Université de Montréal

Classes Alléliques d'Haplotypes et Sélection Positive dans le Génome Humain

par
Julie Hussin

département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de M. Sc.
en Bio-informatique
option Recherche

Août 2008

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Classes Alléliques d'Haplotypes et Sélection Positive dans le Génome Humain

présenté par
Julie Hussin

a été évalué par un jury composé des personnes suivantes :

Nicolas Lartillot
président-rapporteur
Damian Labuda
directeur de recherche
Philip Awadalla
membre du jury

Résumé

L'identification de régions génomiques cibles de la sélection naturelle positive permet de mieux comprendre notre passé évolutif et de trouver des variants génétiques fonctionnels importants. Puisque la fréquence des allèles sélectionnés augmente dans la population, la sélection laisse des traces sur les séquences d'ADN et ces empreintes sont détectées lorsque la variabilité génétique d'une région est différente de celle attendue sous neutralité sélective. Nous proposons une nouvelle approche pour analyser les données de polymorphismes : le calcul des classes alléliques d'haplotypes (HAC), permettant d'évaluer la diversité globale des haplotypes en étudiant leur composition allélique. L'idée de l'approche est de déterminer si un site est sous sélection positive récente en comparant les distributions des HAC obtenues pour les deux allèles de ce site. Grâce à l'utilisation de données simulées, nous avons étudié ces distributions sous neutralité et sous sélection en testant l'effet de différents paramètres populationnels. Pour tester notre approche empiriquement, nous avons analysé la variation génétique au niveau du gène de lactase dans les trois populations inclues dans le projet HapMap.

Mots clés : sélection positive, test statistique, variabilité génétique, génétique des populations, lactase

Abstract

Natural selection eliminates detrimental and favors advantageous phenotypes. This process leaves characteristic signatures in the underlying genomic segments that can be recognized through deviations in the allelic or in haplotypic frequency spectra. We introduce a new way of looking at the genomic single nucleotide polymorphisms : the haplotype allelic classes (HAC). The model combine segregating sites and haplotypic informations in order to reveal useful characteristics of the data, providing an identifiable signature of recent positive selection that can be detected by comparison with the background distribution. We compare the HAC distribution's partition between the haplotypes carrying the selected allele and the remaining ones. Coalescence simulations are used to study the distributions under standard population models assuming neutrality, demographic scenarios and selection models. To test, in practice, the performance of HAC and the derived statistic in capturing deviation from neutrality due to selection, we analyzed the genetic variation in the locus of lactase persistence in the three HapMap populations.

Keywords : positive selection, statistical test, genetic variability, population genetics, lactase

Table des matières

Liste des tableaux

Table des figures

Listes des Sigles et Abréviations

Tous les mots écrits en italique dans le texte de ce mémoire sont dans une autre langue que le français.

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ASI	Population asiatique chinoise et japonaise du projet HapMap
CNV	*Copy-Number Variant*
CEPH	Centre d'Étude du Polymorphisme Humain
CEU	Population européenne américaine du projet HapMap
$\mathbb{E}$	Espérance mathématique
EST	*Expressed Sequence Tag*
EHH	*Extended Haplotype Homozigosity*
FDR	*False discovery rate*
HAC	Classes alléliques d'haplotypes
IAM	Modèle du nombre infini d'allèles (ou d'haplotypes)
ISM	Modèle du nombre infini de sites
iHS	*integrated Haplotype Score*
Kb	Kilobase
LD	Déséquilibre de Liaison
LCT	Gène de Lactase
LRH	*Long-Range Haplotype*
MAF	Fréquence de l'Allèle Mineur
Mb	Mégabase
MCM6	minichromosome de maintenance 6
MRCA	Ancêtre Commun le plus Récent
$\mathbb{P}$	Probabilité
pb	paires de bases

SNP	*Single-Nucleotide Polymorphism*
Svd	*Selection statistic based on HAC Variance Difference*
UA	Ancêtre Ultime
UCSC	University of California Santa Cruz
$\mathbb{V}$	Variance
YRI	Population africaine des Yoruba d'Ibadan du projet HapMap

Remerciements

Je tiens tout d'abord à remercier mon directeur de maîtrise, Damian Labuda, pour m'avoir accueillie au sein de son laboratoire, pour m'avoir guidée sur des problématiques nouvelles, pour m'avoir encouragée lors de présentations orales de mes travaux, et pour m'avoir permis de présenter mes quelques résultats à CSHL, à une conférence internationale exceptionnelle. Finalement, je le remercie pour ses remarques constructives, qui m'ont permis d'améliorer la qualité du contenu de ce mémoire et de l'article qu'il contient, dans sa forme comme dans son fond.

Je remercie Nicolas Lartillot et Philip Awadalla d'avoir accepté de constituer, avec mon présent directeur de maîtrise, le jury de ce mémoire.

Mes remerciements vont également à Jean-François Lefebvre pour les nombreuses discussions, les conseils qu'il m'a donnés, les programmes qu'il a bien voulu m'envoyer ainsi que les corrections importantes qu'il m'a suggérées lors de la rédaction de ce mémoire. Je remercie Philippe Nadeau pour sa collaboration au projet reporté dans ce mémoire et pour les discussions intéressantes que cela a entraîné et je lui souhaite beaucoup de courage pour terminer sa maîtrise qui sera, sans nul doute, brillante.

Une pensée particulière pour Véronique Ladret et Claude Bherer avec qui j'ai partagé de nombreux cafés, plats thaïs et soupes tonkinoises durant ces deux ans, que je voudrais remercier d'abord pour leur amitié sincère et leurs encouragements, mais aussi pour nos innombrables discussions fructueuses qui m'ont été si précieuses.

Je remercie amicalement tous les membres du laboratoire du Dr. Labuda pour les discussions que j'ai eu la chance d'avoir avec eux ainsi que pour leurs remarques et suggestions. Je remercie Daniel Sinnett, Marie-Hélène Roy-Gagnon et Philip Awadalla pour leurs questions pertinentes et leurs suggestions intéressantes lors des réunions de laboratoire informelles où j'ai eu la chance de présenter périodiquement mes travaux.

Merci à tous les professeurs et autres personnes, attachés au programme de Bioinformatique de l'Université de Montréal, qui m'ont beaucoup appris tout au long de cette maîtrise. Je remercie Elaine Meunier pour sa gentillesse, son efficacité et les nombreux conseils administratifs qu'elle m'a donnés pour me permettre de soumettre ce mémoire à distance.

Je remercie le programme stratégique de bourses de formation des IRSC en bioinformatique biT, le Fonds québécois de la recherche sur la nature et les technologies et le Réseau de Médecine Génétique Appliquée pour les financements qui m'ont été attribués durant ces 2 ans et demi de maîtrise.

J'adresse un remerciement tout particulier à Elina, Emy, Jade et Nadine, qui m'ont écoutée me plaindre des aléas de la recherche et ont toujours trouvé les bons mots.

Je remercie très spécialement ma mère pour sa disponibilité et son soutien au jour le jour et pour m'avoir toujours montré le chemin à prendre dans les encouragements et le respect de mes décisions. Je remercie chaleureusement mon père pour ses sages conseils sur le métier de chercheur, pour ses exigences concernant mon orthographe, et pour me bousculer afin de toujours me pousser à me dépasser.

Je remercie également ma grand-mère de m'avoir ouvert les portes de sa maison durant ma période de rédaction, et de m'avoir ainsi offert un cadre simple, calme et agréable, où j'ai été chouchoutée par ses soins. J'adresse une pensée émue à mon grand-père, dont le souvenir continue à vivre dans cette maison et dans mon coeur.

Finalement, je tiens à remercier tout particulièrement Olivier Gandouet pour m'avoir toujours empêchée de baisser les bras, pour son intelligence dans tous les domaines et son indépendance d'esprit, pour son aide déterminante en probabilité et en statistiques, pour sa patience, son attention et son intérêt permanent pour mon travail.

“*I don't think that Evolution is supremely important because it is my specialty, it is my specialty because I think it is supremely important.*”

- George Gaylord Simpson

Introduction

Chez l'Homme, nous le savons, chaque individu est unique : il possède une morphologie, physiologie et psychologie qui lui est propre. C'est à la fois notre exposition à l'environnement et les variations génétiques, héritées de nos parents, qui maintiennent cette variabilité phénotypique. L'étude de ces variations génétiques à l'échelle des populations, chez l'humain mais aussi chez d'autres espèces vivantes, est depuis quelques décennies une discipline à part entière, communément appelée la génétique des populations. C'est depuis 1920 que la théorie de l'Évolution de Charles Darwin a été combinée aux travaux de Gregor Mendel, donnant ainsi naissance à cette application de la génétique, qui étudie les lois, au sein d'une population, de la distribution des gènes et des génotypes qui constituent le patrimoine génétique, héréditaire, de tout individu. Une population regroupe des individus d'une même espèce et ces groupes peuvent se définir de plusieurs façons. Il peut s'agir d'individus provenant d'une même région géographique, comme pour la population québécoise, d'une même origine ethnique, comme pour la population juive ashkénaze ou possédant une même langue maternelle, comme chez les Bantous d'Afrique. Plus précisément, une population est un groupe d'individus au sein duquel s'opère généralement le choix des conjoints pour les actes reproducteurs, assurant le passage du patrimoine génétique d'une génération à l'autre.

La génétique des populations humaines étudie principalement la variabilité génétique de notre espèce ainsi que les interactions entre les populations humaines et d'autres organismes parasitaires. Cette science a également des applications en épidémiologie, où elle permet d'identifier et de comprendre la transmission des gènes responsables de maladies héréditaires chez l'humain.

