, and AfroAmerican individuals from the Southwest of the United States). The resulting data set contains 850 individuals coming from Asia (two Han Chinese and one Japanese populations), Africa (Yoruba and Luhya), and Europe (Finish, British in England and Scotland, Iberian, Toscan, and Utah residents with Northern and Western European ancestry).

Enrichment Analyses

We used Gowinda (Kofler & Schlötterer, 2012) to test for enrichment of GO. A gene is considered as a candidate if there is at least one of the most correlated SNPs (top 1%) that is mapped to the gene (within an interval of 50 kb upstream and downstream of the gene). Enrichment was computed as the proportion of genes containing at least one outlier SNPs among the genes of the given GO category that are present in the data set. In order to sample a null distribution for enrichment, *Gowinda* performs resampling without replacement of the SNPs. We used the `-gene` option of *Gowinda* that assumes complete linkage within genes.

We performed a second enrichment analysis to determine if outlier SNPs are enriched for genic regions. We computed odds ratio (Kudaravalli et al., 2008)

$$\text{OR} = \frac{\Pr(\text{genic|outlier})}{\Pr(\text{not genic|outlier})} \frac{\Pr(\text{not genic|not outlier})}{\Pr(\text{genic|not outlier})} \quad (2.17)$$

We implemented a permutation procedure to test if an odds ratio is significantly larger than 1 (Fagny et al., 2014). The same procedure was applied when testing for enrichment of UTR regions (untranslated regions) and of nonsynonymous SNPs.

Polygenic Adaptation

To test for polygenic adaptation, we determined whether genes in a given biological pathway show a shift in the distribution of the loadings (Daub et al., 2013). We computed the SUMSTAT statistic for testing if there is an excess of selection signal in each pathway (Daub et al., 2013). We applied the same pruning method to take into account redundancy of genes within pathways. The test statistic is the squared loading standardized into a z -score (Daub et al., 2013). SUMSTAT is computed for each gene as the sum of test statistic of each SNP belonging to the gene. Intergenic SNPs are assigned to a gene provided they are situated 50kb up- or downstream. We

downloaded 63,693 known genes from the UCSC website and we mapped SNPs to a gene if a SNP is located within a gene transcript or within 50kb of a gene. A total of 18,267 genes were mapped with this approach. We downloaded 2,681 gene sets from the NCBI Biosystems database. After discarding genes that were not part of the aforementioned gene list, removing gene sets with less than 10 genes and pooling nearly identical gene sets, we kept 1,532 sets for which we test if there was a shift of the distribution of loadings.

Acknowledgments

This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the ANR AGRHUM project (ANR-14-CE02-0003-01). POPRES data were obtained from dbGaP (accession number phs000145.v1.p1).

2.5 Article 2

MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2017) 17, 67–77

doi: 10.1111/1755-0998.12592

SPECIAL ISSUE: POPULATION GENOMICS WITH R

pcadapt: an R package to perform genome scans for selection based on principal component analysis

KEURCIEN LUU,* ERIC BAZIN† and MICHAEL G. B. BLUM*

*Laboratoire TIMC-IMAG, UMR 5525, CNRS, Université Grenoble Alpes, Grenoble, France, †Laboratoire d'Ecologie Alpine UMR 5553, CNRS, Université Grenoble Alpes, Grenoble, France

Abstract

The R package *pcadapt* performs genome scans to detect genes under selection based on population genomic data. It assumes that candidate markers are outliers with respect to how they are related to population structure. Because population structure is ascertained with principal component analysis, the package is fast and works with large-scale data. It can handle missing data and pooled sequencing data. By contrast to population-based approaches, the package handles admixed individuals and does not require grouping individuals into populations. Since its first release, *pcadapt* has evolved in terms of both statistical approach and software implementation. We present results obtained with robust Mahalanobis distance, which is a new statistic for genome scans available in the 2.0 and later versions of the package. When hierarchical population structure occurs, Mahalanobis distance is more powerful than the communality statistic that was implemented in the first version of the package. Using simulated data, we compare *pcadapt* to other computer programs for genome scans (*BayeScan*, *hapflk*, *OutFLANK*, *sNMF*). We find that the proportion of false discoveries is around a nominal false discovery rate set at 10% with the exception of *BayeScan* that generates 40% of false discoveries. We also find that the power of *BayeScan* is severely impacted by the presence of admixed individuals whereas *pcadapt* is not impacted. Last, we find that *pcadapt* and *hapflk* are the most powerful in scenarios of population divergence and range expansion. Because *pcadapt* handles next-generation sequencing data, it is a valuable tool for data analysis in molecular ecology.

Introduction

Looking for variants with unexpectedly large differences of allele frequencies between populations is a common approach to detect signals of natural selection (Lewontin & Krakauer, 1973). When variants confer a selective advantage in the local environment, allele frequency changes are triggered by natural selection leading to unexpectedly

TABLE 2.2 – Summary of the different statistical methods and implementations of *pcadapt*. Pop. structure stands for population structure and dist. stands for distance

Test statistic	Pop. structure	Language	Command line	Versions of the R package	References
Bayes factor	Factor model	C	PCAdapt	NA	Duforet-Frebourg et al. (2014)
Community	PCA	C and R	PCAdapt fast	1. x	Duforet-Frebourg et al. (2016)
Mahalanobis dist.	PCA	R	NA	2. x and 3. x	This study

large differences of allele frequencies between populations. To detect variants with large differences of allele frequencies, numerous test statistics have been proposed, which are usually based on chi-square approximations of F_{ST} -related test statistics (François et al., 2016).

Statistical approaches for detecting selection should address several challenges. The first challenge is to account for hierarchical population structure that arises when genetic differentiation between populations is not identical between all pairs of populations. Statistical tests based on F_{ST} that do not account for hierarchical structure, when it occurs, generate a large excess of false-positive loci (Bierne et al., 2013; Excoffier et al., 2009).

A second challenge arises because approaches based on F_{ST} -related measures require to group individuals into populations, although defining populations is a difficult task (Waples & Gaggiotti, 2006). Individual sampling may not be population based but based on more continuous sampling schemes (Lotterhos & Whitlock, 2015). Additionally assigning an admixed individual to a single population involves some arbitrariness because different regions of its genome might come from different populations (Pritchard, Stephens, & Donnelly, 2000). Several individual-based methods of genome scans have already been proposed to address this challenge and they are based on related techniques of multivariate analysis including principal component analysis (PCA), factor models and non-negative matrix factorization (G. Chen et al., 2016; Duforet-Frebourg et al., 2014, 2015; Galinsky et al., 2016; Hao, Song, & Storey, 2015; Martins, Caye, Luu, Blum, & Francois, 2016).

The last challenge arises from the nature of multilocus data sets generated from next-generation sequencing platforms. Because data sets are massive with a large number of molecular markers, Monte Carlo methods usually implemented in Bayesian statistics may be prohibitively slow (Lange, Papp, Sinsheimer, & Sobel, 2014). Additionally, next-generation sequencing data may contain a substantial proportion of missing data that should be accounted for (B. Arnold, Corbett-Detig, Hartl, & Bomblies, 2013; Gautier et al., 2013).

To address the aforementioned challenges, we have developed the computer program *pcadapt* and the R package *pcadapt*. The computer program *pcadapt* is now deprecated and the R package only is maintained. *pcadapt* assumes that markers excessively related to population structure are candidates for local adaptation. Since its first release, *pcadapt* has substantially evolved in terms of both statistical approach and implementation (Table 2.2).

The first release of *pcadapt* was a command line computer program written in C. It implemented a Monte Carlo approach based on a Bayesian factor model (Duforet-Frebourg et al., 2014). The test statistic for outlier detection was a Bayes factor. Because Monte Carlo methods can be computationally prohibitive with massive NGS data, we then developed an alternative approach based on PCA. The first statistic based on PCA was the communality statistic, which measures the percentage of variation of a single-nucleotide polymorphism (SNP) explained by the first K principal components (Duforet-Frebourg et al., 2015). It was initially implemented with a command line computer program (the *pcadapt fast* command) before being implemented in the *pcadapt* R package. We do not maintain C versions of *pcadapt* anymore. The whole analysis that goes from reading genotype files to detecting outlier SNPs can now be performed in R (Team, 2015).

The 2.0 and following versions of the R package implement a more powerful statistic for genome scans. The test statistic is a robust Mahalanobis distance. A vector containing K z -scores measures to what extent a SNP is related to the first K principal components. The Mahalanobis distance is then computed for each SNP to detect outliers for which the vector of z -scores does not follow the distribution of the main bulk of points. The term robust refers to the fact that the estimators of the mean and of the covariance matrix of z , which are required to compute the Mahalanobis distances, are not sensitive to the presence of outliers in the data set (Maronna & Zamar, 2002). In the following, we provide a comparison of statistical power that shows that Mahalanobis distance provides more powerful genome scans compared with the communality statistic and with the Bayes factor that were implemented in previous versions of *pcadapt*.

In addition to comparing the different test statistics that were implemented in *pcadapt*, we compare statistic performance obtained with the 3.0 version of *pcadapt* and with other computer programs for genome scans. We use simulated data to compare computer programs in terms of false discovery rate (FDR) and statistical power. We consider data simulated under different demographic models including island model, divergence model and range expansion. To perform comparisons, we include programs that require to group individuals into populations : *BayeScan* (Foll & Gaggiotti, 2008), the F_{LK} statistic as implemented in the *hapfk* computer program (Bonhomme et al., 2010), and *OutFLANK* that provides a robust estimation of the null distribution of a F_{ST} test statistic (Whitlock & Lotterhos, 2015). We additionally consider the *sNMF* computer program that implements another individual-based test statistic for genome scans (Frichot, Mathieu, Trouillon, Bouchard, & François, 2014; Martins et al., 2016).

Statistical and computational approach

Input data

The R package can handle different data formats for the genotype data matrix. In the version 3.0 that is currently available on CRAN, the package can handle genotype data files in the *vcf*, *ped* and *lfmm* formats. In addition, the package can

also handle a *pcadapt* format, which is a text file where each line contains the allele counts of all individuals at a given locus. When reading a genotype data matrix with the *read.pcadapt* function, a *.pcadapt* file is generated, which contains the genotype data in the *pcadapt* format.

Choosing the number of principal components

In the following, we denote by n the number of individuals, by p the number of genetic markers and by G the genotype matrix that is composed of n lines and p columns. The genotypic information at locus j for individual i is encoded by the allele count G_{ij} , $1 \leq i \leq n$ and $1 \leq j \leq p$, which is a value in 0,1 for haploid species and in 0,1,2 for diploid species. The current 3.0.2 version of the package can handle haploid and diploid data only.

First, we normalize the genotype matrix columnwise. For diploid data, we consider the usual normalization in population genomics where $\tilde{G}_{ij} = (G_{ij} - p_j)/(2 \times p_j(1 - p_j))^{1/2}$, and p_j denotes the minor allele frequency for locus j (N. Patterson et al., 2006). The normalization for haploid data is similar except that the denominator is given by $(p_j(1 - p_j))^{1/2}$

Then, we use the normalized genotype matrix math formula to ascertain population structure with PCA (N. Patterson et al., 2006). The number of principal components to consider is denoted K and is a parameter that should be chosen by the user. In order to choose K , we recommend to consider the graphical approach based on the scree plot (Jackson, 1993). The scree plot displays the eigenvalues of the covariance matrix Ω in descending order. Up to a constant, eigenvalues are proportional to the proportion of variance explained by each principal component. The eigenvalues that correspond to random variation lie on a straight line whereas the ones corresponding to population structure depart from the line. We recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept (Cattell, 1966).

Test statistic

We now detail how the package computes the test statistic. We consider multiple linear regressions by regressing each of the p SNPs by the K principal components X_1, \dots, X_K

$$G_j = \sum_{k=1}^K \beta_{jk} X_k + \epsilon_j, \quad j = 1, \dots, p, \quad (2.18)$$

where β_{jk} is the regression coefficient corresponding to the j -th SNP regressed by the k -th principal component, and ϵ_j is the residuals vector. To summarize the result of the regression analysis for the j -th SNP, we return a vector of z -scores $z_j = (z_{j1}, \dots, z_{jK})$ where z_{jk} corresponds to the z -score obtained when regressing the j -th SNP by the k -th principal component.

The next step is to look for outliers based on the vector of z -scores. We consider a classical approach in multivariate analysis for outlier detection. The test statistic is a

robust Mahalanobis distance D defined as

$$D_j^2 = (z_j - \bar{z})^T \Sigma^{-1} (z_j - \bar{z}), \quad (2.19)$$

where Σ is the $(K \times K)$ covariance matrix of the z -scores and \bar{z} is the vector of the K z -score means (Maronna & Zamar, 2002). When $K > 1$, the covariance matrix Σ is estimated with the orthogonalized Gnanadesikan–Kettenring method that is a robust estimate of the covariance able to handle large-scale data (Maronna & Zamar, 2002) (*covRob* function of the *robust* R package). When $K = 1$, the variance is estimated with another robust estimate (*cov.rob* function of the *MASS* R package).

Genomic inflation factor

To perform multiple hypothesis testing, Mahalanobis distances should be transformed into P -values. If the z -scores were truly multivariate Gaussian, the Mahalanobis distances D should be chi-square distributed with K degrees of freedom. However, as usual for genome scans, there are confounding factors that inflate values of the test statistic and that would lead to an excess of false positives (François et al., 2016). To account for the inflation of test statistics, we divide Mahalanobis distances by a constant λ to obtain a statistic that can be approximated by a chi-square distribution with K degrees of freedom. This constant is estimated by the genomic inflation factor defined here as the median of the Mahalanobis distances divided by the median of the chi-square distribution with K degrees of freedom (Devlin & Roeder, 1999).

Control of the false discovery rate (FDR)

Once P -values are computed, there is a problem of decision-making related to the choice of a threshold for P -values. We recommend to use the FDR approach where the objective is to provide a list of candidate genes with an expected proportion of false discoveries smaller than a specified value. For controlling the FDR, we consider the q -value procedure as implemented in the *qvalue* R package that is less conservative than Bonferroni or Benjamini–Hochberg correction (Storey & Tibshirani, 2003). The *qvalue* R package transforms the P -values into q -values and the user can control a specified value α of FDR by considering as candidates the SNPs with q -values smaller than α .

Numerical computations

PCA is performed using a C routine that allows to compute scores and eigenvalues efficiently with minimum RAM access (Duforest-Frebourg et al., 2015). Computing the covariance matrix Ω is the most computationally demanding part. To provide a fast routine, we compute the $n \times n$ covariance matrix Ω instead of the much larger $p \times p$ covariance matrix. We compute the covariance Ω incrementally by adding small storable covariance blocks successively. Multiple linear regression is then solved directly by computing an explicit solution, written as a matrix product. Using the fact that the

(n, K) score matrix X is orthogonal, the (p, K) matrix β of regression coefficients is given by $G^T X$ and the (n, p) matrix of residuals is given by $G - XX^T G$. The z -scores are then computed using the standard formula for multiple regression

$$z_{jk} = \hat{\beta}_{jk} \sqrt{\frac{\sum_{i=1}^n x_{ik}^2}{\sigma_j^2}} \quad (2.20)$$

where σ_j^2 is an estimate of the residual variance for the j^{th} SNP, and x_{ik} is the score of the k^{th} principal component for the i^{th} individual.

Missing data

Missing data should be accounted for when computing principal components and when computing the matrix of z -scores. There are many methods to account for missing data in PCA, and we consider the pairwise covariance approach (Dray & Josse, 2015). It consists in estimating the covariance between each pair of individuals using only the markers that are available for both individuals. To compute z -scores, we account for missing data in formula (2.20). The term in the numerator $\sum_{i=1}^n x_{ik}^2$ depends on the quantity of missing data. If there are no missing data, it is equal to 1 by definition of the scores obtained with PCA. As the quantity of missing data grows, this term and the z -score decrease such that it becomes more difficult to detect outlier markers.

Pooled sequence data

When data are sequenced in pool, the Mahalanobis distance is based on the matrix of allele frequency computed in each pool instead of the matrix of z -scores.

Materials and methods

Simulated data

We simulated SNPs under an island model, under a divergence model and we downloaded simulations of range expansion (Lotterhos & Whitlock, 2015). All data we simulated were composed of 3 populations, each of them containing 50 sampled diploid individuals (Table 2.3). SNPs were simulated assuming no linkage disequilibrium. SNPs with minor allele frequencies lower than 5% were discarded from the data sets. The mean F_{ST} for each simulation was comprised between 5% and 10%. Using the simulations based on an island and a divergence model, we also created data sets composed of admixed individuals. We assumed that an instantaneous admixture event occurs at the present time so that all sampled individuals are the results of this admixture event. Admixed individuals were generated by drawing randomly admixture proportions using a Dirichlet distribution of parameter (α, α, α) (α ranging from 0.005 to 1 depending on the simulation).

TABLE 2.3 – Summary of the simulations. The table above shows the average number of individuals, of SNPs, of adaptive markers and the total number of simulations per scenario

	Individuals	SNPs	Adaptive SNPs	Simulations
Island model	150	472	27	35
Divergence model	150	3000	100	6
Island model (hybrids)	150	472	30	27
Divergence model (hybrids)	150	3000	100	9
Range expansion	1200	9999	99	6

Island model

We used *ms* to create simulations under an island model (Fig. B.16). We set a lower migration rate for the 50 adaptive SNPs compared with the 950 neutral ones to mimic diversifying selection (Bazin et al., 2010). For a given locus, migration from population i to j was specified by choosing a value of the effective migration rate that is set to $M_{\text{neutral}} = 10$ for neutral SNPs and to M_{adaptive} for adaptive ones. We simulated 35 data sets in the island model with different strengths of selection, where the strength of selection corresponds to the ratio $M_{\text{neutral}}/M_{\text{adaptive}}$ that varies from 10 to 1000. The *ms* command lines for neutral and adaptive SNPs are given by ($M_{\text{adaptive}} = 0.01$ and $M_{\text{neutral}} = 10$).

```
./ms 300 950 -s 1 -I 3 100 100 100
-ma x 10 10 10 x 10 10 10 x
./ms 300 50 -s 1 -I 3 100 100 100
-ma x 0.01 0.01 0.01 x 0.01 0.01 0.01 x
```

Divergence model

To perform simulations under a divergence model, we used the package *simuPOP*, which is an individual-based population genetic simulation environment (B. Peng & Kimmel, 2005). We assumed that an ancestral panmictic population evolved during 20 generations before splitting into two subpopulations. The second subpopulation then split into subpopulations 2 and 3 at time $T > 20$. All 3 subpopulations continued to evolve until 200 generations have been reached, without migration between them (Figure B.16). A total of 50 diploid individuals were sampled in each population.

Selection only occurred in the branch associated with population 2 and selection was simulated by assuming an additive model (fitness is equal to $1 - 2s$, $1 - s$, 1 depending on the genotypes). We simulated a total of 3000 SNPs comprising of 100 adaptive ones for which the selection coefficient is of $s = 0.1$.

Range expansion

We downloaded in the *Dryad Digital Repository* six simulations of range expansion with two glacial refugia (Lotterhos & Whitlock, 2015). Adaptation occurred during the recolonization phase of the species range from the two refugia. We considered six different simulated data with 30 populations and a number of sampled individuals per location that varies from 20 to 60.

Parameter settings for the different computer programs

When using *hapflk*, we set $K = 1$ that corresponds to the computation of the F_{LK} statistic. When using *BayeScan* and *OutFLANK*, we used the default parameter values. For *sNMF*, we used $K = 3$ for the island and divergence model and $K = 5$ for range expansion as indicated by the cross-entropy criterion. The regularization parameter of *sNMF* was set to $\alpha = 1000$. For *sNMF* and *hapflk*, we used the genomic inflation factor to recalibrate p -values. When using population-based methods with admixed individuals, we assigned each individual to the population with maximum amount of ancestry.

Results

Choosing the number of principal components

We evaluate Cattell's graphical rule to choose the number of principal components. For the island and divergence model, the choice of K is evident (Fig. 2.18). For $K \geq 3$, the eigenvalues follow a straight line. As a consequence, Cattell's rule indicates $K = 2$, which is expected because there are three populations (N. Patterson et al., 2006). For the model of range expansion, applying Cattell's rule to choose K is more difficult (Fig. 2.18). Ideally, the eigenvalues that correspond to random variation lie on a straight line whereas the ones corresponding to population structure depart from the line. However, there is no obvious point at which eigenvalues depart from the straight line. Choosing a value of K between 5 and 8 is compatible with Cattell's rule. Using the package *qvalue* to control 10% of FDR, we find that the actual proportion of false discoveries as well as statistical power is weakly impacted when varying the number of principal components from $K = 5$ to $K = 8$ (Figure B.17).

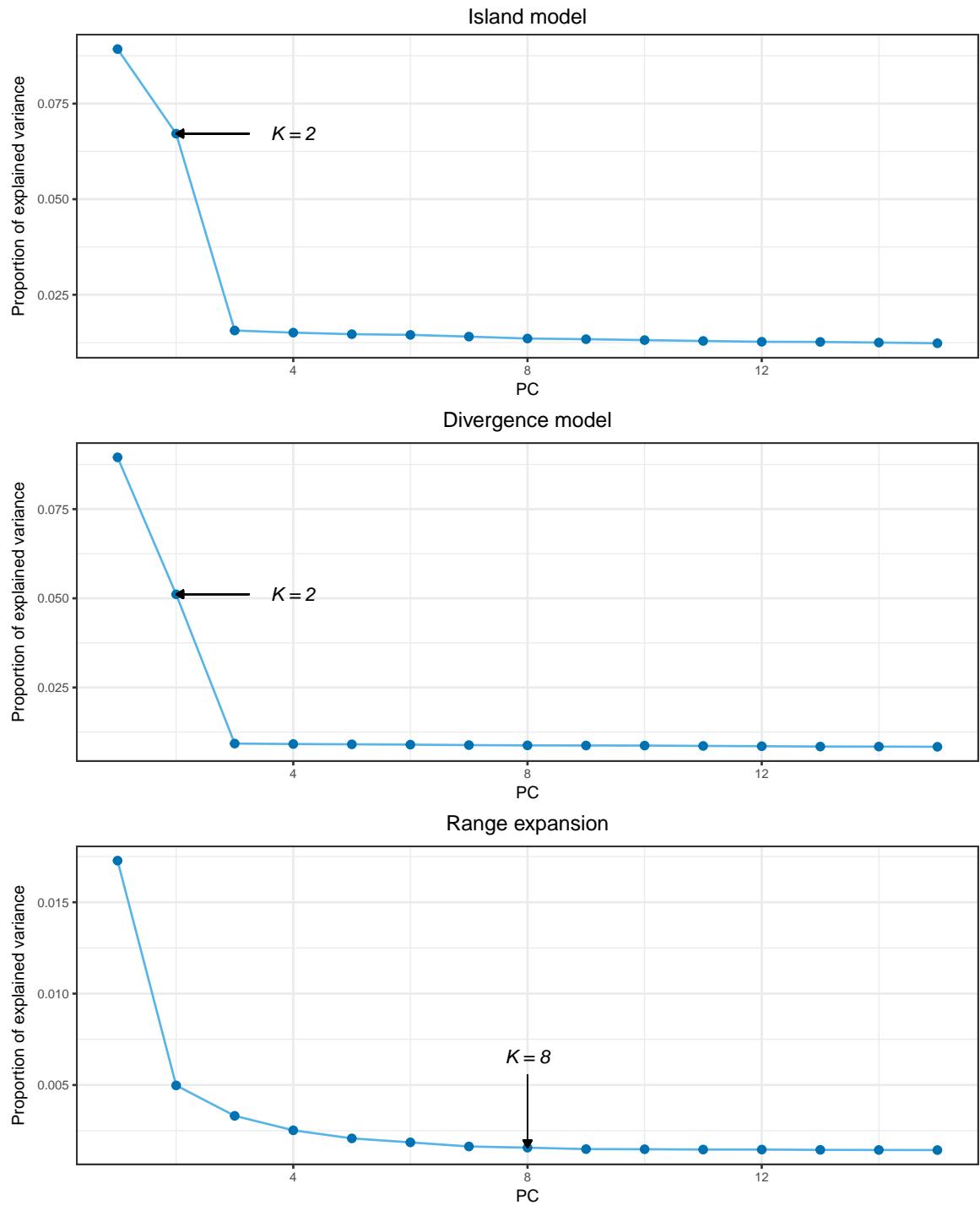


FIGURE 2.18 – Determining K with the scree plot. To choose K , we recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept. According to Cattell's rule, the eigenvalues that correspond to random variation lie on the straight line whereas the ones corresponding to population structure depart from the line. For the island and divergence model, the choice of K is evident. For the model or range expansion, a value of K between 5 and 8 is compatible with Cattell's rule.

An example of genome scans performed with *pcadapt*

To provide an example of results, we apply *pcadapt* with $K = 6$ in the model of range expansion. Population structure captured by the first two principal components is displayed in Fig. 2.19. P -values are well calibrated because they are distributed as a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds to outlier loci (Fig. 2.19). Using a FDR threshold of 10% with the *qvalue* package, we find 122 outliers among 10 000 SNPs, resulting in 23% actual false discoveries and a power of 95%.

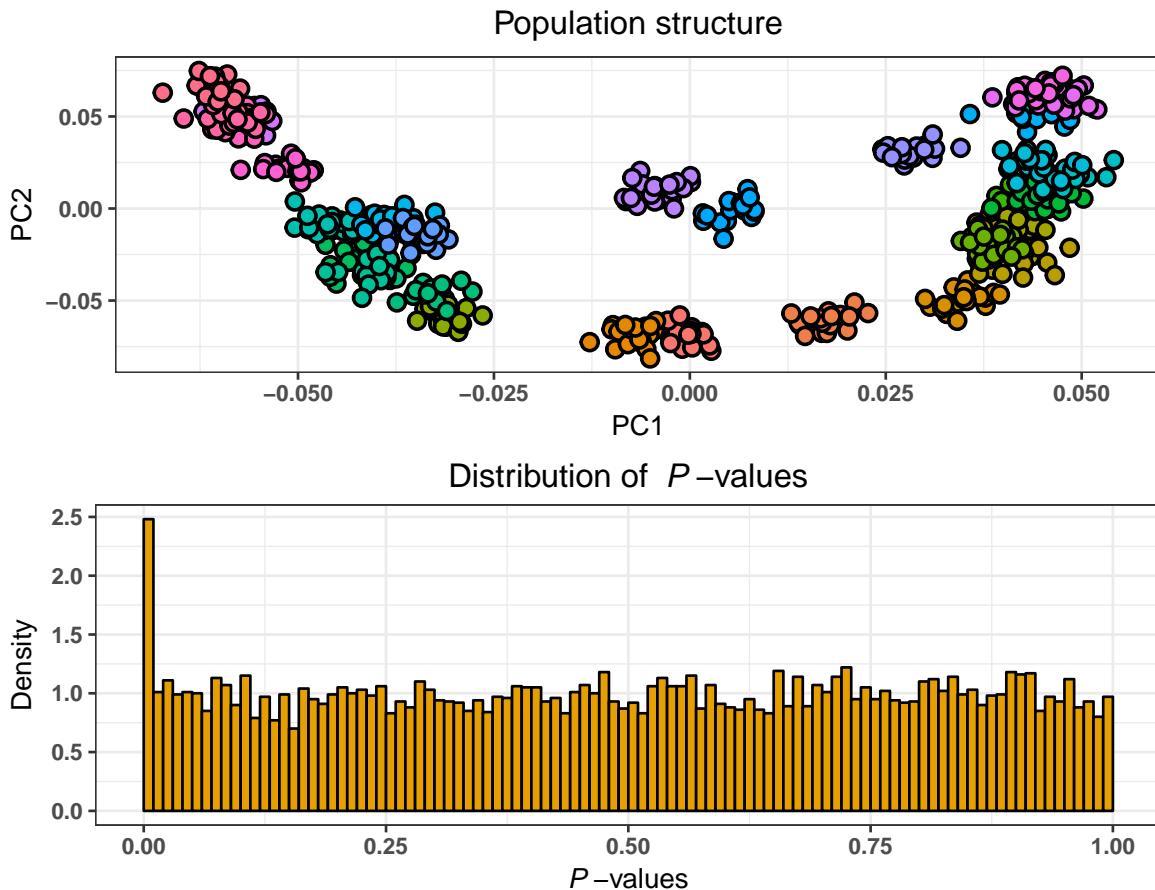


FIGURE 2.19 – Population structure (first 2 principal components) and distribution of P -values obtained with *pcadapt* for a simulation of range expansion. P -values are well calibrated because they are distributed as a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds to outlier loci. In the left panel, each colour corresponds to individuals sampled from the same population.

Control of the false discovery rate

We evaluate to what extent using the packages *pcadapt* and *qvalue* control a FDR set at 10% (Fig. 2.20). All SNPs with a q -value smaller than 10% were considered as

candidate SNPs. For the island model, we find that the proportion of false discoveries is 8% and it increases to 10% when including admixture. For the divergence model, the proportion of false discoveries is 11% and it increases to 22% when including admixture. The largest proportion of false discoveries is obtained under range expansion and is equal to 25%.

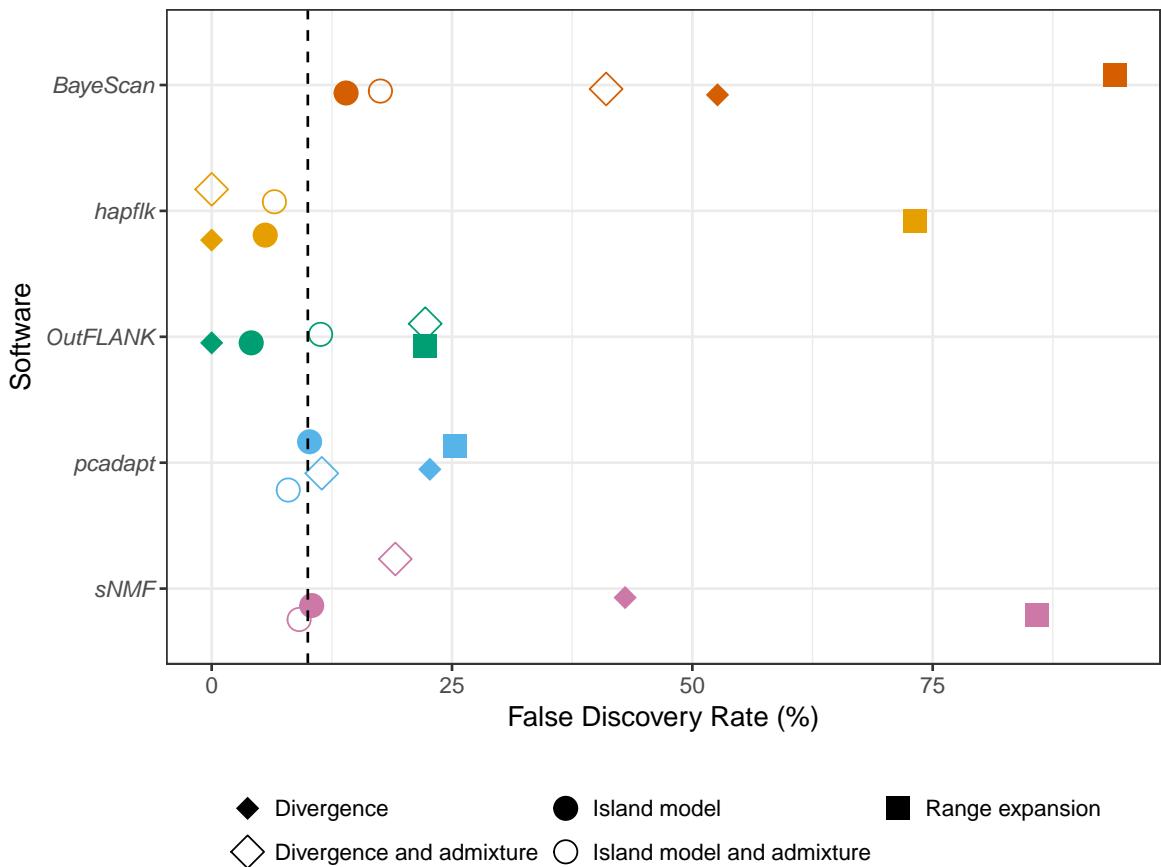


FIGURE 2.20 – Control of the FDR for different computer programs for genome scans. We find that the median proportion of false discoveries is around the nominal FDR set at 10% (6% for *hapflk*, 11% for both *OutFLANK* and *pcadapt* and 19% for *sNMF*) with the exception of *BayeScan* that generates 41% of false discoveries.

We then evaluate the proportion of false discoveries obtained with *BayeScan*, *hapflk*, *OutFLANK* and *sNMF* (Fig. 2.20). We find that *hapflk* is the most conservative approach (FDR = 6%) followed by *OutFLANK* and *pcadapt* (FDR = 11%). The computer program *sNMF* is more liberal (FDR = 19%) and *BayeScan* generates the largest proportion of false discoveries (FDR = 41%). When not recalibrating the *P*-values of *hapflk*, we find that the test is even more conservative (results not shown). For all programs, the range expansion scenario is the one that generates the largest proportion of false discoveries. Proportion of false discoveries under range expansion ranges from 22% (*OutFLANK*) to 93% (*BayeScan*).

