UNIVERSITÉ GRENOBLE-ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE-ALPES

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

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)**

Méthodes statistiques en grande dimension pour l'étude de l'adaptation biologique à l'aide de larges bases de données génomiques

Thèse soutenue publiquement le 31 octobre 2017,
devant le jury composé de :

Remerciements

The preface pretty much says it all.
Second paragraph of abstract starts here.

Préface

blabla

Table des matières

Liste des tableaux

Table des figures

Abstract

** Résumé **

La nécessité de développer des outils d'analyse exploratoire capables de traiter de larges volumes de données. ** Abstract **

Chapitre 1

Introduction

1.1 La génétique des populations

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 de gène et donc de génotype (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 serait 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 qu'elle fut globalement bien accueillie par la communauté scientifique, sa théorie 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 est essentielle. 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.

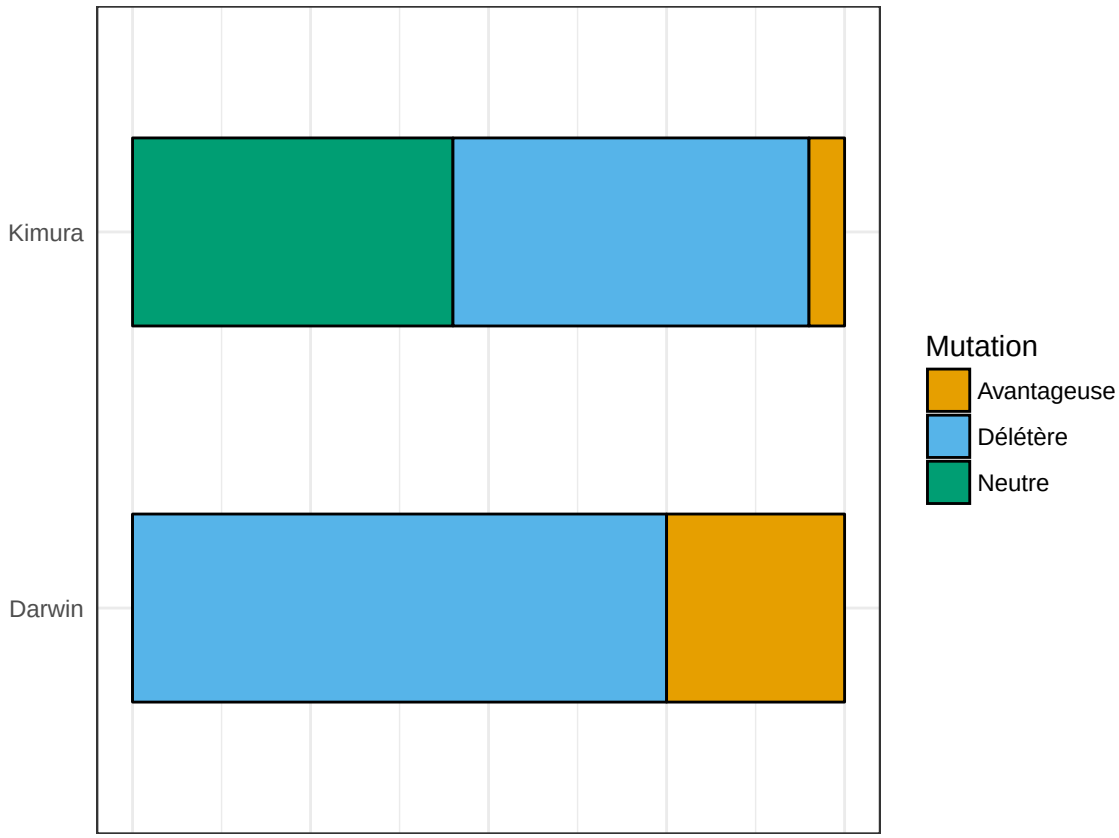


FIGURE 1.1 – Représentation schématique des probabilités d'occurrence pour chaque type de mutation pour la théorie sélectionniste de Darwin et pour la théorie neutraliste de Kimura (Bromham & Penny, 2003).

Une des composantes principales de cette nouvelle théorie consiste à affirmer que les fluctuations aléatoires dans la fréquence des allèles, 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, n'occasionnant pas de changement de phénotype.

1.2.3 Sélection naturelle

1.2.4 Sélection sexuelle

1.2.5 Dérive génétique

La dérive génétique est un mécanisme important en génétique des populations. Les modèles statistiques Nous illustrons ici le principe de dérive génétique à l'aide du modèle de Wright-Fisher tel qu'il est présenté dans l'ouvrage *Population Genetics* (Gillespie, 2010).

Ten simulations of random genetic drift of a single given allele with an initial frequency distribution 0.5 measured over the course of 50 generations, repeated in

three reproductively synchronous populations of different sizes. In these simulations, alleles drift to loss or fixation (frequency of 0.0 or 1.0) only in the smallest population

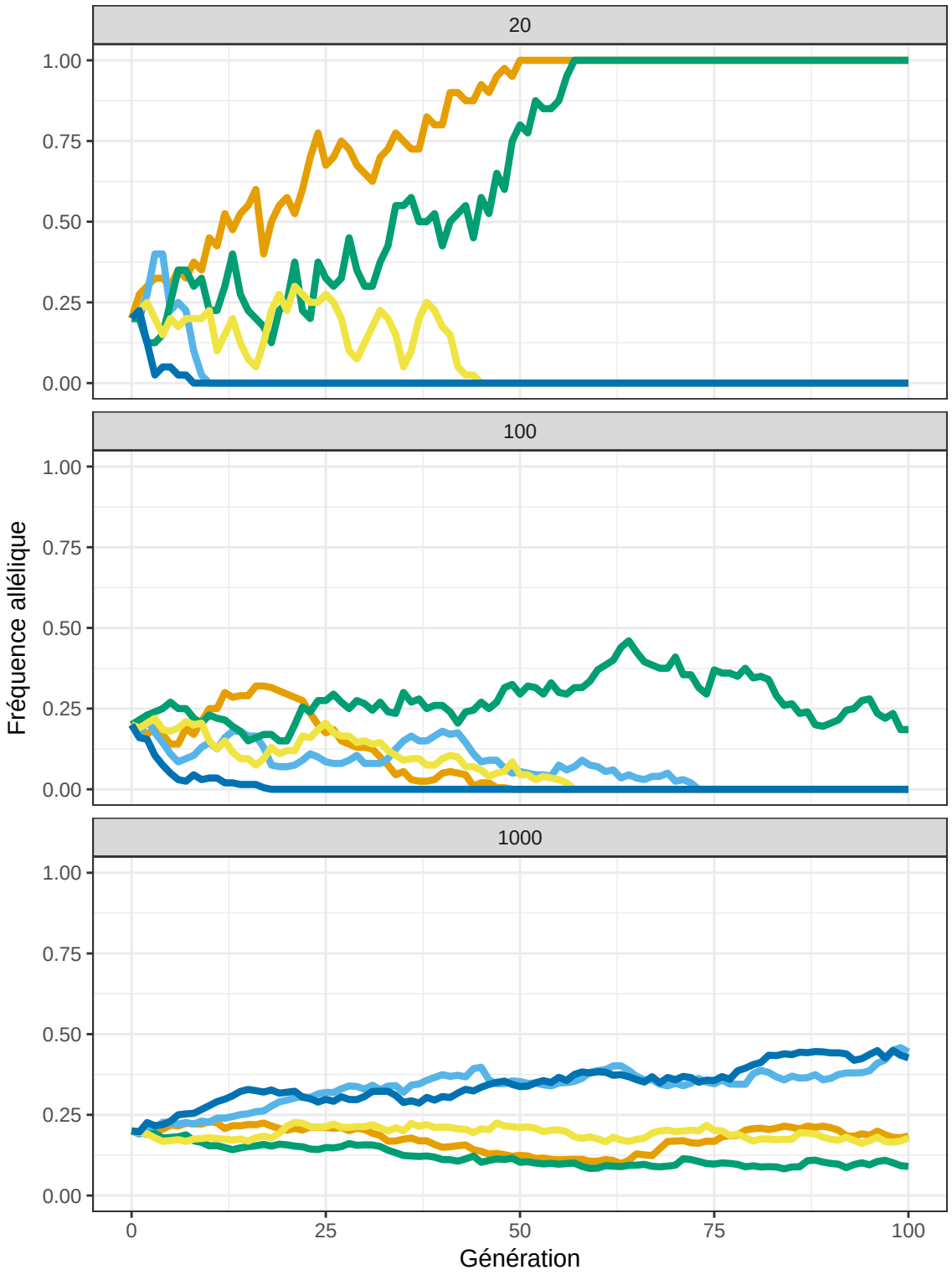


FIGURE 1.2 – Simulation numérique de la dérive génétique. La fréquence de l'allèle étudié est simulée pour 5 populations constituées chacune de 20 individus sur une période de 100 générations. Dans chaque population, la fréquence de l'allèle est initialement de 0.20.

La fréquence d'un allèle au sein d'une population est principalement impactée par deux facteurs :

— Le nombre d'individus composant la population
— Les lois de Mendel

La figure 1.2 met en évidence deux propriétés de la dérive génétique :

— Les fréquences alléliques évoluent de façon indépendante d'une population à une autre.
— La dérive génétique entraîne une perte de diversité allélique au sein des populations de petite taille. Dans le modèle de Wright-Fisher, les fréquences alléliques finissent éventuellement par atteindre les états dits absorbants que sont 0 et 1.

Wiki : "En particulier, l'hypothèse de neutralité, sous laquelle les mutations n'ont aucune influence sur la valeur sélective, est l'hypothèse nulle généralement retenue dans les travaux où une telle hypothèse est nécessaire."

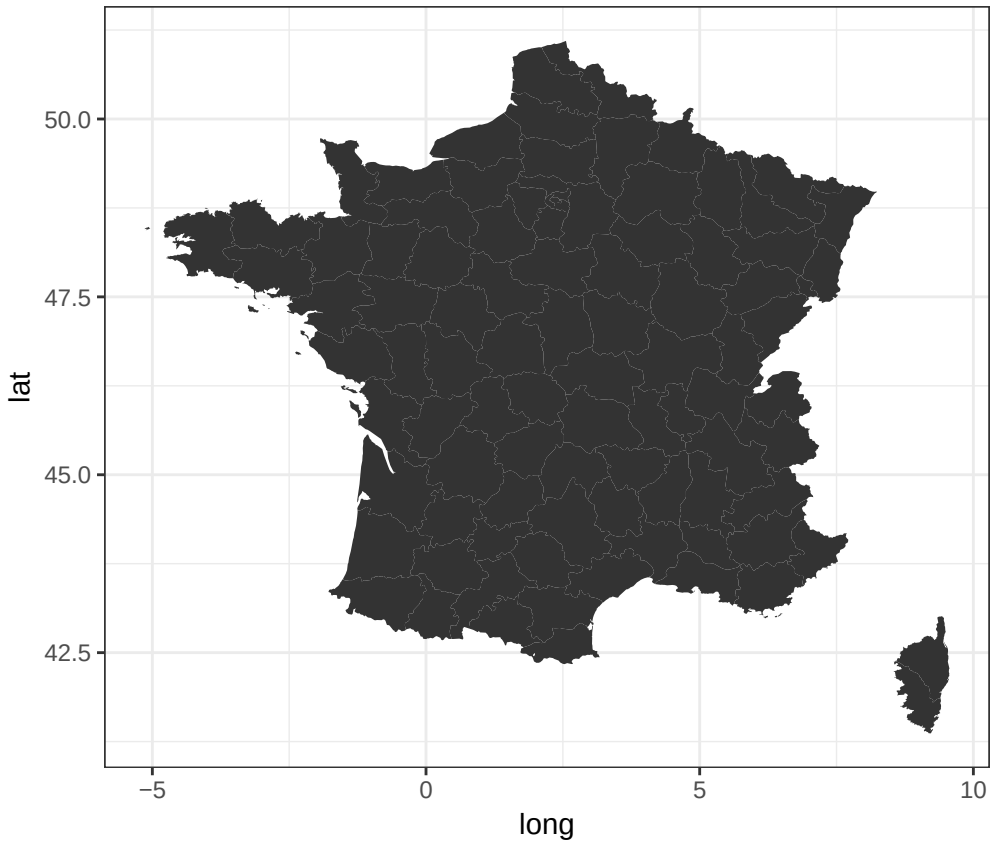
La formulation d'une hypothèse ou d'un ensemble d'hypothèses permettant de décrire un processus évolutif en l'absence de sélection, portant généralement le nom de *modèle neutre*, est souvent de première nécessité dans toute démarche visant à caractériser un mécanisme de sélection. La donnée d'observations mettant en défaut le modèle neutre aura pour conséquences de créditer davantage une hypothèse invoquant un processus de sélection.

1.2.6 Mutations aléatoires

Si la dérive génétique entraîne une perte de diversité allélique, les mutations favorisent quant à elle le maintien des variations génétiques entre les populations (Gillespie, 2010). Une mutation peut survenir à un locus donné avec une probabilité spécifique à chaque espèce, appelée *taux de mutation*.

1.2.7 Flux de gène

Le flux de gène est le résultat d'évènements migratoires, initiés par des individus appartenant à une population donnée, vers une seconde population dont le pool génique diffère éventuellement de la population d'origine.



1.3 Adaptation

1.3.1 Adaptation locale

Haut plateau, Lactase (Jeong & Di Rienzo, 2014)
Sélection naturelle Environnement hétérogène

1.4 Données de séquençage nouvelle génération

1.4.1 Next-Generation Sequencing (NGS)

Exome, Genome, transcriptome, etc. . .

Le séquençage nouvelle génération 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 (Wetterstrand, 2013).