En étudiant les populations, nous cherchons à expliquer et comprendre les facteurs de l'évolution responsables de la diversité génétique : la mutation, la recombi-

naison, la dérive génétique, la migration et finalement, la sélection naturelle. Cette dernière force évolutive, au coeur de la théorie darwinienne, est l'objet principal de ce mémoire. Notre travail porte sur l'identification de régions génomiques ayant été soumises à la sélection naturelle, en recherchant des signatures moléculaires dans les séquences d'ADN grâce à des outils statistiques et informatiques.

L'importance des questions biologiques liées à la sélection naturelle est considérable, d'autant plus que la découverte de régions sous sélection et des mécanismes d'action de cette force évolutive permet l'avancée de la génétique appliquée à la médecine. D'un point de vue méthodologique, la quantité importante de données, générées par les projets de séquençage partout dans le monde oblige les chercheurs à développer de plus en plus de méthodes informatiques afin de traiter efficacement ces informations. De plus, des méthodes mathématiques et statistiques sont nécessaires afin de modéliser les phénomènes étudiés et de donner un sens aux données. En cela, la bio-informatique est grandement sollicitée en génétique des populations.

Une nouvelle méthode d'analyse des données populationnelles a été introduite par le laboratoire de génétique des populations humaines de l'Hôpital Ste-Justine : les classes alléliques d'haplotypes. Il s'agit là du point de départ de ce travail. Des résultats préliminaires suggéraient que cette méthode permettait de détecter des événements de sélection naturelle, cette idée constitue notre hypothèse de recherche. Le projet consiste à valider cette méthode d'analyse en développant une approche permettant de détecter les signatures de sélection positive le long des chromosomes. Les résultats de ce projet ont été rédigés dans un article scientifique original.

Cette contribution va être soumise pour publication, sous le nom de « *Haplotype allelic classes to detect recent and ongoing selection in the human genome* ». La soumission de l'article est prévue pour mars 2009 dans la revue BMC Bioinformatics (BioMed Central). Basée sur différents articles publiés dans le domaine, une méthodologie informatique et statistique a été mise au point, en collaboration avec mes collègues co-auteurs, pour valider l'hypothèse de recherche et pour développer

le test de sélection. L'article à paraître présente les résultats obtenus et j'en ai rédigé toutes les sections, corrigées à la suite de nombreux commentaires et discussions.

Ce mémoire est divisé en quatre parties. Le chapitre 1 est une revue bibliographique permettant à un lecteur non-spécialiste d'avoir une vue d'ensemble sur les concepts et méthodes utiles à la compréhension de mon travail. Le chapitre 2 présente la méthodologie détaillée ayant menée à l'obtention des résultats, présentés dans le chapitre 3 sous forme d'article scientifique. Finalement, ces résultats sont discutés au chapitre 4, qui est suivi de la conclusion générale de ce mémoire.

Chapitre 1

Revue de littérature

Lorsqu'on étudie l'évolution de l'Homme, l'un des grands objectifs est de comprendre comment l'*Homo sapiens* (appellation scientifique désignant les êtres humains que nous sommes) s'est adapté par sélection naturelle à leur environnement au cours de l'Histoire. L'émergence de l'humain remonte à plus de six millions d'années, lorsque la lignée des hominidés diverge de celle du chimpanzé. Depuis cette séparation, de nombreuses espèces ont vu le jour puis se sont éteintes, telle que l'*Homo erectus* ou l'*Homo habilis*, pour ne citer que les plus connues. Le seul représentant du genre Homo de nos jours est l'*Homo sapiens*. Depuis son émergence il y a plus de 200 000 ans, les populations humaines ont vécu un processus de différenciation intra-espèce et se sont adaptées à leur environnement au fur et à mesure de leur colonisation du globe. Ce processus est défini comme étant responsable de l'adaptation des populations humaines à leur milieu. Mais quelle part de cette adaptation doit-on à la sélection naturelle?

Dans ce chapitre, j'exposerai dans une première section les forces évolutives agissant sur les génomes des organismes vivants. Dans une deuxième section, les grandes lignes de certains principes mathématiques utilisés en génétique des populations seront introduits, et plus particulièrement ceux utiles à l'étude de l'action de la sélection naturelle dans les populations. La sélection naturelle positive est le sujet

principal de la troisième section de ce chapitre. Les données et les méthodes utilisées dans l'étude de ce type de sélection en génétique des populations y seront exposées. La dernière section décrit un système génétique modèle de choix dans l'étude des variations génétiques causée par la sélection positive. Il s'agit du trait de l'intolérance au lactose, un trait largement étudié, connu pour être sous sélection positive récente chez certaines populations humaines.

1.1 Forces évolutives

Les forces évolutives sont celles considérées comme responsables de la distribution des variations génétiques dans les populations.

Elles sont, par exemple, responsables de la naissance des grandes populations humaines que l'on observe actuellement. L'*Homo habilis*, notre plus vieil ancêtre identifié, se situait dans le berceau africain il y a trois millions d'années. Vient ensuite l'*Homo erectus* qui a émigré de l'Afrique vers l'Asie il y a deux millions d'années, puis vers l'Europe un million d'années plus tard. Les *Homo erectus* européens et asiatiques commencent alors leur évolution vers l'Homme moderne. Ils se trouvent cependant confrontés à une impasse évolutive que ne connaissent pas les *Homo* d'Afrique, où vers -200 000 à -100 000 ans l'*Homo sapiens* fait son apparition, biologiquement et intellectuellement mieux armé que ses prédécesseurs. Il repartira à la conquête de l'Asie et de l'Europe, supplantant les populations d'hominidés qu'il rencontre. Chaque individu de l'espèce possède une molécule d'ADN différente de celle de ses semblables, ce qui rend chacun de nous unique. Ces différences dans les séquences d'ADN sont causées par le processus de mutation (section 1.1.1) des bases azotées formant la molécule, processus existant chez toutes les espèces du vivant. Si les migrations (section 1.1.4.1) sont l'occasion de transmission d'allèles[1] d'une

[1]On appelle allèles les différentes versions d'un site particulier (appelé variant, site variant ou site polymorphe) sur une séquence d'ADN. Souvent les termes "allèle" et "haplotype" sont utilisés en tant que synonymes pour désigner les différentes versions d'une séquence génomique. Dans ce mémoire, ces différentes versions seront toujours appelés "haplotypes" (voir note 7), "allèle" se référant toujours à un site unique.

population à l'autre, la dérive génétique (section 1.1.3) et la sélection (section 1.1.5) sont responsables des variations de fréquence des allèles au cours des générations dans une population.

Cette théorie sur les origines de l'Homme est connue sous le nom de *Recent Out of Africa*. D'autres théories ont cependant été proposées :

- la théorie multirégionale soutenant que les *Homo erectus* se seraient répandus à travers le monde puis auraient évolué vers l'*Homo sapiens* simultanément et indépendamment;
- la théorie intermédiaire soutenant que plusieurs vagues d'expansion, régulières et progressives dans le temps, auraient provoqué un métissage génétique.

Aujourd'hui, grâce à de nouvelles preuves apportées par une étude génétique et phénotypique [Manica et al., 2007], la théorie monocentriste *Recent Out of Africa* reste largement prépondérante chez les scientifiques.

1.1.1 Mutation

Les mutations génétiques sont des modifications irréversibles de l'information génétique. Elles peuvent survenir de façon aléatoire au cours de la réplication de l'ADN lors de la division cellulaire. Dans la cellule, des mécanismes de contrôle efficaces de réparation de l'ADN corrigent la très grande majorité de ces erreurs, mais une faible proportion est tout de même transmise aux cellules-filles. Notons que les mutations peuvent également être dues à l'exposition à des agents mutagènes présents dans l'environnement (radiations, agents chimiques, virus). Chez les organismes pluricellulaires, ces mutations doivent apparaître dans la lignée germinale[2] pour qu'elles soient transmises à la descendance. Lorsqu'il y a mutation, le site touché présente deux états : l'état ancestral[3], présent avant qu'il y ait mutation, et l'état muté, appelé communément état dérivé. Dans une même population, certains

[2]L'ensemble des cellules germinales sont les cellules qui sont susceptibles de former les gamètes, par opposition aux cellules somatiques.

[3]Chez l'humain, on détermine l'état ancestral par comparaison des séquences génomiques avec des séquences orthologues provenant d'une espèce d'un groupe externe, tel que le chimpanzé ou le macaque.

individus peuvent présenter l'un ou l'autre de ces deux états : le site est alors appelé site polymorphe, ou variant.

1.1.1.1 Mutation et variabilité génétique

En génétique des populations, on ne considère que les mutations qui sont des modifications héréditaires du matériel génétique. En effet, les mutations transmises à la descendance sont à la base de la variabilité génétique présente dans les populations. Elles peuvent avoir, sur l'organisme, un effet neutre (pour la majorité), défavorable (occasionnellement) ou favorable (rarement). Le taux d'élimination et d'accumulation des nouvelles mutations va dépendre de leurs effets.

Les substitutions sont les mutations ponctuelles les plus connues, elles ne modifient qu'un nucléotide de la séquence d'ADN en l'échangeant pour un autre. Cet échange peut être une transition[4] ou une transversion[5]. Elles peuvent se situer dans les séquences codantes ou non-codantes, dans les régions promotrices, dans les sites d'épissage, etc. Dans les régions codantes, ces mutations sont soit synonymes soit non-synonymes. Les mutations synonymes sont des substitutions ne modifiant pas la séquence protéique alors que les substitutions non-synonymes la modifie.

Les substitutions ne sont pas les seules formes de variabilité génétique. On trouve également :

- les indels : insertions ou délétions de 1 ou quelques nucléotides ;
- les microsatellites : motifs de 2 à 10 paires de bases (pb) généralement répétés de 10 à 100 fois. Le nombre de répétitions varie entre les individus ;
- les minisatellites : motifs de 10 à 100 pb répétés un grand nombre de fois. Ces motifs sont généralement riche en guanine(G)-cytosine(C) ;
- les Copy-Number Variant (CNV) : il s'agit de différences entre les individus du nombre de copies d'un gène particulier dans leur génome [Freeman, 2006].