Statistical power

To provide a fair comparison between methods and computer programs, we compare statistical power for equal values of the observed proportion of false discoveries. Then we compute statistical power averaged over observed proportion of false discoveries ranging from 0% to 50%.

We first compare statistical power obtained with the different statistical methods that have been implemented in *pcadapt* (Table 2.2). For the island model, Bayes factor, communality statistic and Mahalanobis distance have similar power (Fig. 2.21). For the divergence model, the power obtained with Mahalanobis distance is 20% whereas the power obtained with the communality statistic and with the Bayes factor is, respectively, 4% and 2% (Fig. 2.21). Similarly, for range expansion, the power obtained with Mahalanobis distance is 46% whereas the power obtained with the communality statistic and with the Bayes factor is 34% and 13%. We additionally investigate to what extent increasing sample size in each population from 20 to 60 individuals affects power. For range expansion, the power obtained with the Mahalanobis distance hardly changes ranging from 44% to 47%. However, the power obtained with the other two statistics changes importantly. The power obtained with the communality statistic increases from 27% to 39% when increasing the sample size and the power obtained with the Bayes factor increases from 0% to 44%.

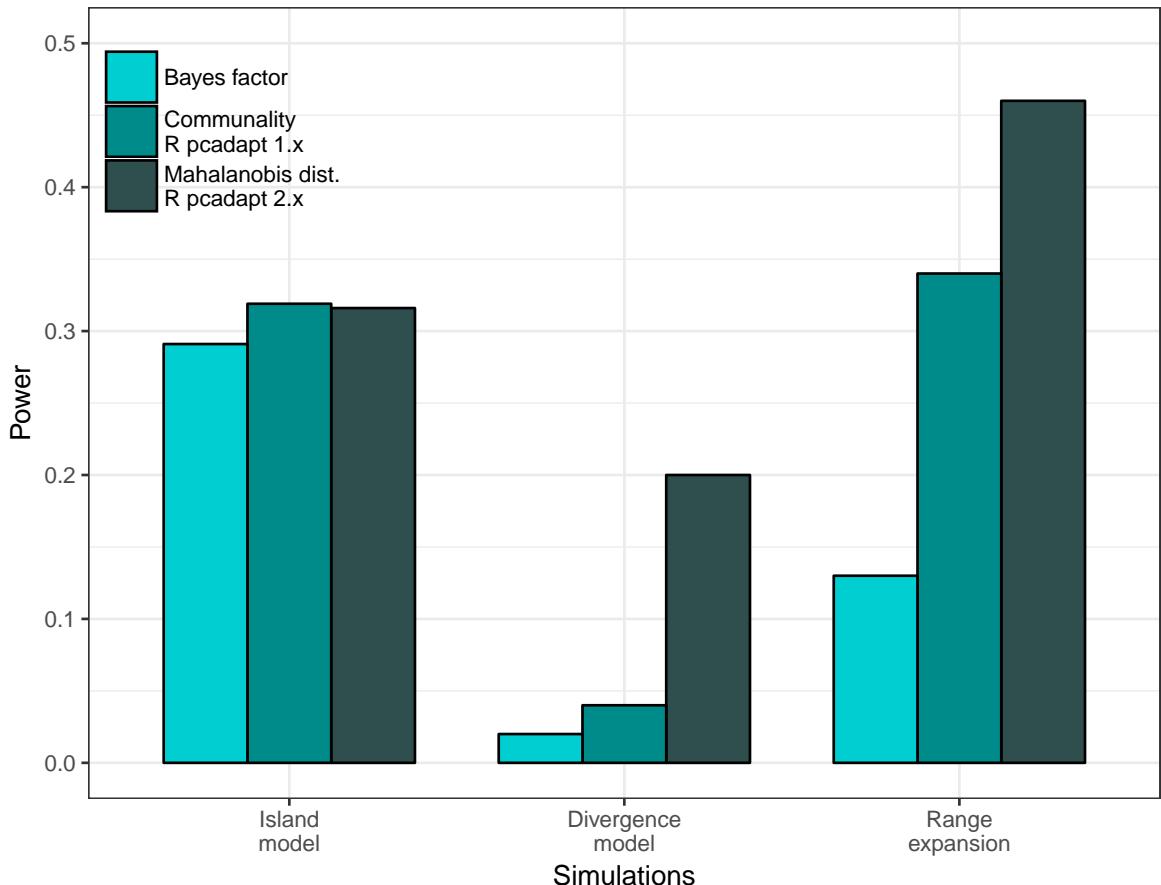


FIGURE 2.21 – Bayes factor corresponds to the test statistic implemented in the Bayesian version of *pcadapt* (Duforet-Frebourg et al., 2014) ; the communality statistic was the default statistic in version 1.x of the R package *pcadapt* (Duforet-Frebourg et al., 2015), and Mahalanobis distances are available since the release of the 2.0 version of the package. When there is hierarchical population structure (divergence model and range expansion), the Mahalanobis distance provides more powerful genome scans compared with the test statistic previously implemented in *pcadapt*. The abbreviation dist. stands for distance. Statistical power is averaged over the observed proportion of false discoveries (ranging between 0% and 50%).

Then we describe our comparison of computer programs for genome scans. For the simulations obtained with the island model where there is no hierarchical population structure, the statistical power is similar for all programs (Figure B.18 and B.19). Including admixed individuals hardly changes their statistical power (Figure B.18).

Then, we compare statistical power in a divergence model where adaptation took place in one of the external branches of the population divergence tree. The programs *pcadapt* and *hapflk*, which account for hierarchical population structure, as well as *BayeScan* are the most powerful in that setting (Fig. 2.22 and Figure B.20). The values of power in decreasing order are of 23% for *BayeScan*, of 20% for *pcadapt*, of

17% for *hapflk*, of 7% for *sNMF* and of 1% for *OutFLANK*. When including admixed individuals, the power of *hapflk* and of *pcadapt* hardly decreases whereas the power of *BayeScan* decreases to 6% (Fig. 2.22).

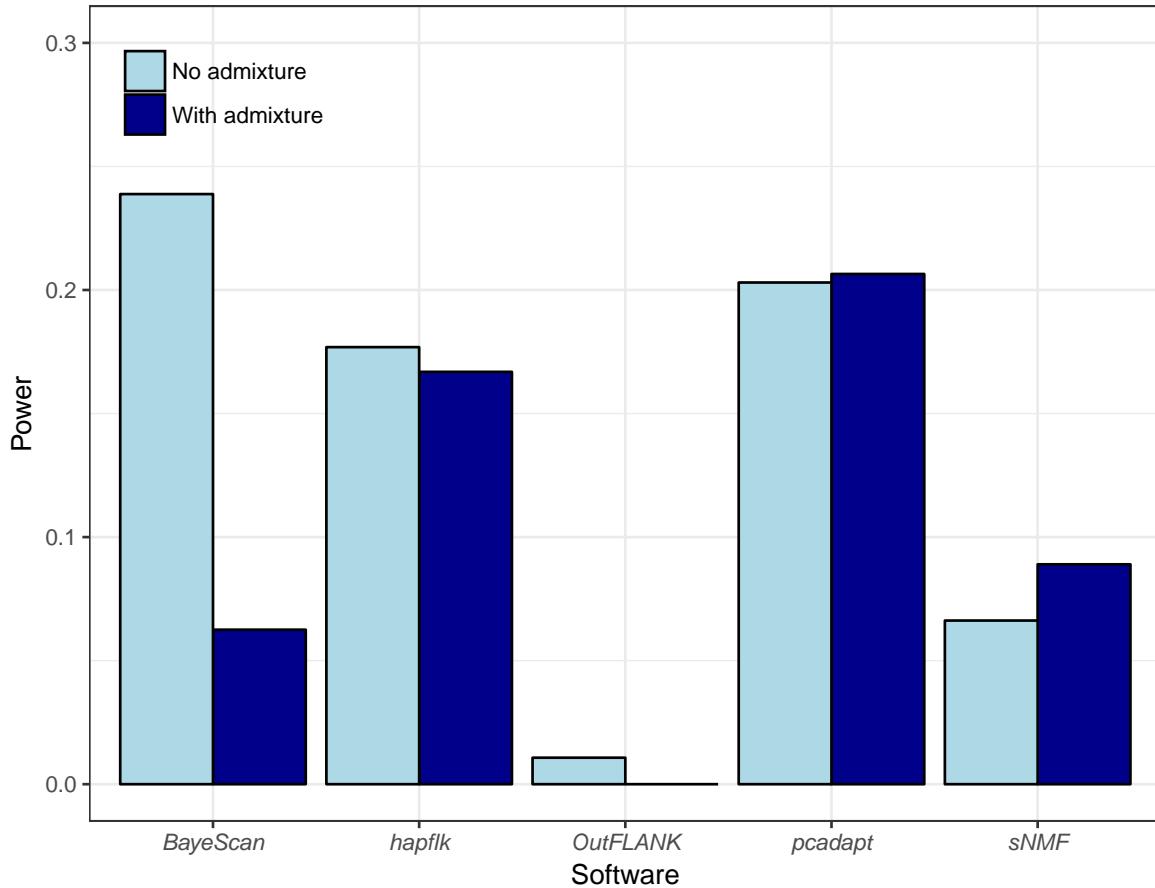


FIGURE 2.22 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the divergence model with three populations. We assume that adaptation took place in an external branch that follows the most recent population divergence event.

The last model we consider is the model of range expansion. The package *pcadapt* is the most powerful approach in this setting (Fig. 2.23 and B.21). Other computer programs also discover many true-positive loci with the exception of *BayeScan* that provides no true discovery when the observed FDR is smaller than 50% (Fig. 2.23 and B.21). The values of power in decreasing order are of 46% for *pcadapt*, of 41% for *hapflk*, of 37% for *OutFLANK*, of 30% for *sNMF* and of 0% for *BayeScan*.

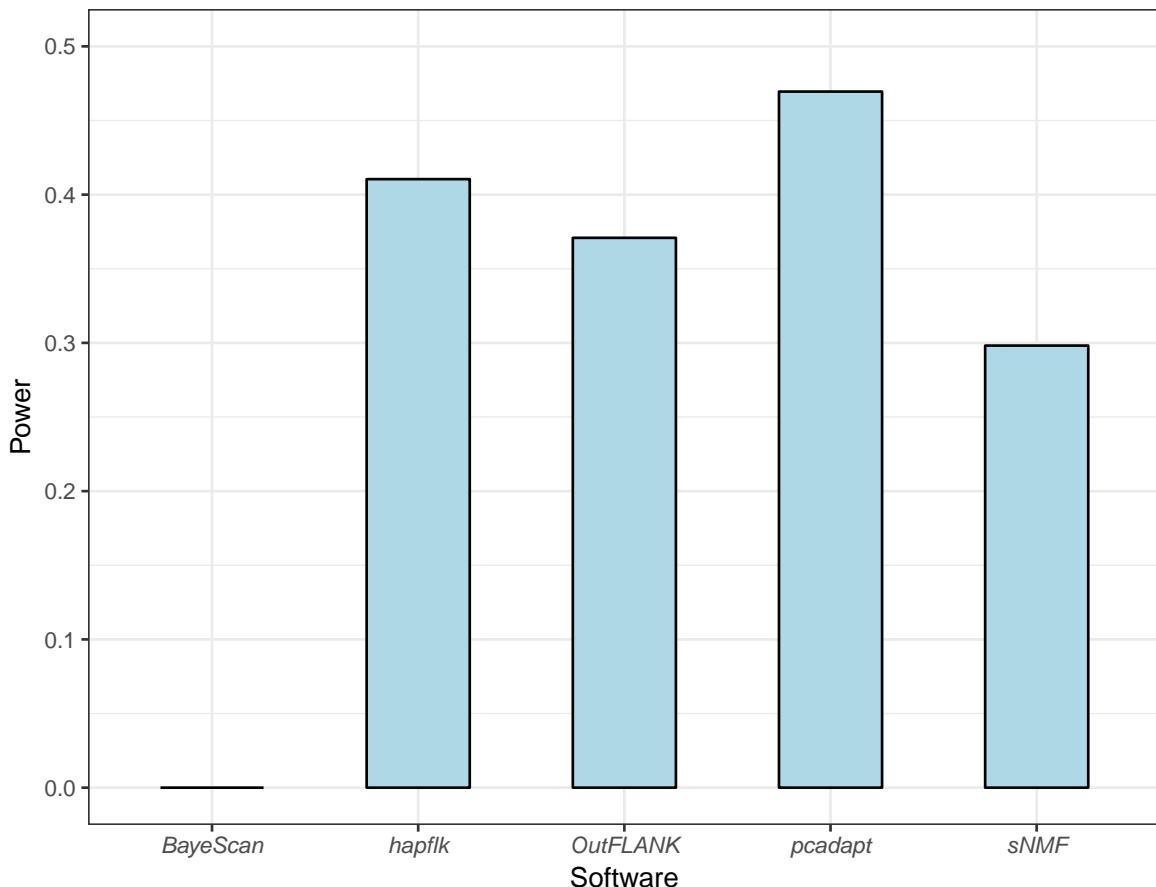


FIGURE 2.23 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for a range expansion model with two refugia. Adaptation took place during the recolonization event.

Running time of the different computer programs

Last, we compare running times. The characteristics of the computer we used to perform comparisons are the following : OSX El Capitan 10.11.3, 2,5 GHz Intel Core i5, 8 Go 1600 MHz DDR3. We discard *BayeScan* as it is too time-consuming. For instance, running *BayeScan* on a genotype matrix containing 150 individuals and 3000 SNPs takes 9h whereas it takes less than one second with *pcadapt*. The different programs were run on genotype matrices containing 300 individuals and from 500 to 50 000 SNPs. *OutFLANK* is the computer program for which the runtime increases the most rapidly with the number of markers. *OutFLANK* takes around 25 min to analyse 50 000 SNPs (Figure B.22). For the other 3 computer programs (*hapflk*, *pcadapt*, *sNMF*), analysing 50 000 SNPs takes less than 3 min.

Discussion

The R package *pcadapt* implements a fast method to perform genome scans with next-generation sequencing data. It can handle data sets where population structure is continuous or data sets containing admixed individuals. It can handle missing data as well as pooled sequencing data. The 2.0 and later versions of the R package implements a robust Mahalanobis distance as a test statistic. When hierarchical population structure occurs, Mahalanobis distance provides more powerful genome scans compared with the communality statistic that was implemented in the first version of the package (Duforet-Frebourg et al., 2015). In the divergence model, adaptation occurs along an external branch of the divergence tree that corresponds to the second principal component. When outlier SNPs are not related to the first principal component, the Mahalanobis distance provides a better ranking of the SNPs compared with the communality statistic.

Simulations show that the R package *pcadapt* compares favourably to other computer programs for genome scans. When data were simulated under an island model, population structure is not hierarchical because genetic differentiation is the same for all pairs of populations. Statistical power and control of the FDR were similar for all computer programs. In the presence of hierarchical population structure (divergence model) where genetic differentiation varies between pairs of populations, the ranking of the SNPs depends on the computer program. *pcadapt* and *hapflk* provide the most powerful scans whether or not simulations include admixed individuals. *OutFLANK* implements a F_{ST} statistic and because adaptation does not correspond to the most differentiated populations, it fails to capture adaptive SNPs (Fig. 2.22) (Bonhomme et al., 2010; Duforet-Frebourg et al., 2015). *BayeScan* does not assume equal differentiation between all pairs of populations, which may explain why it has a good statistical power for the divergence model. However, its statistical power is severely impacted by the presence of admixed individuals because its power decreases from 24% to 6% (Fig. 2.22). Understanding why *BayeScan* is severely impacted by admixture is out of the scope of this study. In the range expansion model, *BayeScan* returns many null q -values (between 376 and 809 SNPs of 9899 neutral and 100 adaptive SNPs) such that the observed FDR is always larger than 50%. Overall, we find that *pcadapt* and *hapflk* provide comparable statistical power. They provide optimal or near optimal ranking of the SNPs in different scenarios including hierarchical population structure and admixed individuals. The main difference between the two computer programs concerns the control of the FDR because *hapflk* is found to be more conservative.

Because NGS data become more and more massive, careful numerical implementation is crucial. There are different options to implement PCA and *pcadapt* uses a numerical routine based on the computation of the covariance matrix Ω . The algorithmic complexity to compute the covariance matrix is proportional to pn^2 where p is the number of markers and n is the number of individuals. The computation of the first K eigenvectors of the covariance matrix Ω has a complexity proportional to n^3 . This second step is usually more rapid than the computation of the covariance because the number of markers is usually large compared with the number of individuals. In brief, computing the covariance matrix Ω is by far the most costly operation when compu-

ting principal components. Although we have implemented PCA in C to obtain fast computations, an improvement in speed could be envisioned for future versions. When the number of individuals becomes large (e.g. $n \geq 10000$), there are faster algorithms to compute principal components (Abraham & Inouye, 2014; Halko, Martinsson, & Tropp, 2011). In addition to running time, numerical implementations also impact the effect of missing data on principal components (Dray & Josse, 2015). Achieving a good trade-off between fast computations and accurate evaluation of population structure in the face of large amount of missing data is a challenge for modern numerical methods in molecular ecology.

Acknowledgements

This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the ANR AGRHUM project (ANR-14-CE02-0003-01). We want to thank two anonymous reviewers and Stéphane Dray for their critical reading of our manuscript.

K.L., E.B. and M.G.B.B. designed and performed the research.

Data accessibility

Island and divergence model data : doi : 10.5061/dryad.8290n.

Range expansion simulated data : doi : 10.5061/dryad.mh67v. Files :

2R_R30_1351142954_453_2_NumPops=30_NumInd=20
2R_R30_1351142954_453_2_NumPops=30_NumInd=60
2R_R30_1351142970_988_6_NumPops=30_NumInd=20
2R_R30_1351142970_988_6_NumPops=30_NumInd=60
2R_R30_1351142986_950_10_NumPops=30_NumInd=20
2R_R30_1351142986_950_10_NumPops=30_NumInd=60

Chapitre 3

Introgression adaptative

Dans la partie précédente nous nous sommes intéressés à la détection d'allèles ayant favorisé l'adaptation d'une population à son environnement. Cependant, les scans génomiques pour la sélection ne permettent pas de comprendre l'origine du mécanisme adaptatif associé. Nous avons vu en introduction que la diversité allélique jouait un rôle important dans les processus d'adaptation locale, en offrant aux populations la possibilité de puiser dans un catalogue d'allèles plus vaste et d'augmenter ainsi leur potentiel adaptatif. Pour cette raison, l'introgression est désormais considérée comme un mécanisme d'adaptation important (Hedrick, 2013). Nous rappelons que *l'introgression adaptative* correspond à la sélection d'allèles transmis par flux de gènes ou par hybridation, ce qui permet de distinguer deux scénarios évolutifs distincts que nous décrivons ici.

Le premier est un scénario de métissage où deux populations apparentées s'hybrident en donnant naissance à une nouvelle population, si bien que certains allèles provenant d'une des populations parentales peuvent être sélectionnés pour l'environnement dans lequel évolue la population métissée (Figure 3.1).

Le second scénario suppose l'entrée en contact d'une population dite donneuse avec une population dite receveuse, permettant à la population receveuse d'intégrer de nouveaux allèles (Figure 3.1).



FIGURE 3.1 – Scénarios d'introgression. À gauche une représentation schématique du modèle de métissage. À droite une représentation schématique du modèle à flux de gènes. Dans les deux cas, la population jaune correspond à la population ayant bénéficié de gènes par introgression de la part de la population bleue.

Nous présentons dans ce chapitre différents outils statistiques (Table 3.1) pour la détection de signaux d'introgression pour chacun des deux scénarios (Figure 3.1), ainsi

TABLE 3.1 – Liste des méthodes et des logiciels présentés dans ce chapitre.

Méthode	Scénario	Référence
Loter	Métissage	Dias-Alves et al. (<i>in prep</i>)
HAPMIX	Métissage	Price et al. (2009)
RFMix	Métissage	Maples et al. (2013)
EILA	Métissage	Yang et al. (2013)
D	Flux de gènes	Durand et al. (2011)
Bd_f	Flux de gènes	Pfeifer et al. (2017)
f_d	Flux de gènes	Martin et al. (2014)
RND_{min}	Flux de gènes	Rosenzweig et al. (2016)

qu'une nouvelle approche basée sur l'Analyse en Composantes Principales valable pour les deux types de scénarios.

3.1 Introgression par métissage

Nous nous intéressons ici à la détection de signaux d'introgression dans une population métissée, à partir de la donnée de coefficients de métissage locaux. Nous avons vu dans le paragrahe 2.2.3 qu'il était possible d'estimer pour un individu, la proportion de son génome provenant d'un ou de plusieurs groupes génétiques. Ces proportions sont connues plus communément sous le nom de *coefficients de métissage* et peuvent servir à mieux comprendre l'histoire démographique des populations métissées. De nombreux logiciels existent pour l'estimation de ces coefficients : STRUCTURE, ADMIXTURE (Alexander, Novembre, & Lange, 2009), LEA (Frichot & François, 2015), tess3r (Caye, Deist, Martins, Michel, & François, 2016). Chaque individu peut de ce fait être vu comme une mosaïque d'allèles qui ont été puisés dans des groupes génétiques distincts. En revanche, les coefficients de métissage ne renseignent pas sur les portions du génome qui proviennent effectivement de ces groupes génétiques.

3.1.1 Coefficients de métissage locaux

Chercher à attribuer à chaque allèle la population ancestrale de laquelle il est originaire constitue la problématique d'estimation des *coefficients de métissage locaux* (Figure 3.2). Avoir à disposition une telle information individuelle et locale permet de quantifier à l'échelle de la population la probabilité qu'un allèle ait été fourni par une population spécifique. Ce type de méthodologie, basé sur l'estimation de coefficients de métissage locaux, a par exemple été employé pour la détection d'introgression adaptative chez les peupliers d'Amérique du Nord (Suarez-Gonzalez, 2016). Encore une fois, plusieurs logiciels ont été proposés dans le but d'estimer ces coefficients locaux : HAPMIX (Price et al., 2009), EILA (J. J. Yang, Li, Buu, & Williams, 2013),

LAMP (Thornton & Bermejo, 2014), Loter (Dias-Alves et al., *in prep*) ou encore RFMix (Maples, Gravel, Kenny, & Bustamante, 2013).

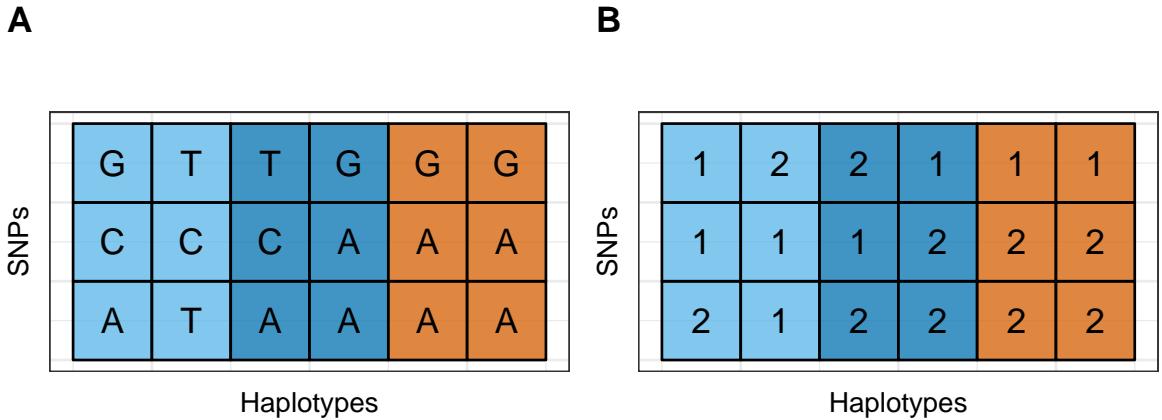


FIGURE 3.2 – Exemple de matrice de coefficients de métissage locaux pour des individus issus du métissage de deux populations sources (numérotées 1 et 2). **A.** Chaque ligne de la matrice correspond à un SNP et chaque haplotype est représenté par une colonne. **B.** Matrice de coefficients de métissage locaux individuels. Pour chaque nucléotide d'un haplotype, les méthodes d'estimation de coefficients de métissage locaux vont chercher à attribuer ce nucléotide soit à la population 1 soit à la population 2.

3.1.2 Statistique de test

À partir de la matrice de coefficients de métissage locaux individuels, il est possible d'estimer pour un SNP donné la proportion d'allèles provenant de chacune des populations sources considérées. Pour estimer la proportion d'allèles provenant de la population i pour le locus j , il suffit de calculer le nombre moyen de fois qu'un allèle observé sur le locus j a été attribué à la population i (un même individu présente deux allèles sur un locus donné (Figure 3.2)). Notant M la matrice de coefficients de métissage locaux individuels (Figure 3.2, panel B), G la matrice de génotypes (précisons que M a deux fois plus de colonnes que G car elle est de même dimension que la matrice des haplotypes) et n le nombre d'individus métissés, la proportion d'allèles p_{jk} provenant de la population P_k au locus j s'écrit :

$$p_{jk} = \frac{1}{2n} \sum_{i=1}^n \left(\delta_{M_{i,2j-1}}^k + \delta_{M_{i,2j}}^k \right), \quad (3.1)$$

où $\delta_x^k = 1$ si $x = k$ et 0 sinon, $j = 1, \dots, p$ où p est le nombre de locus et $k = 1, \dots, K$ où K est le nombre de populations ancestrales. La proportion p_{jk} désigne le coefficient de métissage local au locus j à l'échelle de la population métisse. La recherche de régions d'introgression dans la population hybride se fait ensuite en regardant les coefficients de métissage locaux significativement élevés par rapport aux autres. Il

n'y a pas de valeur standard pour le seuil de significativité, Suarez-Gonzalez (2016) choisissent par exemple un seuil de significativité de 3 écarts-types tandis que B. vonHoldt, Fan, Ortega-Del Vecchyo, Wayne, & others (2017) optent pour un seuil de significativité de 2 écarts-types.

3.2 Introgression par flux de gènes

La seconde classe de méthodes repose sur la comparaison de séquences de nucléotides, nucléotide par nucléotide. Un chromosome est par exemple divisé en un certain nombre de fenêtres (qui peuvent être disjointes ou chevauchantes), et les statistiques sont calculées sur chacune de ces fenêtres. Les séquences de nucléotides sont vues comme des chaînes de caractères et leur dissimilarité est mesurée à l'aide de distances classiques (ou de mesures de comptage) telles que la distance de Hamming.

3.2.1 Diversité nucléotidique par paires de séquences

Soient $x = (x_i)_{1 \leq i \leq n}$ et $y = (y_i)_{1 \leq i \leq n}$ deux séquences de nucléotides. En particulier, $\forall i \in [|1, n|]$, $x_i, y_i \in \{A, C, T, G\}$. La distance de Hamming d_{xy} entre la séquence x et la séquence y est définie par :

$$d_{xy} = \text{Card}(\{i \in [|1, n]| x_i \neq y_i\}). \quad (3.2)$$

Pour étudier les similarités localement sur le génome à l'échelle de populations, deux mesures naturelles peuvent être constituées à partir de d_{xy} , et consistent à calculer la distance moyenne et la distance minimale sur des séquences homologues provenant de différentes populations. Si X (resp. Y) désigne l'ensemble des séquences de la population P_1 (resp. P_2), la distance moyenne et la distance minimale entre X et Y sont données par :

$$\begin{aligned} d_{XY} &= \frac{1}{\text{Card}(X)\text{Card}(Y)} \sum_{x \in X} \sum_{y \in Y} d_{xy}, \\ d_{\min} &= \min_{(x,y) \in X \times Y} d_{xy}. \end{aligned} \quad (3.3)$$

d_{XY} est appelée la *pairwise nucleotide diversity* et estime le nombre moyen de nucléotides différents entre une séquence venant de la population X et une séquence venant de la population Y . d_{\min} estime la distance minimale entre deux séquences tirées de P_1 et de P_2 . Notons que d_{XY} et d_{\min} étant symétriques, X et Y jouent des rôles équivalents, si bien que l'utilisation de ces distances pour détecter l'introgression ne permet pas de conclure sur le sens du flux de gènes que l'on cherche à mettre en évidence. De plus, pour détecter l'introgression entre deux populations X et Y à partir de distances entre séquences nucléotidiques, la seule donnée de d_{XY} ou de d_{\min} ne suffit pas, une distance n'étant une mesure de proximité que si elle est mise en comparaison avec une autre distance. C'est pour cela que les statistiques d'introgression basées sur ce type de distance requièrent généralement la donnée d'une population de contrôle O

(appelée aussi *outgroup*). C'est le cas par exemple des statistiques *RND* (Feder et al., 2005) et *RND min* (Rosenzweig, Pease, Besansky, & Hahn, 2016) définies ci-dessous :

$$\begin{aligned} RND &= \frac{2d_{XY}}{d_{XO} + d_{YO}}, \\ RND \min &= \frac{2d_{\min}}{d_{XO} + d_{YO}}. \end{aligned} \quad (3.4)$$

Dans le cas de *RND min*, une séquence sera donc détectée comme étant introgressée pour de faibles valeurs de *RND min*.

3.2.2 Le modèle ABBA-BABA

L'emploi de ce modèle suppose l'accès à des séquences nucléotidiques provenant de quatre populations :

- une population donneuse.
- deux populations susceptibles de se reproduire avec la population donneuse.
- une population de contrôle avec qui toute possibilité de métissage est exclue.

Ces populations sont généralement présentées selon le schéma suivant (Figure 3.3) :

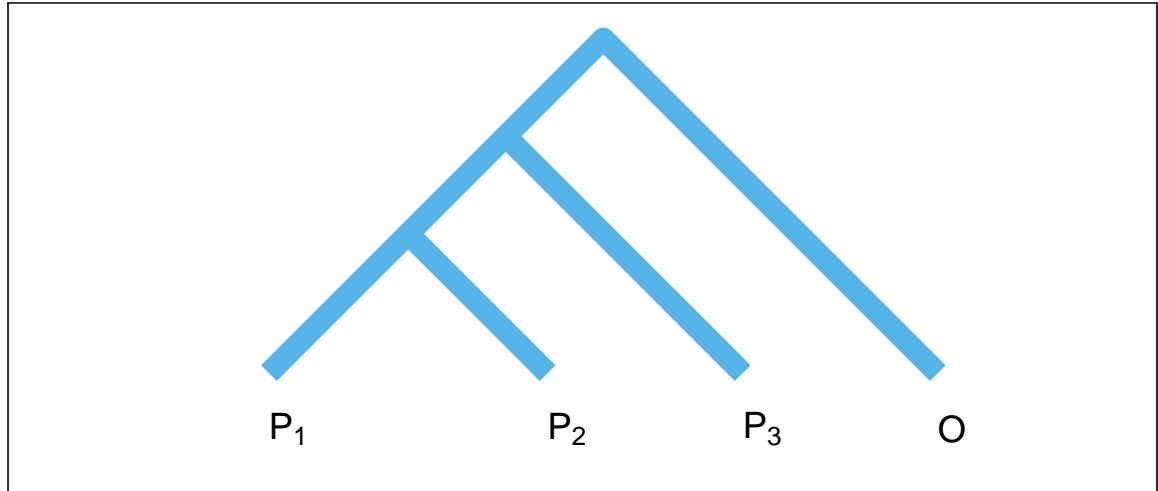


FIGURE 3.3 – Arbre allélique. O désigne la population de contrôle, P_3 la population donneuse, P_1 et P_2 les deux populations susceptibles de recevoir des allèles provenant de P_3 .

Sous l'hypothèse nulle, c'est-à-dire en l'absence d'introgression, la proportion de nucléotides communs à P_3 et P_1 doit être la même que la proportion de nucléotides communs à P_3 et à P_2 . À l'inverse, un événement d'introgression entre P_3 et P_1 devrait s'accompagner de l'augmentation de la fraction de séquences partagées par P_3 et P_1 , relativement à P_2 . Cette différence de proportions est mesurée par la statistique D de Patterson (Durand, Patterson, Reich, & Slatkin, 2011), conçue spécifiquement pour détecter l'introgression de manière globale sur le génome. De façon plus imagée, notant A et B respectivement les allèles portés par O et par P_3 , la statistique D compte le

nombre de motifs *ABBA* (Figure 3.4) et le nombre de motifs *BABA* (Figure 3.4) et teste si les deux quantités sont significativement différentes l'une de l'autre.