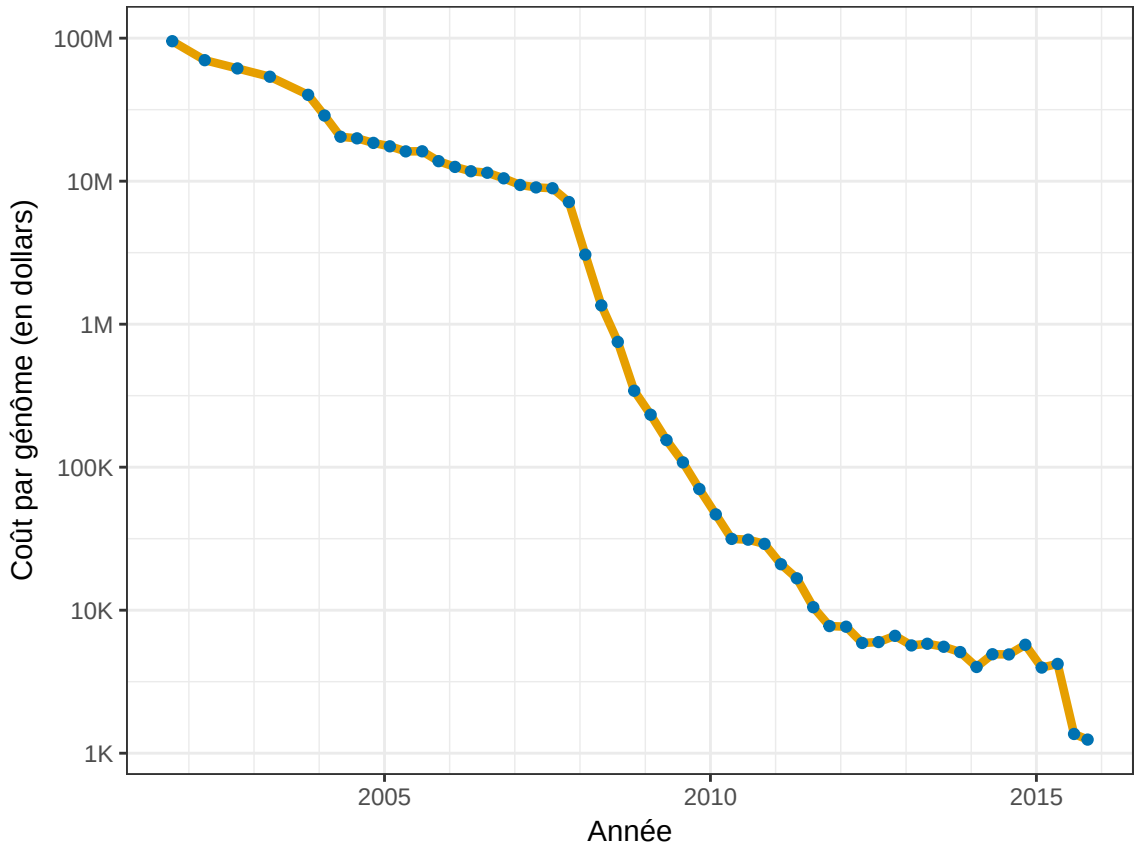


FIGURE 1.3 – Évolution des coûts de séquençage depuis 2001 (Wetterstrand, 2013).

Le séquençage nouvelle génération génèrant de considérables volumes de données, de nouvelles problèmatiques 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). Financement croissant pour l'acquisition de données NGS (Muir et al., 2016).

1.4.2

L'accumulation de données, aussi bien en termes d'observations qu'en termes de variables, laisse à penser que le traitement de celles-ci pourrait permettre de détecter efficacement les variables qui sont responsables ou qui influencent un phénomène particulier. Cela pourrait être par exemple l'utilisation de bases de données automobiles pour prédire la durée de vie de véhicules neufs, ou encore celle de données climatiques pour estimer les variations de température auxquelles pourrait être sujette notre planète. Cette accumulation massive s'accompagne tout de même d'un phénomène bien connu en statistiques, phénomène qui porte le nom de "curse of dimensionality" (Giraud, 2014).

1.4.3 Les marqueurs génétiques

Au concept de génétique est souvent associé l'acronyme ADN, correspondant au nom de la molécule d'Acide désoxyribonucléique.

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder. `https://ghr.nlm.nih.gov/primer/basics/dna` Chaque chromosome est constitué de deux brins d'ADN. La structure spatiale de l'ADN n'étant pas prise en compte dans les travaux présentés ici, nous en garderons une représentation unidimensionnelle.

D'un point de vue statistique, seuls les sites nucléotidiques potentiellement variables d'une observation à une autre présentent un intérêt. Ces variations génétiques peuvent se manifester sous différentes formes, nous amenant à les classer d'autant de façons :

— Microsatellite :

Jusqu'à présent, les microsatellites ont connu un succès important, notamment grâce à la popularisation de techniques telles que la PCR (*Polymerase Chain Reaction*). Leur étude a permis de mettre en évidence l'implication de divers blablabla. Un microsatellite est repérable par la répétition successive de petits motifs chacun composé de 1 à 4 acides aminés.

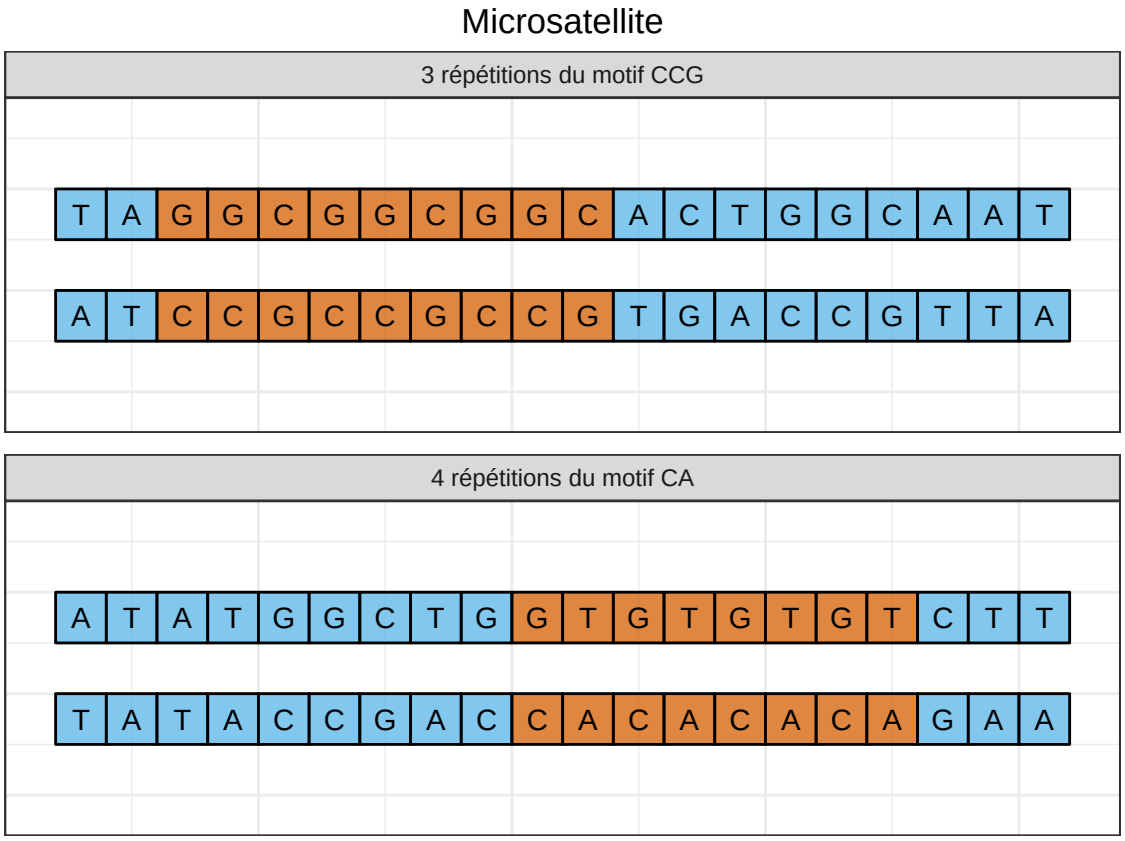
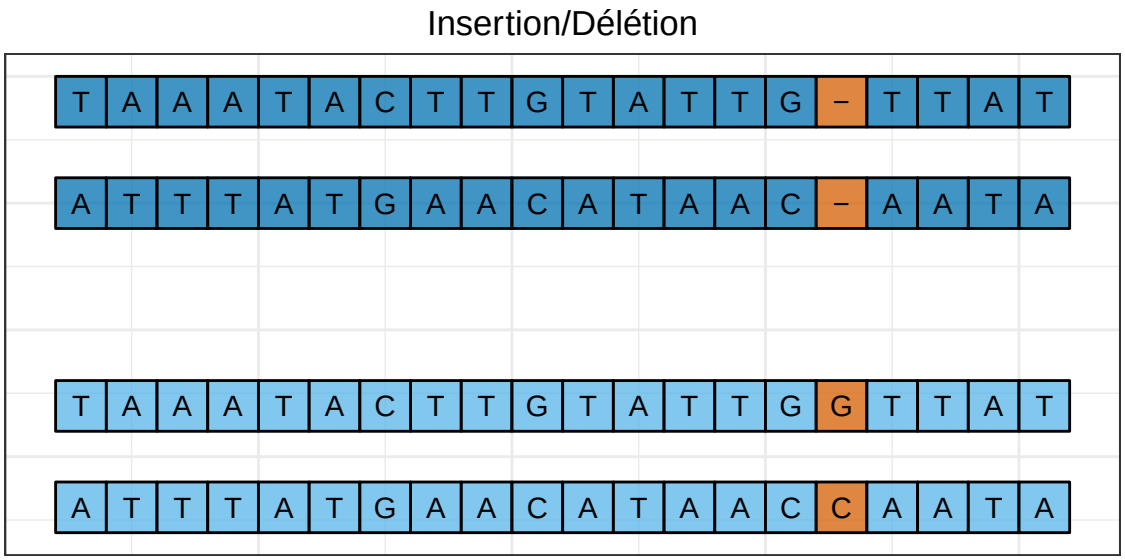


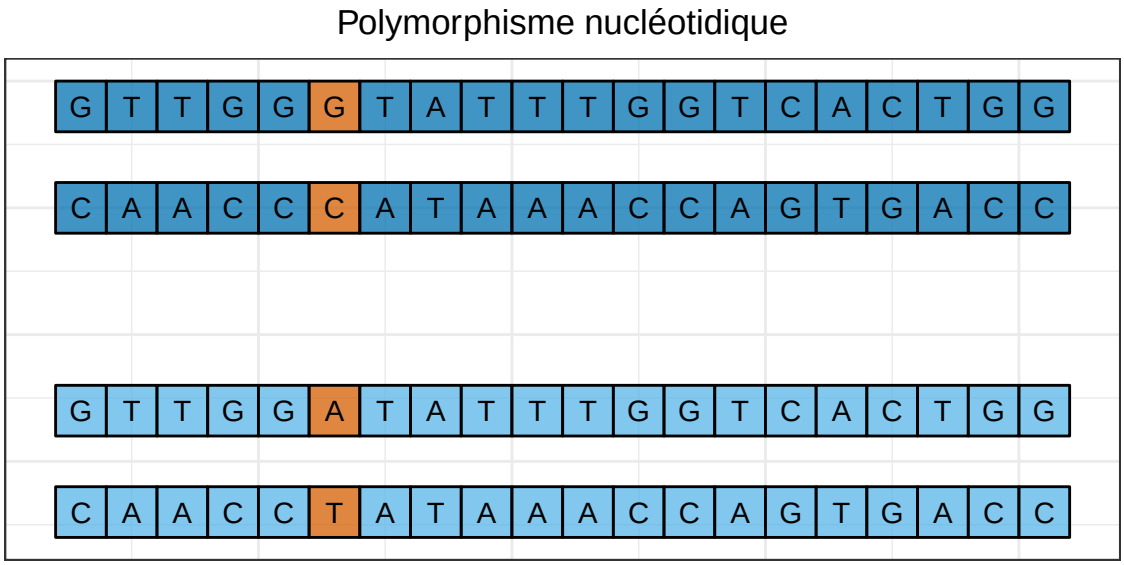
FIGURE 1.4 – Exemples de microsatellites. La première séquence comporte 3 répétitions du motif CCG, tandis que la seconde inclut 4 répétitions du motif CA.

— Indel :

INDEL (INsertion/DELetion) is where a single base has been deleted, or inserted into one genome relative to another. It is a symmetrical relationship, as a deletion in one corresponds to an insertion in another.



— Polymorphisme nucléotidique :



Chapitre 2

État de l'art

2.1 L'indice de fixation

En se différenciant génétiquement, les populations voient leurs fréquences d'allèles évoluer de façon indépendante. L'indice de fixation est une statistique permettant de quantifier pour un allèle donné, l'écart de la fréquence observée dans une sous-population à la fréquence théorique

2.2 Modèle de FLK

Le modèle de FLK estime le modèle neutre d'un SNP bi-allélique lorsque celui-ci est uniquement soumis à la dérive génétique. À l'instant $t = 0$, le SNP a une fréquence p_0. Notant F_t l'indice de fixation de cet allèle, $p(t)$ sa fréquence après t générations, et en supposant que F_t soit suffisamment petit, ce qui devrait être vérifié dans le cas neutre. (Nicholson et al., 2002) :

$$p(t) \sim N(p_0, F_t p_0(1 - p_0)) \tag{2.1}$$

De la loi (2.1) nous tirons $\mathrm{Var}(p(t)) = F_t p_0(1 - p_0)$.

La statistique FLK (Bonhomme et al., 2010) requiert d'estimer au préalable deux paramètres que sont la fréquence allélique initiale p_0 et la matrice d'apparentement V, $V \in M_K(\mathbb{R})$ où K est le nombre de populations observées. Notant :

— $\boldsymbol{p} = (p_1, p_2, \ldots, p_K) \in \mathbb{R}^K$,
— $\boldsymbol{p_0} = (p_0, p_0, \ldots, p_0) \in \mathbb{R}^K$,
— $\boldsymbol{r} = N(0, V)$,

le modèle neutre pour $\boldsymbol{p}$ est donné par la relation suivante :

$$\boldsymbol{p} = \boldsymbol{p_0} + \boldsymbol{r} \tag{2.2}$$

Bonhomme *et al.* proposent pour ce modèle de mesurer une statistique de qualité de l'ajustement pour quantifier la déviance d'un allèle par rapport au modèle neutre :

$$FLK = (\boldsymbol{p} - \hat{\boldsymbol{p}}_{\boldsymbol{0}})^T V (\boldsymbol{p} - \hat{\boldsymbol{p}}_{\boldsymbol{0}}) \tag{2.3}$$

Sous l'hypothèse neutre et suivant (2.3), $FLK \sim \chi^2(K - 1)$.