[4]La substitution d'une base purique par une autre base purique ou d'une base pyrimidique par une autre base pyrimidique

[5]La substitution d'une base purique par une base pyrimidique ou d'une base pyrimidique par une base purique

Puisque nos travaux seront centrés sur l'Homme, nous considérerons toujours des populations diploïdes[6]. Pour chaque site ayant subit une substitution, un individu sera soit homozygote, si ses deux chromosomes présentent le même allèle, soit hétérozygote, si il possède deux allèles différents. Pour une séquence génomique donnée, chaque individu possédera donc deux haplotypes[7] distincts.

1.1.1.2 Taux de mutation

On définit le taux de mutation μ comme la probabilité pour qu'une mutation particulière affecte la séquence d'ADN génomique par gamète et par génération (ou division cellulaire pour les organismes pluricellulaires).

Les taux de mutation peuvent varier grandement en fonction de nombreux paramètres. Chez certaines espèces, le taux de mutation est plus élevé que chez d'autres, et les taux varient aussi le long des séquences génomiques chez une même espèce. Chez l'humain, il y a encore beaucoup de régions génomiques pour lesquelles nous n'avons pas d'estimé fiable du taux de mutation. Cependant, nous savons qu'il se maintient dans une certaine fourchette : une étude a estimé qu'en moyenne, le taux de mutation dans le génome humain était d'environ $2,5 \cdot 10^{-8}$ par pb par génération [Nachman and Crowell, 2000].

En génétique des populations, le paramètre du taux de mutation par séquence par génération que l'on mesure dans une population diploïde est $\theta = 4N_e\mu$, où N_e est l'effectif efficace[8] de la population. Nous reverrons ce paramètre à la section 1.2.2.

1.1.2 Recombinaison

La recombinaison méiotique est un mécanisme de la reproduction sexuée : il s'agit d'un échange de segments homologues entre deux molécules d'ADN. Elle se

[6]Dans une population diploïde, les individus ont deux versions de leur génome : chacun de leurs chromosomes possèdent leur homologue.

[7]Un haplotype est la combinaison des différents allèles des sites variants présents et physiquement liés sur une région génomique.

[8]L'effectif qui contribue réellement à la reproduction de la population.

produit lors de la méiose, lorsque les chromosomes homologues sont appariés sur la plaque métaphasique. Ce processus permet aux descendances de présenter des combinaisons de gènes différentes de celles de leurs parents. La recombinaison entraîne la redistribution des nouvelles mutations. Avec la mutation, elles constituent les principales sources de variabilité génétique et sont à l'origine de la diversité haplotypique du génome d'une espèce qui s'adaptera plus facilement à une modification de l'environnement.

À l'échelle chromosomique, on observe chez l'humain une moyenne de 33 recombinaisons par méiose. Le génome humain comprend environ 3 milliards de pb, la probabilité de recombinaison par mégabase[9] est de 0,01. Cependant, il a été montré que le taux de recombinaison n'est pas uniforme le long des chromosomes.

1.1.2.1 Taux de recombinaison

Le taux de recombinaison r se définit comme la probabilité d'avoir une recombinaison entre deux locus[10] à une génération donnée.

De façon similaire au taux de mutation θ, le paramètre du taux de recombinaison par séquence par génération que l'on mesure pour une population diploïde est $\rho = 4N_e r$, où N_e est l'effectif efficace de la population. Grâce à l'utilisation de nouveaux outils bio-informatiques et statistiques pour analyser les variations haplotypiques dans les populations humaines, il est actuellement possible d'étudier la distribution de ρ le long des séquences d'ADN [Kuhner et al., 2000, Hudson, 2001, McVean et al., 2002, Fearnhead and Donnelly, 2002, Li and Stephens, 2003]. Notre équipe a développé le logiciel `infrec` [Lefebvre and Labuda, 2008], implémentant une méthode heuristique[11] et intuitive permettant d'analyser la densité des recombinaisons à l'échelle de la séquence d'ADN.

[9] 1 Mb = 1 000 000 pb.

[10] Un locus définit un emplacement précis sur un chromosome. On peut se référer à un locus pour désigner un nucléotide unique ou une région génomique composée de plusieurs nucléotides.

[11] Méthode approximative qui fournit rapidement une solution pas nécessairement optimale pour un problème difficile donné.

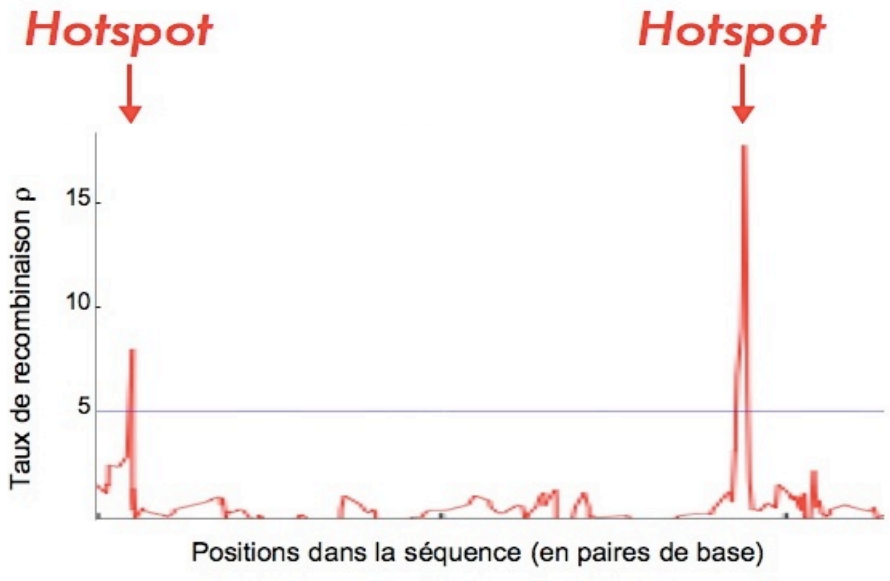

FIG. 1.1 – Distribution du taux de recombinaison le long des séquences
Un hotspot de recombinaison est un court fragment génomique de 1 à 2 Kb très riches en recombinaisons.

1.1.2.2 Hotspots de recombinaison

Les études génomiques d'inférence du taux de recombinaison chez l'humain ont permis de mettre en évidence la présence de courts fragments de 1 à 2 Kb[12], très riches en recombinaison, séparés par de longs fragments qui en sont dépourvus. Ces fragments sont respectivement appelé hotspots et coldspots de recombinaison. Au niveau d'un hotspot, le taux de recombinaison peut être des dizaines, même des centaines de fois plus élevé que pour le reste des séquences (Fig. 1.1). Des études empiriques suggèrent qu'ils s'agissent d'entités omniprésentes dans le génome humain et qu'il y en aurait plus de 25 000. On trouverait en moyenne un hotspot tous les 50 Kb et 80% des événements de recombinaisons dans 10 à 20 % du génome [Myers et al., 2005].

[12]1Kb = 1000 pb

1.1.2.3 Recombinaison et liaison génétique

On dit de deux locus qu'ils sont liés génétiquement lorsqu'ils sont hérités conjointement. Des locus présents sur des chromosomes différents ne sont pas liés génétiquement, car les chromosomes sont transmis indépendamment durant la méiose. Cependant, le processus de recombinaison méiotique peut fragmenter la liaison génétique de locus se trouvant sur un même chromosome. Ceux ci, bien que physiquement liés, seront génétiquement indépendants. Plus la distance physique entre deux locus est importante, plus la probabilité qu'un événement de recombinaison les sépare lors de la méiose est grande.

Pour décrire une situation dans laquelle certaines combinations d'allèles se produisent plus (ou moins) fréquemment que ce qui est attendu lorsque les allèles s'associent indépendamment dans une population, on parle de déséquilibre de liaison (ou LD pour *linkage disequilibrium*). Lorsqu'un nouveau variant apparaît dans une région génétique, il sera en LD complet avec les variants de cette région car il ne sera initialement présent sur un seul haplotype. Avec le temps qui passe, la recombinaison va rompre la liaison génétique et brasser les mutations: le nouveau variant apparu sera alors dissocié de son haplotype initial. Lorsque les séquences évoluent au hasard, on s'attend donc à avoir un déclin dans le temps du LD de ce variant avec d'autres (Fig. 1.2). Les progrès récents faits dans ce domaine sont décrits, entre autres, par J.K. Pritchard et M. Przeworski [2001].

1.1.3 Dérive génétique

Les fréquences des allèles mutés dans les populations changent au cours du temps. Les nouvelles mutations peuvent devenir très fréquentes dans une population et rester très peu présentes dans une autre, par la simple action du hasard : c'est ce qu'on appelle la dérive génétique.