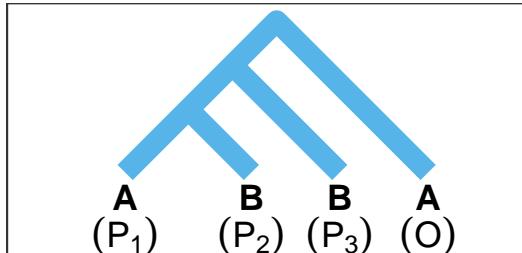
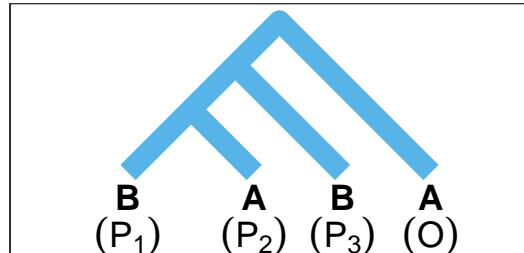
A**B**

FIGURE 3.4 – **A.** À gauche un arbre allélique présentant un motif *ABBA* indiquant que l'allèle *B* présent dans la population *P₂* a été hérité de la population *P₃*. **B.** À droite un arbre allélique présentant un motif *BABA* indiquant que l'allèle *B* présent dans la population *P₁* a été hérité de la population *P₃*.

La statistique *D* de Patterson, définie à partir des motifs *ABBA* et *BABA*, est donnée par la relation suivante :

$$D = \frac{\sum_i C_{ABBA}(i) - C_{BABA}(i)}{\sum_i C_{ABBA}(i) + C_{BABA}(i)} \quad (3.5)$$

où $C_{ABBA}(i)$ (resp. $C_{BABA}(i)$) désigne le nombre de motifs *ABBA* (resp. *BABA*) observés sur le site i . Selon S. H. Martin, Davey, & Jiggins (2014), cette statistique présente le défaut de trouver des outliers dans des régions à faible diversité allélique qui ne sont pas nécessairement introgressées. Partant de ce constat, de nombreuses variantes ont été développées à partir de cette statistique, dont les statistiques Bd_f (Pfeifer & Kapan, 2017) et f_d (S. H. Martin et al., 2014) auxquelles nous nous comparons, afin de corriger les défauts de la statistique D .

3.3 Une nouvelle statistique pour les scans d'introgression

3.3.1 Analyse en Composantes Principales locale

Les méthodes de détection d'introgression que nous avons présentées, aussi bien dans un scénario de métissage que dans un scénario à flux de gènes (Figure 3.1), reposent sur le même principe, à savoir que dans une zone d'introgression les motifs observés diffèrent significativement de ce qui est observé à l'échelle globale (soit en termes de coefficients de métissage, soit en termes de motifs d'arbre allélique).

Nous proposons dans ce paragraphe d'utiliser les scores de l'ACP en tant que motif de comparaison, ceci étant justifié par le fait que le métissage à l'échelle du génome peut être estimé à partir des scores des individus métissés (G. McVean, 2009). Notant B le barycentre des scores des individus d'une population métissée P , et B_1 (resp. B_2) le barycentre des scores des individus appartenant à la population P_1 (resp. P_2), nous pouvons estimer la proportion globale de métissage de P relativement à P_1 et P_2 en calculant les coordonnées barycentriques de B dans le repère (B_1, B_2) , c'est-à-dire en déterminant les coefficients q_1 et q_2 tels que $B = q_1 B_1 + q_2 B_2$ avec $q_1 + q_2 = 1$.

À l'échelle locale, nous pourrions également définir de tels coefficients barycentriques, à la condition de pouvoir définir des *scores locaux*. Dans la littérature, nous trouvons deux façons de définir de tels scores :

- les scores de l'ACP réalisée sur des sous-ensembles de SNPs (H. Li & Ralph, 2016 ; B. vonHoldt et al., 2017). Cette approche revient à fenêtrer le génome et à réaliser l'ACP sur chacune des fenêtres.
- les scores locaux de l'ACP. Cette approche consiste à déterminer les loadings en effectuant une ACP classique sur l'ensemble du génome, puis les scores locaux sont obtenus en ne prenant en compte que les loadings de la fenêtre considérée. Dit autrement, cela revient à calculer les coefficients de la régression de \tilde{G} par un sous-ensemble de loadings ΣV^T où $\tilde{G} \simeq U \Sigma V^T$ est la décomposition en valeurs singulières de la matrice de génotypes normalisée. Ces coefficients de régression sont par ailleurs appelés *scores de métissage* par Brisbin et al. (2012).

Nous choisissons ici d'utiliser la seconde approche basée sur les coefficients de régression, principalement parce que la première n'est pas adaptée au cas de données génétiques denses. Une approche par fenêtre glissante nécessiterait de réaliser un nombre d'ACP locales équivalent au nombre de marqueurs présents, soit une complexité en $O(np^2)$ (où n désigne le nombre d'individus et p le nombre de marqueurs génétiques), contrairement à la seconde méthode qui peut être implémentée en temps linéaire par rapport à n et à p . Notant i un entier compris entre 1 et p , et x_i la position génétique (mesurée en Morgans) ou la position physique (mesurée en paires de bases) du i -ème marqueur génétique. Nous définissons pour cet entier i la fenêtre W_i^T de taille T et centrée en i par :

$$W_i^T = \{j \in [1, p], |x_i - x_j| \leq T/2\} \quad (3.6)$$

Soit $\tilde{G} = U \Sigma V^T$ la décomposition en valeurs singulières de la matrice de génotypes normalisée \tilde{G} . Partant de l'approximation $U \simeq \tilde{G} V \Sigma^{-1}$, les scores de métissage sont définis de la façon suivante (Brisbin et al., 2012) :

$$U_{W_i^T} = \tilde{G}_{:, W_i^T} V_{W_i^T, :} \Sigma^{-1} \quad (3.7)$$

Les proportions de métissage local de la population $B_{W_i^T}$ sont définies de la même façon que dans le cas global, à la différence qu'elles utilisent les scores locaux $U_{W_i^T}$ à la place des scores globaux U . Ainsi, une région sera considérée comme potentiellement introgressée si les proportions de métissage locale calculées sur cette région sont significativement différentes des proportions de métissage calculées à l'échelle du génome.

Calcul des scores locaux

Si nous calculions $U_{W_i^T}$ pour chaque valeur de i de façon naïve, la complexité algorithmique de cette étape serait en $O(nKp^2)$. Pour réduire cette complexité, l'idée est d'exploiter le fait que les fenêtres W_i^T et W_{i+1}^T se chevauchent, ce qui permet de déduire $U_{W_{i+1}^T}$ à partir de $U_{W_i^T}$ en effectuant moins d'opérations que n'en nécessiterait le calcul individuel de $U_{W_{i+1}^T}$.

3.3.2 Résultats principaux

Nous proposons une méthode spécifiquement dédiée à la détection d'introgression. Le développement d'une méthode destinée spécifiquement à la détection d'introgression est motivé par le fait que les méthodes basées sur l'estimation de coefficients de métissage nécessitent de disposer d'informations parfois difficiles à obtenir comme la phase des haplotypes. Le problème de la détection d'introgression étant plus simple que celui de l'estimation des coefficients de métissage locaux, nous proposons une approche qui permet de s'affranchir d'étapes relativement lourdes telles que la détermination des haplotypes et de leur phase, allégeant considérablement le temps de calcul nécessaire. À l'aide de simulations numériques réalisées pour les deux types de scénarios d'introgression présentés, nous montrons que notre statistique produit un taux de fausses découvertes et une puissance comparables aux méthodes de l'état de l'art (Figure 3.7), en utilisant seulement la donnée de génotypes.

Application au jeu de données de peupliers. Nous avons testé notre méthode sur un jeu de données de peupliers d'Amérique du Nord (Suarez-Gonzalez, 2016). Deux populations de peupliers ont été génotypées, *Populus balsamifera* qui est une population adaptée aux conditions boréales et *Populus trichocarpa* qui est une population adaptée au climat doux du Nord-Ouest américain (Suarez-Gonzalez, 2016). Comme Suarez-Gonzalez (2016), nous avons regardé, chez les individus métisses avec une ascendance génétique *Trichocarpa* majoritaire, les régions de leur génome présentant un excès d'ascendance génétique *Balsamifera* qui leur permet de s'adapter à des conditions plus froides et sèches. Nous trouvons la même région d'introgression sur le chromosome 6 que Suarez-Gonzalez (2016) (Figure 3.5). Sur le chromosome 15 nous trouvons également les mêmes régions candidates. En revanche, nous trouvons deux nouvelles régions candidates sur le chromosome 12 (Figure 3.6), qui sont également détectées par RFMix (Maples et al., 2013). L'une d'entre elles est également détectée avec Loter. Cette différence de résultats peut être due au prétraitement des données de génotypes. Suarez-Gonzalez (2016) ont appliqué des filtres enlevant près de 95% du jeu de données initial. De notre côté, nous avons choisi de filtrer les SNPs ayant plus de 5% de données manquantes, ce qui permettait de conserver plus de 70% du jeu de données initial.

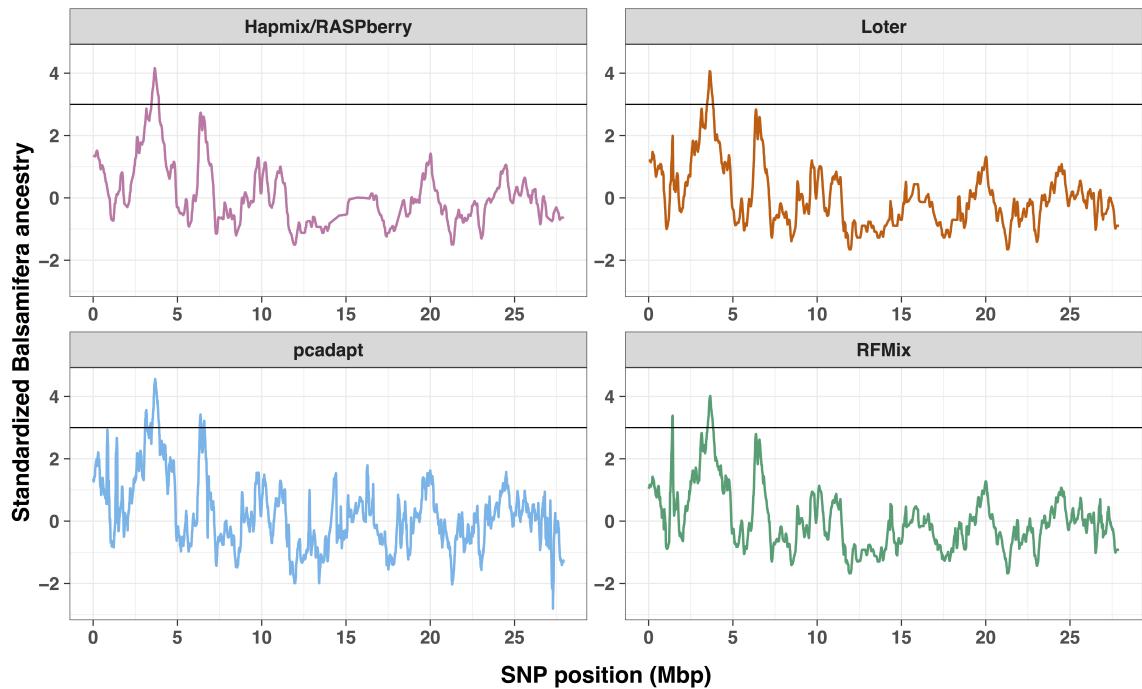


FIGURE 3.5 – Résultats du scan d’introgression obtenus sur le chromosome 6 du jeu de données de peupliers (Suarez-Gonzalez, 2016) avec différentes méthodes basées sur l’estimation des coefficients de métissage locaux.

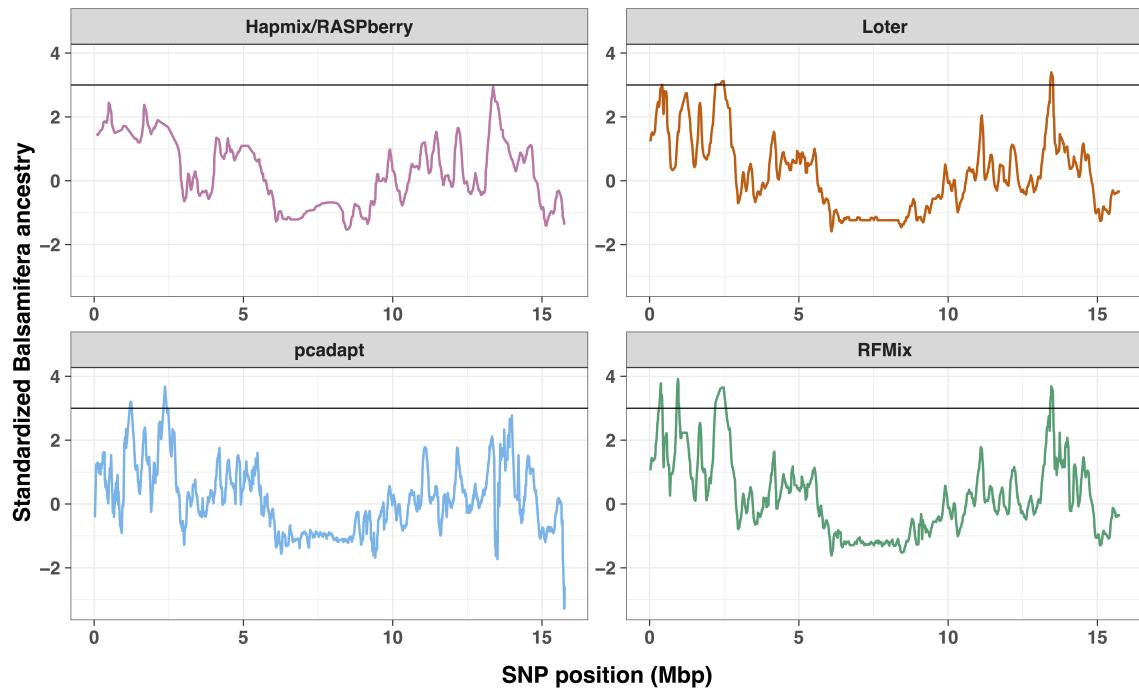


FIGURE 3.6 – Résultats du scan d’introgression obtenus sur le chromosome 12 du jeu de données de peupliers (Suarez-Gonzalez, 2016) avec différentes méthodes basées sur l’estimation des coefficients de métissage locaux.

Résultats sur des simulations de modèle de métissage. Pour quantifier les performances de chaque méthode basée sur l’estimation des coefficients de métissage locaux, nous avons réalisé des simulations à partir du jeu de données de peupliers décrit ci-dessus (Suarez-Gonzalez, 2016). La simulation d’haplotypes d’individus métissés est réalisée à partir des deux populations ancestrales qui y sont présentes. Chacune des simulations est constituée de 50 haplotypes de la souche continentale, de 50 haplotypes de la souche boréale, ainsi que de 50 haplotypes d’individus hybrides générés à partir des haplotypes ancestraux. Ces haplotypes ancestraux ont été estimés à l’aide du logiciel Beagle (S. R. Browning & Browning, 2007). En suivant une procédure que nous ne décrivons pas ici (cf. article 3), nous choisissons pour chaque simulation, de façon aléatoire, une région qui sera sujette à un évènement d’introgression. Nous donnons en figure 3.7 une comparaison des résultats obtenus avec les différents logiciels d’estimation des coefficients de métissage locaux. Nous constatons que notre méthode, basée uniquement sur l’information génotypique, obtient des résultats similaires aux méthodes nécessitant la donnée d’haplotypes.

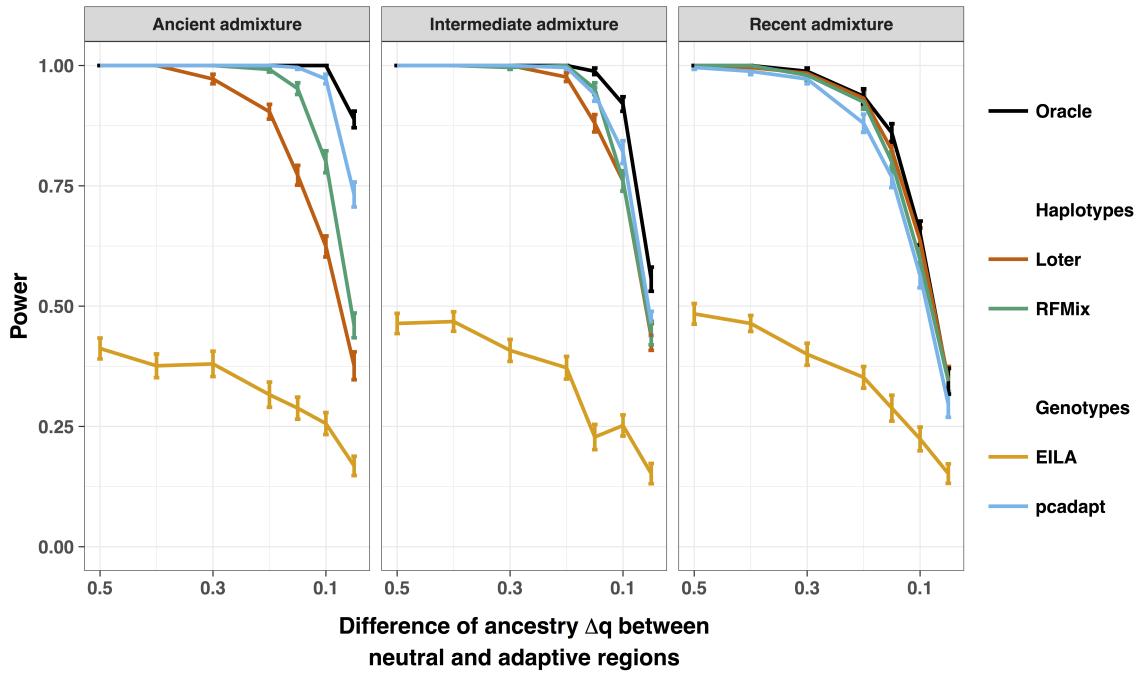


FIGURE 3.7 – Proportion de régions effectivement introgressées sur les 5 régions jugées les plus significatives pour chaque méthode (EILA, Loter, RFMix et pcadapt) dans un scénario de métissage à deux populations. L’intensité de l’introgression est paramétrée par la variable Δ_q donnée en abscisse. De gauche à droite, le résultat de la comparaison pour des épisodes de métissage survenus il y a respectivement 1000, 100 et 10 générations.

Résultats sur des simulations de modèle de flux de gènes. De la même manière, pour évaluer notre méthode dans un scénario à flux de gènes, nous reprenons les modèles de simulation proposés par S. H. Martin et al. (2014). L’idée est la même que précédemment. À l’aide de ces modèles nous générerons des séquences de nucléotides pour quatre populations P_1 , P_2 , P_3 , et O dont l’histoire démographique est celle racontée en figure 3.3. Pour la population receveuse P_2 , nous choisissons une région dans laquelle les séquences de P_2 sont mélangées à des séquences de la population donneuse P_3 , simulant de cette manière un évènement d’introgression. Les résultats de cette comparaison sont donnés en figure 3.8. Nous constatons que dans le cas où le taux de mutation est constant le long du génome, notre méthode obtient des performances similaires aux autres méthodes. En revanche, dans le cas où le taux de mutation est variable, elle semble être la moins robuste de toutes.

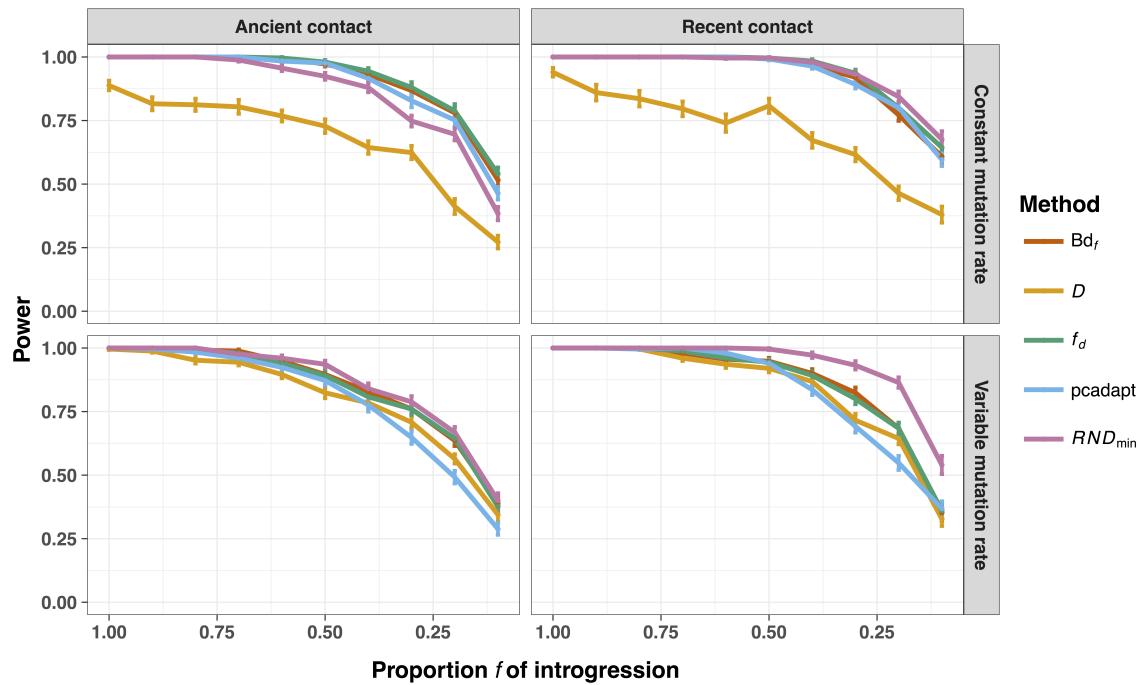


FIGURE 3.8 – Proportion de régions effectivement introgressées sur les 5 régions jugées les plus significatives pour chaque méthode (Bd_f , D , f_d , pcadapt, RND_{min}) dans un scénario à flux de gènes. Ici l'intensité de l'introgression est paramétrée par la variable f .

3.4 Article 3

Scanning genomes for adaptive introgression using principal component analysis

Keurcien Luu, Thomas Dias-Alves & Michael G.B. Blum

Introduction

A potential source of adaptive genetic variation is adaptive introgression. Adaptive introgression occurs when selectively beneficial alleles are transferred between species (M. L. Arnold & Martin, 2009; Dasmahapatra et al., 2012; Fraïsse, Roux, Welch, & Bierne, 2014). Although adaptive genetic variation originates primarily from de novo mutation or standing variation, adaptive introgression is now acknowledged as an important source of genetic variation in plants but also in animals (Hedrick, 2013).

To model adaptive introgression, there are at least two conceptual evolutionary scenarios. In the first “admixture” scenario, two populations or two related species admixed to form a new hybrid population (Figure 3.9). Adaptive introgression occurs when variation from one of the source population is adaptive in the environment of the admixed population. For instance, Tibetans are the result of admixture between Han and Sherpa populations, and alleles that confer adaptation to high altitude have been transmitted by the Sherpa population (Jeong et al., 2014). Evidence of adaptive introgression following admixture has been provided in several admixed populations or species, including Africanized honeybee, *Populus* species, North American canids, or human Bantu-speaking populations to name just a few examples (R. M. Nelson, Wallberg, Simões, Lawson, & Webster, 2017; Patin et al., 2017; Payseur & Rieseberg, 2016; Suarez-Gonzalez, 2016; B. M. vonHoldt, Kays, Pollinger, & Wayne, 2016). The second evolutionary scenario assumes “gene flow” or “introgression” from a donor species to an introgressed species (Figure 3.9). For instance, house mice became resistant to a warfarin pesticide because of selection on VKORC1 polymorphisms that have been acquired from the Algerian mouse (*M. spretus*) through introgression (Y. Song et al., 2011). Another classical example of introgression followed by adaption include colour adaptations in *Heliconius* species (Pardo-Diaz et al., 2012). Whereas detection of loci involved in adaptive introgression are performed with different statistical methods for the admixture and gene flow evolutionary models, we propose a statistical approach that is valid in both evolutionary models.

In the admixture model, outlier loci are found by looking for genomic regions harboring excess of ancestry from the donor population (Buerkle & Lexer, 2008). Excess of ancestry are obtained using local ancestry inference (LAI) that computes for admixed individuals the number of chromosome copies coming from the ancestral populations. There are several software available for LAI including EILA, HAPMIX, LAMP-LD, or RFMix (Baran et al., 2012; Jeong et al., 2014; Maples et al., 2013; Price et al., 2009). Except for EILA, LAI software can require biological information

that can be difficult to obtain especially for non-model species. For instance, HAPMIX or RFMix require phased haplotypes for the ancestral populations and RFMix also require phased haplotypes for the admixed individuals. In addition, a recombination map and an estimate of the admixture date between the ancestral populations should be provided for both HAPMIX and RFMix (Baran et al., 2012; Jeong et al., 2014; Maples et al., 2013; Price et al., 2009).

In the “gene flow” or “introgression” scenario, there have also been attempts to find introgressed regions with outlier detection methods (Rheindt, Fujita, Wilton, & Edwards, 2013; Smith & Kronforst, 2013; W. Zhang, Dasmahapatra, Mallet, Moreira, & Kronforst, 2016). A common approach is to test for an excess of shared derived variants in the donor and recipient populations using the D or ABBA-ABBA test statistic (Durand et al., 2011). However, extreme D values occur disproportionately in genomic regions with lower diversity (S. H. Martin et al., 2014). Because genome scans with the D statistic generate too many false positives, more powerful statistics have been developed such as the f_d or Bd_f statistic (S. H. Martin et al., 2014; Pfeifer & Kapan, 2017). In addition, the D statistic and its extensions cannot be used to identify introgressed regions between sister species (see Figure 3.9); it should be used with three or more lineages to detect the different topologies produced by hybridization (Rosenzweig et al., 2016). To identify introgression between sister species, various alternative statistics have been developed including RND_{\min} , which computes the minimum pairwise distance between the two sister species relative to the divergence to an outgroup. In contrast to the D statistic and its extension, RND_{\min} requires phased data (Rosenzweig et al., 2016).

We propose a statistical method to detect candidates for adaptive introgression that is valid in both the admixture scenario and the introgression scenario. The proposed approach is based on principal component analysis (PCA), which is well-suited to ascertain population structure of large scale genome-wide dataset (G. McVean, 2009; N. Patterson et al., 2006). Numerical solutions to compute principal component scores can be obtained rapidly even with large scale data (Abraham, Qiu, & Inouye, 2017; Duforet-Frebourg et al., 2015; Galinsky et al., 2016). The proposed approach does not require any biological information in addition to the genome-wide genotype data and should therefore be valuable especially for non-model species where recombination map are not available and where haplotype phasing can be a daunting task. The statistical method we propose ascertains population structure locally in the genome by computing local principal components. Candidate SNPs for local introgression correspond to genomic regions for which population structure of the admixed or introgressed population deviates significantly from genome-wide population structure (Figure 3.9). To capture population structure of the admixed or introgressed individuals with respect to the other individuals, the method computes an average ancestry coefficient. In the following, we investigate to what extent the PCA-based approach provides an alternative to LAI and to the aforementioned phylogenetic statistics in the admixture and introgression scenario. We show the potential of the method to detect adaptive introgression in a hybrid *Populus* species resulting from a 2-way admixture model (Suarez-Gonzalez, 2016). The PCA-based method is implemented in the R package *pcadapt* (Luu et al., 2017).

A PCA-based approach to detect local introgression

The objective of PCA is to find a new set of orthogonal variables called the principal components, which are linear combinations of (centered and standardized) allele counts, such that the projections of the data onto these axes lead to an optimal summary of the data. To present the method, we introduce the truncated singular value decomposition (SVD) that approximates the $(n \times p)$ centered and scaled genotype matrix \mathbf{Y} by a matrix of smaller rank

$$\mathbf{Y} \approx \mathbf{U}\Sigma\mathbf{V}^T, \quad (3.8)$$

where \mathbf{U} is a $(n \times K)$ orthonormal matrix, \mathbf{V} is a $(p \times K)$ orthonormal matrix, Σ is a diagonal $(K \times K)$ matrix and K corresponds to the rank of the approximation. The solution of PCA with K components can be obtained using the truncated SVD of equation (3.8) : the K columns of \mathbf{V} contain the loadings, which correspond to the contribution of each SNP to the PCs and the K columns of \mathbf{U} contain the PC *scores* that are usually displayed to visualize population structure (Duforet-Frebourg et al., 2015; N. Patterson et al., 2006).

To provide local measure of population structure, we compute local PCA scores. For a SNP j , we compute the vector of *local scores* \mathbf{U}_j that measure population structure locally in the genome

$$\mathbf{U}_j = \mathbf{Y}_j \mathbf{V}_j \Sigma^{-1}, \quad (3.9)$$

where \mathbf{Y}_j correspond to the genotype restrained to a genomic window around SNP j and \mathbf{V}_j corresponds to the loadings restrained to a genomic window around SNP j .

An average local ancestry coefficient q is then obtained from the local scores \mathbf{U}_j using barycentric coordinates for the barycenter B of the local scores of admixed or introgressed individuals. Barycentric coordinates of B correspond to the K coefficients q_1, \dots, q_K ($q_1 + \dots + q_K = 1$) such that B can be written as a linear combination $B = q_1 B_1 + \dots + q_K B_K$ where B_1, \dots, B_K are the barycenters of local scores for the K source populations. In the admixture scenario, barycentric coordinates are always between 0 and 1 because local scores of admixed individuals are always contained in the simplex determined by the barycenters of the ancestral populations (Figure B.29). By contrast, barycentric coordinates can be negative or larger than 1 for the introgression scenario (B.28). To find outlier regions with excess of ancestry from the donor population j that is one of the K source populations, we compute robust mean μ_j and variance σ_j^2 of the average ancestry coefficients using the median absolute deviation and we consider the standardized average ancestry coefficient $(q_j - \mu_j)/\sigma_j$, $j = 1, \dots, K$.

Materials and Methods

Parameter settings for local ancestry and adaptive introgression software

For the admixture scenario, we consider three software of local ancestry inference, which are EILA, Loter and RFMix (Maples et al., 2013; J. J. Yang et al., 2013). EILA considers genotypes as input and we consider a regularization parameter of $\lambda = 0.1$. Loter uses haplotypes as input data and has no tuning parameter. RFMix uses haplotypes as input data and we consider the default parameter except the window size that was set at 0.002cM. Using the default window size generates an error and as recommended by the manual of RFMix, we reduce window size to overcome this issue. When analyzing *Populus* data, we consider the same averaging strategy as implemented in the initial analysis with RASPBerry and average local ancestry coefficients obtained with RFMix and Loter using 100kbp windows (Suarez-Gonzalez, 2016).

For the introgression scenario, we compared pcadapt to the Bd_f , D , f_d , and RND_{\min} test statistics (Durand et al., 2011; S. H. Martin et al., 2014; Pfeifer & Kapan, 2017; Rosenzweig et al., 2016). The D , Bd_f and f_d statistics require genotype data for the three related species as well as from an outgroup species, for which we also simulate genotypes. Because RND_{\min} tests for introgression between sister species, we provide sequence (haplotype) data from 3 species, which are the introgressed, and donor species as well as an outgroup species (Figure 3.9). When using pcadapt, we provide genotype data from three species, which are introgressed, donor and sister species (Figure 3.9).