2.3 Modèle de OutFLANK

En reprenant le modèle proposé par Lewontin et Krakauer (Lewontin & Krakauer, 1973) et en y apportant les corrections nécessaires afin de prendre en compte les erreurs d'échantillonnage, Whitlock *et al.* proposent une méthode permettant de détecter les allèles sous sélection en environnement hétérogène (Whitlock & Lotterhos, 2015). Ainsi, la quantité

$$k\frac{F'_{ST}}{\bar{F}'_{ST}} \tag{2.4}$$

où k représente le nombre de degrés de libertés.

2.4 Modèle du logiciel Bayescan

Bayescan est aujourd'hui encore un des logiciels les plus utilisés pour détecter l'adaptation locale. Le modèle employé suppose que les sous-populations observées proviennent toutes d'une même population ancestrale. Pour une sous-population donnée et un SNP donné, la statistique de F_{ST} peut être estimée en utilisant la vraisemblance d'un modèle multinomial-Dirichlet (Beaumont & Balding, 2004). $F_{ST} \in [0, 1]$ est une quantité qui peut être interprété comme proportionnel à la probabilité que deux individus aient un ancêtre commun dans la sous-population

$$\log\left(\frac{F_{ST}}{1 - F_{ST}}\right) = \alpha_j + \beta_i + \gamma_{ij} \tag{2.5}$$

2.5 Fast PCA

2.5.1 L'ACP en génétique des populations

L'utilisation de l'Analyse en Composantes Principales en génétique des populations a été popularisée par Cavalli-Sforza (Menozzi, Piazza, & Cavalli-Sforza, 1978).

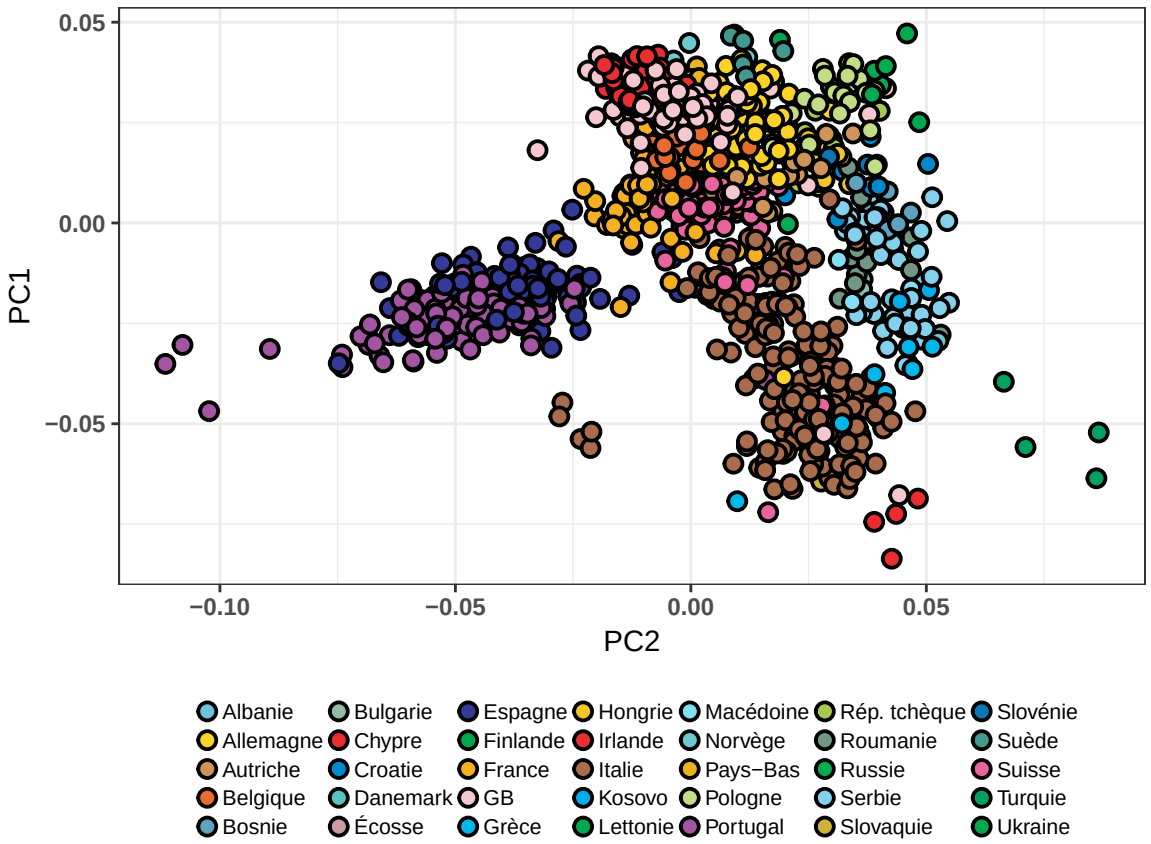


FIGURE 2.1 – ACP réalisée sur le jeu de données POPRES (Novembre et al., 2008).

2.6 Analyse en Composantes Principales parcimonieuse

2.7 Bootstrap ACP

2.8 Contexte

2.9 Tests multiples

2.10 Contrôle du taux de fausse découverte

Le taux de fausse découverte, correspond à la proportion de faux positifs parmi les positifs. En notant FP le nombre de faux positifs, FP le nombre de vrais positifs, on définit le taux de fausse découverte FDR par :

$$FDR = \mathbb{E}\left[\frac{FP}{TP + FP} 1_{FP+TP>0}\right]$$

- Référence cours de Christophe Giraud

q-value, bonferroni, benjamini-hochberg La figure suivante donne les comparaisons entre les différentes procédures de correction :

Chapitre 3

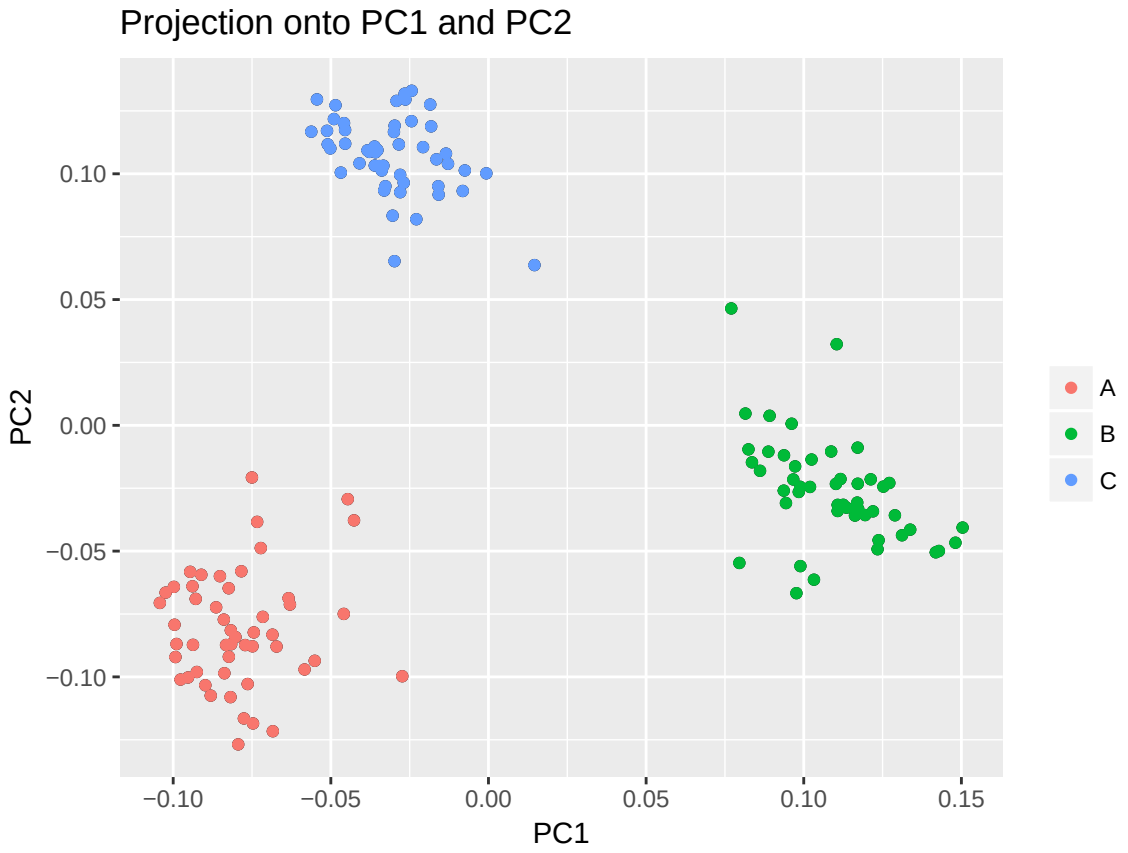
Adaptation locale

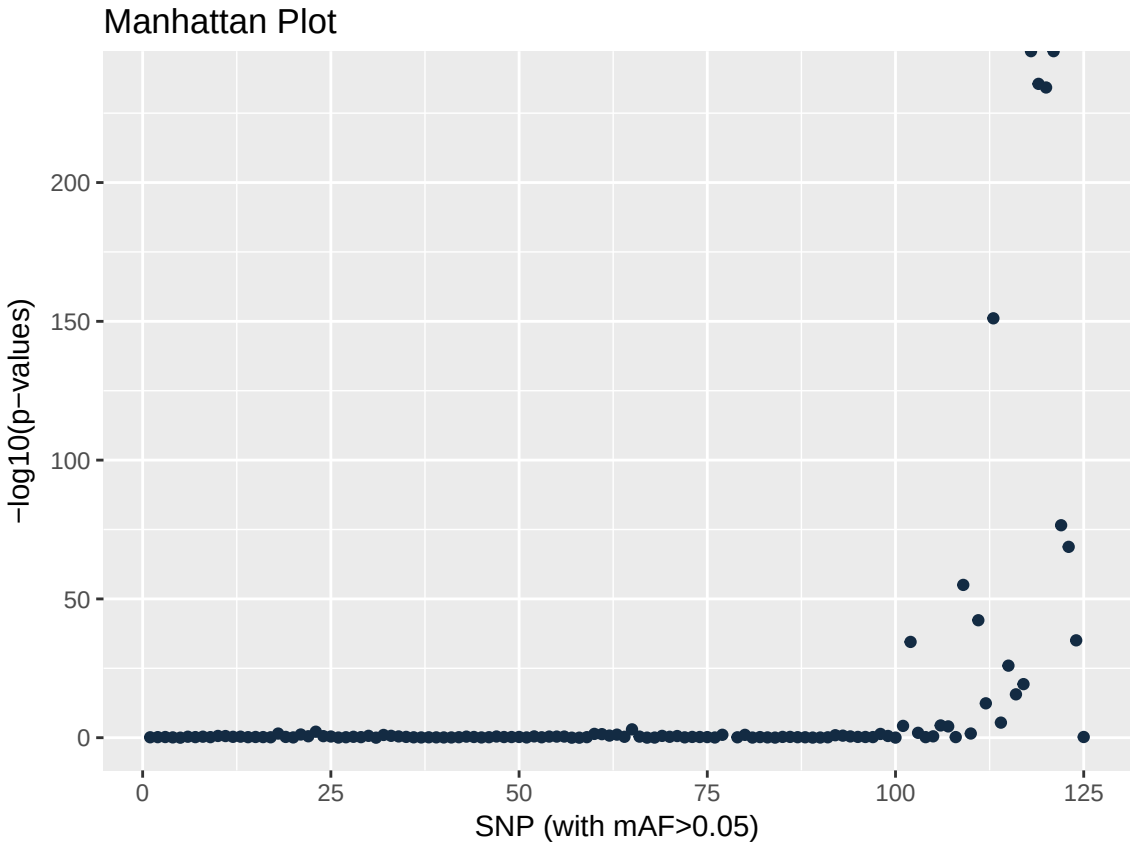
Une population est dite localement adaptée à son environnement si elle a connu une évolution différente de celles qu'ont connu les autres populations de la même espèce, et ce, en réponse aux pressions sélectives auxquelles elle peut être confrontrée.