Cette force évolutive est due, avant tout, à la taille finie de la population. Si les populations étaient de taille infinie, les fréquences alléliques seraient stables au

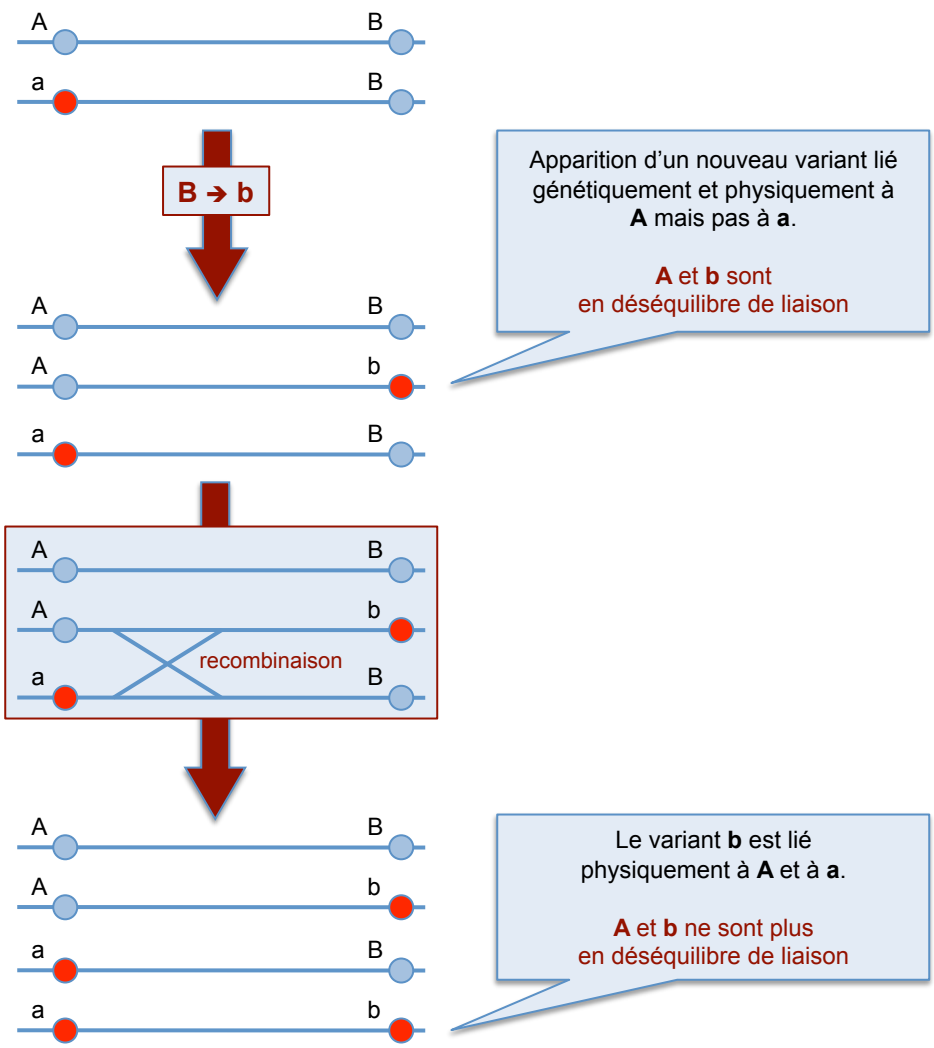

FIG. 1.2 – La recombinaison effrite le déséquilibre de liaison

Initialement, un seul site polymorphe est présent sur les séquences et possède les allèles A et a. La mutation d'un site voisin, transformant l'allèle B en b, ne survient que sur un seul chromosome, portant soit l'allèle A ou a au premier site (A dans cet exemple). De ce fait, lorsque cette mutation B/b est jeune, seulement trois haplotypes sur les 4 possibles (AB, aB, Ab, ab) existent. L'allèle b sera toujours retrouvé sur le même chromosome que l'allèle A. Cependant, l'association entre les allèles aux deux sites sera graduellement rompue par la recombinaison entre les sites, ce qui va permettre l'apparition du quatrième haplotype ab. Plus la fréquence du chromosome recombinant (portant l'haplotype ab) va croître dans la population, plus le déséquilibre de liaison entre les deux sites mutés va décroître. source : [Ardlie et al., 2002].

cours des générations, en l'absence de nouvelles mutations et de sélection naturelle. Dans une population de taille finie, un nombre limité de ses membres participe au processus reproductif, et ceux-ci ne produisent pas nécessairement le même nombre de descendants. Ainsi, les fréquences alléliques varient aléatoirement : cela est dû à la variabilité du tirage aléatoire des gènes d'une génération à l'autre.

Ces variations de fréquence sont plus fortes dans les petites populations. En effet, si la population est grande, la perte d'une copie d'un variant présent chez un individu sans descendance sera compensée par le fait qu'un autre individu, possédant le même variant, aura plusieurs descendants. Par contre, dans les petites populations, cet effet de moyenne n'agit pas et les fréquences alléliques fluctueront grandement. Cela conduit à la différenciation génétique progressive des populations filles issues d'une même population mère.

La théorie neutraliste de l'évolution moléculaire [Kimura, 1983] attribue un très grand rôle à la dérive pour expliquer la diversité génétique. Cette théorie décrit la façon dont le hasard contribue à l'évolution des populations. La théorie neutraliste n'est pas incompatible avec la théorie darwinienne de sélection naturelle, qui décrit comment l'environnement influe sur l'évolution adaptative des populations.

1.1.4 Facteurs démographiques

1.1.4.1 La migration

La première force démographique qui a été étudiée en génétique des populations est le processus de migration. En effet, les différentes populations d'une même espèce ne sont pas complètement isolées les unes des autres. Chacune évolue dans des conditions qui lui sont propres, mais peut échanger des individus avec les autres.

Les migrations ont pour conséquence de modifier les fréquences alléliques dans la population qui reçoit les nouveaux individus. Elles peuvent y introduire de nouveaux mutants. Le flux migratoire, aussi appelé flux génique, contribue au brassage génétique et limite la diversification des populations d'une même espèce.

1.1.4.2 L'expansion démographique et le *bottleneck*

L'expansion démographique est une augmentation importante de l'effectif efficace d'une population en un court laps de temps. Ce processus démographique provoque une augmentation de la variabilité génétique de la population, car plus il y a d'individus, plus les forces évolutives vont agir pour créer de la diversité. Le *bottleneck*, ou goulot d'étranglement en français, est un rétrécissement soudain de l'effectif efficace d'une population. En général, cette réduction est importante et provoque une perte considérable de diversité génétique dans la population en question, car elle cause l'extinction de nombreux variants génétiques.

Puisque l'effectif de la population est réduit lors d'un *bottleneck*, l'effet de la dérive génétique augmente, étant donné que cet effet est plus fort dans les populations de petite taille (voir section 1.1.3). Puisqu'un *bottleneck* réduit le nombre d'individus aptes à se reproduire, il va également entraîner une augmentation de la consanguinité au sein de la population.

1.1.5 Selection naturelle

Charles Darwin, père de la théorie de l'Évolution, écrivait : « J'ai nommé sélection naturelle le principe selon lequel toute petite variation est conservée lorsqu'elle est utile » [Darwin, 1859]. Ce concept désigne le fait que des variants génétiques qui favorisent la survie et la reproduction voient leurs fréquences augmenter d'une génération à l'autre. Les porteurs de ces mutations avantageuses ont plus de descendants, et ceux-ci auront eux même plus de descendants porteurs.

La sélection naturelle ne résulte pas toujours en une adaptation évolutive directionnelle, elle permet souvent la maintenance du *status quo* en éliminant, par exemple, les variants les moins adaptés. Trois grands types de sélection ont été précisément défini : la sélection positive, la sélection négative et la sélection balancée.

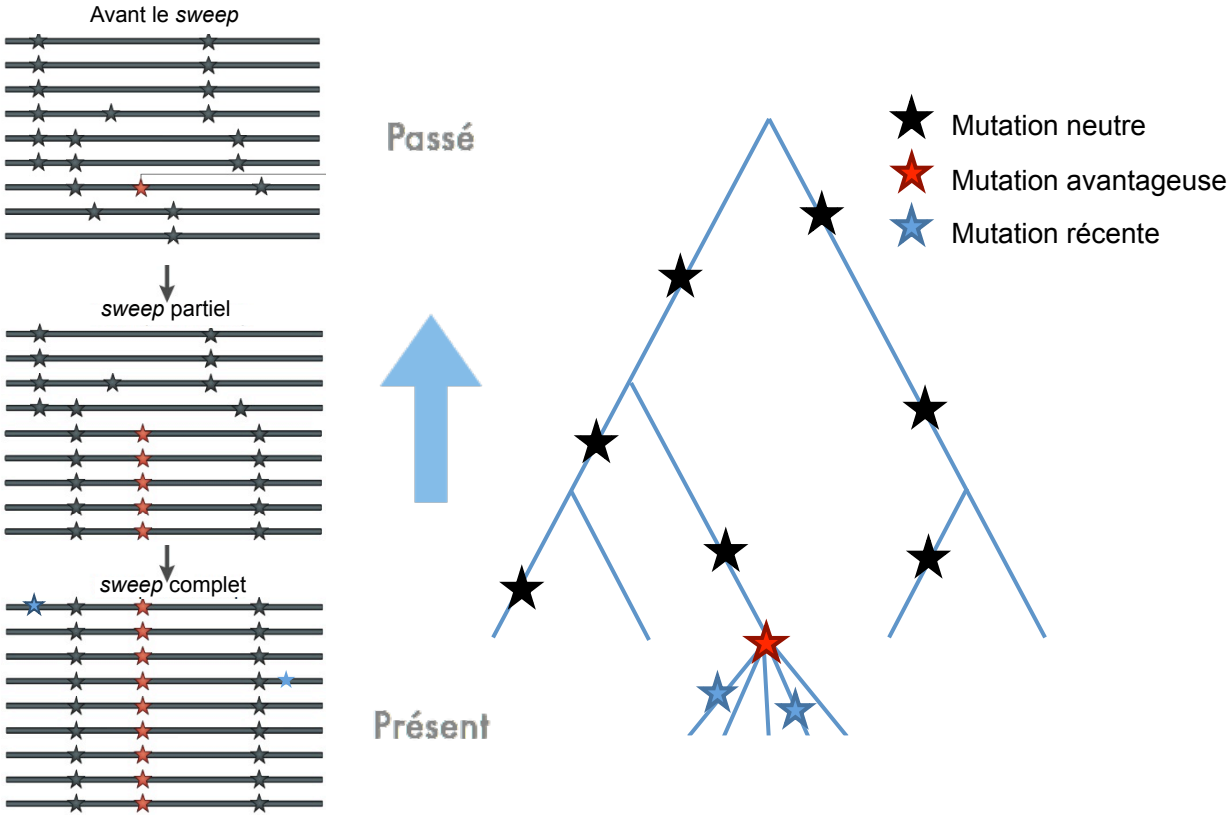

FIG. 1.3 – Empreinte de la sélection positive : le *selective sweep*

Une nouvelle mutation avantageuse apparaît initialement sur un chromosome. Après plusieurs générations, une réduction de la variabilité dans la région va apparaître. Les variants neutres situés sur le même fragment que le site sélectionné vont voir leur fréquence augmenter jusqu'à fixation, lorsque le *sweep* est completé (*sweep* complet). Les autres variants seront perdus : un *selective sweep* complet élimine la variabilité génétique dans la région concernée pour la population concernée par la sélection positive, sauf pour les nouvelles mutations à faible fréquence qui continueront à s'accumuler lentement à un taux de mutation de θ. La forme de l'arbre généalogique sera transformée par le *selective sweep* en généalogie en étoile pour la branche portant la mutation avantageuse. Les autres branches se terminent, ce qui cause la perte des mutations survenues sur ces lignages.