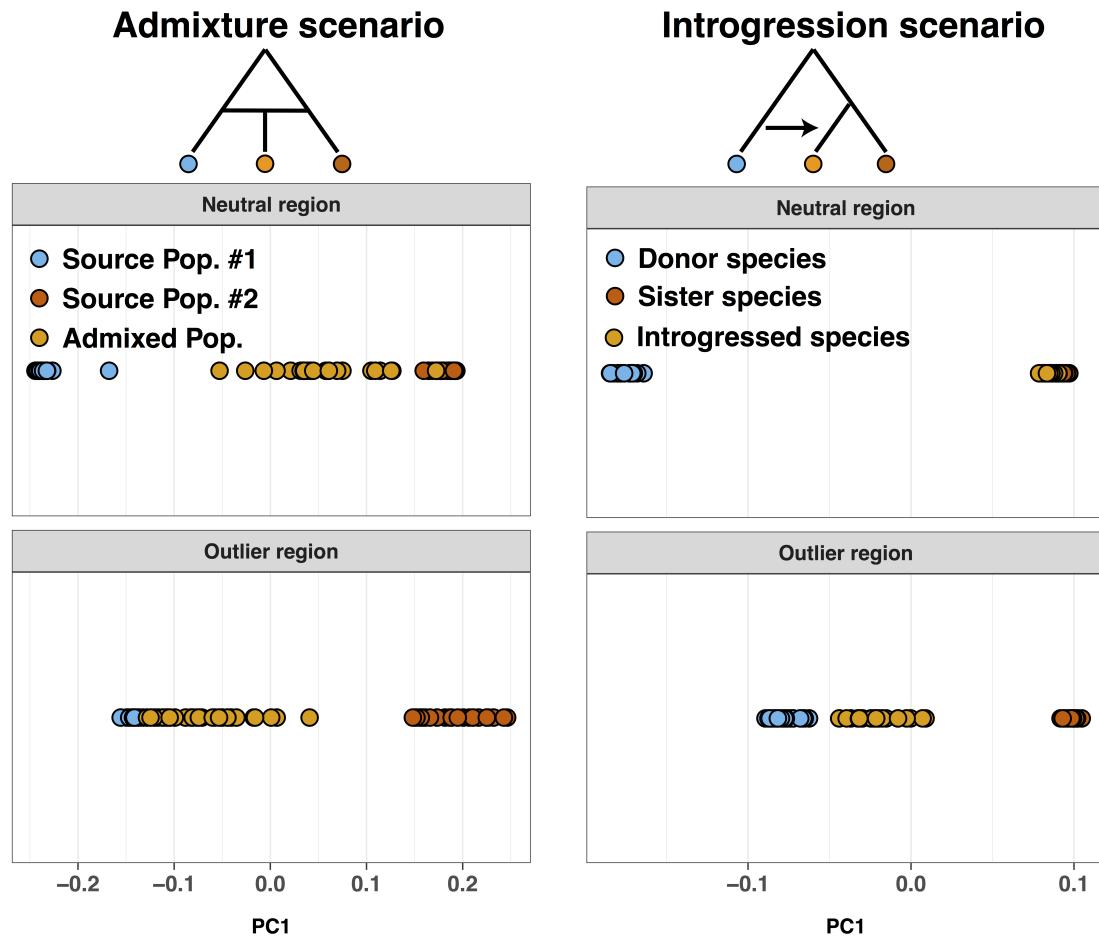


FIGURE 3.9 – Principal components obtained for a neutral region and an outlier region in the admixture scenario (left panels) and in the introgression scenario (right panels). For the admixture scenario, simulations use phased genotype data from *P. balsamifera* and *P. trichocarpa* individuals and we assume that the difference Δ_q of ancestry between outlier and neutral regions is of 50% and that admixture took place $\lambda = 100$ generations ago. For the introgression scenario, the gene flow parameter $f = 0.5$.

Simulations under an admixture scenario

We consider the example of admixture of two *Populus* species in North America to simulate admixed individuals (Suarez-Gonzalez, 2016). We considered genotypes from chromosome 6 of 25 individuals from the species *Populus balsamifera* (balsam poplar) and 25 individuals from the species *Populus trichocarpa* (black cottonwood) (Suarez-Gonzalez, 2016). In order to generate admixed individuals, we phase the 50 genotypes using Beagle (S. R. Browning & Browning, 2007). To construct admixed genomes, we began at the first marker on each chromosome and sampled *Populus balsamifera* ancestry with probability $\alpha = 70\%$ and *Populus trichocarpa* ancestry with

probability $(1 - \alpha) = 30\%$. Haplotype of the admixed genome was form by sampling at random one of the haplotype coming from the *P. balsamifera* or *P. trichocarpa* species. Ancestry was resampled based on an exponential distribution with weight λ , which corresponds to the number of generations since admixture (Price et al., 2009). A new ancestry tract was sampled with probability $1 - e^{-\lambda g}$ when traversing a genetic distance of g Morgans. Each time ancestry was resampled, we sampled *P. balsamifera* ancestry with probability $\alpha = 70\%$ and *P. trichocarpa* ancestry with probability $(1 - \alpha) = 30\%$. In the simulations, we consider the first 100,000 SNPs of the chromosome 6 of *Populus* individuals. To model adaptive introgression, we assume that there is a region containing 1,000 SNPs where *P. balsamifera* ancestry was equal to $70\% - \Delta_q$ instead of 70%. We consider different values for Δ_q which are equal to 5%, 10%, 15%, 20%, 30%, 40%, or 50%. Simulations correspond to a model of soft sweep where different alleles of *P. balsamifera* ancestry, which span the 500 SNPs genomic regions, are adaptive (Messer & Petrov, 2013). Because there is no recombination map available in *Populus*, we assume a constant recombination rate of 0.05cM/Mbp (Suarez-Gonzalez, 2016). We consider three different values for the time since admixture λ , which is equal to 10, 100, or 1000 generations. For each value of λ and of Δ_q , we generate 50 replicates containing 25 admixed individuals each. When looking for adaptive introgression in admixed *P. trichocarpa* individuals, we consider SNPs with less than 5% of missing values typed in 36 admixed individuals (Suarez-Gonzalez, 2016).

Simulations under an introgression scenario

We consider simulations of introgression as described by (S. H. Martin et al., 2014). To simulate genomes, we concatenate 45 5,000 bp regions without introgression and 5 5,000 bp regions where each individuals from the introgressed species is drawn from the donor population with a probability f (Supplementary Figure B.28). The parameter f measures the extent of introgression for introgressed regions. When $f = 1$ all individuals from the introgressed species come from the donor population and $f = 0$ correspond to the genomic windows without introgression. The *ms* command lines for simulating neutral and introgressed ($f = 20\%$) regions for 25 individuals in each population are given as follows (Hudson, 2002)

Sequence data were then generated using the software Seq-Gen using the Hasegawa-Kishino-Yano substitution model and a branch scaling factor of 0.01 (`./seq-gen -mHKY -l 1 5000 -s 0.01`) (Rambaut & Grass, 1997).

Evaluation rules for simulations

For admixture scenarios, we consider the 50,000 first SNPs of chromosome 6 when evaluating different software. To evaluate statistical power, we consider 50 windows containing 1,000 SNPs each. To compute a score for each window, we consider two methods : averaging ancestry coefficients over all SNPs within a window or considering the maximum of ancestry coefficients within a window. For a fixed number of regions considered as outliers, we report statistical power defined as the average

over simulations of the number of detected outlier regions divided by five, which is the true number of outlier regions.

For introgression scenarios, we simulated 50 windows of 5,000 bp. All statistics but pcadapt are computed using 5,000 bp windows. Statistical power is returned when considering the 5 top hits, which corresponds to the 5 largest values of the test statistics.

Results

Simulations under an admixture scenario

For a simulated admixture scenario between two *Populus* species, we display standardized ancestry coefficients computed by pcadapt and LAI software (Supplementary Figure B.23). When simulating admixed individuals, we assume that there are five outlier regions where the amount of *P. trichocarpa* ancestry is increased by an amount of Δ_q compared to other regions. In the genome scans shown in figure B.23, the top 5 peaks obtained with all software correspond to the peaks that should be detected. The power of LAI software based on haplotypes (Loter and RFMix), when varying values of Δ_q and of λ , is comparable to the power obtained by an ideal method called *oracle*, which would know ancestry chunks for each admixed individual (Figure 3.10). For Loter, the loss of power—averaged over Δ_q —is comprised between 1% and 19% depending on the number of generations since admixture (Supplementary Figure B.24). For RFMix, the loss of power is comprised between 3% and 11%. When considering methods based on genotypes, we find that the power obtained with EILA is considerably reduced compared to an oracle because loss of power is always larger than 55%. When considering other values of the regularization parameter λ for EILA, statistical power was further reduced. By contrast, the power of pcadapt is comparable to the power obtained with haplotype-based methods. Compared to an oracle, the loss of power of pcadapt is comprised between 3% and 7%. The power of pcadapt is increased compared to haplotype-based methods for ancient admixture events ($\lambda = 1,000$ generations) whereas it is decreased for recent admixture events ($\lambda = 10$ generations) (Figure 3.10). Because we evaluate statistical power on a genomic region-by-region basis, we had to define a compound ancestry measure for each genomic region and figure 3.10 corresponds to results obtained when considering the average ancestry coefficient within each genomic region. If we rather compute the maximum of ancestry coefficients within each window, the power of pcadapt is again increased compared to haplotype-based methods for ancient admixture events whereas results are more balanced for recent admixture events (Supplementary Figure B.25). Although pcadapt considers genotypes only, we find that it provides statistical power comparable to LAI software that make use of haplotypes. When considering additional hits to the five top hits, the statistical power of pcadapt is again comparable to the one obtained with haplotype-based methods (Supplementary Figure B.26).

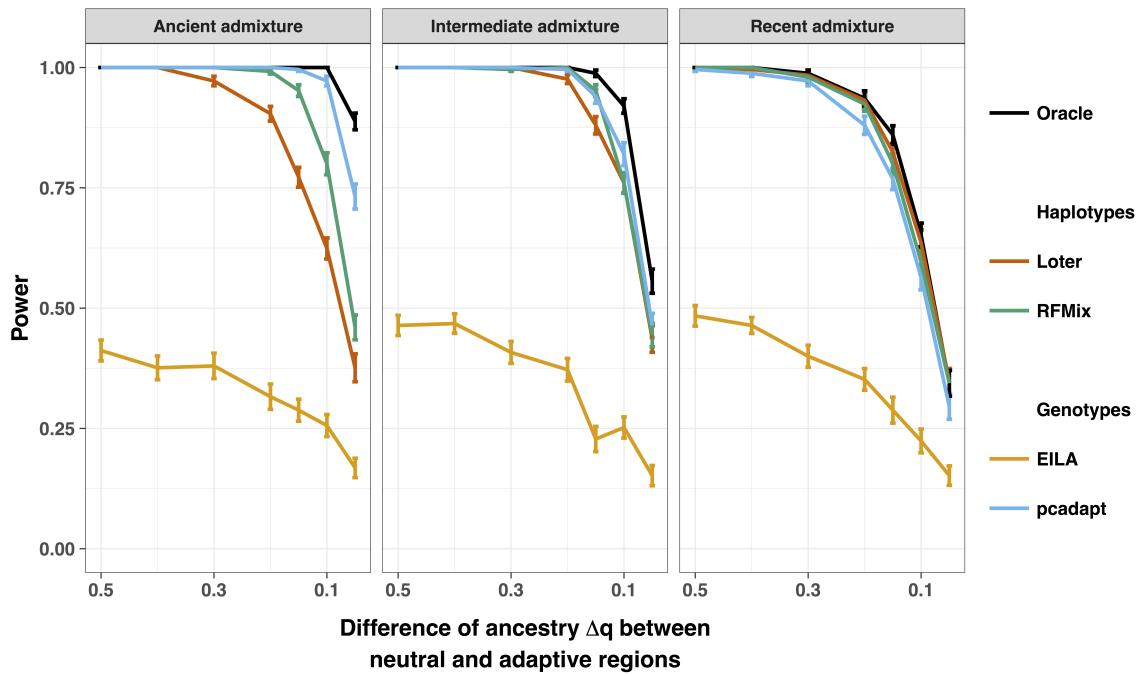


FIGURE 3.10 – Proportion of true outlier peaks among the five top peaks found with pcadapt and different LAI methods (EILA, Loter and RFMix) in a scenario where 2 *Populus* populations experienced admixture. Proportion of true outlier peaks is displayed as a function of the difference Δ_q of ancestry between outlier and neutral regions. The three panels correspond to the three different possible values ($\lambda = 10$ or 100 or 1000) of the number of generations since admixture.

Adaptive introgression in *Populus* admixed individuals

We replicated the search for adaptive introgression in admixed *Populus* species (Suarez-Gonzalez, 2016). We searched for genomic regions with excess of *P. balsamifera* ancestry (boreal species) in admixed *P. trichocarpa* individuals (temperate species). We compare outlier regions found with RFMix, pcadapt and Loter to the regions already found by Suarez-Gonzalez (2016) with the software RASPberry, which implements the HAPMIX model for local ancestry. All software found the peak in chromosome 6 that is located 3.36Mb away from the start of the chromosome (Suarez-Gonzalez, 2016).

Simulations under an introgression scenario

We compared power of several introgression statistics in introgression scenarios. When the proportion f of introgression for introgressed region is equal to 1, all individual from the introgressed population descend from the donor population. In this situation of massive introgression, the power of all statistics is equal to 100% except for the D statistics, which has a power of 94% (Figure B.27). The power of all statistics remains at 100% when $f \geq 0.7$ except the D statistic whose power decreases

to 80% for $f = 80\%$. When f further decreases, powers of all statistics also decrease with the D statistics being the less powerful statistic. The more powerful statistics is RND_{min} , followed by f_d , Bd_f , and $pcadapt$. For the most difficult scenarios where $f = 5\%$, the power of RND_{min} is 68%, the power of f_d is 64%, the power of Bd_f is 60%, the power of $pcadapt$ is 56% and the power of D is 38%.

We consider *Populus* simulated admixed individuals to evaluate if the PCA-based approach manages to identify true outliers for local ancestry. By considering the 5 peaks with highest values of local ancestry, we compare statistical power obtained with the PCA-based approach and with LAI software. For each software, statistical power decreases as a function of Δ_q as it was expected (Figure 3.10). Among all software we consider, RFMix and Loter maximize statistical power whatever the value of the time since admixture λ and of the difference of allele frequency between introgressed and non-introgressed regions (Figure 3.10). Although the power of the PCA approach is reduced, the loss of power is very small when compared to RFMix or Loter. By contrast, the reduction of power obtained with EILA, which is another method that use genotype only, is considerable.

Chapitre 4

Aspect computationnel

Dans cette partie, nous nous intéresserons brièvement à l'aspect computationnel des méthodes qui ont été présentées dans les chapitres précédents. Le développement d'outils logiciels destinés à l'exploration de données génétiques volumineuses requiert qu'une attention particulière soit portée à l'utilisation des ressources de calcul. Étant donné que nous nous intéressons à la possibilité de réaliser des scans génomiques pour des données génétiques de grande taille, il semblait intéressant de préciser les points sur lesquels des améliorations ont été faites d'un point de vue computationnel.

4.1 Du langage C au langage R

La première version de pcadapt a été implémentée par Nicolas Duforet-Frebourg. Le logiciel a été initialement développé en C mais nous avons décidé de poursuivre le développement du logiciel en R/Rcpp, afin de simplifier son utilisation et de bénéficier des performances du langage C++, mais surtout de la portabilité, des outils de visualisation et de documentation du langage R. La librairie pcadapt est disponible sur CRAN.

4.2 Du calcul de la matrice de covariance à l'algorithme IRAM

Nous rappelons ici que nous nous intéressons seulement aux premières composantes principales obtenues avec l'ACP, ce qui signifie qu'il n'est pas nécessaire d'effectuer la décomposition en valeurs singulières (SVD) complète de la matrice d'apparentement génétique $G_{RM} \in \mathcal{M}_n(\mathbb{R})$ (où n est le nombre d'individus). Il s'agit de ce que l'on appelle une décomposition en valeurs singulières tronquée. Pour effectuer l'ACP à partir d'une matrice de génotypes $\tilde{G} \in \mathcal{M}_{np}(\mathbb{R})$, il est donc nécessaire de calculer la matrice d'apparentement génétique G_{RM} pour ensuite calculer la SVD de G_{RM} . En suivant cette procédure, le calcul de la matrice G_{RM} est l'étape la plus coûteuse d'un point de vue algorithmique, car de complexité quadratique en le nombre d'individus.

Cependant de nouvelles méthodes permettent d'effectuer la SVD tronquée de G_{RM} sans qu'il n'y ait besoin de calculer explicitement G_{RM} . La méthode que nous avons choisie d'utiliser pour le calcul de l'ACP est basée sur l'algorithme IRAM (Implicitly Restarted Arnoldi Method) (Calvetti, Reichel, & Sorensen, 1994 ; Lehoucq & Sorensen, 1996).

L'algorithme IRAM repose sur l'idée que les vecteurs propres d'une matrice carrée $A \in \mathcal{M}_n(\mathbb{R})$ peuvent être estimés en construisant une base orthogonale de l'espace vectoriel $\mathcal{K}_k(x) = \text{Vect}(x, Ax, A^2x, \dots, A^kx)$ engendré par les itérations de A où $x \in \mathbb{R}^n$ et $k \in \mathbb{N}$. $\mathcal{K}_k(x)$ est connu sous le nom d'espace de Krylov. Dans notre cas, la matrice carrée que l'on souhaite décomposer est la matrice G_{RM} . Pour construire les espaces de Krylov associés à G_{RM} , il faut donc calculer les produits $G_{RM}x$. Or, pour ne pas avoir à calculer G_{RM} lors du calcul de $G_{RM}x$ (nous rappelons que $G_{RM} = \frac{1}{p}\tilde{G}\tilde{G}^T$), il suffit de calculer d'abord $y = \frac{1}{\sqrt{p}}\tilde{G}^Tx$ puis $\frac{1}{\sqrt{p}}\tilde{G}y$, résultant en un coût algorithmique linéaire en n et en p (Figure 4.1).

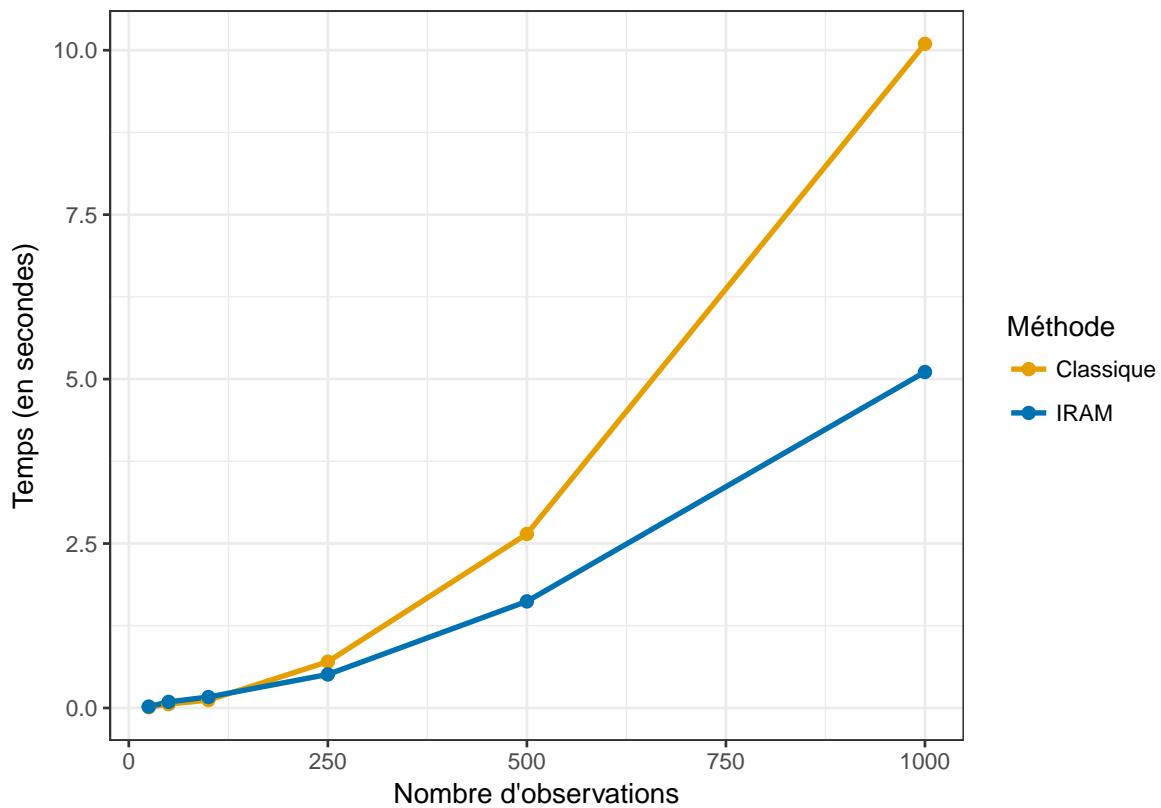


FIGURE 4.1 – Comparaison des temps de calcul pour la SVD tronquée de rang 2 obtenue avec la méthode classique et la méthode IRAM. Le nombre de variables est fixé à 10000 et le nombre d'observations varie de 25 à 1000. Nous constatons que pour un nombre d'observations élevé, l'algorithme IRAM est plus efficace que la méthode nécessitant de calculer la matrice de covariance.

4.3 ACP et valeurs manquantes

Pour tenir compte de la présence de données manquantes dans le calcul de l'ACP, plusieurs stratégies peuvent être envisagées (Dray & Josse, 2015) :

- Imputer : chaque entrée manquante est remplacée par une valeur et l'ACP est réalisée à partir de la matrice de génotypes complétée. Dans le logiciel flashpca par exemple, si la valeur G_{ij} est manquante, elle complétée par la valeur moyenne observée sur le j -ème locus (Abraham & Inouye, 2014). D'autres suggèrent encore d'utiliser des logiciels spécifiquement conçus pour l'imputation de données génétiques comme Beagle ou SHAPEIT (Browning & Browning, 2016 ; Delaneau, Marchini, & Zagury, 2012).
- Tenir compte des données manquantes dans le calcul de G_{RM} : lors du calcul de la corrélation entre deux individus i et j , sont exclus du calcul tout marqueur génétique manquant chez l'individu i ou j .

$$G_{RM,ij} = \frac{1}{\sum_{k=1}^p \delta_{ik} \delta_{jk}} \sum_{k=1}^p \frac{(G_{ik} - 2p_k) \times (G_{jk} - 2p_k)}{2p_k(1 - p_k)} \delta_{ik} \delta_{jk} \quad (4.1)$$

où $\delta_{ik} = 0$ si G_{ik} est manquant et $\delta_{ik} = 1$ sinon. Dans le cas où il n'y a pas de valeur manquante, nous retrouvons bien l'expression donnée par l'équation (2.7). Un exemple d'implémentation en Rcpp du calcul de G_{RM} tenant compte des données manquantes est donné ci-dessous.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericMatrix PairGRM(const NumericMatrix &G,
                      const NumericVector &p) {

    // In our algorithms, individuals are stored in columns
    // and SNPs are stored in rows.

    int nSNP = G.nrow(); // number of SNPs
    int nIND = G.ncol(); // number of individuals
    NumericMatrix GRM(nIND, nIND); // Genetic Relationship Matrix

    for (int i = 0; i < nIND; i++) {
        for (int j = 0; j < nIND; j++) {

            // value = GRM(i, j)
            double value = 0;
            double tmp = 0;
            // number of missing values for each pair
            // of individuals (i, j)
            // nbmv = \sum_{k = 1}^{nSNP} \delta_{ik} \delta_{jk}
```

```

int nbmv = 0;

// Loop over the SNPs to compute the dot product
for (int k = 0; k < nSNP; k++) {
    if ((!NumericVector::is_na(G(k, i))) &&
        (!NumericVector::is_na(G(k, j)))) {
        tmp = (G(k, i) - 2 * p[k]) * (G(k, j) - 2 * p[k]);
        value += tmp / (2 * p[k] * (1 - p[k]));
    } else {
        nbmv++;
    }
}
// Divide by the number of non-missing values for (i, j)
GRM(i, j) = value / (nSNP - nbmv);
}