3.1 Cas d'étude utilisant pcadapt

3.2 Simulations et modèles démographiques

3.2.1 Modèle en île





3.2.2 Modèle de divergence

Nous adaptons une version du script Python utilisé dans (Roux et al., 2012), basé sur le module de simulation simuPOP (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 immediatly
            # 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')
```

3.2.3 Modèle d'expansion spatiale

3.3 La communalité

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

Nicolas Duforet-Frebourg,[1,2,3] Keurcien Luu,[1,2] Guillaume Laval,[4,5] Eric Bazin,[6] and Michael G.B. Blum*,[1,2]

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

[2]CNRS, TIMC-IMAG, Grenoble, France

[3]Department of Integrative Biology, University of California, Berkeley

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

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

[6]CNRS, Laboratoire D'ecologie 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 (3×). 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 (APPBPP2, 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.

Key words: **FST, principal component analysis, population structure, population genomics, landscape genetics, selection scan, local adaptation, 1000 genomes.**

Article

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 (Sabeti et al. 2007; Barreiro et al. 2008; Pickrell et al. 2009; Grossman et al. 2013). Local positive selection tends to increase genetic differentiation, which can be measured by difference of allele frequencies between populations (Nielsen 2005; Sabeti et al. 2006; Colonna et al. 2014). 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 et al. 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 (Luikart et al. 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 et al. 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 et al. 2013; Günther and Coop 2013). Grouping individuals into populations can be subjective, and important signals of selection may be missed with an inadequate choice of populations (Yang et al. 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 (Patterson et al. 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 (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 (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 and Gaggiotti 2008; Riebler et al. 2008; Günther and 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 $\mathbf{Y}$ 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 $\mathbf{Y}$ by a matrix of smaller rank

$$\mathbf{Y} \approx U \Sigma \mathbf{V}^T,$$

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: the K columns of $\mathbf{V}$ contain the coefficients of the new orthogonal variables, the K columns of $\mathbf{U}$ contain the projections (called "scores") of the original variables onto the principal components and capture population structure (supplementary fig. S1, Supplementary Material online), and the squares of the elements of Σ are proportional to the proportion of variance explained by each principal component (Jolliffe 2005). We denote the diagonal elements of Σ by $\sqrt{\lambda_k}$, $k = 1, \ldots, K$ where the λ_k's are the ranked eigenvalues of the matrix $\mathbf{YY}^T$. Denoting by $\mathbf{V}_{jk}$, the entry of $\mathbf{V}$ at the j^{th} line and k^{th} column, then the correlation ρ_{jk} between the jth SNP and the kth principal component is given by $\rho_{jk} = \sqrt{\lambda_k} V_{jk} / \sqrt{n-1}$ (Cadima and 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

$$h_j^2 = \sum_{k=1}^{K} \rho_{jk}^2.$$

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^2, the statistic h'^2

should theoretically give the same importance to each PC because the normalized loadings are on the same scale as we have $\sum_{j=1}^{p} V_{jk}^2 = 1$, for $k = 1 \ldots K$.

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 $\Omega = \mathbf{Y}\mathbf{Y}^T/(p-1)$. 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 $\mathbf{U}$. Eigenanalysis is performed with the *dsyevr* routine of the linear algebra package LAPACK (Anderson et al. 1999). The matrix $\mathbf{V}$, which captures the relationship between each SNPs and population structure, is obtained by the matrix operation $\mathbf{V}^T = \Sigma^{-1}\mathbf{U}^T\mathbf{Y}$. 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. S2, Supplementary Material online).

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. 1 and 2). The first simulation scheme assumes that local adaptation took place in the lineages corresponding to the environments of populations A and B_1 (fig. 1). 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. S1, Supplementary Material online). 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. S3, Supplementary Material online).

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 1 and 2 (supplementary fig. S4, Supplementary Material online). 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. 1). 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. S5, Supplementary Material online). 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). 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 1, we find the same pattern of population structure (supplementary fig. S1, Supplementary Material online). The squared loadings ρ^2_{j1} 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^2_{j2} 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^2_1 (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. S1, Supplementary Material online), 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).

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. S6, Supplementary Material online). The main difference being that the FDR increases considerably when considering the top 5% instead of the top 1% (supplementary fig. S6, Supplementary Material online). 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. 1 and 2). As expected, the FDR of the different statistics based on PCA is considerably reduced when the selection coefficient increases (supplementary fig. S7, Supplementary Material online).

In the divergence model of figure 1, 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. S8, Supplementary Material online). 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. S8, Supplementary Material online).

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. 1), 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. S9, Supplementary Material online). 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. S10, Supplementary Material online) but by a more clinal or continuous pattern (supplementary fig. S10, Supplementary Material online) 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. 3 and supplementary fig. S11, Supplementary Material online). The first and second PCs ascertain population structure between Africa, Asia, and Europe (fig. 3) and the third principal component separates the Yoruba from the Luhya population (supplementary fig. S11, Supplementary Material online). The decay of eigenvalues suggests to use $K = 2$ because the

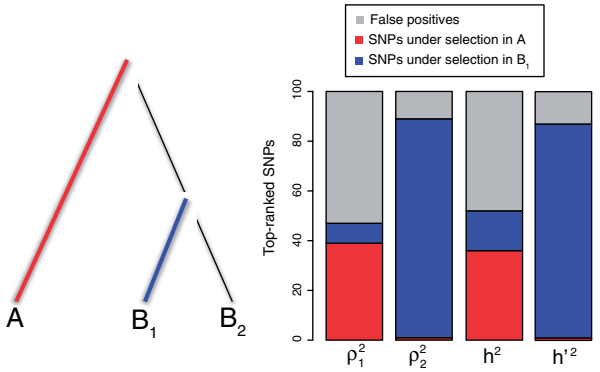


Fig. 1. Repartition of the 1% top-ranked SNPs for each PCA-based statistic under a divergence model with two types of adaptive constraints. Thicker and colored lineages correspond to lineages where adaptation took place. The squared loadings with PC1 ρ^2_{j1} pick a large proportion of SNPs involved in selection in population A whereas the squared loadings with PC2 ρ^2_{j2} 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.

eigenvalues drop between $K = 2$ and $K = 3$ where a plateau of eigenvalues is reached (supplementary fig. S3, Supplementary Material online).

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^2_{j1} corresponding to PC1 and 20% are in the top 0.1% of the squared loadings ρ^2_{j2} 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^2_{j1} corresponding to PC1 and 85% are in the top 0.1% of the squared loadings ρ^2_{j2} corresponding to PC2. The h'^2 statistic is mostly influenced by the second PC because the distribution of the V^2_{2j} (normalized squared loadings) has a longer tail than the corresponding distribution for PC1 (supplementary fig. S5, Supplementary Material online). 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^2_{j1} and ρ^2_{j2}.

The two Manhattan plots based on the squared loadings for PC1 and PC2 are displayed in figures 4 and 5 (supplementary table S1, Supplementary Material online, contains the loadings for all variants). 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. S12, Supplementary Material online). Additionally, the clustering is larger for the outliers related to the second PC as they cluster in fewer windows (supplementary fig. S12, Supplementary Material online). As the genome scan for PC2 captures more recent adaptive events, it reveals larger genomic windows that experienced fewer recombination events.

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 (supplementary table S2, Supplementary Material online). 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. 4). Confirming a larger clustering for PC2 outliers, the 100 SNPs that are the most correlated with PC2 cluster in fewer genomic regions (supplementary table S3, Supplementary Material online). 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. 5).

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) (supplementary table S4, Supplementary Material online). For PC2, we find enrichment (FDR $\leq 5\%$) related to olfactory receptors, keratinocyte and epidermal cell differentiation, and ethanol metabolism (supplementary table S5, Supplementary Material online). 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 (supplementary table S6, Supplementary Material online). 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 $\leq 5\%$, supplementary table S7, Supplementary Material online). 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/

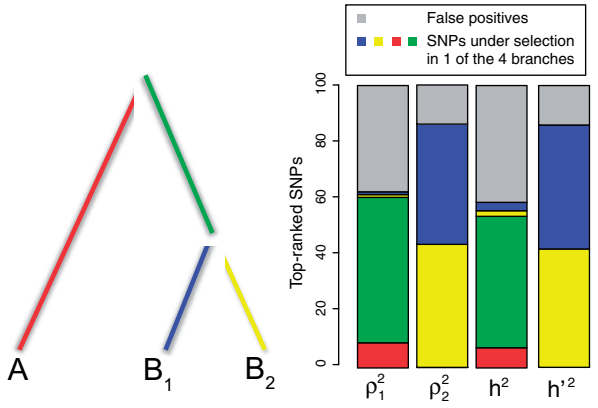


Fig. 2. Repartition of the 1% top-ranked SNPs of each PCA-based statistic under a divergence model with four types of adaptive constraints. Thicker and colored lineages correspond to lineages where adaptation occurred. The different types of SNPs picked by the squared loadings ρ^2_{j1} and ρ^2_{j2} is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

gluconeogenesis and fatty acid omega oxidation for PC2 (supplementary table S8, Supplementary Material online).

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 (Kudaravalli et al. 2009; Fagny et al. 2014). For PC1, we do not find significant enrichments (table 1) except when measuring the enrichment of genic regions compared with nongenic regions (OR = 10.18 for the 100 most correlated SNPs, $P < 5\%$ using a permutation procedure). For PC2, we find an enrichment of genic regions among outliers as well as an enrichment of nonsynonymous SNPs (table 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 1).

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 et al. 2013). Population-differentiation based methods such as genome scans based on F_{ST} look for marked differences in allele frequencies between population (Holsinger and 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 SNPSs typed

Table 1. Enrichment Measured with Odds Ratio (OR) of the Variants Most Correlated with the Principal Components Obtained from the 1000 Genomes Data.

	Top 0.1%	Top 0.01%	Top 0.005%	Top 100 SNPs
pc1–genic/nogenic	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/nogenic	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

Note.—Enrichment significant at the 1% (respectively, 5%) level are indicated with ** (resp. *).

for 1,385 European individuals (Nelson et al. 2008). The scree plot indicates that there are $K = 2$ relevant clusters (supplementary fig. S3, Supplementary Material online). The first principal component corresponds to a Southeast–Northwest gradient and the second one discriminates individuals from Southern Europe along a East–West gradient (Novembre et al. 2008; Jay et al. 2013) (fig. 6). 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. 7). When considering the 100 SNPs most correlated with the second PC, we find less clustering than for PC1 with more peaks (supplementary fig. S13, Supplementary Material online). 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.

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 et al. 2013; Lotterhos and 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 (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; Günther and Coop 2013; Foll et al. 2014), we proposed an alternative statistic h'^2, which should give equal weights to the different PCs. However, analyzing simulations and the 1000 Genomes data show that h'^2 do not properly account for hierarchical population structure because outliers identified by h'^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. 2015). 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 and Bentler 2010)

$$h^2 \times K/c \rightsquigarrow \chi^2_K,$$

where $c = \sum_{i=1}^{K} \lambda_K/(p \times (n-1))$ is the proportion of variance explained by the first K PCs. The chi-square approximation of equation (3) bears similarity with the approximation of Lewontin and Krakauer (1973) that states that $F_{ST} \times (n_{pops} - 1)/\bar{F}_{ST}$ follows a chi square approximation with $(n_{pops} - 1)$ degrees of freedom where $\bar{F}_{ST}$ is the mean F_{ST} over loci and n_{pops} 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 (3) (supplementary fig. S14, Supplementary Material online). 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. S15, Supplementary Material online). 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 and 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 that 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, supplementary table S9, Supplementary Material online). 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 (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 Eurasiatic population 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

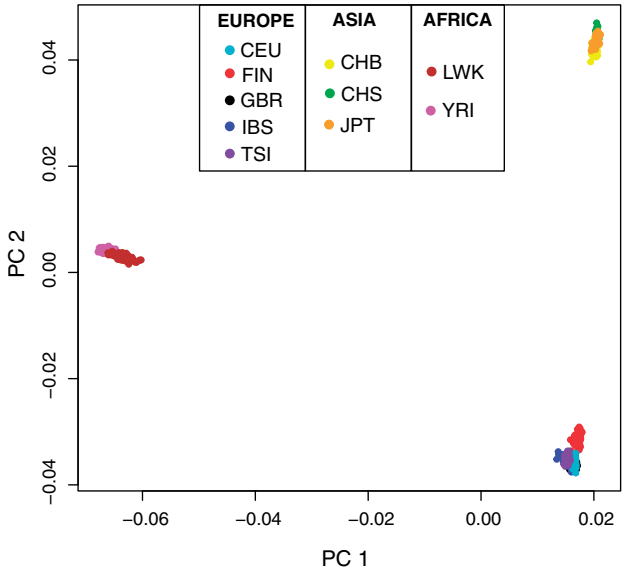


Fig. 3. PCA with $K = 2$ applied to the 1000 Genomes data. The sampled populations are the following: British in England and Scotland (GBR), Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), Iberian populations in Spain (IBS), Toscani in Italy (TSI), Han Chinese in Bejing (CHB), Southern Han Chinese (CHS), Japanese in Tokyo (JPT), Luhya in Kenya (LWK), Yoruba in Nigeria (YRI).

substrate for human adaptation (Li et al. 2012; Grossman et al. 2013). 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 has been detected (Carlson et al. 2005; Barreiro et al. 2008), and it is involved in the development of the human skeletal system (Wu and 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 (Chen et al. 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 and 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 and 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 (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 (Hancock et al. 2008; Barreiro and Quintana-Murci 2009; Fumagalli et al. 2011; Daub et al. 2013).

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 (3×), 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).

```
./ms 300 1400 -s 1 -I 3 100 100 100 -ma x 4 4 4 x
4 4 4 x #neutral.
```

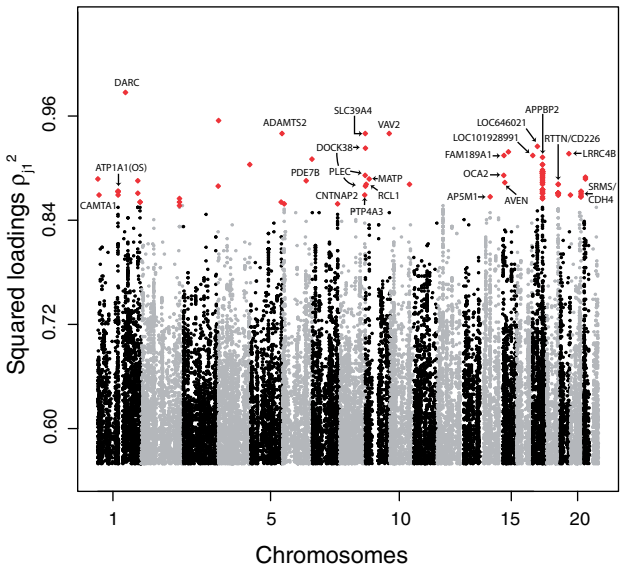


FIG. 4. Manhattan plot for the 1000 Genomes data of the squared loadings ρ_{j1}^2 with the first principal component. For sake of presentation, only the top-ranked SNPs (top 0.1%) are displayed and the 100 top-ranked SNPs are colored in red.

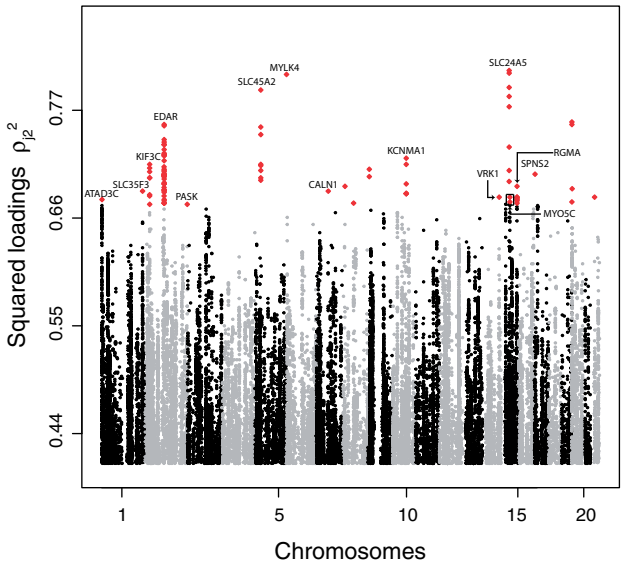


FIG. 5. Manhattan plot for the 1000 Genomes data of the squared loadings ρ_{j2}^2 with the second principal component. For sake of presentation, only the top-ranked SNPs (top 0.1%) are displayed and the 100 top-ranked SNPs are colored in red.