1.1.5.1 La sélection positive

Ce type de sélection est aussi appelé sélection directionnelle. Elle survient lorsque l'état dérivé d'un variant génétique présente un avantage sélectif par rapport à l'état ancestral. La fréquence de l'allèle dérivé va ainsi croître dans la population. Durant cette montée en fréquence, l'allèle sélectionné entraîne avec lui les allèles qui lui sont génétiquement liés : leur fréquence va également croître, jusqu'à fixation. La région génomique concernée par cet effet d'auto-stop génétique (ou *hitch-hiking*) va subir une réduction de variation génétique. Ce phénomène, connu sous le nom de *selective sweep*, va laisser une empreinte spécifique sur les chromosomes (Fig. 1.3).

1.1.5.2 La sélection négative

Ce type de sélection est aussi appelé sélection purificatrice. Il s'agit de la perte, par sélection, d'allèles fortement délétères. La sélection négative est qualifiée de stabilisante car elle participe à la diminution de la diversité génétique. Il s'agit probablement du mécanisme d'action de sélection naturelle le plus commun. Lorsque les niveaux de sélection purificatrice sont bas, comme dans les séquences de *junk DNA*[13], le taux d'accumulation des nouvelles mutations est supérieur. Cela ne signifie cependant pas que le taux de mutation dans ces séquences soit plus élevé.

L'identification de cible de la sélection négative peut mener à la détection des régions ou variants fonctionnellement importants. Par exemple, les gènes codant pour une protéine impliquée dans une maladie sont fréquemment sous sélection négative, lorsque la maladie en question entraîne une réduction de la capacité reproductive et/ou de la durée de vie.

1.1.5.3 La sélection balancée

Contrairement à la sélection positive, ce type de sélection favorise deux (ou plusieurs) allèles : elle maintient le caractère polymorphe d'un trait héréditaire dans une population. La diversité génétique de la région sous sélection balancée est maintenue

[13]Portions de séquences d'ADN pour lesquelles aucune fonction n'a été identifiée

dans la population. La sélection balancée est assez commune chez certaines espèces et différents mécanismes biologiques peuvent y mener.

Un premier mécanisme propose que la sélection soit sur-dominante, lorsque dans un milieu l'individu hétérozygote a un *fitness*[14] plus grand que l'individu homozygote. Un exemple est celui de la drépanocytose[15]. La maladie est causée par un allèle récessif : les homozygotes recessifs sont atteint de la maladie et leur durée de vie est réduite. Un individu hétérozygote ne sera pas atteint de drépanocytose mais la forme de ses cellules sanguines sera altérée, permettant une résistance à la malaria : cette résistance est donc soumise à la sélection naturelle dans les zones tropicales. Les deux allèles sont conservés, causant la persistance de la maladie dans ces populations.

La sélection fréquence-dépendante est un autre mécanisme de sélection balancée. Le *fitness* relatif d'un phénotype spécifique baisse lorsque la fréquence d'un phénotype augmente. Ce mécanisme est souvent le résultat d'interactions entre différents organismes. L'interaction entre l'homme et le virus de la grippe est un exemple de ce type de sélection. Lorsque qu'une souche devient fréquente dans une population humaine, la plupart des individus développe une réponse immunitaire contre la souche. Par contre, une nouvelle souche pourra se répandre rapidement parmi les individus. Cet avantage permet l'évolution constante des souches virales et des nouveau virus de la grippe chaque année! Lorsque les conditions environnementales fluctuent, dans le temps ou dans l'espace, il est possible qu'un phénotype qui est négativement sélectionné sous certaines conditions présente un avantage sélectif sous d'autres. Il s'agit là d'un troisième mécanisme pouvant mener à la sélection balancée.

La sélection fréquence-dépendante est cependant considérée comme de la sélection positive dans certaines études. Il s'agit en effet de la sélection directionnelle d'un caractère, par contre elle n'intervient pas nécessairement dés l'apparition d'une

[14]Mesure relative de la contribution génétique d'un état génétique aux générations futures. Appelé aussi valeur adaptative.

[15]Maladie héréditaire responsable d'une anomalie de l'hémoglobine contenue dans les globules rouges.

mutation. Les empreintes moléculaires laissées par ce type de mécanisme seront différentes de celles laissées par une *selective sweep*, nous allons donc considérer qu'il s'agit bel et bien d'un cas à part.

1.2 Mathématique génétique

La génétique des populations est une science jeune, débutée au début du XXième siècle avec les travaux de génétique théorique de Hardy [1908] et Weinberg [1908], suivit dans les années 1930 par ceux de Fisher [1930] et Wright [1931], ansi que de Haldane [1932]. En tenant compte des forces évolutives présentées à la section 1.1, ces chercheurs se sont intéressés aux lois régissant l'évolution des gènes dans les populations. Ils ont considéré ce problème, central en science de la vie, d'un point de vue essentiellement mathématique.

1.2.1 Le modèle de Wright-Fisher

Il s'agit d'un modèle de base décrivant la reproduction d'une population de taille finie et la façon dont les fréquences alléliques évoluent en présence de dérive génétique, de mutation, de forces démographiques et de sélection.

Définition du modèle : on considère une population haploïde[16] de taille 2N où N est le nombre d'individus diploïdes. On suppose que les générations sont indépendantes, c'est-à-dire qu'elles ne se chevauchent pas. Le passage d'une génération à l'autre se fait comme suit : la génération g+1 se constitue grâce à 2N tirages au hasard avec remise dans les 2N copies de séquences de la génération g. La figure 1.4 montre un exemple de l'évolution des copies de séquences par le processus de Wright-Fisher.

[16]Population dont les individus n'ont qu'une seule version de leur génome, par opposition à "diploïde" (voir note 6).

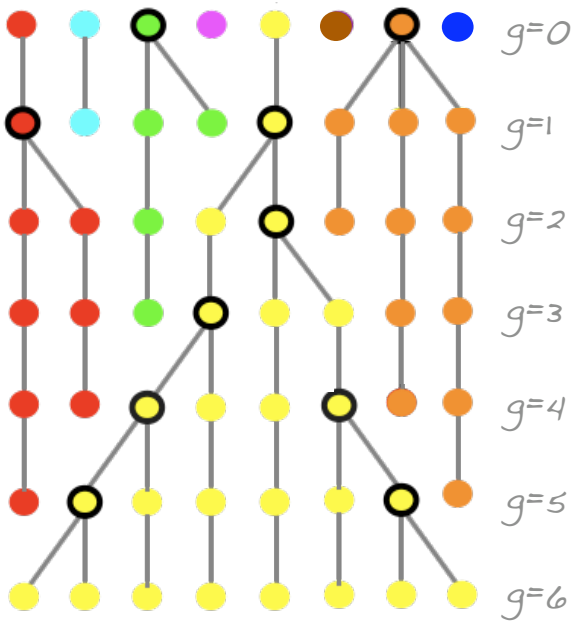

FIG. 1.4 – Processus d'évolution simple de Wright-Fisher

Illustration du processus de Wright-Fisher, avec 2N = 8, au cours de 7 générations. Notons que toutes les copies de séquences de la génération 6 sur l'exemple sont issus d'une seul et même séquence de la génération 0.

Résultats dérivés de ce modèle : Soit $z_{A,g}$ le nombre de séquences de type A à la génération g : la probabilité que l'on ait j séquences de type A à la génération g+1, sachant qu'on en a i à la génération g suit une loi binomiale $B(2N, \frac{i}{2N})$:

$$\mathbb{P}(z_{A,g+1} = j | z_{A,g} = i) = \binom{2N}{j} (\frac{i}{2N})^j (1 - \frac{i}{2N})^{2N-j} \tag{1.1}$$

Aussi, on peut montrer que :

$$\mathbb{E}(z_{A,g+1} | z_{A,g} = i) = i, \tag{1.2}$$

$$\mathbb{P}(\text{fixation de A} | z_{A,0} = i) = \frac{i}{2N}. \tag{1.3}$$

Ce dernier résultat est intéressant car il signifie que la probabilité de fixation de A est égale à la fréquence initiale de A dans la population.

De même,

$$\mathbb{P}(\text{extinction de A} | z_{A,0} = i) = \frac{2N - i}{2N}. \tag{1.4}$$

Ainsi, une évolution génétique peut se produire en ne faisant jouer que le hasard. Ce modèle, très simple, peut être compliqué avec des critères d'accouplements, l'influence de forces évolutives, etc.

Équilibre de Hardy-Weinberg : Lorsque l'on utilise le modèle de Wright-Fisher, on suppose souvent que la population est à l'équilibre de Hardy-Weinberg, modèle théorique central en génétique des populations. Il s'agit d'une théorie qui postule qu'il y a un équilibre entre la fréquence des allèles et des génotypes[17] au cours des générations. La notion d'équilibre dans le modèle de Hardy-Weinberg est soumise aux conditions (ou hypothèses) suivantes :

– le modèle concerne les espèces diploïdes qui se reproduisent de façon sexuée;

[17]ici, génotype signifie la combinaison des deux allèles sur les chromosomes homologues d'un variant. B et b sont deux allèles d'un variant, BB, Bb et bb sont les trois génotypes possibles.