return GRM;
}
}

```

Algorithme IRAM et données manquantes

Cependant, comme cela a été dit dans le paragraphe précédent, nous souhaitons nous détacher du calcul de la matrice G_{RM} . Nous avons donc cherché à adapter l'algorithme IRAM pour qu'il puisse tenir compte de la présence de données manquantes, à la manière de la fonction `PairGRM` décrite ci-dessus. Nous avons donc adopté la même démarche, mais en l'appliquant aux produits $\tilde{G}^T x$ et $\tilde{G}y$ où $x \in \mathbb{R}^n$ et $y \in \mathbb{R}^p$. Plus précisément, plutôt que de calculer $\tilde{G}^T x$ et $\tilde{G}y$, nous calculons les produits $\frac{1}{n}\tilde{G}^T x$ et $\frac{1}{p}\tilde{G}y$, ce qui nous permet de tenir compte des données manquantes de la façon suivante :

$$\begin{aligned} \forall j \in [|1, p|], \left(\frac{1}{n} \tilde{G}^T x \right)_j &= \frac{1}{\sum_{i=1}^n \delta_{ij}} \sum_{i=1}^n \tilde{G}_{ij} \delta_{ij} x_i, \\ \forall i \in [|1, n|], \left(\frac{1}{p} \tilde{G}y \right)_i &= \frac{1}{\sum_{j=1}^p \delta_{ij}} \sum_{j=1}^p \tilde{G}_{ij} \delta_{ij} y_j, \end{aligned} \quad (4.2)$$

où $\delta_{ij} = 0$ si G_{ij} est manquant et $\delta_{ij} = 1$ sinon. En l'absence de données manquantes, cette méthode donne une estimation du produit $\frac{1}{np}\tilde{G}\tilde{G}^T x = \frac{1}{n}G_{RM}x$, ce qui signifie que si l'on applique l'algorithme IRAM en utilisant les produits définis à l'équation (4.2), nous obtenons la SVD de $\frac{1}{n}G_{RM}$ et non pas celle de G_{RM} . Pour déduire la SVD de G_{RM} à partir de la SVD de $\frac{1}{n}G_{RM}$, il suffit simplement de multiplier les valeurs propres de $\frac{1}{n}G_{RM}$ par n , les vecteurs propres étant les mêmes pour les deux décompositions.

Précision de la SVD en présence de données manquantes

Nous comparons ici les différentes méthodes dont nous avons parlé dans le paragraphe précédent, à savoir flashpca, PairGRM et notre nouvelle méthode. La comparaison est réalisée sur un échantillon du jeu de données POPRES. Nous évaluons la sensibilité des différentes méthodes à la présence de données manquantes en comparant les valeurs singulières et les scores de l'ACP obtenus par chacune des méthodes en présence de données manquantes aux valeurs singulières et scores de l'ACP obtenus avec flashpca en l'absence de données manquantes. Les résultats de cette comparaison sont donnés en figure 4.2.

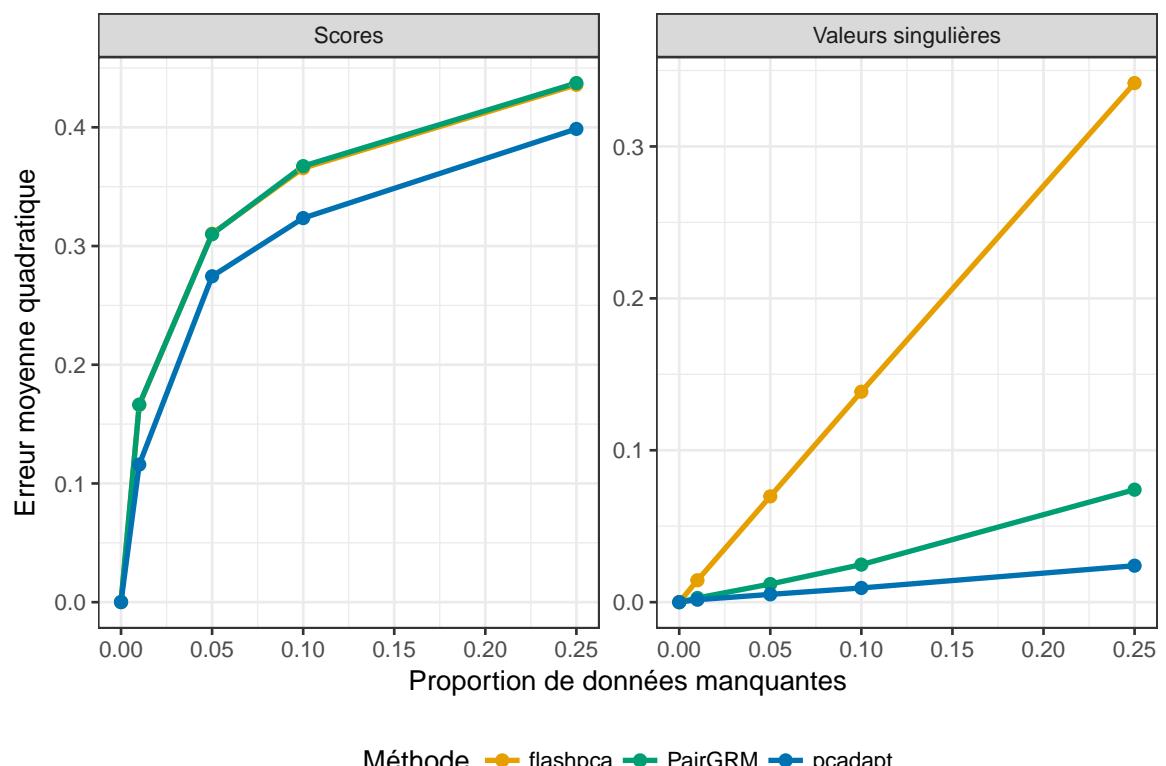


FIGURE 4.2 – Comparaison des scores d'ACP obtenus avec flashpca, PairGRM et pcadapt pour différentes proportions de valeurs manquantes distribuées uniformément. Nous calculons l'erreur moyenne quadratique pour le calcul des 5 premiers axes de l'ACP et des 5 premières valeurs singulières.

En conclusion, la possibilité de déduire la SVD d'une matrice A à partir de la SVD d'une matrice proportionnelle à celle-ci nous a permis d'adapter la méthode IRAM au cas de matrices contenant des données manquantes. Par ailleurs, d'après nos comparaisons numériques, notre nouvelle méthode s'avère être moins sensible à la présence de données manquantes.

TABLE 4.1 – Encodage des génotypes dans le format .bed. Les génotypes sont encodés sur 2 bits ce qui permet de stocker 4 génotypes sur un octet, contrairement au format .pcadapt où chaque génotype est encodé sur un octet. Le quadruplet 2 1 NA 2 sera par exemple encodé par 11100111.

Génotype	Encodage
0	00
NA	01
1	10
2	11

4.4 Du format .pcadapt au format .bed

Le format .pcadapt stocke la matrice de génotypes dans un fichier texte où chaque ligne représente un marqueur génétique. Les caractères sont séparés par un espace et les valeurs manquantes sont encodées par des 9. Ce format a été utilisé car le calcul de l'ACP pouvait être effectué sans qu'il n'y ait besoin de charger la matrice de génotypes dans la mémoire vive. La matrice G_{RM} était alors calculée de façon incrémentale, en parcourant le fichier ligne par ligne (et donc SNP par SNP). Ce format, bien que très simple, présente quelques inconvénients :

- les espaces n'encodent aucune information. La moitié de l'espace mémoire occupée par le fichier est donc essentiellement vide d'un point de vue informatif.
- les génotypes 0, 1, 2 et 9 sont encodés en ASCII et donc chaque valeur occupe un octet (ou 8 bits).

Le logiciel PLINK (Purcell et al., 2007) dispose quant à lui du format de fichier .bed qui semble plus adapté au développement logiciel :

- le format .bed est un format de fichier binaire, l'accès à un fichier binaire est plus rapide que l'accès à un fichier texte.
- sachant qu'il n'y a que 4 valeurs possibles pour un génotype, chaque génotype peut être encodé sur 2 bits (Table 4.1).

Pour résumer, un fichier .bed et un fichier .pcadapt contiennent exactement la même information. Un fichier .bed occupe exactement 8 fois moins d'espace mémoire physique qu'un fichier au format .pcadapt. Nous avons donc décidé de développer nos algorithmes afin qu'ils puissent être directement utilisés sur des fichiers .bed sans qu'il n'y ait besoin de les convertir au format .pcadapt.

4.5 Interface Shiny

Nous proposons également une interface graphique, basée sur la librairie Shiny, pour une utilisation simplifiée de nos outils statistiques. Cette interface se présente sous la forme d'une application web, et nécessite donc l'utilisation d'un navigateur.

Chapitre 5

Perspectives et conclusions

Nous discutons dans ce chapitre des extensions ou alternatives possibles à ce qui a été présenté jusqu'ici.

5.1 Substitution de l'Analyse en Composantes Principales

La méthode de scan à sélection basée sur la distance de Mahalanobis peut être vue comme la succession de trois étapes :

- une première étape visant à déterminer la structure de populations à partir de l'ACP.
- une deuxième étape consistant à effectuer la régression linéaire multiple de la matrice de génotypes normalisée par les scores de l'ACP.
- une troisième étape calculant la distance de chaque marqueur génétique à la moyenne de ces marqueurs pour une métrique donnée, ce qui permet d'identifier les marqueurs génétiques dont les coefficients de régression sont excessivement corrélés avec une ou plusieurs composantes principales. Dans le cas de la distance de Mahalanobis, la métrique est donnée par l'estimateur robuste de la matrice de covariance.

En réalité, on retrouve ce schéma également pour le calcul de la F_{ST} :

- structure de populations définie par la matrice de scores U_δ telle qu'elle est définie dans le chapitre 1.
- régression linéaire multiple de la matrice de \tilde{G} par les scores U_δ .
- calcul de la distance euclidienne $||\tilde{G}^T U_\delta||_2$ de chaque marqueur à la moyenne des marqueurs (ici la métrique est la matrice identité et la moyenne des marqueurs est la fréquence).

De la même façon, le calcul de la statistique T_{F-LK} peut être décomposé suivant le schéma décrit ci-dessus. Ces trois étapes communes permettent donc de définir un schéma général pour le développement de nouvelles méthodes de scan génomique. La première étape pourrait très bien être remplacée par des variantes de l'ACP telles qu'une ACP régularisée, une ACP à noyau ou une ACP pondérée.

5.2 Utilisation du déséquilibre de liaison pour améliorer l'inférence de la structure

Lorsque deux marqueurs génétiques sont physiquement proches l'un de l'autre, il y a de fortes chances qu'ils soient corrélés entre eux. Ce phénomène est connu sous le nom de déséquilibre de liaison. Lors de l'inférence de la structure de populations à l'aide de l'ACP, il est généralement recommandé de filtrer ce déséquilibre de liaison en effectuant une procédure de *pruning* (Abdellaoui et al., 2013 ; Prive, Aschard, & Blum, 2017). À l'inverse, D. J. Lawson, Hellenthal, Myers, & Falush (2012) suggèrent quant à eux d'inclure cette information et montrent qu'en utilisant leur logiciel fineSTRUCTURE sur des jeux de données relativement denses, la structure de populations est bien mieux estimée qu'avec l'ACP telle qu'elle est implémentée dans EIGENSTRAT (Price et al., 2006). L'utilisation de fineSTRUCTURE ou de méthodes qui prennent en compte le déséquilibre de liaison pour inférer la structure de populations peut donc s'avérer intéressante pour deux raisons. Cela permettrait de ne pas avoir à prétraiter les données pour le déséquilibre de liaison puisqu'il est directement pris en compte. Et surtout, nous nous attendons à ce qu'une meilleure estimation de la structure de population améliore la puissance des test statistiques et réduise le nombre de fausses découvertes. En pratique, pour définir un scan à sélection prenant en compte le déséquilibre de liaison pour l'estimation de la structure, il suffit d'utiliser les scores de l'ACP obtenus avec fineSTRUCTURE à la première étape.

5.3 Utilisation de variables environnementales