```
/ms 300 100 -s 1 -I 3 100 100 100 -ma x 0.1 0.1
0.1 x 4 0.1 4 x #outlier
```

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 $N_0 = 1,000$ 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* (Peng and 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 1, 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_1. We consider an additive model for selection with a selection coefficient of $s = 1.025$ for heterozygotes. For the simulation of figure 2, 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 and Cockerham (1984).

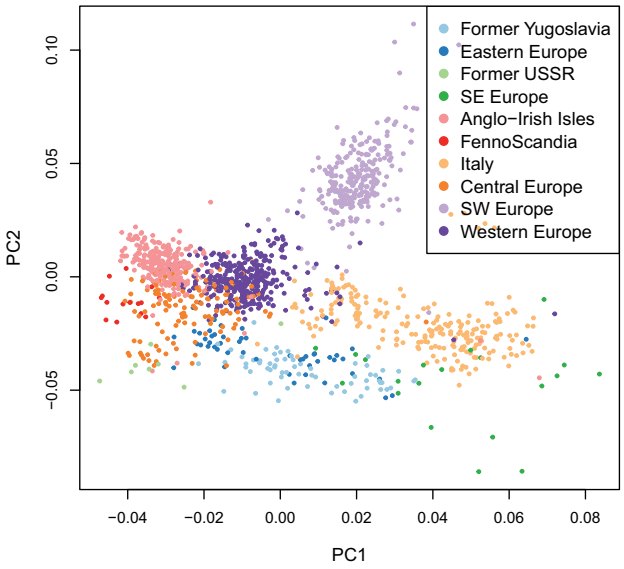


FIG. 6. PCA with $K = 2$ applied to the POPRES data.

1000 Genomes Data

We downloaded the 1000 Genomes data (phase 1 v3) (The 1000 Genomes Project Consortium 2012). We kept low-coverage genome data and excluded exomes and triome data to minimize variation in read depth. Filtering the data resulted in

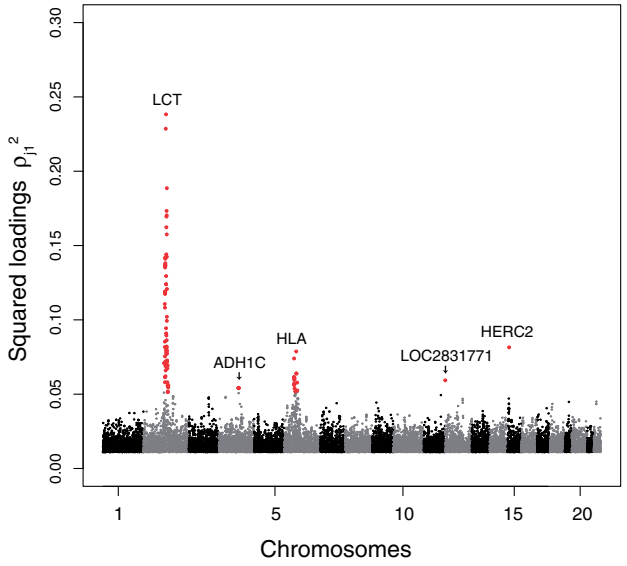


FIG. 7. Manhattan plot for the POPRES data of the squared loadings ρ_{j1}^2 with the first principal component. For sake of presentation, only the top-ranked SNPs (top 5%) are displayed and the 100 top-ranked SNPs are colored in red.

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 and 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. 2009)

$$\mathrm{OR} = \frac{\Pr(\text{genic} \mid \text{outlier})}{\Pr(\text{not genic} \mid \text{outlier})} \frac{\Pr(\text{not genic} \mid \text{not outlier})}{\Pr(\text{genic} \mid \text{not outlier})}.$$

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 50 kb 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 50 kb 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.

Supplementary Material

Supplementary figures S1–S15 and tables S1–S9 *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

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)

References

Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, et al. 2009. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet.* 41:585–590.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D. 1999. LAPACK users' guide. 3rd edn. Philadelphia (PA): Society for Industrial and Applied Mathematics.

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

Barreiro LB, Quintana-Murci L. 2009. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 11:17–30.

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

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 57:289–300.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of

strong recent positive selection at the lactase gene. *Am J Hum Genet*. 74:1111–1120.
Bierne N, Roze D, Welch JJ. 2013. Pervasive selection or is it . . . ? why are F_{ST} outliers sometimes so frequent? *Mol Ecol* 22:2061–2064.
Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M. 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186:241–262.
Cadima J, Jolliffe IT. 1995. Loading and correlations in the interpretation of principal components. *J Appl Stat*. 22:203–214.
Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M. 2008. The signature of long-standing balancing selection at the human defensin beta-1 promoter. *Genome Biol*. 9:R143.
Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res*. 15:1553–1565.
Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res*. 20:393–402.
Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-Smith C, the 1000 Genomes Project Consortium. 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol*. 15:R88.
Daub JT, 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. *Mol Biol Evol*. 30:1544–1558.
Duforet-Frebourg N, Bazin E, Blum MGB. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Mol Biol Evol*. 31:2483–2495.
Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Mol Biol Evol*. 31:1850–1868.
Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193:929–941.
Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977–993.
Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. 2014. Widespread signals of convergent adaptation to high altitude in Asia and America. *Am J Hum Genet*. 95:394–407.
Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, Sironi M. 2010. Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. *PLoS Genet*. 6:e1000849.
Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*. 7:e1002355.
Galinsky KJ, Bhatia G, Loh P, Georgiev R, Mukherjee SS, Patterson NJ, Price AL. 2015. Fast principal components analysis reveals convergent evolution of ADH1B gene in Europe and East Asia. *bioRxiv* 018143.
Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152:703–713.
Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195:205–220.
Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African genome variation project shapes medical genetics in Africa. *Nature* 517:327–332.
Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet*. 70:369–383.
Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. 2007. Evidence of positive selection on a class I *ADH* locus. *Am J Hum Genet*. 80:441–456.
Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, et al. 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A*. 107:8924–8930.
Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet*. 4:e32.
Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.
Hollox EJ, Armour JA. 2008. Directional and balancing selection in human beta-defensins. *BMC Evol Biol*. 8:113.
Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat Rev Genet*. 10:639–650.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Jay F, Sjödin P, Jakobsson M, Blum MGB. 2013. Anisotropic isolation by distance: the main orientations of human genetic differentiation. *Mol Biol Evol*. 30:513–525.
Jiao H, Arner P, Hoffstedt J, Brodin D, Dubern B, Czernichow S, van't Hooft F, Axelsson T, Pedersen O, Hansen T, et al. 2011. Genome wide association study identifies KCNMA1 contributing to human obesity. *BMC Med Genomics*. 4:51.
Jolliffe I. 2005. Principal component analysis. Springer: New-York.
Kofler R, Schlötterer C. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* 28:2084–2085.
Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2009. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol*. 26:649–658.
Lewontin R, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175–195.
Li J, Liu Y, Xin X, Kim TS, Cabeza EA, Ren J, Nielsen R, Wrana JL, Zhang Z. 2012. Evidence for positive selection on a number of MicroRNA regulatory interactions during recent human evolution. *PLoS Genet*. 8:e1002578.
Lotterhos KE, Whitlock MC. 2014. Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Mol Ecol*. 23:2178–2192.
Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet*. 4:981–994.
McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*. 5:e1000686.
Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, et al. 2008. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*. 83:347–358.
Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. *Nature* 456:98–101.
Oeggerli M, Tian Y, Ruiz C, Wijker B, Sauter G, Obermann E, Güth U, Zlobec I, Sausbier M, Kunzelmann K, et al. 2012. Role of KCNMA1 in breast cancer. *PLoS One* 7:e41664.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.
Peng B, Kimmel M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687.

Peng Y, Shi H, Qi Xb, Xiao Cj, Zhong H, Run-lin ZM, Su B. 2010. The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evol Biol*. 10:15.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*. 19:826–837.

Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178:1817–1829.

Sabeti P, Schaffner S, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen T, Altshuler D, Lander E. 2006. Positive natural selection in the human lineage. *Science* 312:1614–1620.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet*. 47:97–120.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 1358–1370.

Whitlock MC, Lotterhos KE. 2015. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *Am Nat*. 186:S24–S36.

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, et al. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci U S A*. 111:4832–4837.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet*. 3:e90.

Wu DD, Zhang YP. 2010. Positive selection drives population differentiation in the skeletal genes in modern humans. *Hum Mol Genet*. 19:2341–2346.

Yang WY, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*. 44:725–731.

Yuan KH, Bentler PM. 2010. Two simple approximations to the distributions of quadratic forms. *Br J Math Stat Psychol*. 63:273–291.

3.4 La distance robuste de Mahalanobis

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 handle 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.

Keywords: R package, Mahalanobis distance, outlier detection, population genetics, principal component analysis

Received 31 May 2016; revision received 29 July 2016; accepted 1 August 2016

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 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 (Excoffier *et al.* 2009; Bierne *et al.* 2013).

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 *et al.* 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 (Duforet-Frebourg *et al.* 2014; Chen *et al.* 2016; Duforet-Frebourg *et al.* 2016; Galinsky *et al.* 2016; Hao *et al.* 2016; Martins *et al.* 2016).

Correspondence: Michael G. B. Blum, Fax: +33 (0) 4 56 52 00 55; E-mail: michael.blum@imag.fr

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 *et al.* 2014). Additionally, next-generation sequencing data may contain a substantial proportion of missing data that should be accounted for (Arnold *et al.* 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 1).