Tab. 1.1 – Carré de Punnett pour l'équilibre de Hardy-Weinberg

		♀ A (p)	♀ a (q)
♂	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

- la population est panmictique: les couples se forment au hasard (panmixie), et leurs gamètes se recontrent au hasard (pangamie);
- la population est de taille infinie (ce qui signifie en vérité qu'elle est très grande) pour minimiser les variations d'échantillonnage;
- il ne doit y avoir ni sélection, ni mutation, ni migration. De ce fait, il n'y a ni perte ni gain d'allèle;
- les générations successives sont non-chevauchantes : il n'y a pas de croisement entre les différentes générations;

Soit A et a, deux allèles d'un même locus, de fréquence p et q, respectivement, à la génération g. D'après le tableau 1.1, à la génération suivante, on a que la fréquence d'un génotype homozygote AA est de p^2, la fréquence d'un génotype homozygote aa est de q^2 et la fréquence d'un génotype hétérozygote Aa est de $2pq$. De cet équilibre de Hardy-Weinberg, découle la loi de distribution génotypique :

$$p^2 + 2pq + q^2 = 1 \tag{1.5}$$

Sans perturbations, ce système permet aux fréquences alléliques de rester constantes d'une génération à l'autre. Les fréquences génotypiques se déduisent directement des fréquences alléliques et restent également constantes.

1.2.2 Le processus de mutation

Dans ce qui est présenté ci-dessus, on suppose qu'il n'y a pas de mutation. Si l'on suppose que l'allèle A peut muter en l'allèle a (ou inversement) à un taux μ faible, comme chez l'Homme, les résultats de Wright [1931] et Haldane [1932] indiquent que les changements de fréquences causé par le processus de mutation uniquement se feront très lentement au fil des génération. Cependant, c'est ce processus qui va mener au maintient de ces deux allèles dans la population.

1.2.2.1 Modèles de mutation

Dans les populations diploïdes, on s'intéresse au taux de substitutions survenues dans les séquences qui se transmettent d'une génération à l'autre : il s'agit du paramètre $\theta = 4N\mu$.

Pour estimer ce taux de mutation à partir des données populationnelles, il est important de modéliser le processus mutationel. Les modèles de mutation du nombre infini d'allèles et de sites (Fig. 1.5), introduit par M. Kimura [1968] pour étudier la dérive génétique (voir section 1.1.3), sont des simplification communément utilisées et permettent de réduire la complexité du phénomène étudié. Ces modèles sont décrit dans ce qui suit.

Infinite alleles model (IAM): D'un point de vue moléculaire, un gène consiste en une séquence d'un grand nombre de bases azotées, appelées nucléotides. Une mutation se produisant à un nucléotide donné va très probablement créer un nouvel haplotype, n'existant pas dans la population. Pour cela, Kimura et Crow [1964] proposent de considérer que le processus de mutation provoque toujours l'apparition d'un nouvel haplotype (qu'ils dénomment *allele*) non-représenté dans la population. Ainsi, on suppose que le nombre d'haplotypes générés par le processus de mutation est infini.

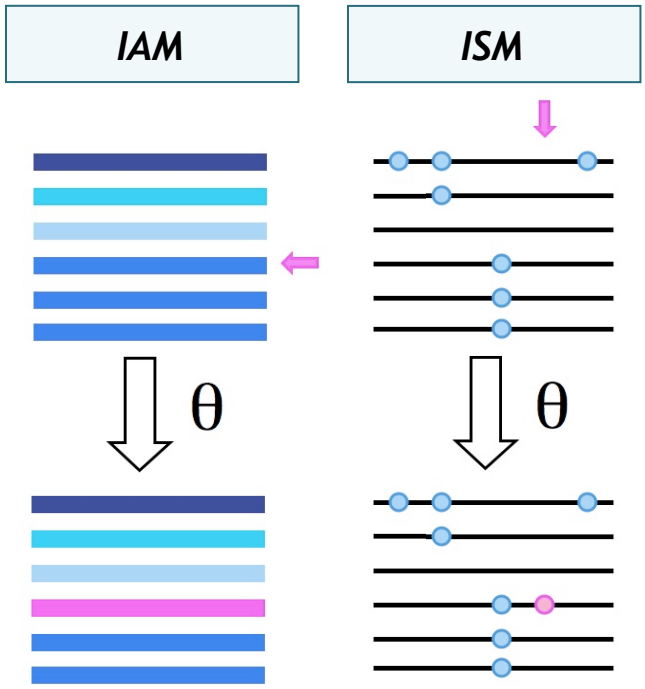

FIG. 1.5 – Modèles du nombre infini d'allèles et de sites

Infinite site model (ISM) : chaque nouvelle mutation apparaît à un taux θ à un site qui n'a été jamais touché par le processus de mutation. Infinite allele model (IAM) : chaque mutation provoque l'apparition d'un nouvel haplotype (appelé ici *allele*).

Infinite sites model (ISM): Selon ce modèle, chaque nouvelle mutation se produit à un nouveau site, jamais touché par le processus de mutation. Ceci implique que chaque position de la séquence a muté au plus une fois, et donc, que chaque site polymorphe est bi-allélique, c'est-à-dire qu'il ne présente que deux allèles, celui de l'état ancestral et celui de l'état dérivé. Lorsque la séquence considérée est très longue (ce qui est le cas des chromosomes humains) et que le taux de mutation est faible, ce modèle constitue une très bonne approximation de la réalité. Ce modèle est l'équivalent du IAM, cependant les séquences sont considérées site par site et non plus de façon globale. Considérer ce modèle implique donc que le même site ne peut muter qu'une seule fois. Ainsi, les méthodes utilisant ce modèle ne permettent pas ce qu'on appelle les *back mutations*[18].

1.2.2.2 Spectres de fréquences

Les spectres de fréquences sont construit pour représenter les patrons de variation génétique dans les séquences (Fig. 1.6). Ces représentations s'inspirent directement des modèles de mutation IAM (pour spectre de fréquences par haplotype) et ISM (pour le spectre de fréquences par site) et se calculent sur des échantillons de séquences provenant de la population étudiée, considérés comme étant représentatifs de celle ci.

Le spectre de fréquences par haplotype se construit en comptant le nombre de fois qu'apparaît chacun des haplotypes présent dans l'échantillon. Dans l'exemple de la figure 1.6, quatre haplotypes sont présent dans l'échantillon, l'un d'eux est représenté trois fois et les autres une seule fois. Le spectre de fréquences par haplotype est représenté par un histogramme des comptes pour chaque haplotype présent dans les donnée, classés du plus fréquent au moins fréquent. La formule d'échantillonnage d'Ewens [Ewens, 1972] nous permet de retrouver la distribution théorique du spectre de fréquences par haplotypes, attendue sous le modèle de mutation IAM. Notons cependant que ceci est vrai uniquement lorsque le type de l'allèle mutant de la

[18]Processus qui permettrait de revenir de l'état dérivé à l'état ancestral par mutation.

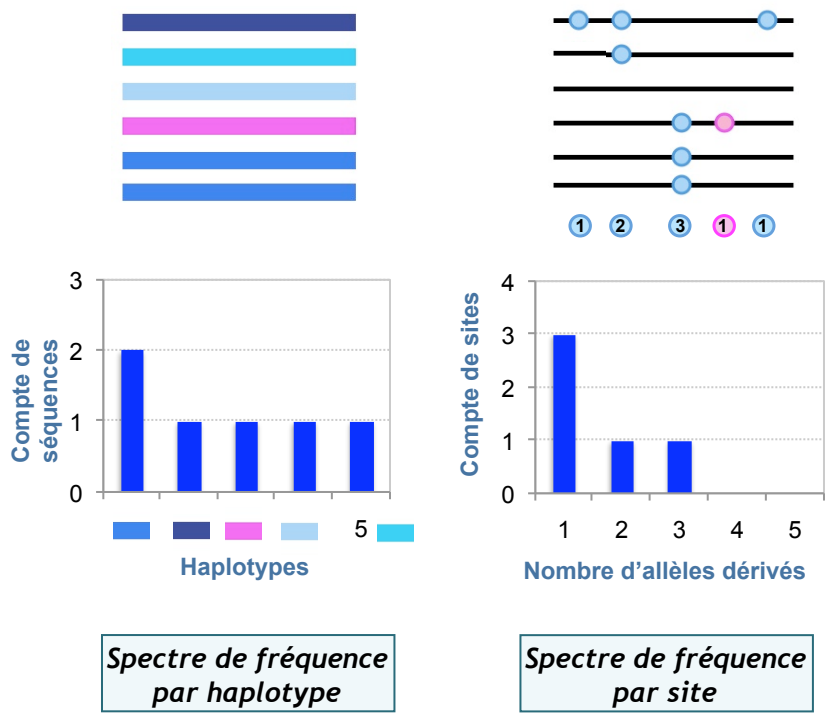

FIG. 1.6 – Spectres de fréquences

Exemple des spectres de fréquences par site et par haplotypes, pour les données de polymorphismes fictives obtenues par le processus de mutation à la figure 1.5, selon les modèles ISM et IAM respectivement. Les petits chiffres en dessous du jeu de données de gauche représentent le nombre d'états dérivés à chacun des cinq sites.

progéniture est indépendant de l'allèle que possède le parent, c'est-à-dire en présence d'un modèle PIM pour *parent-independant mutations*. Le modèle IAM classique est un modèle PIM.