Une autre alternative possible serait la prise en compte de variables environnementales pour la détection d'adaptation locale. Notre méthode actuelle suppose que les locus sous adaptation locale devraient être excessivement corrélés avec la structure de populations, qui est représentée par les composantes principales. Dans le cas où la sélection n'est pas liée à la structure de populations, notre méthode n'est plus adaptée. Pour illustrer cela, il suffit de considérer une simulation où les locus sous sélection ne sont pas du tout corrélés à la structure de populations, mais à des variables environnementales qui ne sont pas non plus corrélés à la structure de populations. Nous observons sur la figure 5.1 que notre méthode ne détecte aucun des locus sous adaptation locale. Pour détecter des signaux de ce type, une approche consiste à décorreler les variables environnementales à l'aide d'une analyse des redondances (RDA) (Lasky et al., 2012). Cette analyse produit des nouveaux scores qui sont des combinaisons linéaires des variables environnementales et qui peuvent être utilisées en lieu et place des scores de l'ACP réalisée sur les génotypes. Cette étape est suivie des étapes 2 (régression linéaire multiple) et 3 (distance robuste de Mahalanobis) décrites ci-dessus. Cette procédure fait l'objet d'un travail de recherche de Thibaut Capblancq, post-doctorant au Laboratoire d'Ecologie Alpine à Grenoble.

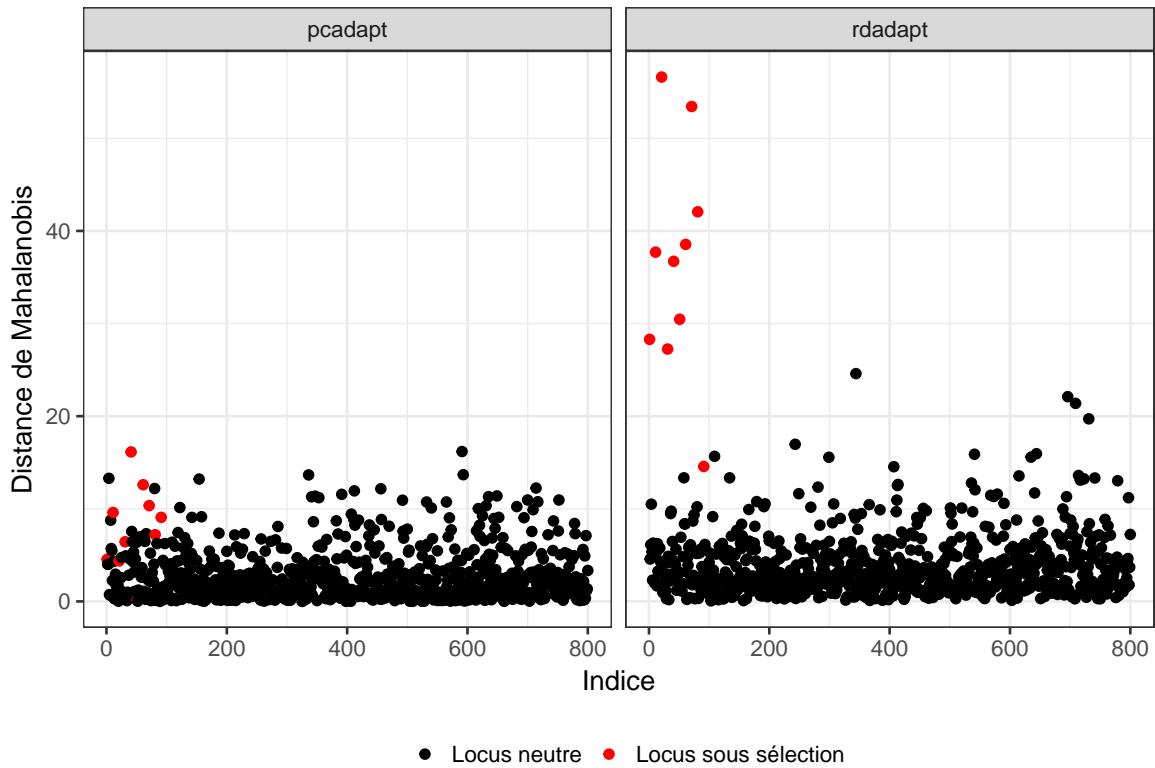


FIGURE 5.1 – Exemple de scan à sélection réalisé avec pcadapt sur une simulation réalisée par Éric Bazin contenant des locus sous sélection qui ne sont pas corrélés à la structure de la population. Les points rouges correspondent aux locus sous sélection pour la simulation.

5.4 Scans pour l'introgression et données manquantes

Dans notre étude portant sur l'introgression, nous n'avons utilisé que des jeux de données complets, puisqu'ils ont été soit simulés, soit imputés à l'aide de Beagle (S. R. Browning & Browning, 2007). Le recours à des logiciels d'imputation est souvent nécessaire car la plupart des méthodes développées ne sont pas utilisables en présence de valeurs manquantes, ce qui est le cas pour notre méthode de scan d'introgression. Il convient toutefois de rappeler que les logiciels d'imputation doivent être utilisés avec précaution, l'imputation de valeurs manquantes très localisées peut créer des motifs particuliers localement sur le génome susceptibles d'être détectés par les différentes méthodes présentées (Figure 5.2). Une possibilité intéressante serait d'associer aux valeurs manquantes les degrés de confiance avec lesquels celles-ci ont été complétées, ce qui pourrait par exemple permettre de pondérer les statistiques de détection par ces degrés de confiance.

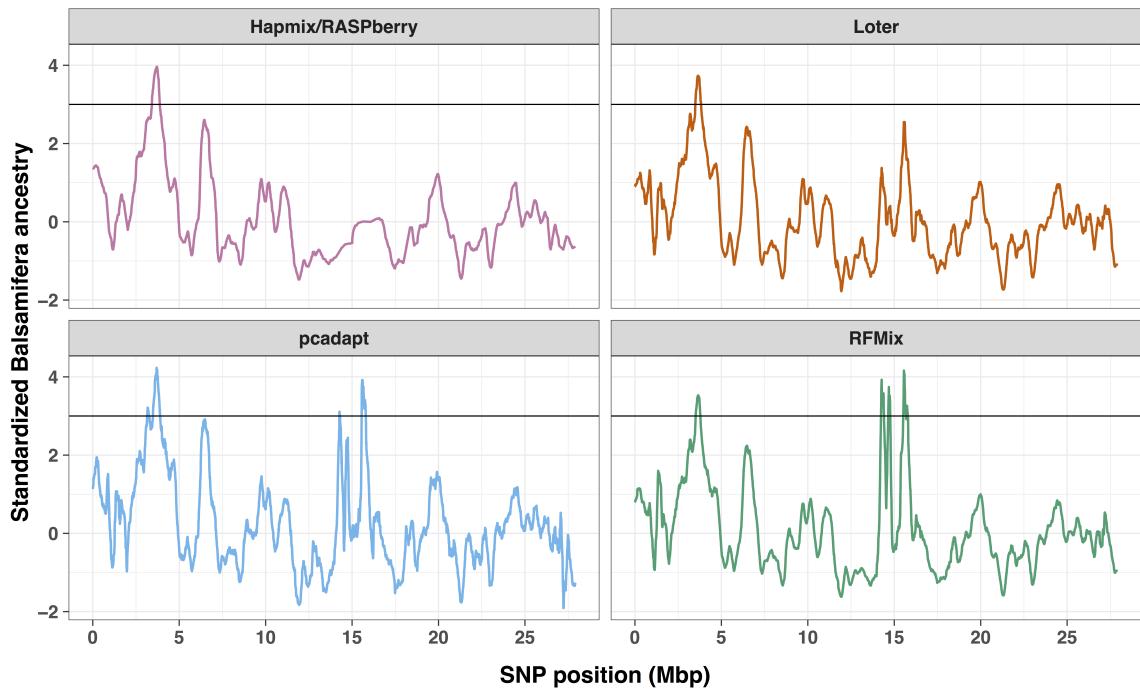


FIGURE 5.2 – Exemple de scans d’introgression à partir de données imputées avec Beagle (S. R. Browning & Browning, 2007) sans prétraitement des données. Les régions uniquement détectées par pcadapt et RFMix (Maples et al., 2013) sont des régions qui en réalité présentent une très grande proportion de valeurs manquantes. Cette figure illustre l’importance de filtrer les marqueurs ayant une proportion trop élevée de valeurs manquantes.

5.5 L’approche IRAM pour les données manquantes

Dans le chapitre précédent, nous avons proposé un moyen d’adapter tout algorithme dérivé de la méthode des puissances itérées à des jeux de données contenant des données manquantes. Notre souhait était de disposer d’une méthode d’ACP de complexité linéaire en n et p tout en étant capable d’estimer raisonnablement bien les éléments propres de la matrice d’apparentement génétique, même en présence de données manquantes. La plupart des méthodes d’ACP utilisables sur des jeux de données incomplets affichent souvent des complexités en $O(\min(n^2 p, np^2))$ (Severson, Molaro, & Braatz, 2017), ce que l’on préfère éviter. Notre méthode a été comparée à deux approches. La première est celle utilisée par le logiciel flashpca (Abraham et al., 2017), qui consiste à remplacer les valeurs manquantes par les valeurs moyennes de chaque SNP. La seconde est celle qui était implémentée dans les premières versions de pcadapt. Cette méthode tient compte de la présence de valeurs manquantes dans le calcul de la matrice d’apparentement génétique. Les résultats de notre comparaison, bien que très

spécifique (distribution de valeurs manquantes uniforme, données présentant beaucoup de structure), nous confortent dans l'idée de pouvoir intégrer la prise en compte des données manquantes dans des méthodes telles que IRAM, sans pour autant modifier la complexité de l'algorithme. Nous pensons que cette méthode peut toutefois s'avérer efficace pour d'autres types de données ou d'autres types de distributions de valeurs manquantes.

5.6 Conclusion

Alors que le développement technologique des ressources de calcul semble s'esouffler, du fait de la difficulté croissante liée à la miniaturisation des composants électroniques, les données ne cessent quant à elles de s'accumuler, et ce, quelque soit le domaine d'activité. Cette tendance suggère l'utilisation d'algorithmes à faible complexité pour le traitement et l'analyse de ces données. L'Analyse en Composantes Principales se révèle donc être un outil de premier choix pour traiter ces nouveaux volumes de données. En génétique des populations, elle présente l'avantage supplémentaire de très bien s'interpréter par l'intermédiaire de la structure de populations. Pour ces raisons, nous avons développé au cours de cette thèse des méthodes statistiques reposant exclusivement sur l'Analyse en Composantes Principales.

Nous avons tout d'abord montré comment l'utilisation de l'ACP permettait d'étendre les tests classiques de différenciation au cas de populations continues. D'abord parce que l'indice de fixation correspond à la proportion de variance expliquée par un modèle à facteurs discrets, alors que la statistique de communalité renvoie à un modèle à facteurs qui peuvent être discrets ou continus. Ensuite parce que, de même que notre nouvelle statistique, la statistique T_{F-LK} correspond essentiellement à une distance de Mahalanobis dans le cas de populations discrètes. La principale différence entre ces méthodes réside dans la façon dont sont estimés les moments d'ordre 1 et 2 dans le calcul de la distance de Mahalanobis.

Nous avons ensuite développé une nouvelle approche statistique pour la détection de régions introgressées, basée sur l'utilisation de scores de métissage locaux calculés à partir de l'ACP. Les avantages de cette méthode, par rapport à celles qui sont présentées, sont essentiellement pratiques. La détection de l'introgression via l'estimation des coefficients de métissage locaux requiert souvent l'utilisation conjointe d'une méthode de phasage ou bien la connaissance de certains paramètres biologiques qui ne sont pas forcément bien définis (comme le taux de recombinaison moyen par exemple). De plus, notre méthode est de complexité linéaire par rapport au nombre d'individus ainsi que par rapport au nombre de SNPs, ce qui la rend très rapide à utiliser même pour des jeux de données qui sont amenés à être de plus en plus denses.

Nous avons développé nos méthodes en veillant à proposer un outil simple d'utilisation tout en cherchant à optimiser l'utilisation des ressources de calcul. Malgré ces précautions, il est certain que la librairie pcadapt pourra encore bénéficier d'améliorations, à la fois en termes de méthodologie statistique et en termes d'implémentation. Bien que notre librairie soit destinée à l'analyse de données volumineuses, elle reste à ce jour principalement utilisée sur des matrices de génotypes relativement petites (de

l'ordre du million de SNPs et du millier d'individus).

Annexe A

Détails

Rapport entre la communalité et l'indice de fixation

Pour faire le parallèle avec la communalité, il est nécessaire de trouver une matrice colonne $U'_\delta \in \mathcal{M}_{n,1}(\mathbb{R})$ telle que $F_{ST} = \|\tilde{G}^T U'_\delta\|_2^2$. Pour ce faire, plaçons-nous dans le cas $N = 2$ et cherchons une matrice de rotation telle que $U_\delta R$ ait sa première colonne constante, c'est-à-dire telle que :

$$\begin{pmatrix} \frac{1}{\sqrt{2n_1}} & 0 \\ 0 & \frac{1}{\sqrt{2n_2}} \end{pmatrix} R = \begin{pmatrix} a & x \\ a & y \end{pmatrix} \quad (\text{A.1})$$

où x, y, a sont des réels à déterminer. Soit $R \in M_2(\mathbb{R})$ une matrice de rotation :

$$R = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

En injectant R dans (A.1), on obtient :

$$\begin{pmatrix} \frac{\cos \theta}{\sqrt{2n_1}} & -\frac{\sin \theta}{\sqrt{2n_1}} \\ \frac{\sin \theta}{\sqrt{2n_2}} & \frac{\cos \theta}{\sqrt{2n_2}} \end{pmatrix} = \begin{pmatrix} a & x \\ a & y \end{pmatrix} \quad (\text{A.2})$$

(A.2) implique que l'angle de la rotation vérifie la relation $\frac{\cos \theta}{\sqrt{2n_1}} = \frac{\sin \theta}{\sqrt{2n_2}}$, d'où $\theta = \arctan(\frac{n_2}{n_1})$. Nous en déduisons ainsi les valeurs de x et de y :

$$\begin{aligned} x &= -\frac{\sin(\arctan(\frac{n_2}{n_1}))}{n_1} \\ y &= \frac{\cos(\arctan(\frac{n_2}{n_1}))}{n_2} \end{aligned} \quad (\text{A.3})$$

Or :

$$\begin{aligned} \sin(\arctan(x)) &= \frac{x}{\sqrt{1+x^2}} \\ \cos(\arctan(x)) &= \frac{1}{\sqrt{1+x^2}} \end{aligned} \quad (\text{A.4})$$

Notant R la rotation d'angle $\arctan(\frac{n_2}{n_1})$, on a finalament :

$$U_\delta R = \begin{pmatrix} a & -\delta_{11} \frac{n_2}{\sqrt{2(n_1^2+n_2^2)}} + \delta_{21} \frac{n_1}{\sqrt{2(n_1^2+n_2^2)}} \\ a & -\delta_{12} \frac{n_2}{\sqrt{2(n_1^2+n_2^2)}} + \delta_{22} \frac{n_1}{\sqrt{2(n_1^2+n_2^2)}} \\ \vdots & \vdots \\ a & -\delta_{1n} \frac{n_2}{\sqrt{2(n_1^2+n_2^2)}} + \delta_{2n} \frac{n_1}{\sqrt{2(n_1^2+n_2^2)}} \end{pmatrix}$$

Puisque R est une rotation, $\|\tilde{G}^T U_\delta\|_2 = \|\tilde{G}^T U_\delta R\|_2$. En développant $\tilde{G}^T U_\delta R$, on obtient :