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.* 2016). 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 (R Core 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 2012). 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 *hapflk* 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 *et al.* 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.

Table 1 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)
Communality	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

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 (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 $\tilde{G}$ to ascertain population structure with PCA (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 $X_1, , X_K$

$$G_j = \sum_{k=1}^{K} \beta_{jk} X_k + \epsilon_j,\ j = 1, ..., p, \tag{1}$$

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 $z_j = (z_{j1}, ..., z_{jK})$ 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

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

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 2012). 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 2012) (*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 (Duforet-Frebourg *et al.* 2016). 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 $\hat{\beta}$ 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

$$z_{jk} = \hat{\beta}_{jk} \sqrt{\frac{\sum_{i=1}^{n} x_{ik}^2}{\sigma_j^2}}, \quad (3)$$

where σ_j^2 is an estimate of the residual variance for the j^{th} SNP, and x_{ik} is the score of 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 (3). 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). 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).

Island model

We used *ms* to create simulations under an island model (Fig. S1). 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 (Peng &

Table 2 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

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 S1). 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 individual 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 *FLK* 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. 1). 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 (Patterson *et al.* 2006). For the model of range expansion, applying Cattell's rule to choose K is more difficult (Fig. 1). 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 S2).

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. *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). 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%.

Control of the false discovery rate

We evaluate to what extent using the packages *pcadapt* and *qvalue* control a FDR set at 10% (Fig. 3). 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%.

We then evaluate the proportion of false discoveries obtained with *BayeScan*, *hapflk*, *OutFLANK* and *sNMF* (Fig. 3). 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%.

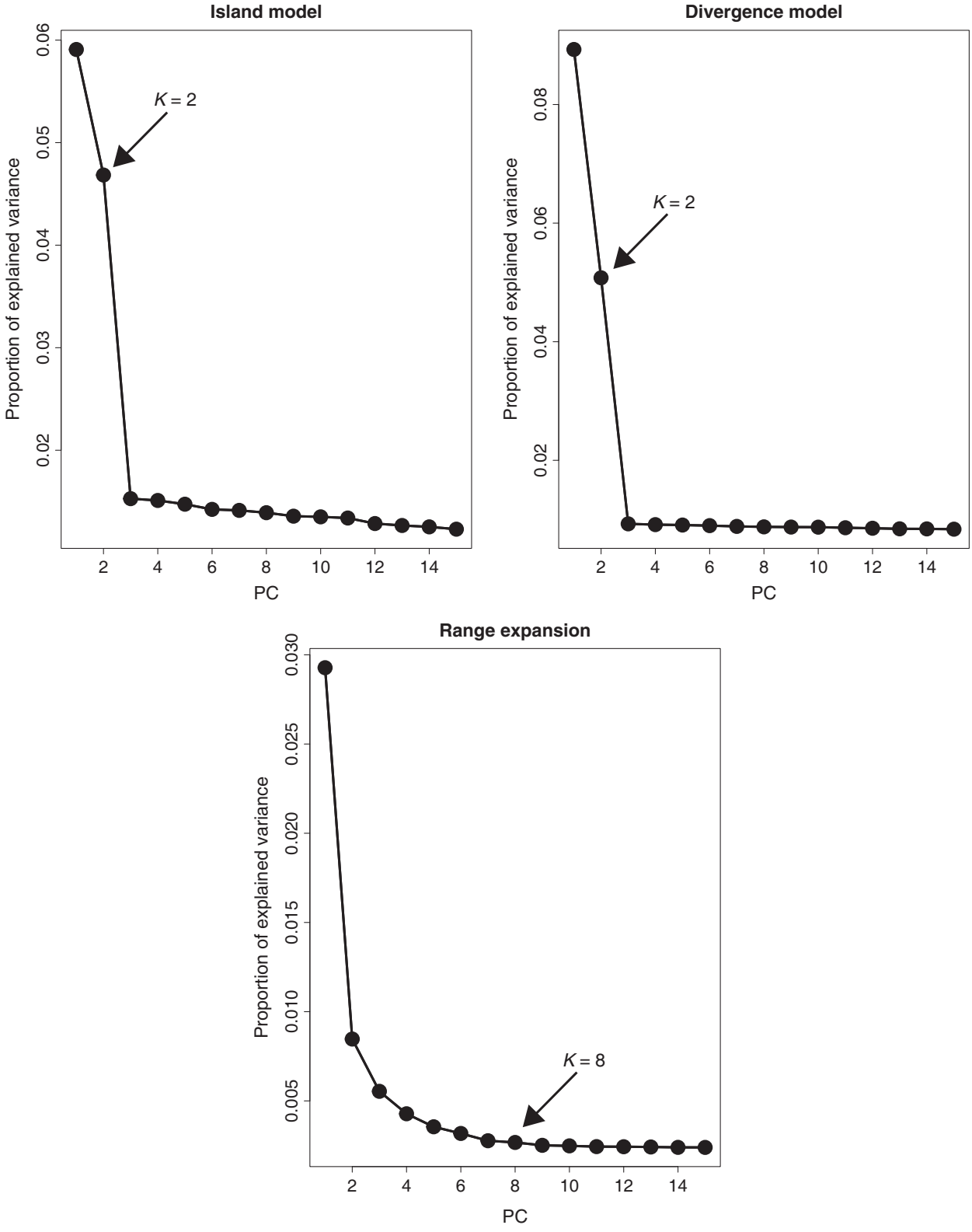


Fig. 1 Determining *K* with the scree plot. To choose *K*, we recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept. According to Cattell's rule, the eigenvalues that correspond to random variation lie on the straight 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.

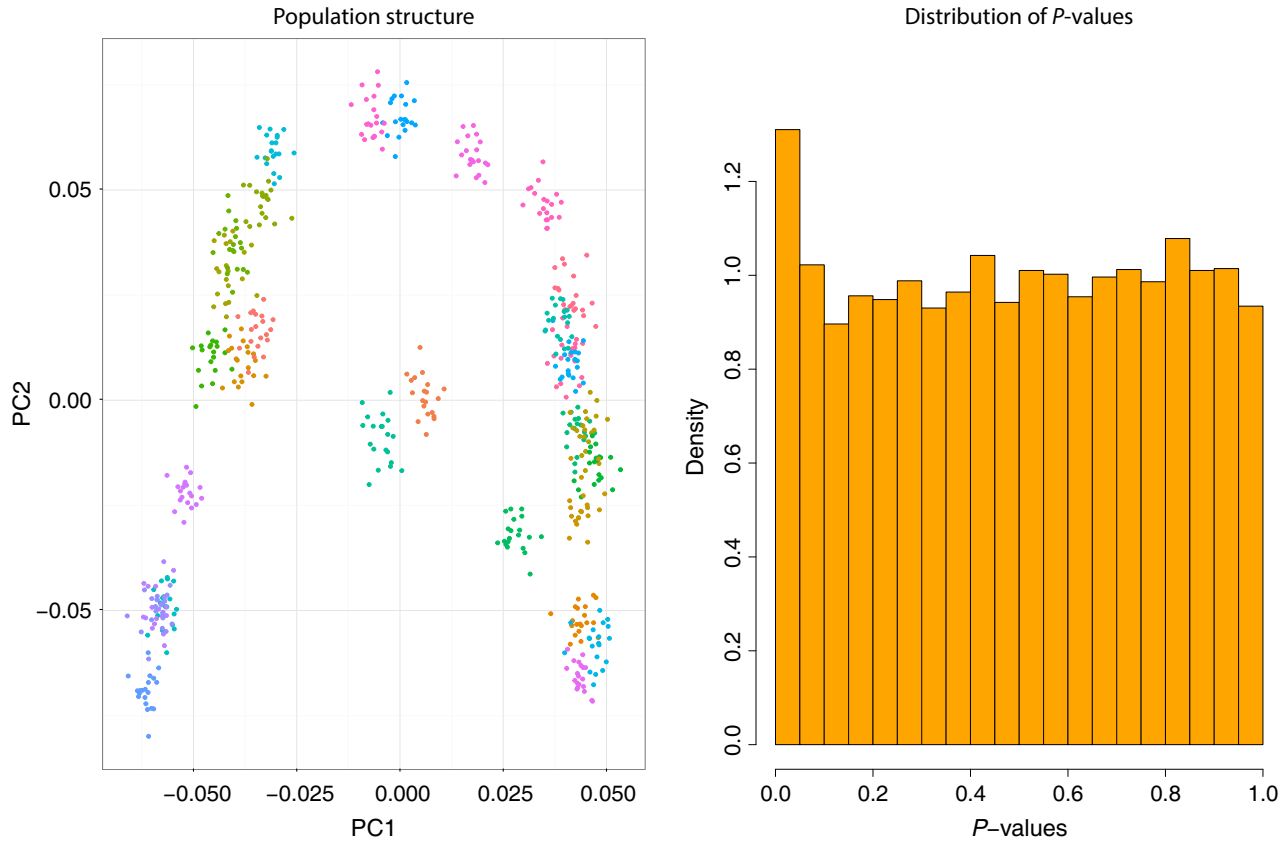


Fig. 2 Population structure (first 2 principal components) and distribution of *p*-value obtained with *pcadapt* for a simulation of range expansion. *P*-values are well calibrated because they are distributed as a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds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

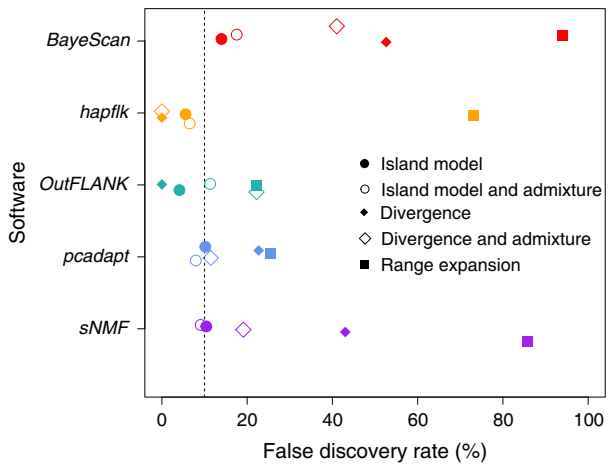


Fig. 3 Control of the FDR for different computer programs for genome scans. We find that the median proportion of false discoveries is around the nominal FDR set at 10% (6% for *hapflk*, 11% for both *OutFLANK* and *pcadapt* and 19% for *sNMF*) with the exception of *BayeScan* that generates 41% of false discoveries. Comparison of statistical power for the different test statistics that have been implemented in *pcadapt* (Table 1).

We first compare statistical power obtained with the different statistical methods that have been implemented in *pcadapt* (Table 1). For the island model, Bayes factor, communality statistic and Mahalanobis distance have similar power (Fig. 4). 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. 4). 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%.

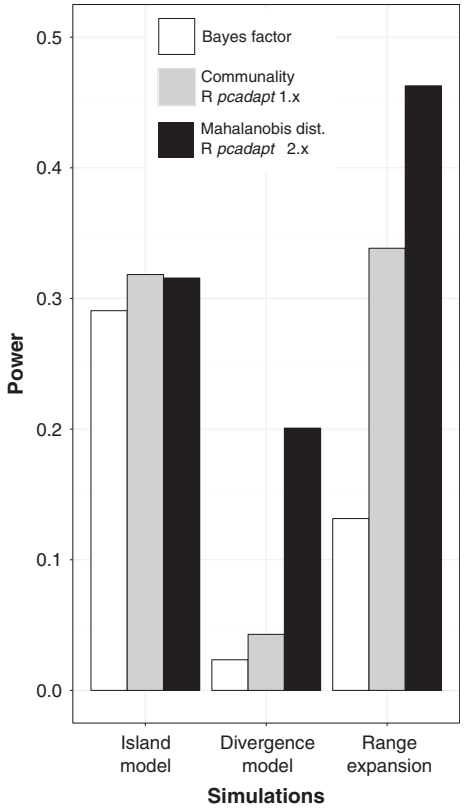


Fig. 4 Bayes factor corresponds to the test statistic implemented in the Bayesian version of *pcadapt* (Duforet-Frebourg *et al*. 2014); the communality statistic was the default statistic in version 1.x of the R package *pcadapt* (Duforet-Frebourg *et al*. 2016), and Mahalanobis distances are available 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%).

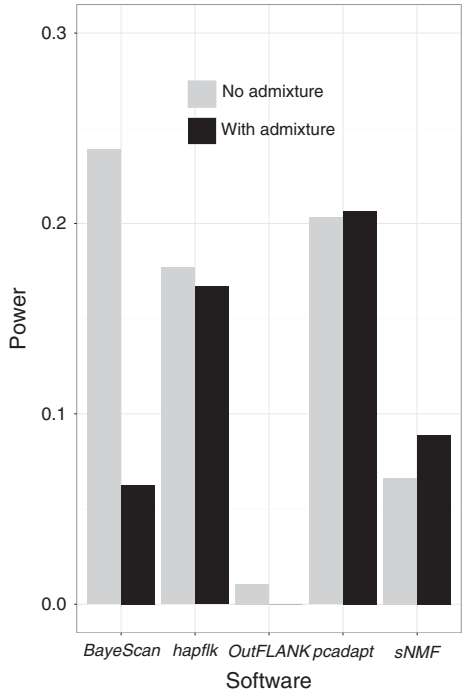


Fig. 5 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the divergence model with three populations. We assume that adaptation took place in an external branch that follows the most recent population divergence event.

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 S3 and S4). Including admixed individuals hardly changes their statistical power (Figure S3).

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. 5 and Figure S5). 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. 5).

The last model we consider is the model of range expansion. The package *pcadapt* is the most powerful approach in this setting (Fig. 6 and S6). 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. 6 and S6). 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*.

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 9 h 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

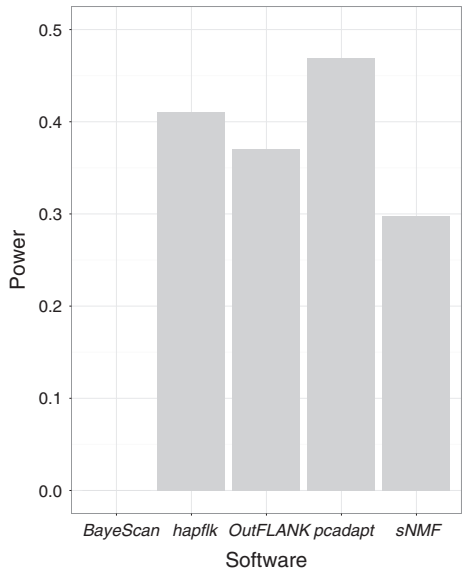


Fig. 6 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for a range expansion model with two refugia. Adaptation took place during the recolonization event.

which the runtime increases the most rapidly with the number of markers. *OutFLANK* takes around 25 min to analyse 50 000 SNPs (Figure S7). For the other 3 computer programs (*hapflk*, *pcadapt*, *sNMF*), analysing 50 000 SNPs takes <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.* 2016). 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. 5) (Bonhomme *et al.* 2010; Duforet-Frebourg *et al.* 2016). *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. 5). 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* provides 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 computing 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. $n \geq 10\,000$), there are faster algorithms to compute principal components (Halko *et al.* 2011; Abraham & Inouye 2014). 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 Stephane Dray for their critical reading of our manuscript.

References

Abraham G, Inouye M (2014) Fast principal component analysis of large-scale genome-wide data. *PloS One*, **9**, e93766.

Arnold B, Corbett-Detig R, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

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

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

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

Cattell RB (1966) The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245–276.

Chen G-B, Lee SH, Zhu Z-X, Benyamin B, Robinson MR (2016) EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, **117**, 51–61.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Dray S, Josse J (2015) Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, **216**, 657–667.

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, **31**, 2483–2495.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular Biology and Evolution*, **33**, 1082–1093.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

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

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

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

Galinsky KJ, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson NJ, Price AL (2016) Fast principal components analysis reveals independent evolution of adh1b gene in Europe and East Asia. *American Journal of Human Genetics*, **98**, 456–472. 018143

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Halko N, Martinsson P-G, Tropp JA (2011) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**, 217–288.

Hao W, Song M, Storey JD (2016) Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, **32**, 713–721.

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, **74**, 2204–2214.

Lange K, Papp JC, Sinsheimer JS, Sobel EM (2014) Next generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application*, **1**, 279.

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

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Maronna RA, Zamar RH (2012) Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**, 307–317.

Martins H, Caye K, Luu K, Blum MG, François O (2016) Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *bioRxiv*, 054585.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet*, **2**, e190.

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

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Waples RS, 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**, 1419–1439.

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. *The American Naturalist*, **186**, S24–S36.

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
```