Le spectre de fréquences par site est basé sur les classes de mutations, introduite par Fu [1995] : soit ξ_i le nombre de sites polymorphes dans un échantillon où le type mutant est présent i fois. ξ_i est le nombre de sites présents dans la classe de mutations i. Dans l'exemple de la figure 1.6, on a 5 sites polymorphes dans l'échantillon dont trois sites dans la classe de mutations $i = 1$, un site dans la classe de mutations $i = 2$ et un site dans la classe de mutations $i = 3$. Le spectre de fréquences par site est représenté par un histogramme des comptes ξ_i pour chaque valeur de i. Il a été

montré que

$$\mathbb{E}(\xi_i) = \frac{\theta}{i} \text{ pour } i = 1, ..., n-1 \tag{1.6}$$

où n est le nombre de séquences dans l'échantillon. Ce résultat nous permet d'obtenir la distribution théorique du spectre de fréquences par site [Fu, 1995].

1.2.2.3 Estimateurs du taux de mutation θ

Une fois les modèles mis en place, un grand effort a été fait pour estimer le paramètre de mutation $\theta = 4N\mu$ à partir des échantillons de séquences. Les estimateurs de θ sans biais[19] développés sont principalement des fonctions linéaires de ξ_i. Ci dessous, les estimateurs les plus utilisés dans l'analyse de données de polymorphismes sont listés.

θ_W **[Watterson, 1975]**

$$\theta_W = \frac{1}{a_n} \sum_{i=1}^{n-1} \xi_i \tag{1.7}$$

avec $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$. Notons que $\sum_{i=1}^{n-1} \xi_i = S$ où S est le nombre de sites polymorphes dans l'échantillon. Ainsi, θ_W est uniquement influencé par S.

θ_π **[Tajima, 1983]**

$$\theta_\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i \tag{1.8}$$

θ_π représente le nombre moyen de différences entre deux séquences dans l'échantillon. Il est influencé principalement par les fréquences intermédiaires.

Nous n'avons pas nécessairement besoin de connaître l'état ancestral pour calculer θ_W et θ_π. Cette information est cependant nécessaire pour les estimateurs suivants :

[19] Un estimateur est dit sans biais lorsque son espérance mathématique est égale à la valeur vraie du paramètre.

θ_F **[Fu and Li, 1993]**

$$\theta_F = \xi_1 \tag{1.9}$$

Notons que ξ_1 est le nombre de singleton[20] dans l'échantillon.

θ_H **[Fay and Wu, 2000]**

$$\theta_H = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i^2 \xi_i \tag{1.10}$$

θ_H est pondéré par l'homozygotie des allèles dérivés.

θ_L **[Zeng et al., 2006]**

$$\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\xi_i \tag{1.11}$$

θ_L représente le nombre moyen de mutations sur chaque séquence depuis le MRCA.

Ces estimateurs de θ seront utilisés pour construire des tests de neutralité, développés entre autre pour identifier des locus sous sélection naturelle.

1.2.3 Résultats sur la sélection

Le modèle mathématique le plus simple de sélection basée sur la viabilité (par opposition à la sélection basée sur la fertilité) suppose qu'elle affecte la survie des individus diploïdes entre leur état de zygote[21] et leur état adulte. On suppose que chaque génotype a un *fitness* fixé et spécifique.

Dans le cas d'un locus avec deux allèles, A_1 et A_2, trois génotypes sont possibles : A_1A_1, A_1A_2 et A_2A_2, avec leur *fitness* respectifs w_{11}, w_{12} et w_{22}.

[20] Variant pour lequel une seule séquence de l'échantillon porte l'état dérivé. De même, les variants dont deux ou trois séquences portent l'état dérivé sont appelés doubleton et tripleton.

[21] L'oeuf, cellule provenant de la fusion des deux gamètes.

L'étude de la dynamique génétique des populations s'est faite grâce à deux approches de philosophie différente :

- l'approche classique, prospective, étudiant les populations entières;
- l'approche par coalescence, rétrospective, pour étudier un échantillon observé d'une population donnée.

Ces deux approches utilisent les mêmes modèles de base présentés précédemment.

1.2.3.1 Approche prospective en génétique des populations

L'approche classique prospective cherche à prédire l'évolution du polymorphisme génétique dans une population sous l'influence de différentes forces évolutives. On peut la qualifier aussi d'approche conditionnelle : sachant l'état d'une génération, on cherche à prédire ce qui va se passer dans la prochaine. Elle nous permet de modéliser les phénomènes nécessaires pour déterminer comment l'évolution agit sur la diversité génétique.

Nous supposons le modèle simple de sélection (décrit ci-dessus) pour une population à l'équilibre de Hardy-Weinberg. On suppose qu'à la génération g, il y a i séquences de type A_1 et $2N - i$ séquences de type A_2. Soit $p(g)$ la fréquence allélique de A_1 avant sélection ($p(g) = i/2N$), alors la fréquence allélique de A_1 après sélection sera :

$$\phi(g) = \frac{p(g)[p(g)w_{11} + (1 - p(g)))w_{12}]}{\bar{w}(g)} \tag{1.12}$$

avec $\bar{w}(g) = p(g)^2 w_{11} + 2p(g)(1 - p(g))w_{12} + (1 - p(g))^2 w_{22})$, qui est le *fitness* moyen. Ainsi, la probabilité qu'il y ait j séquences de type A_1 à la génération suivante est :

$$\binom{2N}{j} \phi(g)^j (1 - \phi(g))^{2N-j}. \tag{1.13}$$

Notons que, dans ce processus, le futur des fréquences alléliques ne dépend pas des états passés, mais uniquement de l'état présent. Cette propriété fait de ce processus stochastique un processus de Markov discret (voir le chapitre 1 du livre de

R. Durrett [1999] pour une introduction sur les processus markoviens). L'expression 1.13 est appelée probabilité de transition d'un processus de Markov. Même si cette probabilité de transition paraît simple, il reste difficile d'en déduire de façon exacte des quantités d'intérêt biologique comme, par exemple, le temps attendu jusqu'à fixation de A_1. Le processus de Markov doit donc être approché par un processus de diffusion (pour une introduction sur les processus de diffusion voir le chapitre 15 de Karlin et Taylor [1981]).

1.2.3.2 Approche rétrospective : la coalescence

La plupart des résultats obtenus par approche prospective font l'hypothèse que la population que l'on considère est à un état d'équilibre entre différentes forces évolutives et sont valables au niveau de la population dans son entiereté. Lors des applications empiriques, un problème de coût computationnel se pose avec ce genre d'approche : puisque le matériel que l'on observe en réalité sont des échantillons tirés de la population, il faut donc développer la théorie qui les concerne ce qui demande un travail fastidieux. Supposons que l'on dispose d'un échantillon de 10 séquences parmi une population estimée à 10^6 individus. Avec l'approche prospective, il faudrait d'abord simuler les 10^6 individus, puis en choisir 10 au hasard. Cette approche est peu adéquate pour l'analyse statistique des données observées.

Avec l'approche par coalescence, nous allons échantillonner une généalogie conditionnellement aux 10 individus, ce qui est moins coûteux. On ne va en fait que regarder les relations généalogiques entre les copies d'une séquence dans l'échantillon observé. Pour reprendre l'exemple présenté à la figure 1.4, en regardant le processus du présent vers le passé, on constate que deux lignages peuvent provenir d'une même copie ancestrale : on dit que ces lignages coalescent (Fig. 1.7).

Le modèle de coalescence a été introduit dans les années 1980 [Kingman, 1982]. Il s'agit d'échantillonner une généalogie conditionnellement aux nombres d'individus que l'on observe. La théorie se base sur deux idées fondamentales:

1. le processus généalogique est séparé du processus de mutation neutre.

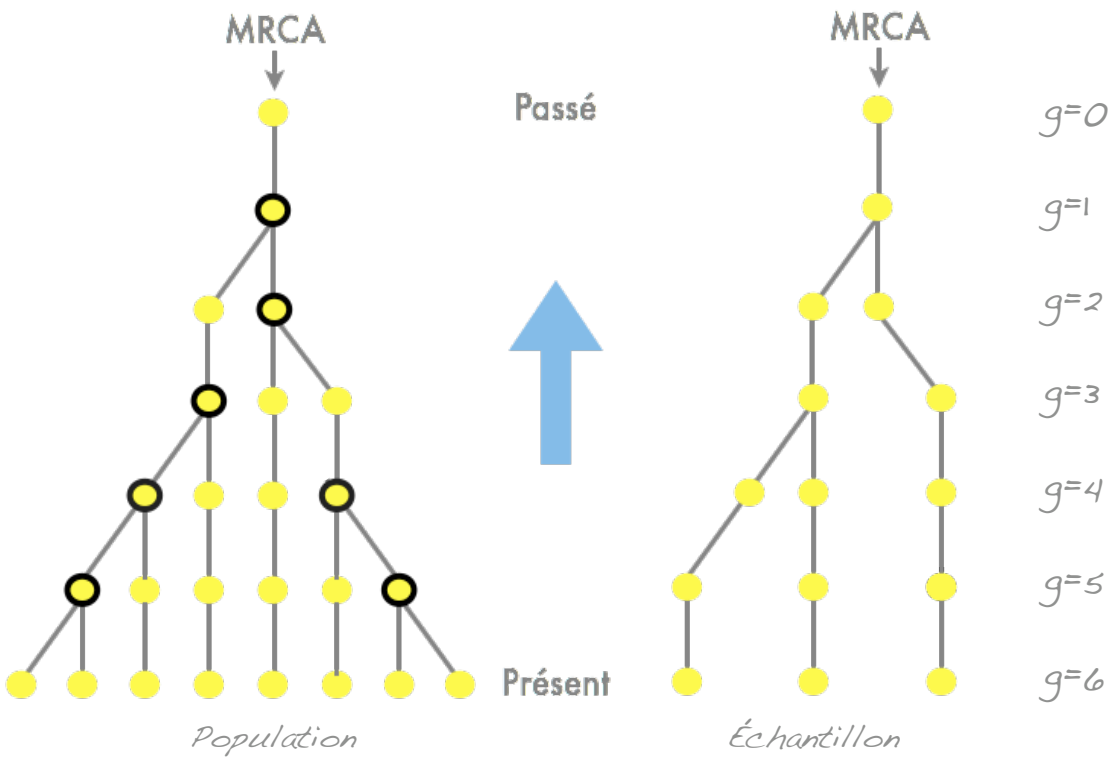

FIG. 1.7 – Processus d'évolution par coalescence

Illustration du processus d'évolution par coalescence, avec 2N = 8, au cours de 7 générations. Chaque événement de coalescence est représenté par un cercle noir dans la généalogie de gauche. Les lignages vont coalescer jusqu'à l'ancêtre commun le plus récent (ou MRCA pour *most recent common ancestor*). Les autres copies de séquences (voir Fig. 1.4), n'existant plus dans la population, ne sont plus représentées. Les relations de coalsescence de l'échantillon observé se résument par un sous-arbre de celui ayant servi à générer la population entière (trois copies de séquences, donc deux événements de coalescence avant le MRCA).