$$\tilde{G}^T U_\delta R = \left(\sum_{i=1}^n a \tilde{G}_i, \sum_{i=1}^n \left(-\delta_{1i} \frac{n_2}{\sqrt{2(n_1^2+n_2^2)}} + \delta_{2i} \frac{n_1}{\sqrt{2(n_1^2+n_2^2)}} \right) \tilde{G}_i \right) \quad (\text{A.5})$$

Or $\sum_{i=1}^n \tilde{G}_i = 0$ par définition de \tilde{G} , ce qui permet d'écrire, en posant $U'_\delta \in \mathcal{M}_{n,1}(\mathbb{R})$ la matrice colonne correspondant à la deuxième colonne de $U_\delta R$:

$$\begin{aligned} F_{ST} &= \|\tilde{G}^T U_\delta\|_2^2 \\ &= \|\tilde{G}^T U_\delta R\|_2^2 \\ &= \|\tilde{G}^T U'_\delta\|_2^2 \end{aligned} \quad (\text{A.6})$$

Remarquons que U'_δ a une expression similaire à celle des scores de l'ACP exprimée dans G. McVean (2009).

Une généralisation de la statistique T_{F-LK}

Soit $U\Sigma V^T$ la décomposition en valeurs singulières tronquée de rang K de \tilde{G} . Notons \mathcal{F} la matrice d'apparentement génétique interpopulationnel. En utilisant l'estimateur usuel de la matrice de covariance, nous avons :

$$\mathcal{F} = \frac{1}{p} U_\delta^T \tilde{G} \tilde{G}^T U_\delta \quad (\text{A.7})$$

où U_δ est la matrice définie en proposition 2.1. Par définition de la matrice d'apparentement génétique interindividuel, $G_{RM} = \frac{1}{p} \tilde{G} \tilde{G}^T$, si bien que :

$$\mathcal{F} = U_\delta^T G_{RM} U_\delta \quad (\text{A.8})$$

Or $\tilde{G} \simeq U\Sigma V^T$, d'où $G_{RM} \simeq U\Sigma^2 U^T$. De la même manière qu'en proposition 2.2, nous pouvons réécrire la statistique T_{F-LK} en un locus j de la façon suivante :

$$\begin{aligned} T_{F-LK} &= \tilde{G}_{.,j}^T U_\delta \mathcal{F}^{-1} U_\delta^T \tilde{G}_{.,j} \\ &= \tilde{G}_{.,j}^T U_\delta (U_\delta^T G_{RM} U_\delta)^{-1} U_\delta^T \tilde{G}_{.,j} \end{aligned} \quad (\text{A.9})$$

Encore une fois, si l'on considère U plutôt que U_δ , l'expression $\tilde{G}_{.,j}^T U_\delta (U_\delta^T G_{RM} U_\delta)^{-1} U_\delta^T \tilde{G}_{.,j}$ se simplifierait en $\tilde{G}_{.,j}^T U \Sigma^{-2} U^T \tilde{G}_{.,j}$ étant donnée l'approximation $U^T G_{RM} U \simeq \Sigma^2$, ce qui permet de faire le lien entre la statistique T_{F-LK} et la distance de Mahalanobis calculée à partir des loadings.

Annexe B

Informations supplémentaires

Article 1

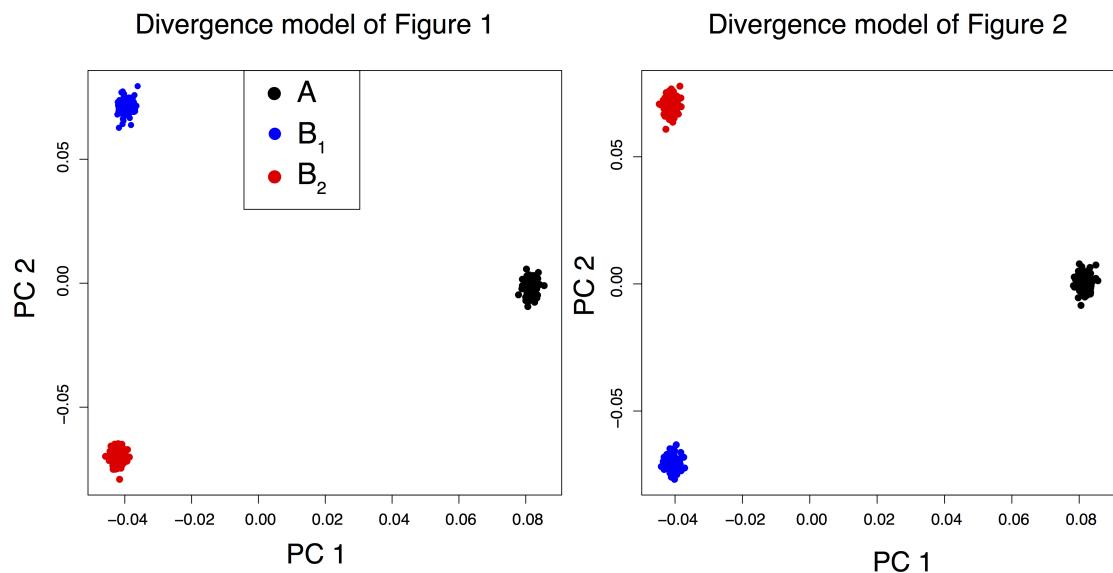


FIGURE B.1 – Principal component analysis for SNP data simulated under the divergence model depicted in Figures 1 and 2 of the main text.

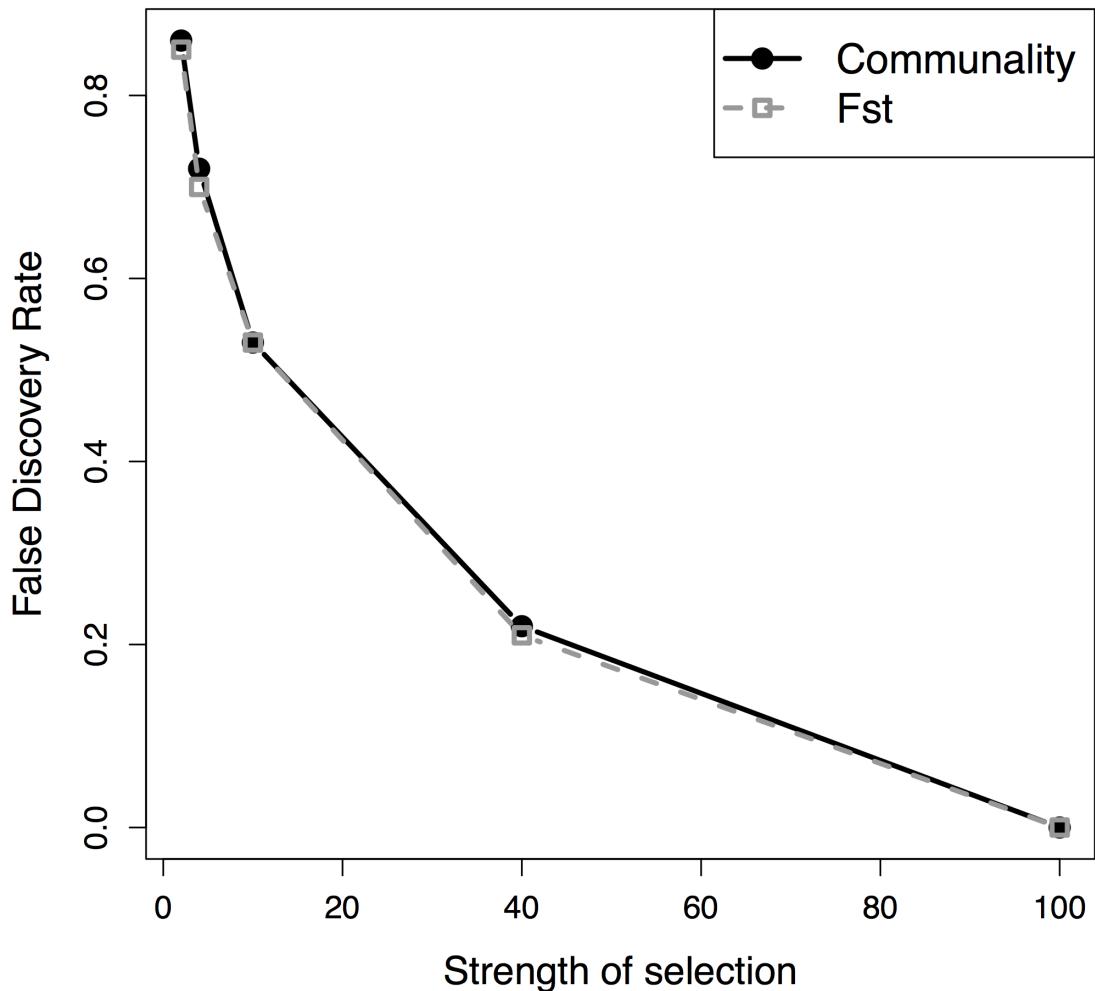


FIGURE B.2 – False discovery rate of the 1% top-ranked SNPs obtained with h^2 and with F_{ST} under an island model.

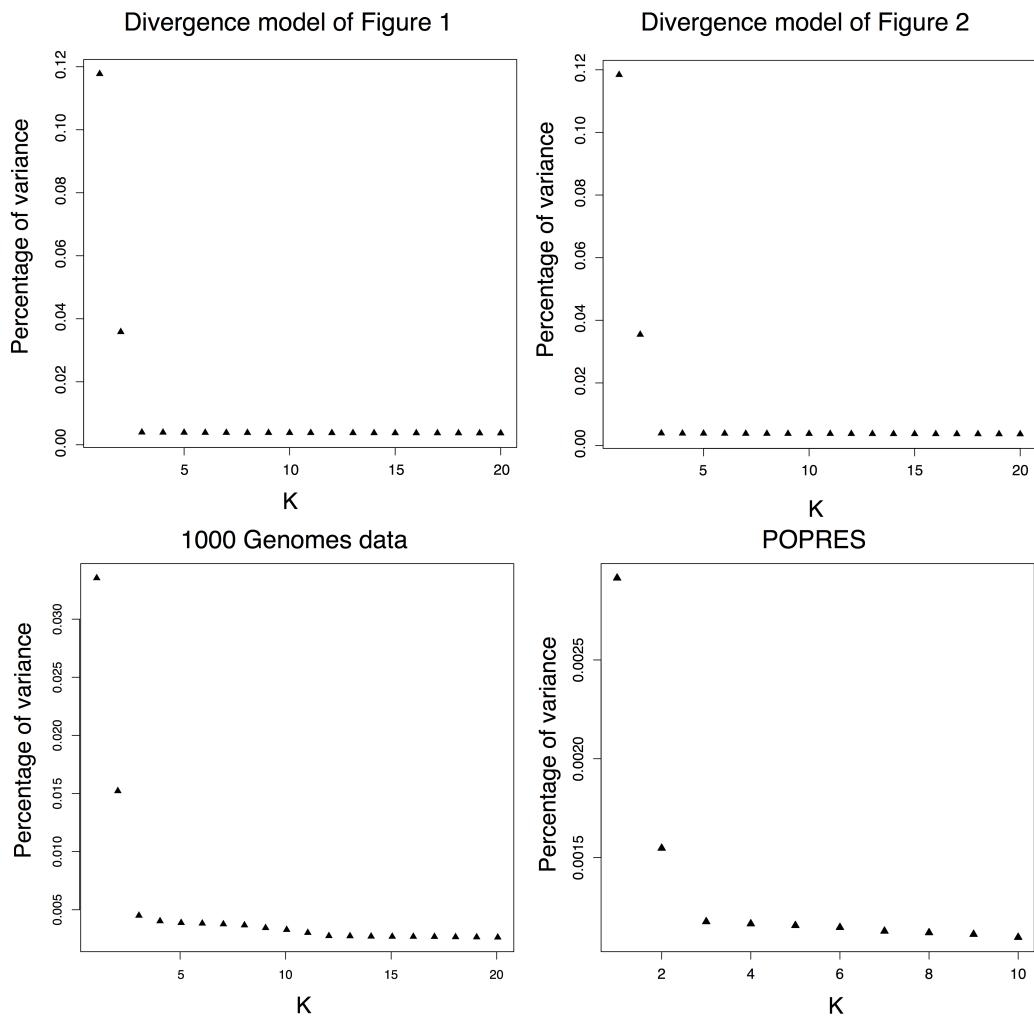


FIGURE B.3 – Decay of eigenvalues of the covariance matrix for divergence models, the 1000 Genome data, and POPRES.

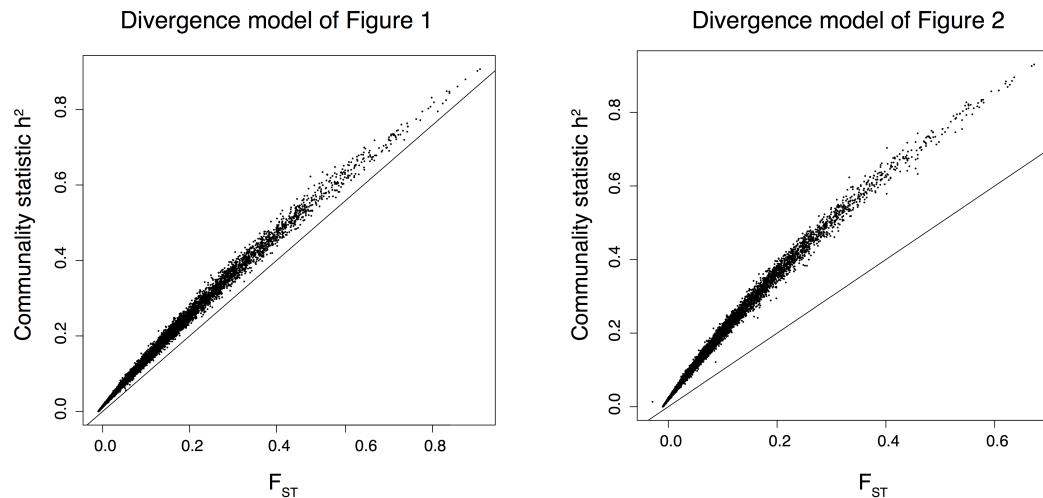


FIGURE B.4 – Communality statistic as a function of F_{ST} for SNPs simulated with a divergence model.

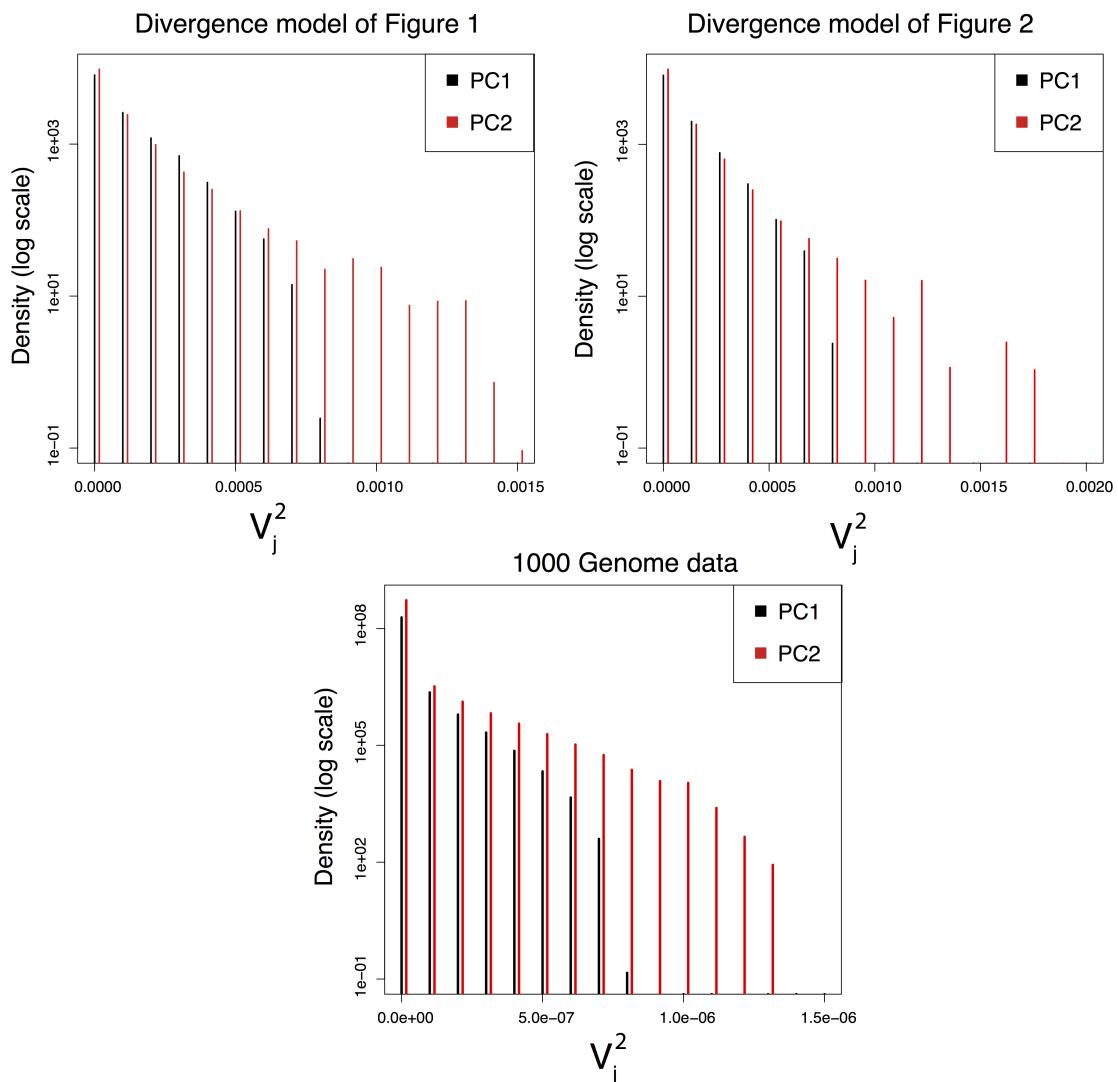


FIGURE B.5 – Histograms of the squared normalized loadings V_{jk}^2 , $k = 1, 2$, obtained for SNPs simulated with divergence models and for SNPs of the 1000 Genomes data.

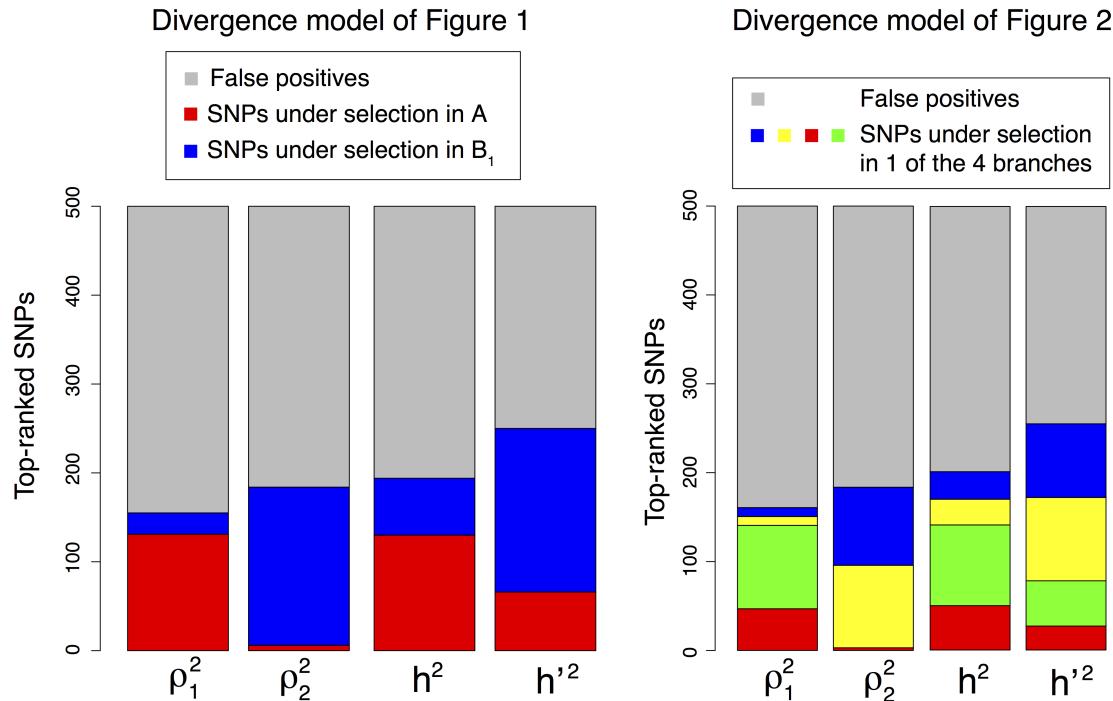


FIGURE B.6 – Repartition of the 5% top-ranked SNPs of each PCA-based statistic under the two divergence models.

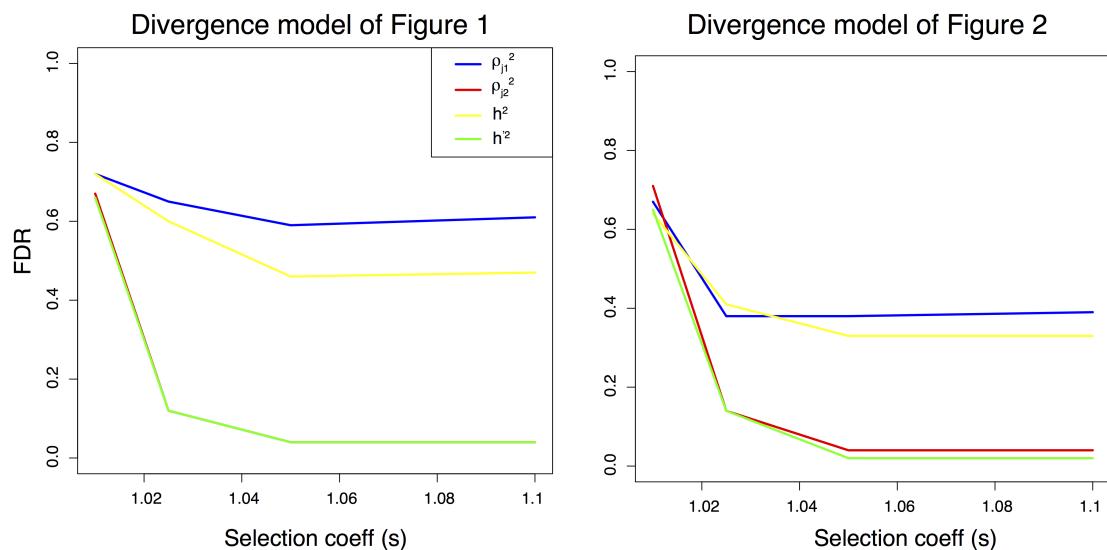


FIGURE B.7 – False discovery rate as a function of the selection coefficient under the two divergence models.

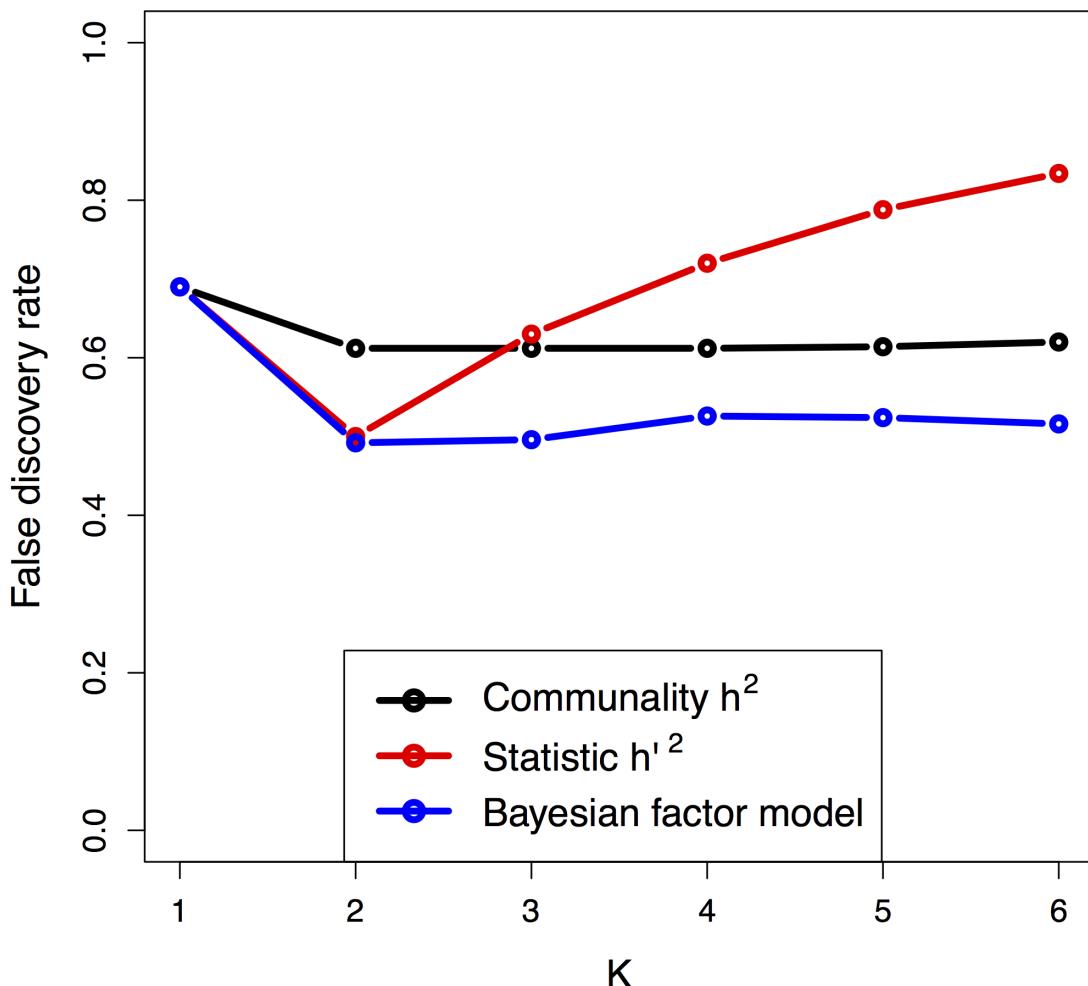


FIGURE B.8 – False discovery rate as a function of the number K of principal components under the two divergence models.

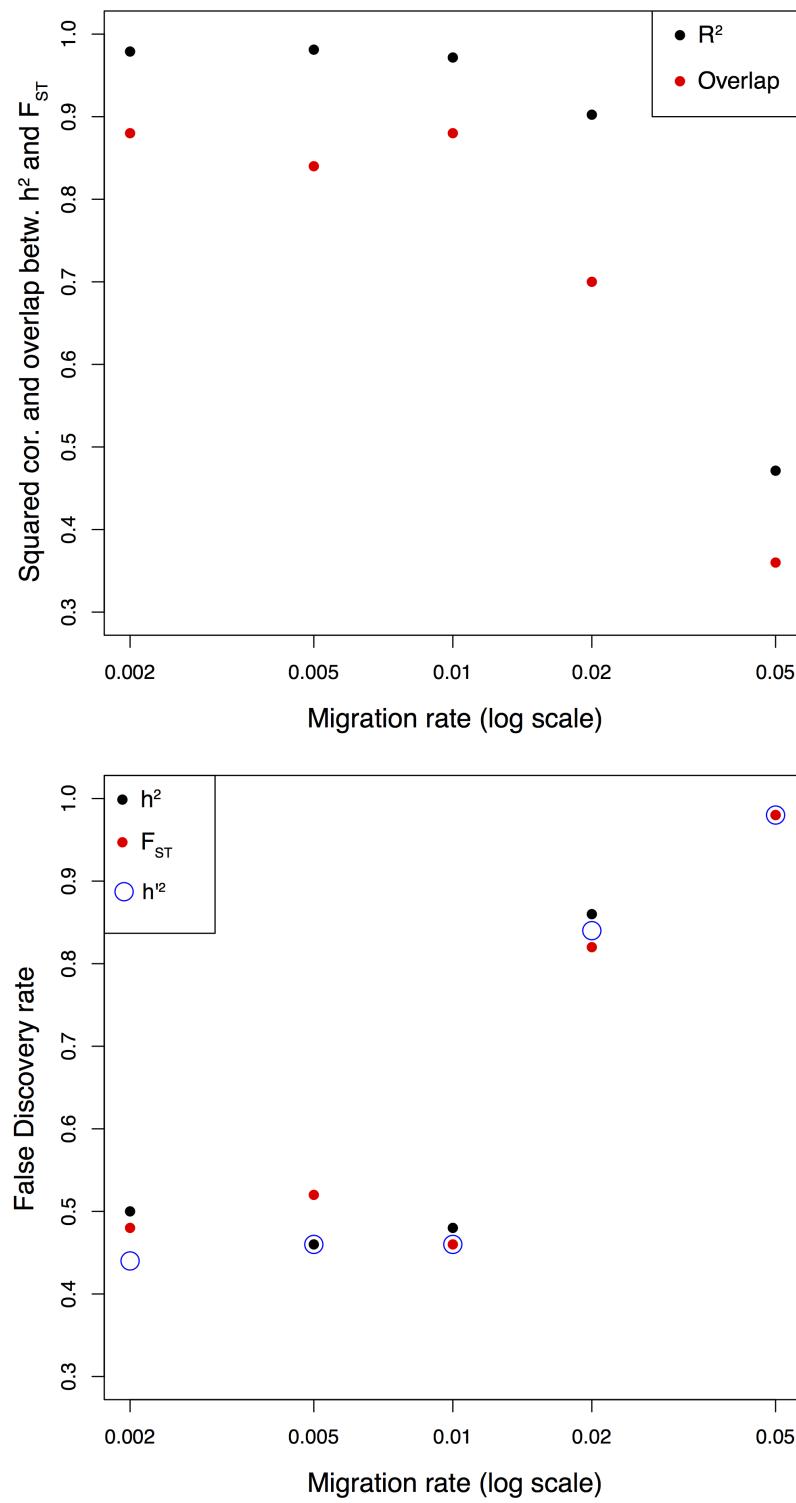


FIGURE B.9 – Comparison between h^2 and F_{ST} in a isolation-with-migration model.

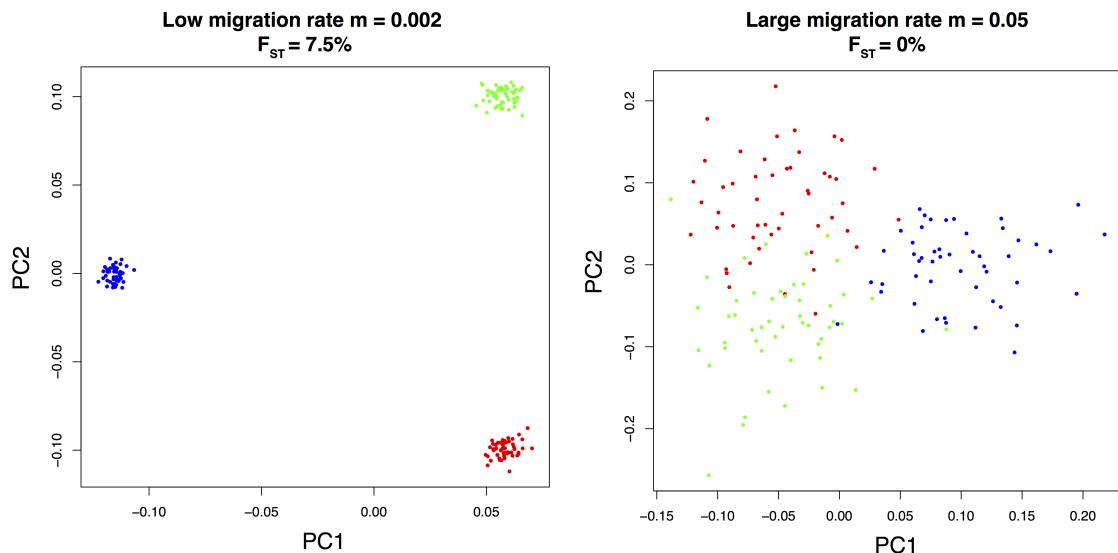


FIGURE B.10 – Principal component analysis of SNPs simulated with an isolation-with-migration model.

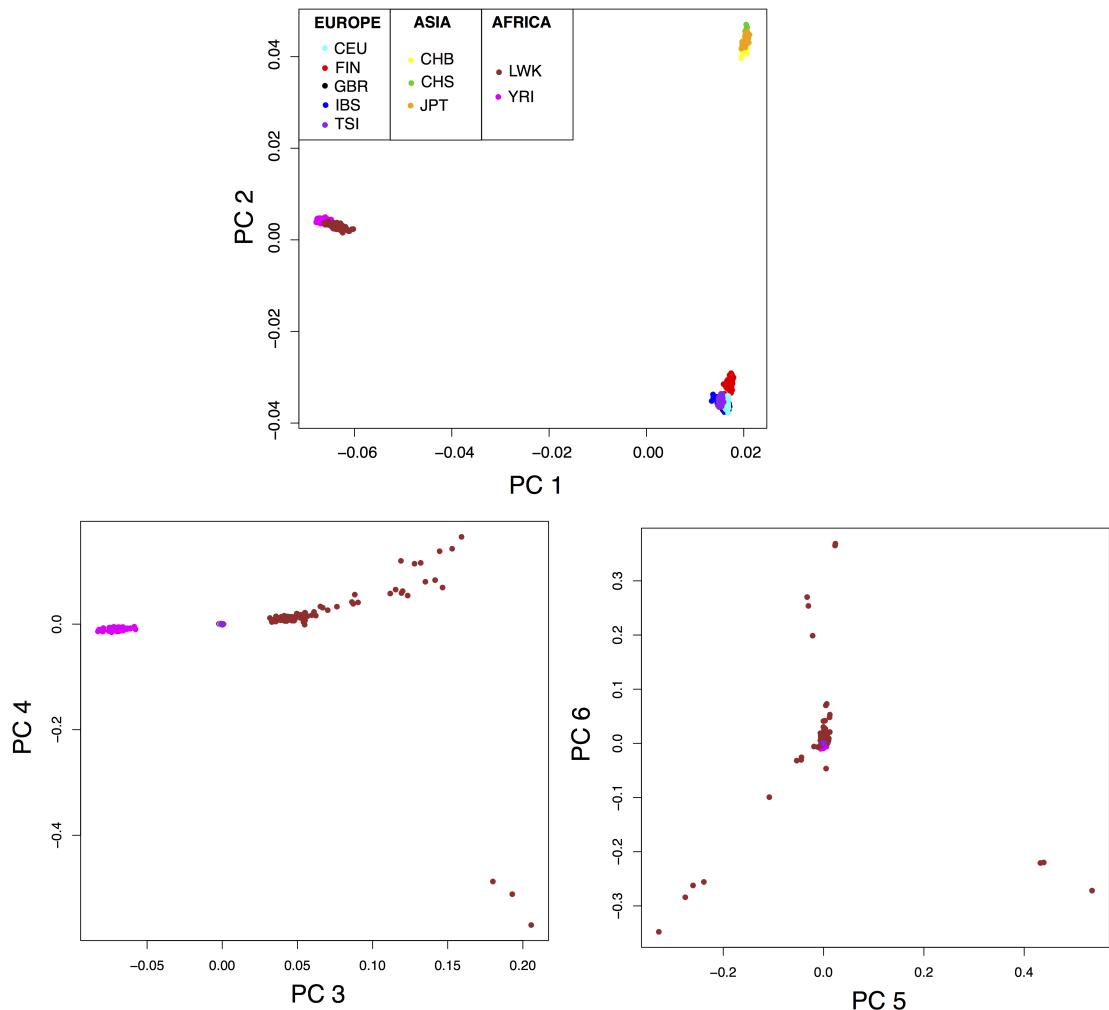


FIGURE B.11 – Principal component analysis of the 1000 Genome data.

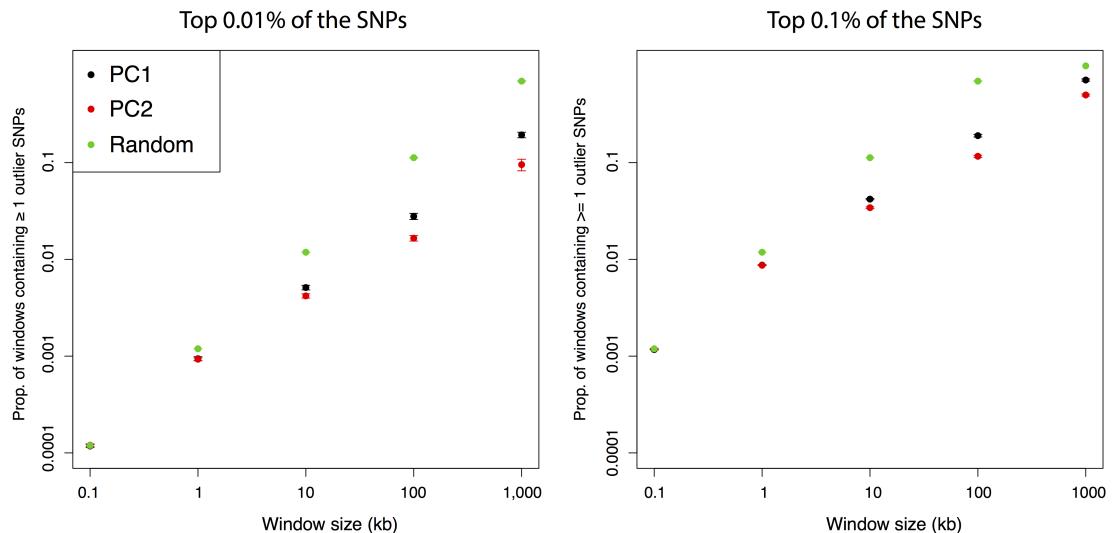


FIGURE B.12 – Number of contiguous windows containing one or more outlier SNPs.

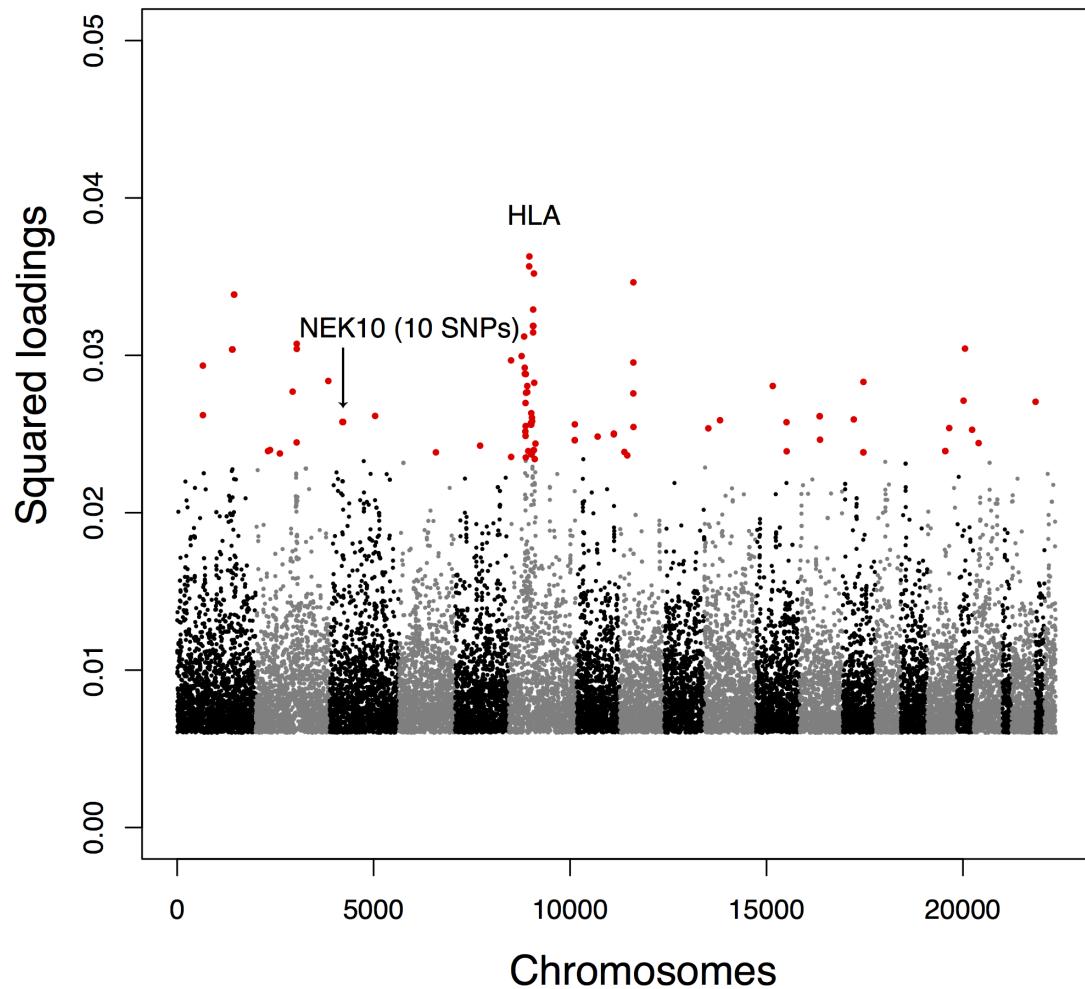


FIGURE B.13 – Manhattan plot for the POPRES data of the squared loadings ρ_{j2}^2 with the second principal component.

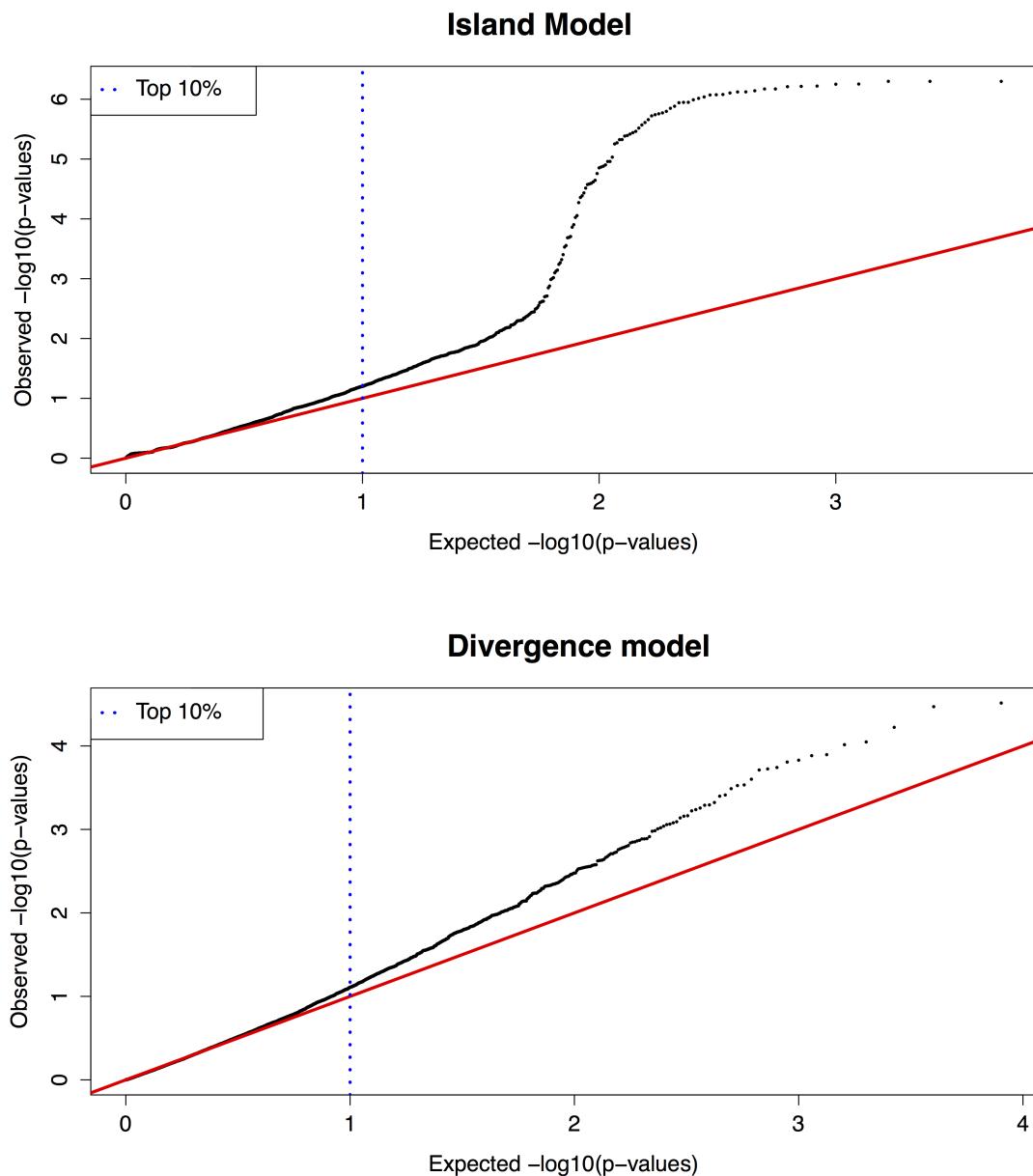


FIGURE B.14 – Q-Q plots of the P -values, which are based on the communality h^2 statistic, under an island and a divergence model.

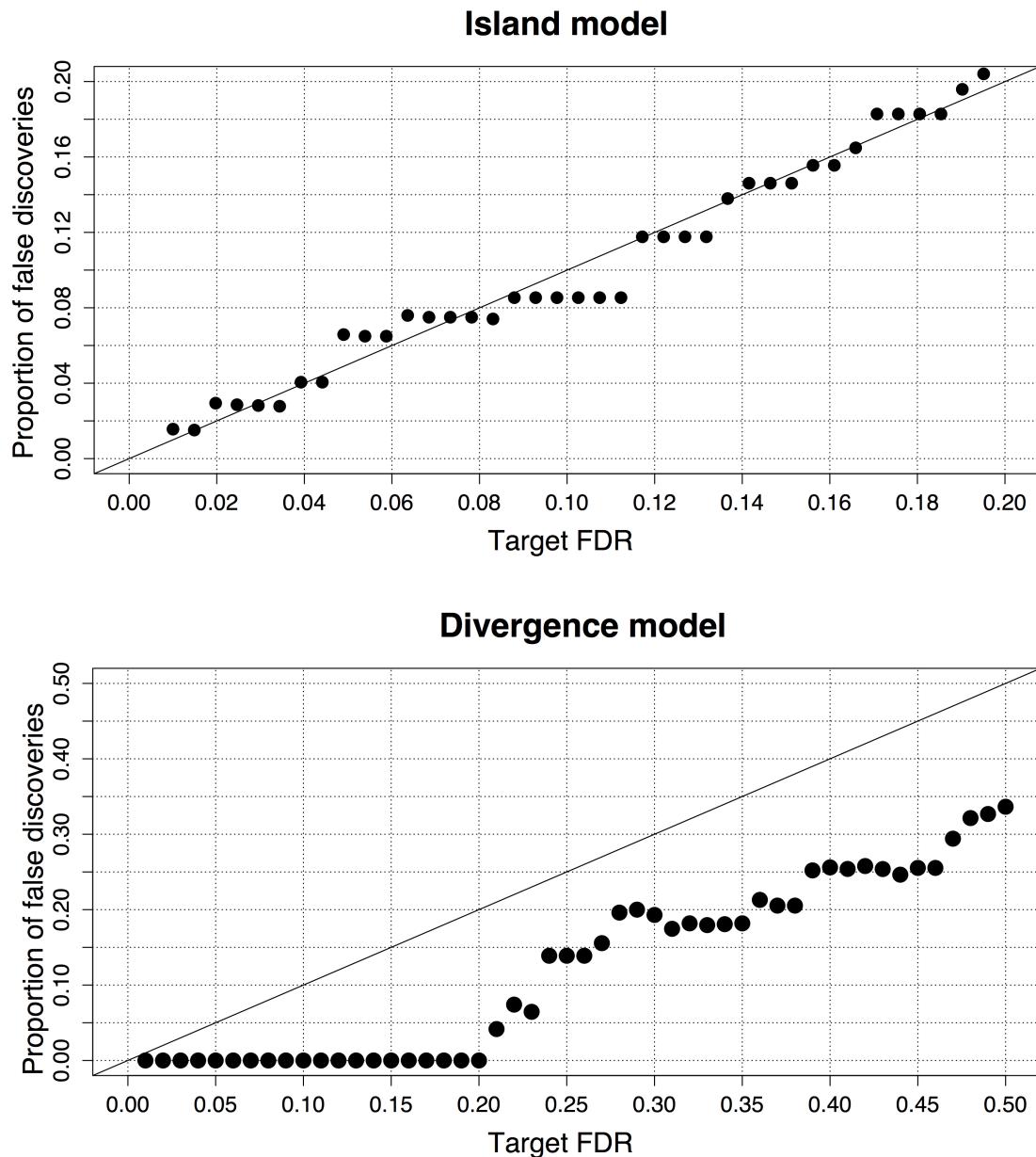
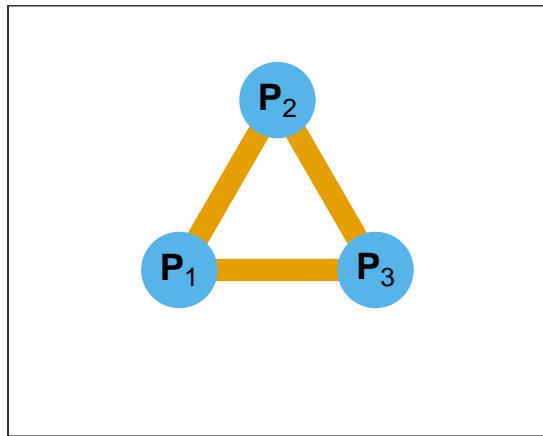


FIGURE B.15 – Control of the false discovery rate for SNPs simulated under an island and a divergence model.

Article 2

Island model



Divergence model

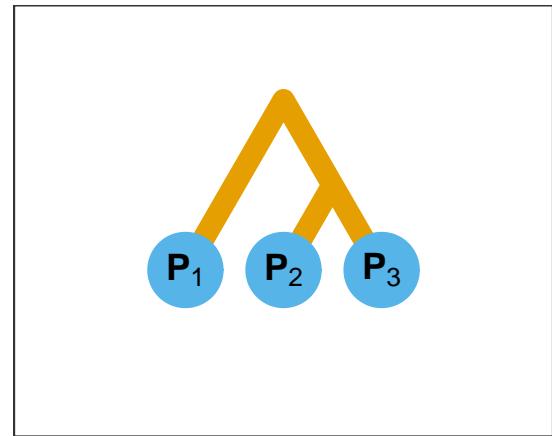


FIGURE B.16 – Schematic description of the island and divergence model. For the island model, adaptation occurs simultaneously in each population. For the divergence model, adaptation takes place in the branch leading to the second population.

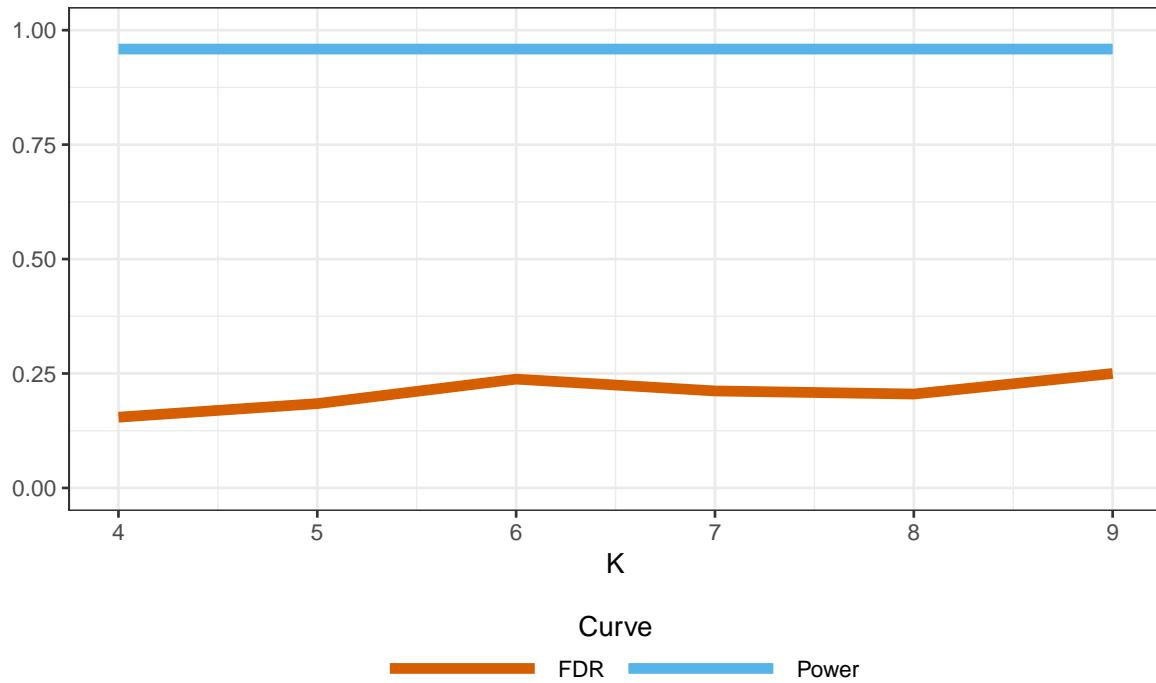


FIGURE B.17 – Proportion of false discoveries and statistical power as a function of the number of principal components in a model of range expansion.

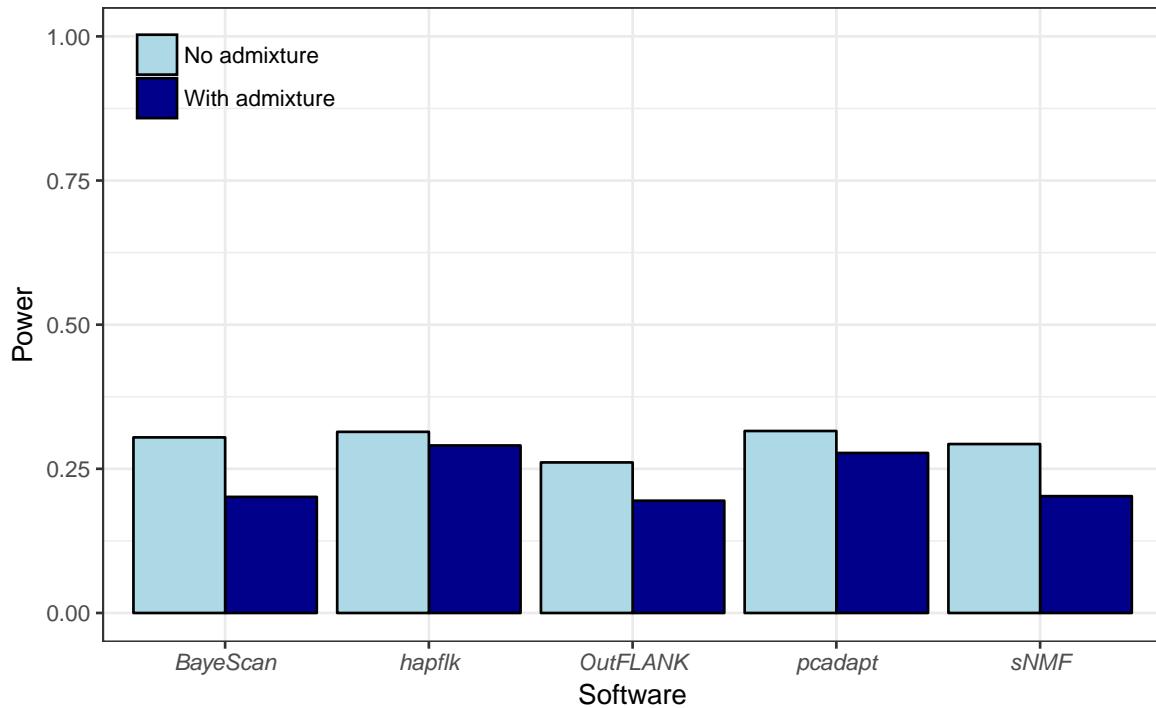


FIGURE B.18 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the island model.

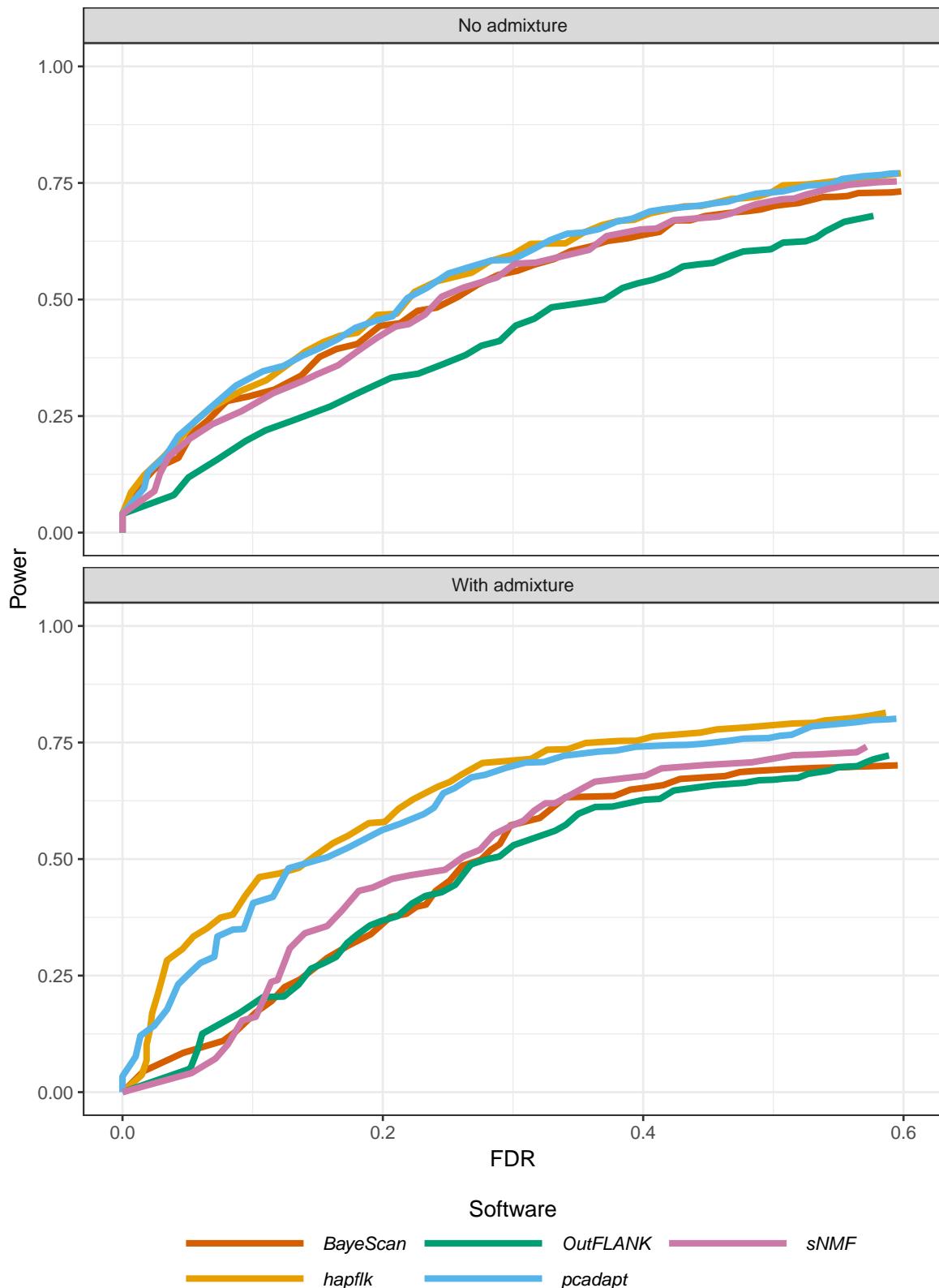


FIGURE B.19 – Statistical power as a function of the proportion of false discoveries for the island model.

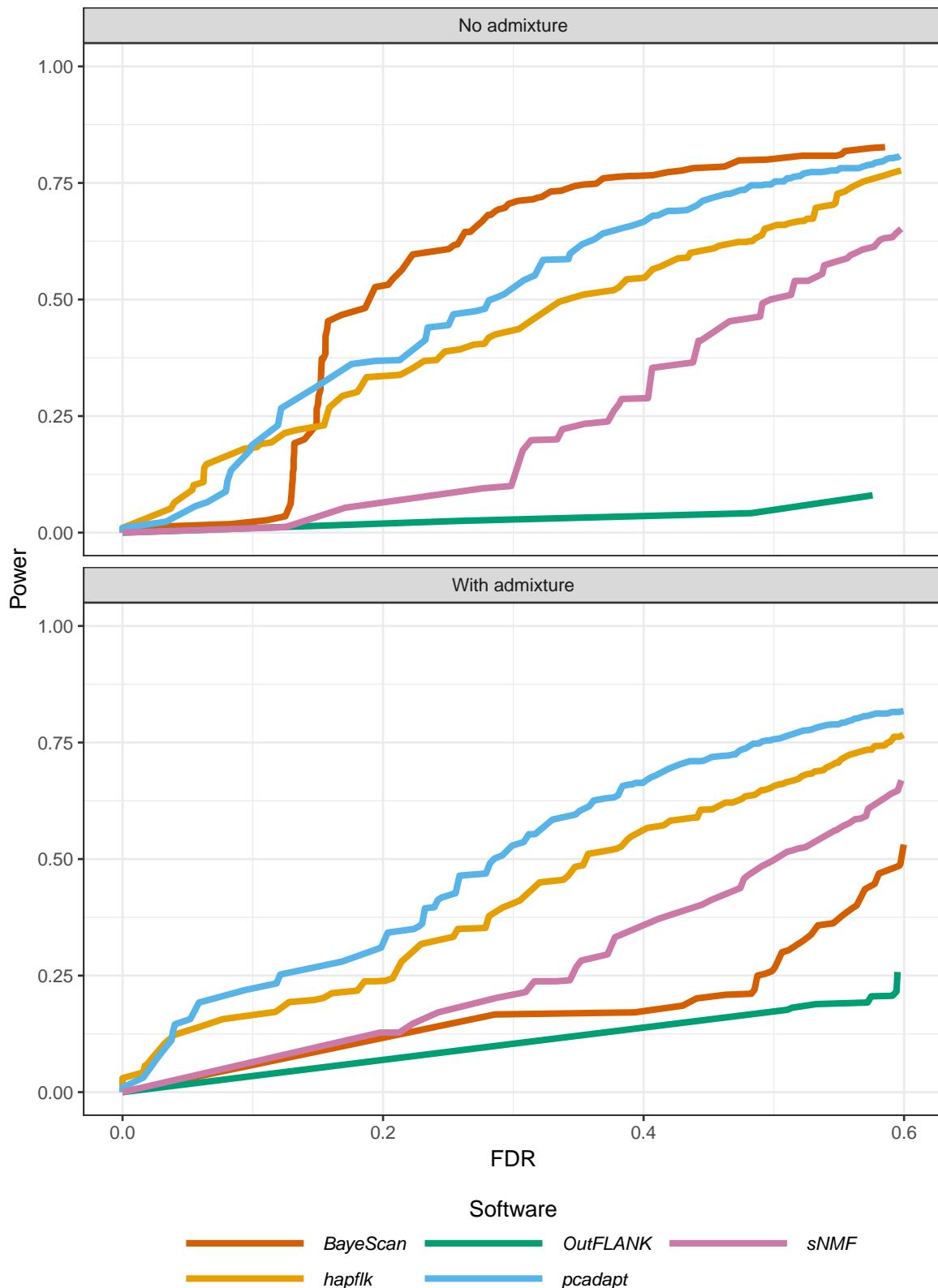


FIGURE B.20 – Statistical power as a function of the proportion of false discoveries for the divergence model.

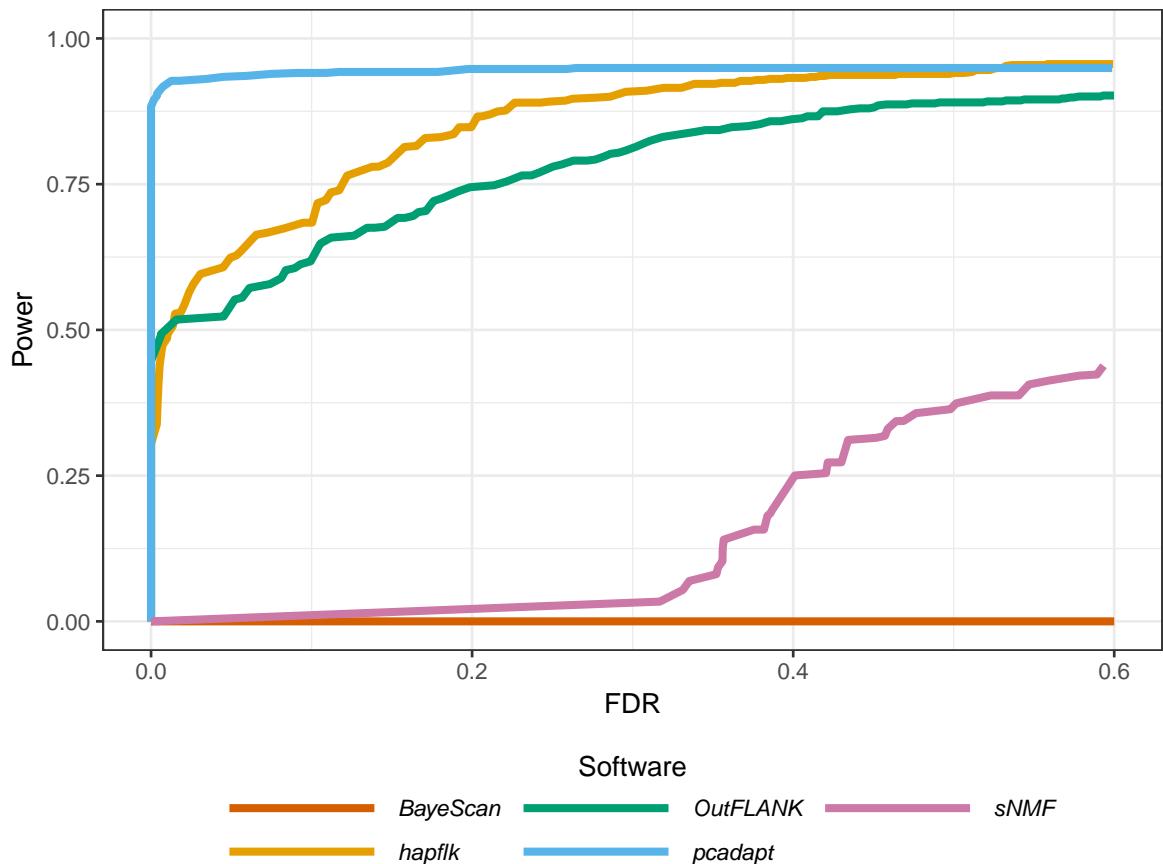


FIGURE B.21 – Statistical power as a function of the proportion of false discoveries for the model of range expansion.

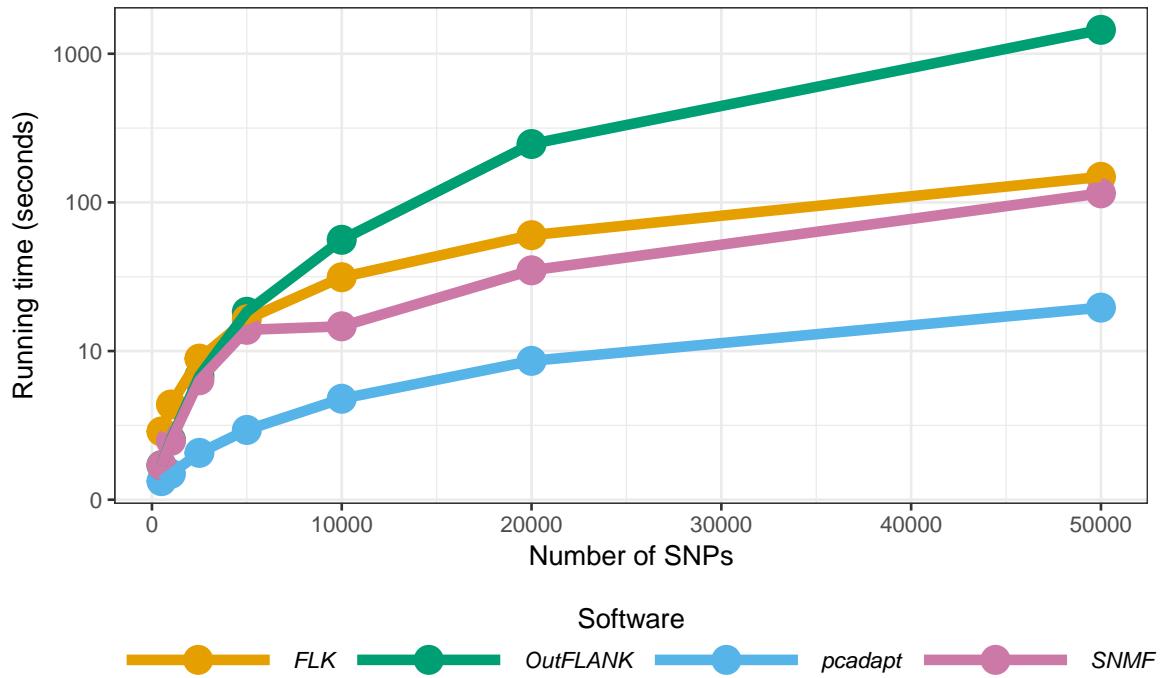


FIGURE B.22 – Running times of the different computer programs. The different programs were run on genotype matrices containing 300 individuals and from 500 to 50,000 SNPs. The characteristics of the computer we used to perform comparisons are the following : OSX El Capitan 10.11.3, 2,5 GHz Intel Core i5, 8 Go 1600 MHz DDR3.

Article 3

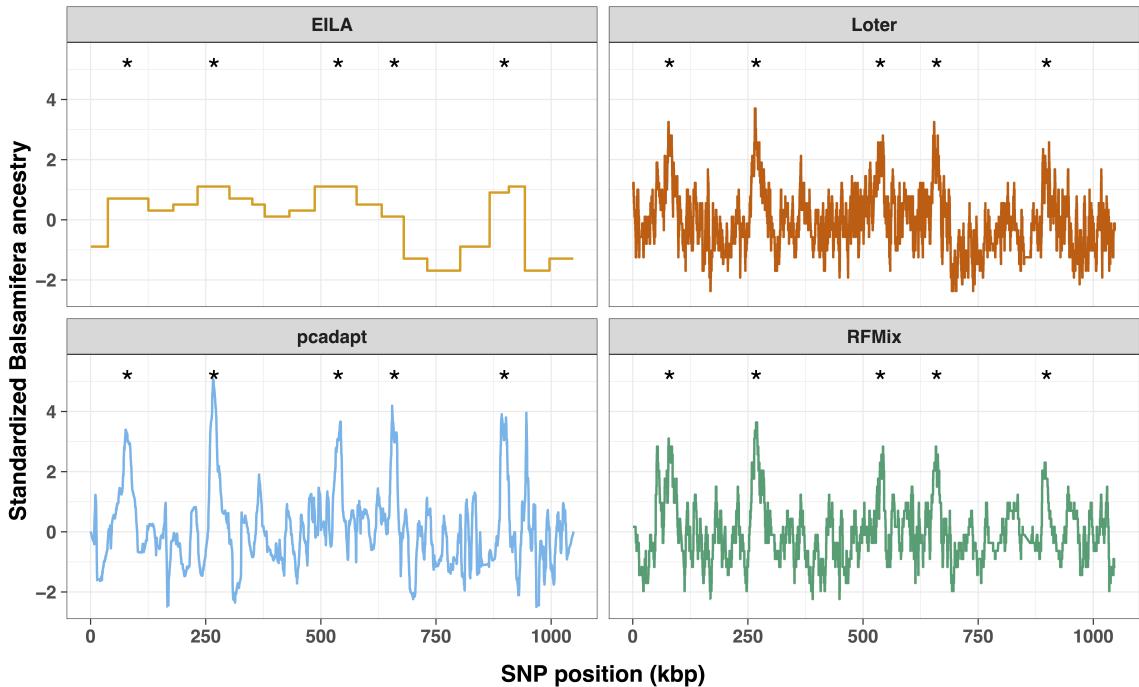


FIGURE B.23 – Standardized average ancestry coefficient computed with pcadapt and different LAI methods (EILA, Loter and RFMix) for a population of simulated admixed individuals. Simulations use phased genotype data from 25 *Populus balsamifera* individuals and 25 *Populus trichocarpa* individuals and assume that admixture took place $\lambda = 100$ generations ago. A total of 25 admixed individuals were generated assuming 30% of *Populus balsamifera* ancestry except from the 5 outlier 500 SNP regions where *Populus balsamifera* ancestry is of 50%.

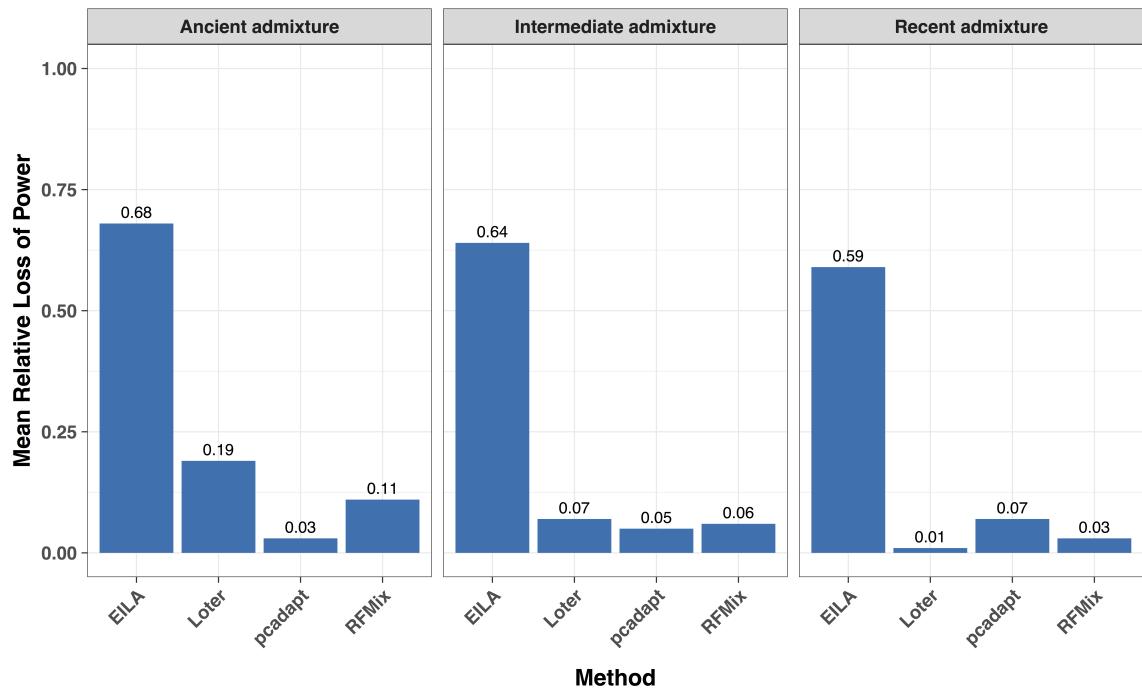


FIGURE B.24 – Relative loss of power of the different methods when compared to an ideal method called *oracle*, which would know ancestry chunks for each admixed individual. The relative power is averaged over the difference Δ_q of ancestry between neutral and outlier regions. Simulations use phased genotype data from 25 *Populus balsamifera* individuals and 25 *Populus trichocarpa* individuals and assume that admixture took place λ generations ago. A total of 25 admixed individuals were generated assuming 70% of *Populus balsamifera* ancestry except from the 5 outlier 500 SNP regions where *Populus balsamifera* ancestry is of $70\% - \Delta_q$ where Δ_q is equal to 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, or 0.50.

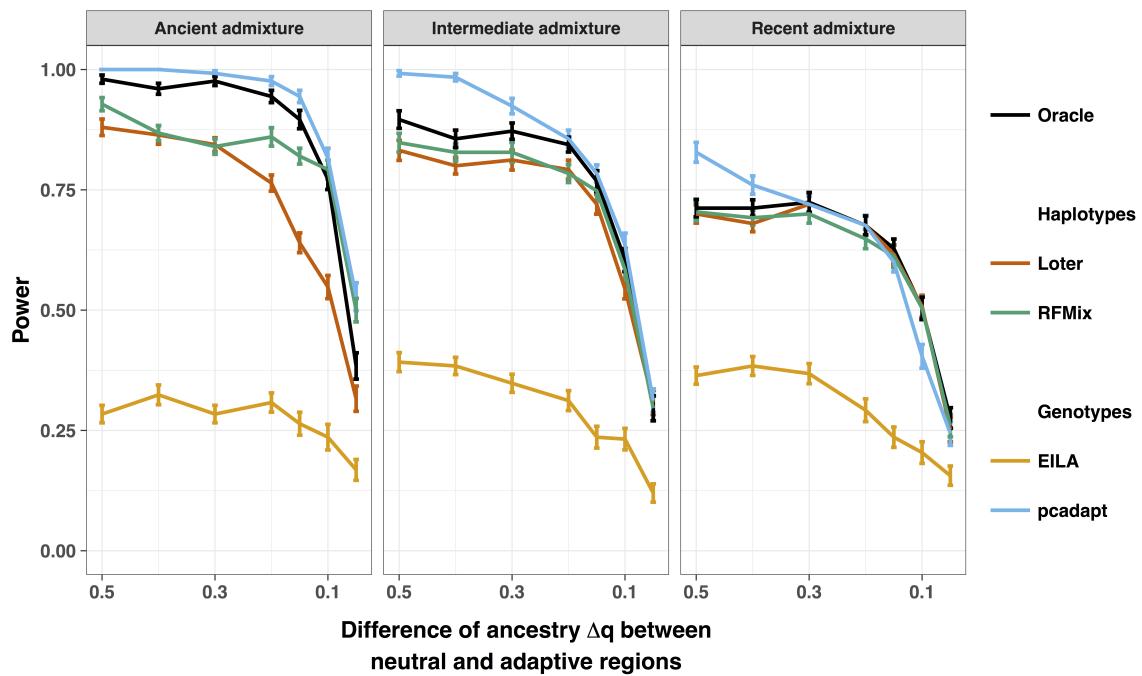


FIGURE B.25 – Proportion of true outlier peaks among the five top peaks found with pcadapt and different LAI methods (EILA, Loter and RFMix) in a scenario where 2 *Populus* populations experienced admixture. Compared to Figure 3.10, we compute maximum of ancestry coefficients within each genomic region instead of mean of ancestry coefficients. Proportion of true outlier peaks are displayed as a function of the difference Δ_q of ancestry between outlier and neutral regions. The three panels correspond to the three different possible values ($\lambda = 10$ or 100 or 1000) of the number of generations since admixture.

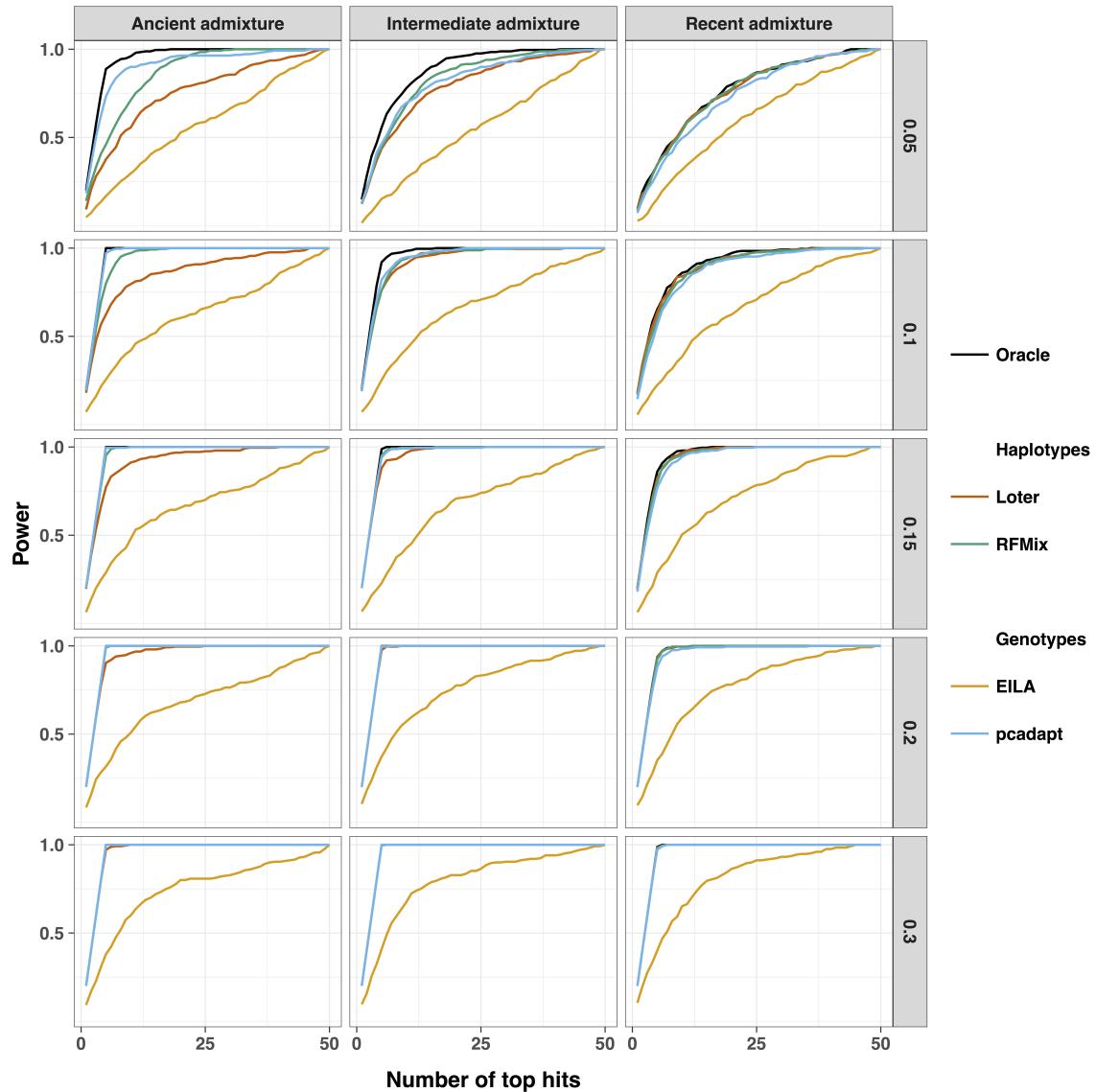


FIGURE B.26 – Proportion of true outlier peaks as a function of the number of top peaks found with pcadapt and different LAI methods (EILA, Loter and RFMix) in a scenario where 2 *Populus* populations experienced admixture. The different panels correspond to the different possible values for the number of generations since admixture occurred ($\lambda = \{10, 100, 1000\}$) and to the different values of the difference of ancestry Δ_q between neutral and outlier regions.

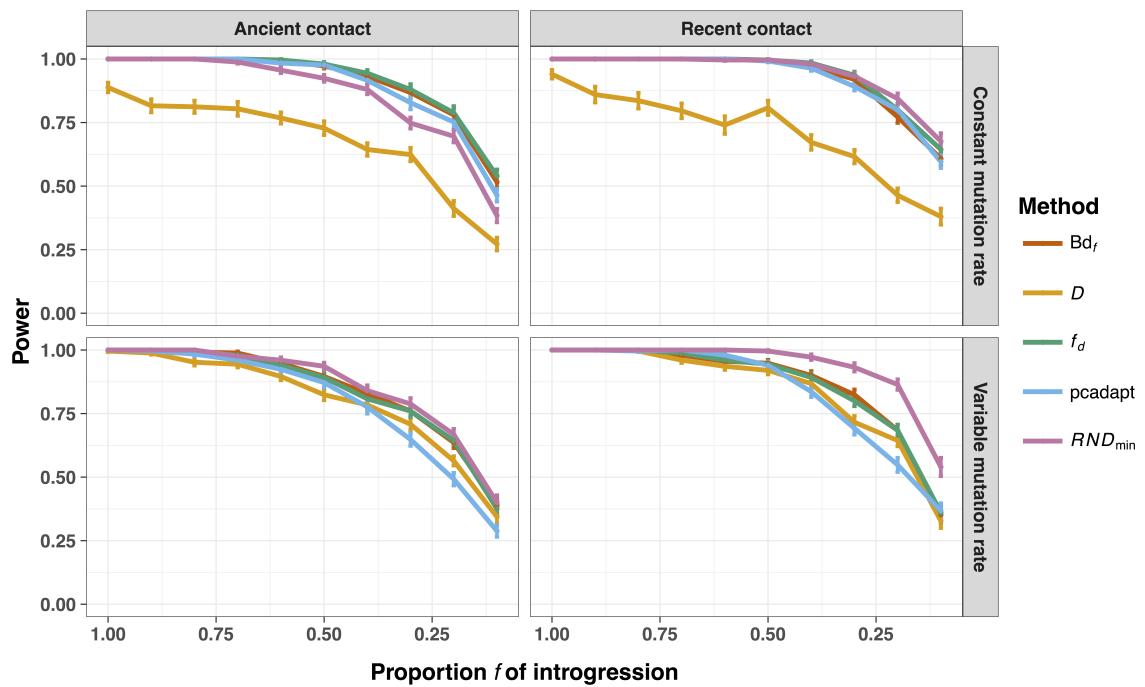
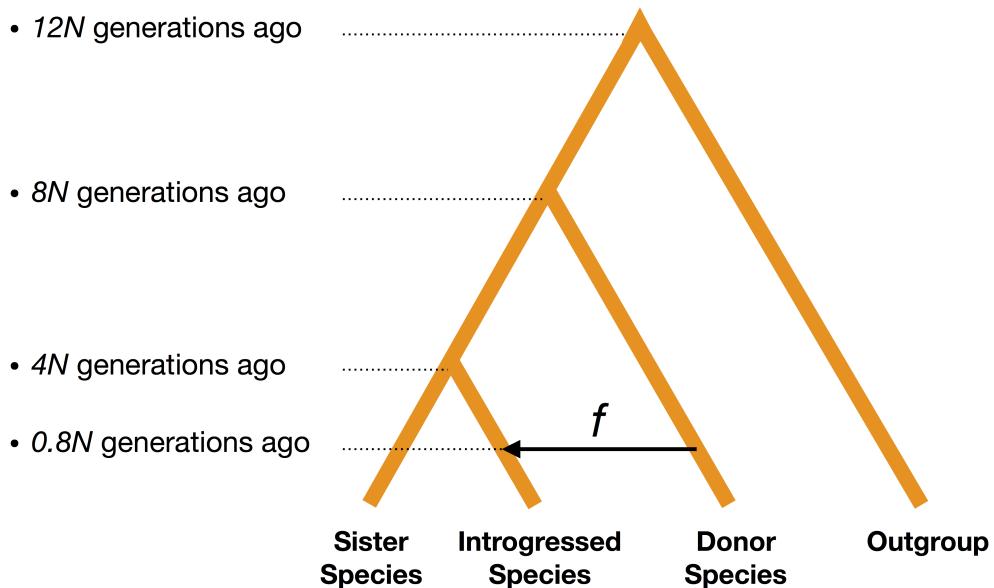


FIGURE B.27 – Power as a function of the proportion of introgression.



Neutral regions: $f = 0$

Introgressed regions: $f > 0$

FIGURE B.28 – Schematic description of simulations under an introgression scenario.

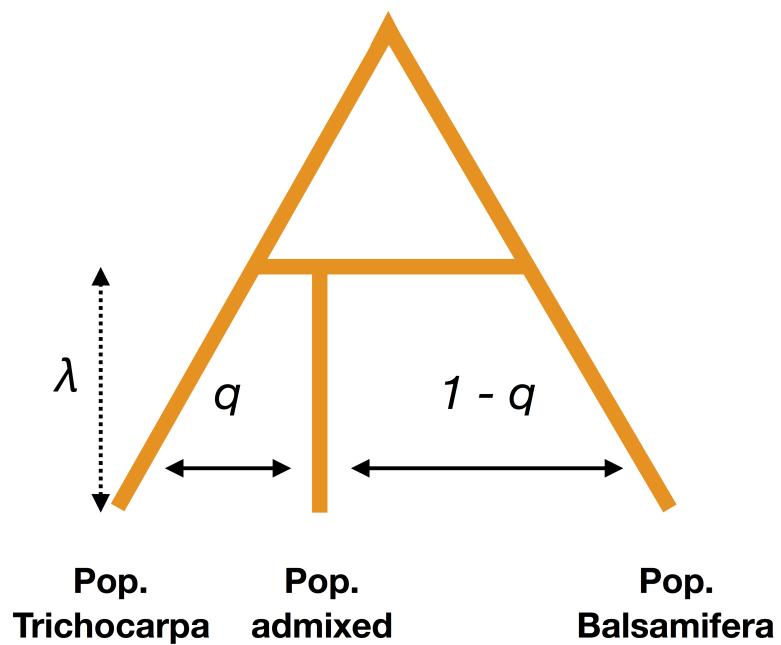


FIGURE B.29 – Schematic description of simulations under an admixture scenario.

Annexe C

R & Python

C.1 Simulations et modèles démographiques

Nous mettons ici à disposition une partie des codes qui ont servi à produire les simulations.

C.1.1 Modèle en îles