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Schematic description of the island and divergence model.

Fig. S2 Proportion of false discoveries and statistical power as a function of the number of principal components in a model of range expansion.

Fig. S3 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the island model.

Fig. S4 Statistical power as a function of the proportion of false discoveries for the island model.

Fig. S5 Statistical power as a function of the proportion of false discoveries for the divergence model.

Fig. S6 Statistical power as a function of the proportion of false discoveries for the model of range expansion.

Fig. S7 Running times of the different computer programs.

Chapitre 4

Introgression adaptative

4.1 Qu'est-ce que l'introgression ?

Avant de s'intéresser à la notion d'introgression, intéressons-nous d'abord à celle d'hybridation. L'hybridation peut être définie comme la reproduction entre deux individus appartenant à deux espèces ou à deux populations différentes. Cette définition nous amène à nous poser deux questions. La première, relative à la notion d'espèce, est souvent sujette à controverse. La seconde concerne quant à elle la désignation de populations différentes. Qu'est-ce qui fait que deux groupes d'individus sont différents ? Harrison suggère en 1990 que deux individus issus de populations différentes doivent chacun posséder des traits héritables qui les différencient (Harrison & others, 1990).

Nous parlons d'introgression lorsqu'un certain nombre de gènes est transféré d'une population à une autre.

L'étude de régions génomiques présentant des caractéristiques d'introgression ou de divergence peut se révéler intéressante pour plusieurs raisons.

4.2 Coefficients de métissage globaux et locaux

Étant données des populations ancestrales, il est possible d'estimer pour un individu donné, la proportion de son génôme provenant de chacune des populations ancestrales. Ces proportions sont connues plus communément sous le nom de *coefficients de métissage globaux*. 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). En complément à cette information globale, il peut être intéressant de déterminer sur des portions plus petites du génôme, de la même manière que dans le cas global, les proportions venant de telle ou telle population ancestrale pour chacune de ces portions. Nous parlons dans ce cas de *coefficients de métissage locaux*. Encore une fois, plusieurs logiciels ont été proposés dans le but d'estimer ces coefficients : Hapmix (Price et al., 2009), EILA (Yang, Li, Buu, & Williams, 2013), LAMP (Thornton & Bermejo, 2014), loter ou encore RFmix (Maples, Gravel, Kenny, & Bustamante, 2013).

4.3 Introgression

L'introgression peut être détectée de différentes façons. Une première approche consiste à utiliser les *coefficients de métissage locaux*. Les méthodes mentionnées plus haut estiment ces coefficients pour chaque individu, permettant de calculer à partir de ceux-ci des coefficients de métissage locaux pour chaque population.

4.4 Lien entre Analyse en Composantes Principales et métissage global.

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 de Gil McVean (McVean, 2009) :

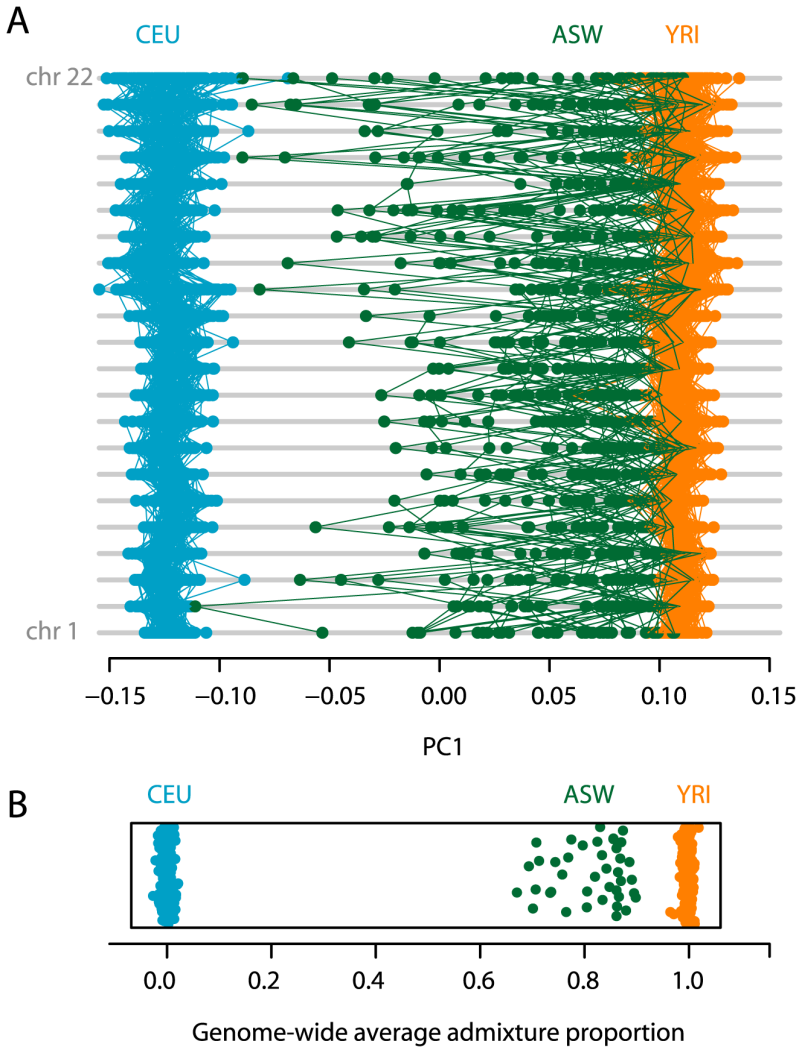


FIGURE 4.1 – Coefficients de métissage et ACP (McVean, 2009).

Pour chacun des 22 chromosomes,

4.5 Analyse en Composantes Principales locale

Notant p le nombre de marqueurs génétiques, i un entier compris entre 1 et p, et x_i la position génétique (en Morgans) ou la position physique (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 :

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

4.6 Sensibilité à l'imputation des données manquantes

4.6.1 Méthodes de détection

Etat de l'art

Scénario à flux de gènes

La statistique D de Patterson La statistique D de Patterson (Durand, Patterson, Reich, & Slatkin, 2011) demeure aujourd'hui la méthode la plus utilisée pour détecter la trace de flux de gènes dans une population. La méthode repose sur l'observation de motifs portant les noms *ABBA* et *BABA*, en référence aux différents types de généalogie possible pour un site nucléotidique.

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

RNDmin [Descriptif de RNDmin]

Pour comprendre la récente méthode proposée par (Rosenzweig, Pease, Besansky, & Hahn, 2016), il est nécessaire de définir un certain nombre de statistiques dont $RND_{\min}$ dérive.

— d_{xy} est la distance de Hamming entre la séquence X de la population 1 et la séquence Y de la population 2. Ainsi, si $x = (x_i)_{1 \leq i \leq n}$ et $y = (y_i)_{1 \leq i \leq n}$, alors :

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

Cette statistique est ainsi définie pour une paire de séquences (x, y). Pour quantifier la dissimilarité entre deux ensembles de séquences, deux approches sont possibles. La première consiste à calculer la distance moyenne d_{XY}. De par sa définition, cette distance présente néanmoins le défaut d'être peu sensible aux épisodes récents d'introgression (Geneva, Muirhead, Kingan, & Garrigan, 2015). En effet, les faibles valeurs de d_{xy} correspondant à des évènements de divergence récents peuvent voir leur influence diminuée en cas de présence d'évènements de divergence plus anciens. Pour pallier à ce problème, considérer la distance minimale entre les deux ensembles de séquences (Joly, McLenachan, & Lockhart, 2009) constitue une solution intéressante.

[Expliquer pourquoi on définit d_{out} et RND]

En définissant $d_{out} = \frac{1}{2}(d_{XO} + d_{YO})$, il est possible de définir de la même façon RND_{min} :

$$RND_{\min} = \frac{d_{\min}}{d_{\text{out}}}$$

Pour récapituler, $RND_{\min}$ est une statistique robuste aux variations de taux de mutation et qui reste sensible aux récents évènements d'introgression. En pratique, l'introduction de d_{out} requiert ainsi la donnée d'une population ancestrale commune aux deux populations d'intérêt.

Bdf

Analyse Linéaire Discriminante

Régression linéaire, régression logistique, forêts aléatoires et importance des variables

Régression locale, package mgcv, locfit, Backward selection strategy

ACP locale et espace de formes

4.7 Simulations

4.7.1 Données de peupliers

Le premier jeu de données est issu d'une étude d'introgression adaptative chez les peupliers d'Amérique du Nord (Suarez-Gonzalez, 2016). La simulation d'haplotypes d'individus admixés est effectuée à partir des deux populations ancestrales qui y sont présentes. La première, *Populus Balsamifera*, est une espèce de peupliers qui peuple le nord du continent nord-américain, d'Est en Ouest, et se trouve exposée à des conditions climatiques peu clémentes.La seconde, *Populus Trichocarpa*, est principalement localisée en Californie, et bénéficie d'un climat continental.

Chacune des simulations est constituée de 50 haplotypes de la souche continentale, de 50 haplotyêpes 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. A partir des positions en paires de base, une carte de recombinaison génétique est générée en utilisant le taux de recombinaison moyen chez le peuplier. Le taux de recombinaison, noté τ_r, correspond au nombre moyen de paires de bases à parcourir pour qu'ait lieu un épisode de recombinaison génétique, *i.e.*, notant L la longueur du chromosome en Morgans (M), et N_{bp} le nombre de paires de bases le constituant, le taux de recombinaison génétique pour ce chromosome est donné par la relation :

$$\tau_r = \frac{L}{N_{bp}}$$

Dans ce scénario, les simulations ont été produites en utilisant un taux de recombinaison génétique moyen τ_r de 0.05 centiMorgans par million de paire de bases, correspondant à la valeur utilisée par les auteurs de l'étude avec le logiciel RASPberry (*Recombination via Ancestry Switch Probability*). A partir de la donnée de la position physique en paires de bases ainsi que du taux de recombinaison moyen, nous générons une carte de recombinaison génétique adaptée à nos simulations.

4.7.2 Génération aléatoire d'individus hybrides

Pour simuler un individu métissé, il est d'abord nécessaire de simuler l'emplacement des évènements de recombinaison. Pour ce faire, nous utilisons le modèle décrit dans (Price et al., 2009), en parcourant

4.7.3 Simulations à partir de ms et Seq-Gen

Dans le scénario d'introgression via flux de gènes, nous nous inspirons des modèles de simulation décrits dans (Martin & Jiggins, 2015). Ces modèles sont largement repris dans la littérature pour l'évaluation de statistiques telles que RND_{min} (Rosenzweig et al., 2016) et Bd_f (Pfeifer & Kapan, 2017). Chaque simulation est constituée de 100 individus. Un individu est généré en concaténant un certain nombre de séquences de nucléotides, d'une longueur fixée à 5000 paires de bases par séquence. Chacune de ces séquences est elle-même simulée suivant un modèle neutre ou alternatif. Le modèle neutre décrit un scénario démographique classique de populations divergentes. Le modèle alternatif décrit quant à lui un scénario légèrement différent, et servira à caractériser les séquences *introgressées*. Les lignes de commande *ms* permettant de générer les séquences de nucléotides pour le modèle neutre ainsi que pour le modèle alternatif sont données ci-dessous :

— Modèle neutre :

```
./ms 200 1 -I 4 50 50 50 50 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1
-r 50 5000 -T
```

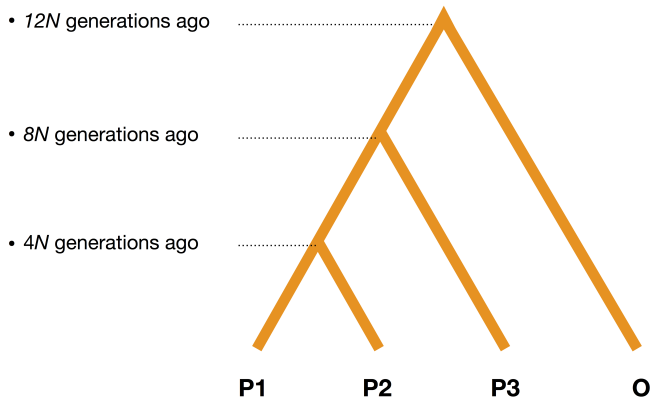


FIGURE 4.2 – Modèle neutre. $12N$ générations auparavant, premier épisode de divergence donnant naissance à P1 et à O. $8N$ générations auparavant, deuxième épisode de divergence voyant l'apparition de P3. $4N$ générations auparavant, dernier épisode de divergence et apparition de P2.

— Modèle alternatif :

```
./ms 200 1 -I 4 50 50 50 50 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1
-es 0.1 2 0.8 -ej 0.1 5 3 -r 50 5000 -T
```

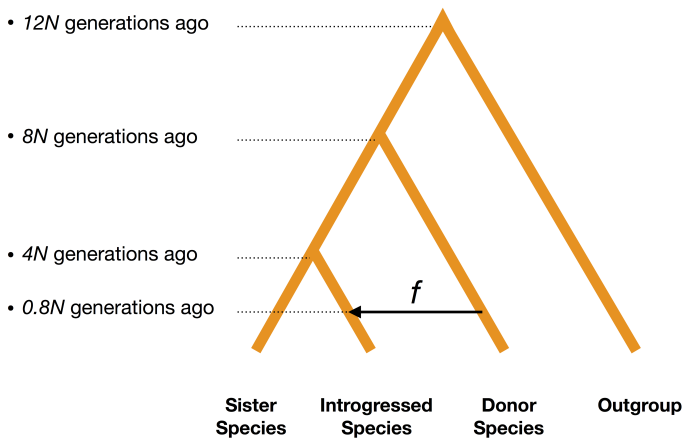


FIGURE 4.3 – Modèle alternatif. $12N$ générations auparavant, premier épisode de divergence donnant naissance à P1 et à O. $8N$ générations auparavant, deuxième épisode de divergence voyant l'apparition de P3. $4N$ générations auparavant, dernier épisode de divergence et apparition de P2. t unités de temps auparavant, épisode de flux de gènes de P3 vers la population P2.

La variable f présente en figure 4.3 représente le taux d'introgression, elle quantifie la proportion d'haplotypes présents dans la population P2 et qui proviennent de la population P3. Ainsi, une valeur de f égale à 1 reviendrait à simuler un épisode de divergence de la population P3, duquel découlerait la naissance de la population P2. La valeur de f utilisée ci-dessus est 0.2, signifiant que 20% des haplotypes présents dans la population 2 sont issus de la population P3.

4.7.4 Résultats de la comparaison des logiciels

Scénario de métissage Nous comparons ici notre méthode au logiciel RFMix destiné à la détermination de coefficients de métissage local.