```
if (!file.exists("ms/ms")) {  
    system("gcc -o ms/ms ms/ms.c ms/streec.c ms/rand1.c -lm")  
}  
  
### ms : list of parameters ###  
nb_demes <- 3  
nCHR_per_POP <- 50  
nCHR <- nb_demes * nCHR_per_POP  
nIND <- nCHR / 2  
nb_neutral <- 100  
nb_adaptive <- 25  
mig_rate_neutral <- 10  
mig_rate_adaptive <- 0.1  
#####  
  
nCHR_per_POP_string <- nCHR_per_POP  
for (k in 1:(nb_demes - 1)) {  
    nCHR_per_POP_string <- paste(nCHR_per_POP_string, nCHR_per_POP)  
}  
  
mig_rate_neutral_string <- "x"  
mig_rate_adaptive_string <- "x"  
for (k in 1:nb_demes){  
    for (j in 1:nb_demes){
```

```

if ((k == j) && (k > 1)){
  mig_rate_neutral_string <- paste(mig_rate_neutral_string, "x")
  mig_rate_adaptive_string <- paste(mig_rate_adaptive_string, "x")
} else if (k != j) {
  mig_rate_neutral_string <- paste(mig_rate_neutral_string,
                                    mig_rate_neutral)
  mig_rate_adaptive_string <- paste(mig_rate_adaptive_string,
                                    mig_rate_adaptive)
}
}

cmd_neutral <- paste("ms ./ms",
                     nCHR,
                     nb_neutral,
                     "-s 1 -I",
                     nb_demes,
                     nCHR_per_POP_string,
                     "-ma",
                     mig_rate_neutral_string,
                     ">",
                     "data/neutral.txt")

cmd_adaptive <- paste("ms ./ms",
                      nCHR,
                      nb_adaptive,
                      "-s 1 -I",
                      nb_demes,
                      nCHR_per_POP_string,
                      "-ma",
                      mig_rate_adaptive_string,
                      ">",
                      "data/adaptive.txt")

system(cmd_neutral)
system(cmd_adaptive)

file.neutral <- scan(file = "data/neutral.txt",
                     what = "character",
                     sep = "\n",
                     skip = 2)

g.neutral <- NULL

for (locus in 1:nb_neutral){

```

```

res.locus1 <- file.neutral[4:(nCHR + 3)]
file.neutral <- file.neutral[-(1:(nCHR+3))]
g.neutral <- cbind(g.neutral, as.numeric(as.factor(res.locus1)))
}

file.adaptive <- scan(file = "data/adaptive.txt",
                       what = "character",
                       sep = "\n",
                       skip = 2)

g.adaptive <- NULL

for (locus in 1:nb_adaptive){
  res.locus1 <- file.adaptive[4:(nCHR + 3)]
  file.adaptive <- file.adaptive[-(1:(nCHR+3))]
  g.adaptive <- cbind(g.adaptive, as.numeric(as.factor(res.locus1)))
}

g <- cbind(g.neutral, g.adaptive)

x <- pcadapt::pcadapt(t(g), K = 2)
pop <- c(rep("A", 50), rep("B", 50), rep("C", 50))
plot(x, option = "scores", pop = pop)

```

C.1.2 Modèle de divergence

Nous adaptons une version du script Python utilisé dans (Roux et al., 2012), basé sur le module de simulation simuPOP (B. Peng & Kimmel, 2005).

```

#!/usr/bin/env python
from __future__ import division
import simuOpt, types, os, sys, time
simuOpt.setOptions(alleleType = 'long')
from operator import itemgetter
import numpy as np
from simuPOP import *
from simuPOP.utils import *
from simuPOP.sampling import drawRandomSample

def simulate(Ne, Nsam, T1, T2, T3, s10, s11):
    pop = Population(size = Ne,
                      ploidy = 2,
                      loci = [1],
                      infoFields = ['fitness', 'migrate_to'])

```

```

def getfitness10(geno):
    if geno[0] + geno[1] == 0 :
        return 1 - 2 * s10
    if geno[0] + geno[1] == 1 :
        return 1 - s10
    else :
        return 1

def getfitness11(geno):
    if geno[0] + geno[1] == 0 :
        return 1 - 2 * s11
    if geno[0] + geno[1] == 1 :
        return 1 - s11
    else :
        return 1

pop.evolve(
    initOps = [
        InitSex(),
        InitGenotype(loci = ALL_AVAIL,
                      freq = [0.5, 0.5],
                      begin = 0,
                      end = 1)
    ],
    preOps = [
        # resize the ancestral population at the time immediately
        # before the split
        ResizeSubPops([0],
                      sizes = [Ne + Ne],
                      at = T1 - 1),
        ResizeSubPops(["S1_1"],
                      sizes = [Ne + Ne],
                      at = T1 + T2 - 1),
        # split populations in 2 subpopulations
        SplitSubPops(subPops = [0],
                      sizes = [Ne, Ne],
                      names = ["S1_0", "S1_1"],
                      at = T1),
        SplitSubPops(subPops = ["S1_1"],
                      sizes = [Ne, Ne],
                      at = T1 + T2,

```

```

        names = ["S2_0", "S2_1]),

    # apply selection by invoking function getfitness
    PySelector(loci = [0],
                func = getfitness11,
                begin = T1 + T2,
                subPops = ["S2_1"]),

    PySelector(loci = [0],
                func = getfitness10,
                begin = T1,
                subPops = ["S1_0"],
                end = T1 + T2 + T3 - 1)
] ,

matingScheme = RandomMating(ops =
                            Recombinator(intensity = 1)
                        ]),

gen = T1 + T2 + T3

)

sample = drawRandomSample(pop, sizes = [Nsam, Nsam, Nsam])

return sample

Ne = 1000
Nsam = 25
T1 = 10
T2 = 100
T3 = 100
s = 0.1
nSNP = 10

G = np.zeros([3 * Nsam, nSNP])

for i in range(nSNP):
    if i < 1:
        s10 = 2 * s
        s11 = 0.0
    elif i < 2:
        s10 = 0.0
        s11 = s

```

```
else:  
    s10 = 0.0  
    s11 = 0.0  
res = simulate(Ne, Nsam, T1, T2, T3, s10, s11)  
for j in range(3):  
    Sj = res.genotype(j)  
    for k in range(int(len(Sj) / 2)):  
        idx = j * int(len(Sj) / 2) + k  
        G[idx][i] = Sj[2 * k] + Sj[2 * k + 1]  
  
np.savetxt('data/simuPOP.pcadapt', G, fmt = '%i')
```

Bibliographie

- Abdellaoui, A., Hottenga, J.-J., De Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., ... others. (2013). Population structure, migration, and diversifying selection in the netherlands. *European Journal of Human Genetics*, 21(11), 1277–1285.
- Abraham, G., & Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. *PloS One*, 9(4), e93766.
- Abraham, G., Qiu, Y., & Inouye, M. (2017). FlashPCA2 : Principal component analysis of biobank-scale genotype datasets. *Bioinformatics*.
- Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., ... others. (2009). Newly discovered breast cancer susceptibility loci on 3p24 and 17q23. 2. *Nature Genetics*, 41(5), 585–590.
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–1664.
- Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., ... others. (1999). LAPACK users' guide 3rd edn (philadelphia, pa : Society for industrial and applied mathematics).
- Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11), 3179–3190.
- Arnold, M. L., & Martin, N. H. (2009). Adaptation by introgression. *Journal of Biology*, 8(9), 82.
- Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63(3), 221–230.
- Balding, D. J., Bishop, M., & Cannings, C. (2008). *Handbook of statistical genetics*. John Wiley & Sons.
- Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., ... others. (2012). Fast and accurate inference of local ancestry in latino populations. *Bioinformatics*, 28(10), 1359–1367.
- Barreiro, L. B., & Quintana-Murci, L. (2010). From evolutionary genetics to human

- immunology : How selection shapes host defence genes. *Nature Reviews. Genetics*, 11(1), 17.
- Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 40(3), 340–345.
- Bateson, W., & Mendel, G. (1913). *Mendel's principles of heredity*. University press.
- Bazin, E., Dawson, K. J., & Beaumont, M. A. (2010). Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. *Genetics*, 185(2), 587–602.
- Beall, C. M. (2007). Two routes to functional adaptation : Tibetan and andean high-altitude natives. *Proceedings of the National Academy of Sciences*, 104(suppl 1), 8655–8660.
- Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4), 969–980.
- Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., ... Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6), 1111–1120.
- Bierne, N., Roze, D., & Welch, J. J. (2013). Pervasive selection or is it... ? Why are fst outliers sometimes so frequent ? *Molecular Ecology*, 22(8), 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(1), 241–262.
- Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., ... Bustamante, C. D. (2012). PCAdmix : Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, 84(4), 343–364.
- Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews. Genetics*, 4(3), 216.
- Browning, B. L., & Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1), 116–126.
- Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5), 1084–1097.
- Buerkle, C. A., & Lexer, C. (2008). Admixture as the basis for genetic mapping. *Trends in Ecology & Evolution*, 23(12), 686–694.
- Cadima, J., & Jolliffe, I. T. (1995). Loading and correlations in the interpretation of

- principle components. *Journal of Applied Statistics*, 22(2), 203–214.
- Cagliani, R., Fumagalli, M., Riva, S., Pozzoli, U., Comi, G. P., Menozzi, G., ... Sironi, M. (2008). The signature of long-standing balancing selection at the human defensin β -1 promoter. *Genome Biology*, 9(9), R143.
- Calvetti, D., Reichel, L., & Sorensen, D. C. (1994). An implicitly restarted lanczos method for large symmetric eigenvalue problems. *Electronic Transactions on Numerical Analysis*, 2(1), 21.
- Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., & Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11), 1553–1565.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- Cavalli-Sforza, L. (1994). Francesco. qui sommes-nous ? Une histoire de diversité humaine. *Trans. Brun, Françoise. Flammarion Ed. Paris : Centre National Des Lettres*.
- Caye, K., Deist, T. M., Martins, H., Michel, O., & François, O. (2016). TESS3 : Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, 16(2), 540–548.
- Charlesworth, B., & Charlesworth, D. (2009). Darwin and genetics. *Genetics*, 183(3), 757–766.
- Chen, G., Lee, S., Zhu, Z., Benyamin, B., & Robinson, M. (2016). EigenGWAS : Finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, 117(1), 51.
- Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, 20(3), 393–402.
- Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., ... Tyler-Smith, C. (2014). Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15(6), R88.
- Consortium, 1. G. P., & others. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56.
- Darwin, C. (1980). L'Origine des espèces, trad. *Edmond Barbier (1876)*, Paris, Masspero.
- Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., ... others. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405), 94.
- Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-

- Rechavi, M., & Excoffier, L. (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*, 30(7), 1544–1558.
- Delaneau, O., Marchini, J., & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2), 179–181.
- Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4), 997–1004.
- Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148(4), 1667–1686.
- Dray, S., & Josse, J. (2015). Principal component analysis with missing values : A comparative survey of methods. *Plant Ecology*, 216(5), 657–667.
- Duforet-Frebourg, N. (2014). *Statistiques bayésiennes en génétique des populations : Modèle à facteurs et processus gaussiens pour étudier la variation génétique neutre et adaptative* (PhD thesis). Grenoble.
- Duforet-Frebourg, N., Bazin, E., & Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using a bayesian factor model. *Molecular Biology and Evolution*, 31(9), 2483–2495.
- Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. (2015). Detecting genomic signatures of natural selection with principal component analysis : Application to the 1000 genomes data. *Molecular Biology and Evolution*, 33(4), 1082–1093.
- Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8), 2239–2252.
- Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4), 285.
- Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes : Application to human mitochondrial dna restriction data. *Genetics*, 131(2), 479–491.
- Fagny, M., Patin, E., Enard, D., Barreiro, L. B., Quintana-Murci, L., & Laval, G. (2014). Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular Biology and Evolution*, 31(7), 1850–1868.
- Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3), 929–941.
- Feder, J. L., Xie, X., Rull, J., Velez, S., Forbes, A., Leung, B., ... Aluja, M. (2005). Mayr, dobzhansky, and bush and the complexities of sympatric speciation in rhagoletis. *Proceedings of the National Academy of Sciences*, 102(suppl 1), 6573–

- 6580.
- 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(2), 977–993.
- Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., & Excoffier, L. (2014). Widespread signals of convergent adaptation to high altitude in asia and america. *The American Journal of Human Genetics*, 95(4), 394–407.
- Fraïsse, C., Roux, C., Welch, J. J., & Bierne, N. (2014). Gene-flow in a mosaic hybrid zone : Is local introgression adaptive ? *Genetics*, 197(3), 939–951.
- François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, 25(2), 454–469.
- Frichot, E., & François, O. (2015). LEA : An r package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6(8), 925–929.
- Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4), 973–983.
- Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G. P., Bresolin, N., Clerici, M., & Sironi, M. (2010). Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. *PLoS Genetics*, 6(2), e1000849.
- Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, 7(11), e1002355.
- Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2016). Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3), 456–472.
- Gautier, M., Gharbi, K., Cezaud, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... Estoup, A. (2013). The effect of rad allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, 22(11), 3165–3178.
- Gayon, J. (1992). Darwin et l'après-darwin : Une histoire de l'hypothèse de sélection dans la théorie de l'évolution. Kimé.
- Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., ... others. (2003). The international hapmap project.
- Gillespie, J. H. (2010). *Population genetics : A concise guide*. JHU Press.
- Gogol-Döring, A., & Chen, W. (2012). An overview of the analysis of next generation sequencing data. *Next Generation Microarray Bioinformatics : Methods and*

- Protocols*, 249–257.
- Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., ... others. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4), 703–713.
- Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., ... others. (2015). The african genome variation project shapes medical genetics in africa. *Nature*, 517(7534), 327.
- Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1), 205–220.
- Haasl, R. J., & Payseur, B. A. (2016). DETECTING selection in natural populations : MAKING sense of genome scans and towards alternative solutions : Fifteen years of genomewide scans for selection : Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1), 5.
- Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2), 217–288.
- Hamblin, M. T., Thompson, E. E., & Di Rienzo, A. (2002). Complex signatures of natural selection at the duffy blood group locus. *The American Journal of Human Genetics*, 70(2), 369–383.
- Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., ... Kidd, K. K. (2007). Evidence of positive selection on a class i adh locus. *The American Journal of Human Genetics*, 80(3), 441–456.
- Hancock, A. M., Witonsky, D. B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., ... others. (2010). Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 8924–8930.
- Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2008). Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, 4(2), e32.
- Hao, W., Song, M., & Storey, J. D. (2015). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5), 713–721.
- Harrison, R. G., & others. (1990). Hybrid zones : Windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, 7, 69–128.
- Hedrick, P. W. (2013). Adaptive introgression in animals : Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular*

- Ecology*, 22(18), 4606–4618.
- Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., ... others. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019), 920–924.
- Hollox, E. J., & Armour, J. A. (2008). Directional and balancing selection in human beta-defensins. *BMC Evolutionary Biology*, 8(1), 113.
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations : Defining, estimating and interpreting fst. *Nature Reviews. Genetics*, 10(9), 639.
- Hu, H., Petousi, N., Glusman, G., Yu, Y., Bohlender, R., Tashi, T., ... others. (2017). Evolutionary history of tibetans inferred from whole-genome sequencing. *PLoS Genetics*, 13(4), e1006675.
- Hudson, R. R. (2002). Generating samples under a wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2), 337–338.
- Hufford, M. B., Lubinsky, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, 9(5), e1003477.
- Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in europe. *PLoS Computational Biology*, 5(8), e1000491.
- Jackson, D. A. (1993). Stopping rules in principal components analysis : A comparison of heuristical and statistical approaches. *Ecology*, 74(8), 2204–2214.
- Jay, F., Sjödin, P., Jakobsson, M., & Blum, M. G. (2012). Anisotropic isolation by distance : The main orientations of human genetic differentiation. *Molecular Biology and Evolution*, 30(3), 513–525.
- Jeong, C., & Di Rienzo, A. (2014). Adaptations to local environments in modern human populations. *Current Opinion in Genetics & Development*, 29, 1–8.
- Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., ... Di Rienzo, A. (2014). Admixture facilitates genetic adaptations to high altitude in tibet. *Nature Communications*, 5, 3281.
- Jiao, H., Arner, P., Hoffstedt, J., Brodin, D., Dubern, B., Czernichow, S., ... others. (2011). Genome wide association study identifies kcnma1 contributing to human obesity. *BMC Medical Genomics*, 4(1), 51.
- Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis* (pp. 115–128). Springer.
- Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology*

- Letters*, 7(12), 1225–1241.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.
- Kofler, R., & Schlötterer, C. (2012). Gowinda : Unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15), 2084–2085.
- Kudaravalli, S., Veyrieras, J.-B., Stranger, B. E., Dermitzakis, E. T., & Pritchard, J. K. (2008). Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution*, 26(3), 649–658.
- Landguth, E., Cushman, S., Murphy, M., & Luikart, G. (2010). Relationships between migration rates and landscape resistance assessed using individual-based simulations. *Molecular Ecology Resources*, 10(5), 854–862.
- Lange, K., Papp, J. C., Sinsheimer, J. S., & Sobel, E. M. (2014). Next-generation statistical genetics : Modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application*, 1, 279–300.
- Lasky, J. R., Des Marais, D. L., McKAY, J., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of arabidopsis thaliana : The roles of geography and climate. *Molecular Ecology*, 21(22), 5512–5529.
- Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), e1002453.
- Lehoucq, R. B., & Sorensen, D. C. (1996). Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4), 789–821.
- Lewontin, R., & 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, H., & Ralph, P. (2016). Local pca shows how the effect of population structure differs along the genome. *bioRxiv*, 070615.
- Li, J., Liu, Y., Xin, X., Kim, T. S., Cabeza, E. A., Ren, J., ... Zhang, Z. (2012). Evidence for positive selection on a number of microrna regulatory interactions during recent human evolution. *PLoS Genetics*, 8(3), e1002578.
- Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of fst outlier tests. *Molecular Ecology*, 23(9), 2178–2192.
- Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, 24(5), 1031–1046.
- Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics : From genotyping to genome typing. *Nature*

- Reviews. Genetics*, 4(12), 981.
- Luu, K., Bazin, E., & Blum, M. G. (2017). Pcadapt : An r package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17(1), 67–77.
- Ma, J., & Amos, C. I. (2012). Principal components analysis of population admixture. *PloS One*, 7(7), e40115.
- Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix : A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2), 278–288.
- Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307–317.
- Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of abba–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32(1), 244–257.
- Martins, H., Caye, K., Luu, K., Blum, M. G., & Francois, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology*, 25(20), 5029–5042.
- McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10), e1000686.
- Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358), 786–792.
- Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11), 659–669.
- Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., ... others. (2016). The real cost of sequencing : Scaling computation to keep pace with data generation. *Genome Biology*, 17(1), 53.
- Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., ... others. (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.
- Nelson, R. M., Wallberg, A., Simões, Z. L. P., Lawson, D. J., & Webster, M. T. (2017). Genome-wide analysis of admixture and adaptation in the africanized honeybee. *Molecular Ecology*.
- Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39, 197–218.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... others.

- (2008). Genes mirror geography within europe. *Nature*, 456(7218), 98.
- Ochoa, A., & Storey, J. D. (2016). F st and kinship for arbitrary population structures i : Generalized definitions. *bioRxiv*, 083915.
- Oeggerli, M., Tian, Y., Ruiz, C., Wijker, B., Sauter, G., Obermann, E., ... others. (2012). Role of kcnma1 in breast cancer. *PLoS One*, 7(8), e41664.
- Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., ... Jiggins, C. D. (2012). Adaptive introgression across species boundaries in heliconius butterflies. *PLoS Genetics*, 8(6), e1002752.
- Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., ... others. (2017). Dispersals and genetic adaptation of bantu-speaking populations in africa and north america. *Science*, 356(6337), 543–546.
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), e190.
- Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, 25(11), 2337–2360.
- Peng, B., & Kimmel, M. (2005). SimuPOP : A forward-time population genetics simulation environment. *Bioinformatics*, 21(18), 3686–3687.
- Peng, Y., Shi, H., Qi, X.-b., Xiao, C.-j., Zhong, H., Run-lin, Z. M., & Su, B. (2010). The adh1b arg47his polymorphism in east asian populations and expansion of rice domestication in history. *BMC Evolutionary Biology*, 10(1), 15.
- Petry, D. (1983). The effect on neutral gene flow of selection at a linked locus. *Theoretical Population Biology*, 23(3), 300–313.
- Pfeifer, B., & Kapan, D. D. (2017). Estimates of introgression as a function of pairwise distances. *bioRxiv*, 154377.
- Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., ... others. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5), 826–837.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904.
- Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., ... Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), e1000519.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Prive, F., Aschard, H., & Blum, M. G. (2017). Efficient management and analysis of

- large-scale genome-wide data with two r packages : Bigstatsr and bigsnpr. *bioRxiv*, 190926.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... others. (2007). PLINK : A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559–575.
- Rambaut, A., & Grass, N. C. (1997). Seq-gen : An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3), 235–238.
- Rheindt, F. E., Fujita, M. K., Wilton, P. R., & Edwards, S. V. (2013). Introgression and phenotypic assimilation in zimmerius flycatchers (tyrannidae) : Population genetic and phylogenetic inferences from genome-wide snps. *Systematic Biology*, 63(2), 134–152.
- Riebler, A., Held, L., & Stephan, W. (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178(3), 1817–1829.
- Roll-Hansen, N. (2014). The holist tradition in twentieth century genetics. wilhelm johannsen's genotype concept. *The Journal of Physiology*, 592(11), 2431–2438.
- Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, 25(11), 2387–2397.
- Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8, 283–297.
- Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., & Vekemans, X. (2012). Recent and ancient signature of balancing selection around the s-locus in arabidopsis halleri and a. lyrata. *Molecular Biology and Evolution*, 30(2), 435–447.
- Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., ... Lander, E. (2006). Positive natural selection in the human lineage. *Science*, 312(5780), 1614–1620.
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., ... others. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164), 913.
- Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.
- Severson, K. A., Molaro, M. C., & Braatz, R. D. (2017). Principal component analysis of process datasets with missing values. *Processes*, 5(3), 38.
- Smith, J., & Kronforst, M. R. (2013). Do heliconius butterfly species exchange mimicry alleles ? *Biology Letters*, 9(4), 20130503.
- Song, Y., Endepols, S., Kleemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., ... Kohn, M. H. (2011). Adaptive introgression of anticoagulant rodent poison resistance

- by hybridization between old world mice. *Current Biology*, 21(15), 1296–1301.
- Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era : Is it still useful ? *Nature Reviews Genetics*, 16(1), 33–44.
- Stevison, L. (2008). Hybridization and gene flow. *Nature Education*, 1(1), 111.
- Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.
- Suarez-Gonzalez, et a., Adriana. (2016). Genomic and functional approaches reveal a case of adaptive introgression from *populus balsamifera* (balsam poplar) in *p. trichocarpa* (black cottonwood). *Molecular Ecology*, 2427–2442.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3), 585–595.
- Team, R. C. (2015). R : A language and environment for statistical computing [internet]. vienna, austria : R foundation for statistical computing ; 2014.
- Thornton, T. A., & Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genetic Epidemiology*, 38(S1).
- Villemereuil, P., & Gaggiotti, O. E. (2015). A new fst-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11), 1248–1258.
- Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47, 97–120.
- vonHoldt, B. M., Kays, R., Pollinger, J. P., & Wayne, R. K. (2016). Admixture mapping identifies introgressed genomic regions in north american canids. *Molecular Ecology*, 25(11), 2443–2453.
- vonHoldt, B., Fan, Z., Ortega-Del Vecchyo, D., Wayne, R. K., & others. (2017). EPAS1 variants in high altitude tibetan wolves were selectively introgressed into highland dogs. *PeerJ*, 5, e3522.
- Waples, R. S., & Gaggiotti, O. (2006). INVITED review : What is a population ? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15(6), 1419–1439.
- Weir, B. S., & Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6), 1358–1370.
- Wetterstrand, K. A. (2013). DNA sequencing costs : Data from the nhgri genome sequencing program (gsp).
- Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation : Inference of a null model through trimming the distribution of f

- st. *The American Naturalist*, 186(S1), S24–S36.
- Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., ... others. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, 111(13), 4832–4837.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, 3(6), e90.
- Wright, S. (1943). Isolation by distance. *Genetics*, 28(2), 114.
- Wu, D.-D., & Zhang, Y.-P. (2010). Positive selection drives population differentiation in the skeletal genes in modern humans. *Human Molecular Genetics*, 19(12), 2341–2346.
- Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., ... others. (2010). A genome-wide search for signals of high-altitude adaptation in tibetans. *Molecular Biology and Evolution*, 28(2), 1003–1011.
- Yang, J. J., Li, J., Buu, A., & Williams, L. K. (2013). Efficient inference of local ancestry. *Bioinformatics*, 29(21), 2750–2756.
- Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, 44(6), 725–731.
- Yuan, K.-H., & Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology*, 63(2), 273–291.
- Zhang, W., Dasmahapatra, K. K., Mallet, J., Moreira, G. R., & Kronforst, M. R. (2016). Genome-wide introgression among distantly related heliconius butterfly species. *Genome Biology*, 17(1), 25.