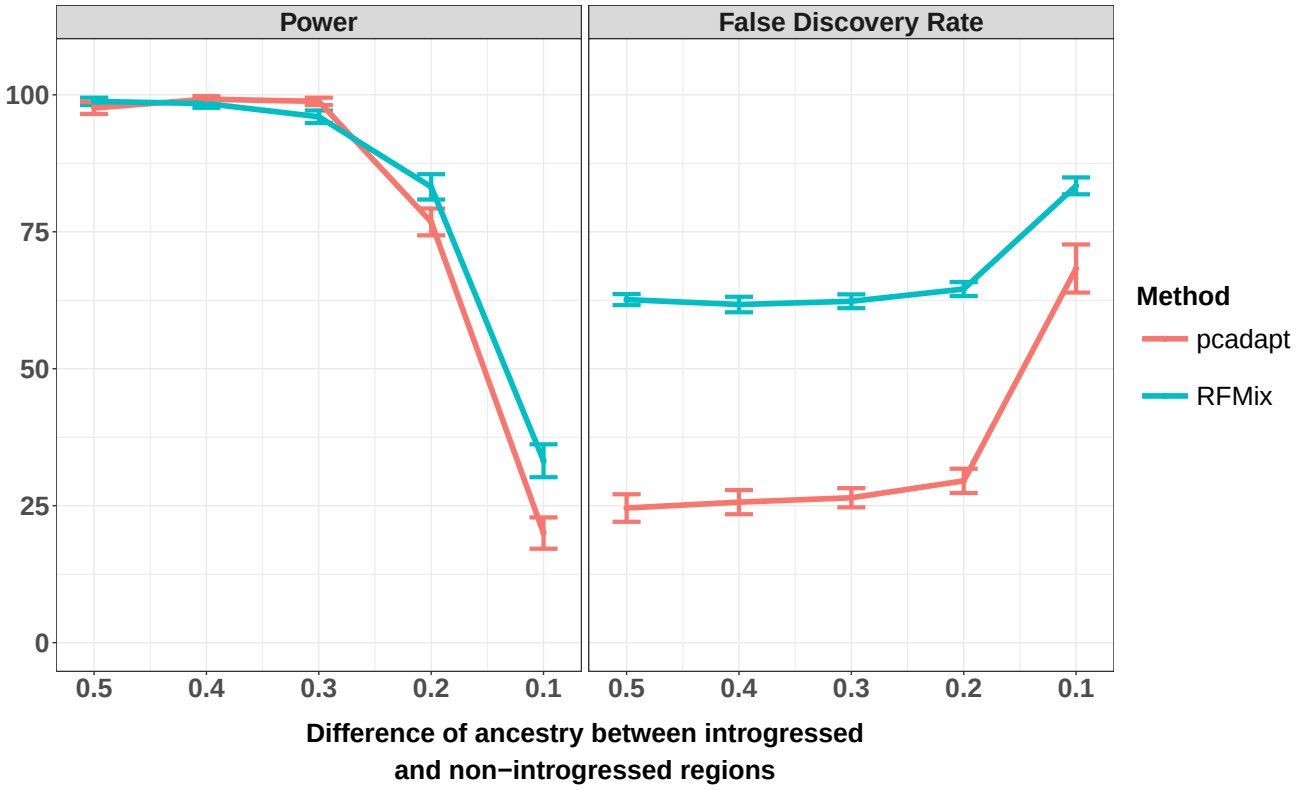


FIGURE 4.4 – 10 generations

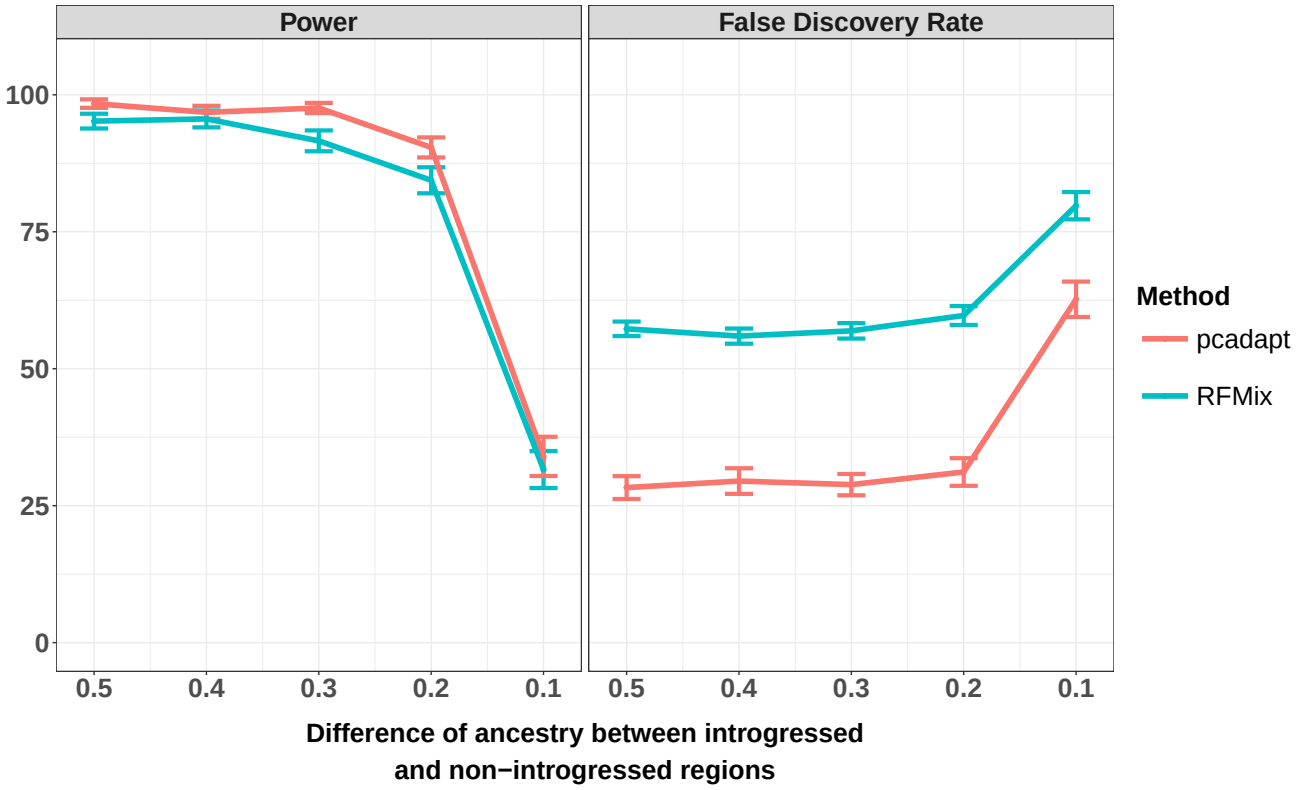


FIGURE 4.5 – 100 generations

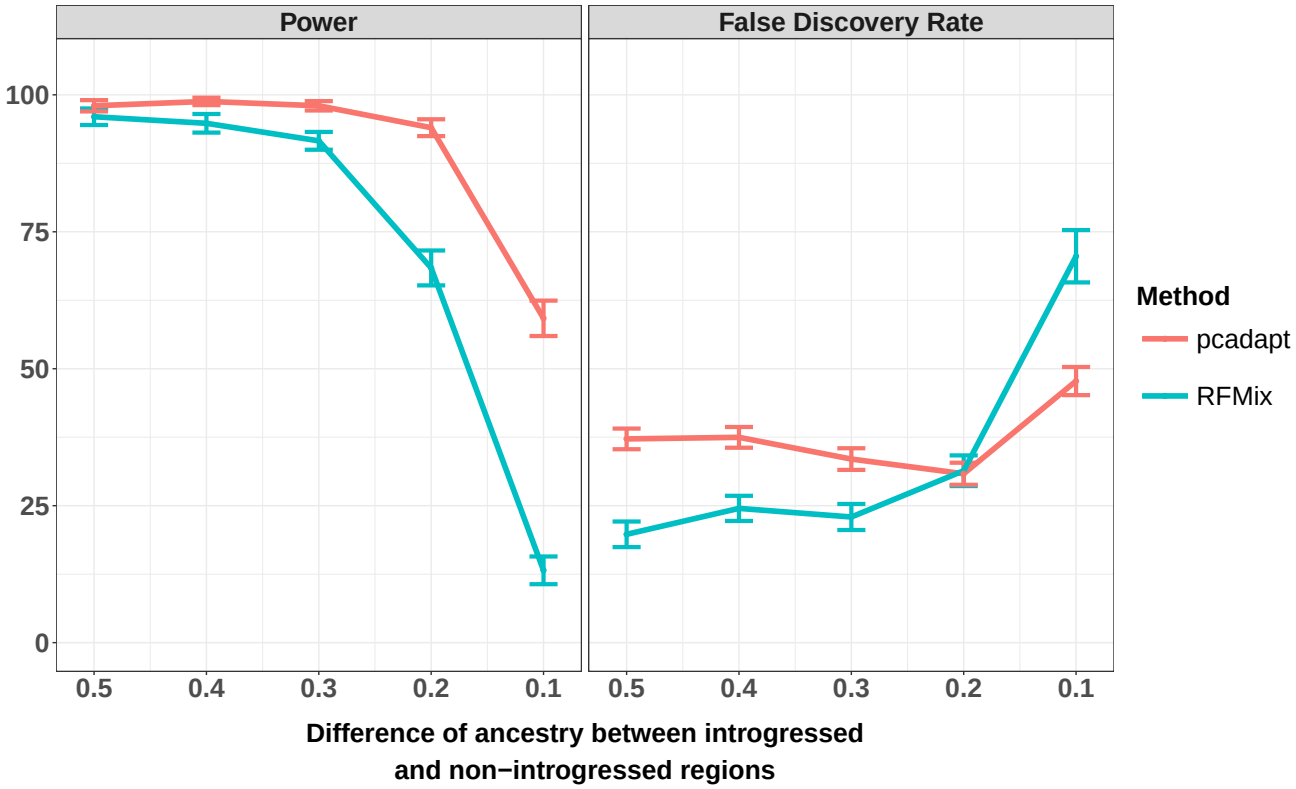


FIGURE 4.6 – 1000 generations

Scénario à flux de gènes Dans ce paragraphe, nous comparons notre statistique de test à un ensemble de statistiques implémentées dans le package R *PopGenome* : la statistique D de Patterson, RNDmin (Rosenzweig et al., 2016) et BDF (Pfeifer & Kapan, 2017).

Chapitre 5

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 toute particulière soit portée à l'utilisation des ressources de calcul.

Blockwise computation of covariance matrix Random SVD Storage in binary format Memory-mapping

Pairwise-Cor (Dray & Josse, 2015)

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info : the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Annexe A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
if(!require(devtools)) {
  install.packages("devtools", repos = "http://cran.rstudio.com")
}

if(!require(dplyr)) {
  install.packages("dplyr", repos = "http://cran.rstudio.com")
}

if(!require(ggplot2)) {
  install.packages("ggplot2", repos = "http://cran.rstudio.com")
}

if(!require(bookdown)) {
  install.packages("bookdown", repos = "http://cran.rstudio.com")
}

if(!require(thesisdown)) {
  devtools::install_github("ismayc/thesisdown")
}

if(!require(data.table)) {
  install.packages("data.table", repos = "http://cran.rstudio.com")
}

if(!require(pcadapt)) {
  devtools::install_github("bcm-uga/pcadapt")
}
```

```
if(!require(simulate)) {
  devtools::install_github("keurcien/simulate")
}

if(!require(kableExtra)){
  install.packages("kableExtra")
}

if(!require(maps)){
  install.packages("maps")
}

knitr::opts_chunk$set(echo = FALSE,
                      fig.align = 'center',
                      fig.width = 6,
                      results = 'hide')

# knitr::opts_chunk$set(cache = TRUE)

# The palette with black:
cbbPalette <- c("#000000",
                "#E69F00",
                "#56B4E9",
                "#009E73",
                "#F0E442",
                "#0072B2",
                "#D55E00",
                "#CC79A7")
```

Annexe B

The Second Appendix, for Fun

Bibliographie

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Bateson, W., & Mendel, G. (1913). *Mendel's principles of heredity*. University press.

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*(4), 969–980.

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.

Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews. Genetics*, *4*(3), 216.

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.

Darwin, C. (1980). L'Origine des espèces, trad. *Edmond Barbier (1876), Paris, Maspero*.

Dray, S., & Josse, J. (2015). Principal component analysis with missing values : A comparative survey of methods. *Plant Ecology*, *216*(5), 657–667.

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.

Frichot, E., & François, O. (2015). LEA : An r package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929.

Gayon, J. (1992). Darwin et l'après-darwin : Une histoire de l'hypothèse de sélection

dans la théorie de l'évolution. Kimé.

Geneva, A. J., Muirhead, C. A., Kingan, S. B., & Garrigan, D. (2015). A new method to scan genomes for introgression in a secondary contact model. *PloS One*, *10*(4), e0118621.

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

Giraud, C. (2014). *Introduction to high-dimensional statistics* (Vol. 138). CRC 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.

Harrison, R. G., & others. (1990). Hybrid zones : Windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, *7*, 69–128.

Jeong, C., & Di Rienzo, A. (2014). Adaptations to local environments in modern human populations. *Current Opinion in Genetics & Development*, *29*, 1–8.

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, *174*(2), E54–E70.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

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.

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.

Martin, J. W. D., Simon H., & Jiggins, C. D. (2015). Evaluating the use of abba–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 244–257.

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.

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.

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society : Series B (Statistical*

Methodology), *64*(4), 695–715.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... others. (2008). Genes mirror geography within europe. *Nature*, *456*(7218), 98.

Peng, B., & Kimmel, M. (2005). SimuPOP : A forward-time population genetics simulation environment. *Bioinformatics*, *21*(18), 3686–3687.

Pfeifer, B., & Kapan, D. D. (2017). Estimates of introgression as a function of pairwise distances. *BioRxiv*, 154377.

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.

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.

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.

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.

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).

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.

Yang, J. J., Li, J., Buu, A., & Williams, L. K. (2013). Efficient inference of local ancestry. *Bioinformatics*, *29*(21), 2750–2756.