2. on peut modéliser la généalogie d'un groupe d'individus sans s'occuper du reste de la population.

Cette approche nous permet de développer des algorithmes très efficaces et d'utiliser des méthodes statistiques modernes pour comprendre ce que nous observons (voir le chapitre 7 du livre de Balding et collaborateurs [2001]).

Pour respecter la première idée fondamentale, les premiers résultats dérivés de la théorie de coalescence se basent sur l'hypothèse de neutralité sélective. En effet, l'état allélique d'un lignage influence son succès reproducteur, il est impossible de séparer le processus généalogique du processus mutationel. Il a été possible de contourner ce problème de deux façons différentes pour incorporer la sélection dans la théorie de la coalescence. La première approche est une extension élégante du processus de coalescence, connue sous le nom de *ancestral selection graph* [Neuhauser and Krone, 1997] et la deuxième approche est connue sous le nom de *conditional structured coalescent* [Hudson and Kaplan, 1988].

Brièvement, l'idée de l'*ancestral selection graph* est de générer la généalogie de façon rétrospective, en y incluant des événements de ramification (ou *branching*) en plus des événements de coalescence. Ces événements de ramification sont des branches potentielles supplémentaires dans l'arbre de coalescence permettant de tenir compte du succès reproductif qui diffère parmi les individus. La figure 1.8.a explique pourquoi ces ramifications se produisent en présence de sélection. La figure 1.8.b montre un *ancestral selection graph*. Les généalogies peuvent être obtenues à partir de ce graphe : l'ancêtre ultime (UA à la figure 1.8.b) peut être de type A_1 ou A_2. Si A_2 est l'allèle possédant un avantage sélectif, nous allons alors obtenir les généalogies présentées à la figure 1.8.c . Notons que le MRCA ne sera pas toujours l'ancêtre ultime.

Cette approche est utilisée lorsque la sélection est faible : s, l'avantage sélectif de l'allèle sous sélection appelé aussi coefficient de sélection, doit être dans $O(\frac{1}{2N})$ ou plus précisément, $\lim_{N\to\infty} 2Ns = \sigma$. Le paramètre σ est appelé *scaled selection coefficient*.

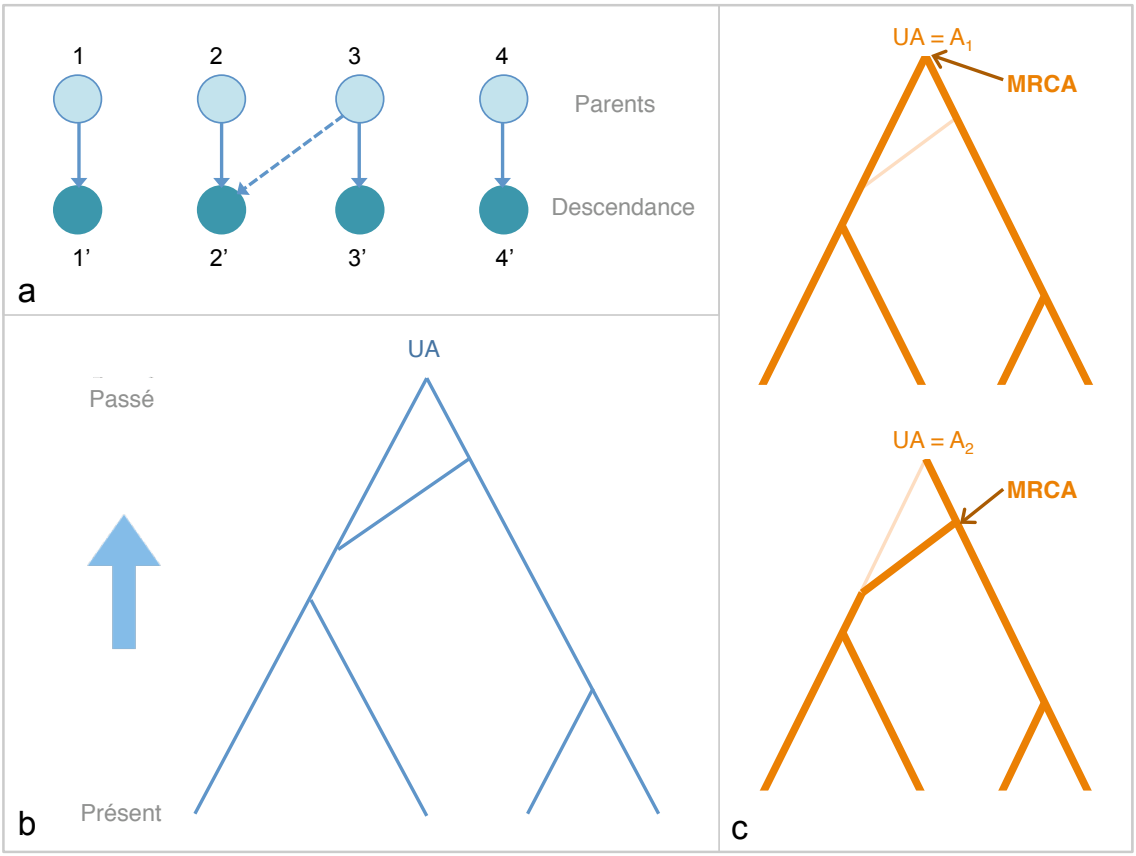

FIG. 1.8 – Ancestral selection graph

(a) Les ramifications proviennent de succès reproductifs différents parmi les individus. Si l'individu 3 possède un avantage sélectif, 2' et 3' proviennent d'un ancêtre commun (flèche pointillée). Sinon, 2 et 3 ont chacun un descendant (flèches pleines) (b) *ancestral selection graph* pour un échantillon de taille 4, avec un événement de ramification et deux événements de coalescence (c) Extraction de généalogies à partir de l'*ancestral selection graph* (source : [Neuhauser and Krone, 1997, Balding et al., 2001])

1.3 L'étude de la sélection positive en génétique des populations

Il est encore difficile de quantifier le rôle de la sélection naturelle dans les variations génétiques observées entre les espèces et populations d'organismes vivants.

Afin de trouver et valider des cibles génétiques de la sélection positive, nous profitons des avancées en génétique des populations, mais aussi en génomique comparative et en biologie moléculaire. Grâce aux nouvelles technologies de séquençage, il est maintenant possible de recueillir de grands jeux de données de séquences d'ADN de plusieurs individus de la même espèce ou population et de génotyper[22] les sites polymorphes. Ceci permet de mettre en évidence les patrons de variation génétique existant au sein des populations ou entre elles. L'abondance de données générées dans la dernière décennie a motivé le développement de méthodes informatiques et mathématiques pour traiter l'information issue de ces données. Ces méthodes permettent entre autres de valider certaines hypothèses sur les mécanismes d'action de la sélection et nous aident à identifier ses cibles de façon de plus en plus fiable. Auparavant, afin de proposer qu'un trait était sous sélection, au moins l'un de ces trois points devaient être vérifiés [Harris and Meyer, 2006] :

- le trait génétique cause une différence dans le taux de fertilité ou de mortalité;
- il y a une différence fonctionnelle entre deux génotypes, qui affecte la capacité de reproduction des individus;
- il existe une concordance géographique entre la distribution du trait et un facteur environnemental qui pourrait être une pression sélective.

Les approches récentes nous permettent maintenant de balayer des génomes entiers à la recherche de signatures moléculaires de sélection, et ainsi d'identifier de nouveaux gènes et régions génomiques à investiguer pour les caractéristiques mentionnées ci-dessus. Par le biais d'un tel travail, nous pouvons identifier des gènes candidats

[22] Déterminer les deux allèles que porte un individu, à un site polymorphe donné dans sa séquence d'ADN

fonctionnellement importants, impliqués d'une part dans l'adaptation de l'humain à son environnement, et d'autre part dans des maladies multifactorielles et complexes dont les causes et le fonctionnement sont encore non-élucidés.

1.3.1 Données

1.3.1.1 Polymorphisme

Lors de la réalisation de ce travail, les données utilisées proviennent de molécules d'ADN de plusieurs individus. Les sites qui diffèrent ponctuellement d'un individu à l'autre sont dit polymorphes (par opposition à des sites dit divergents, lorsqu'ils diffèrent entre les espèces) et sont communément nommés SNPs (*single-nucleotide polymorphisms*). Ces SNPs sont issus du processus de mutation (voir section 1.1.1.1) et peuvent résulter de transitions ou de transversions. Ils peuvent se situer dans les séquences codantes ou non-codantes, dans les régions promotrices, dans les sites d'épissage, etc. Lorsque l'on compare des séquences alignées d'individus d'une même espèce, les SNPs sont les seuls sites informatifs. De ce fait, les données brutes utilisées dans les analyses des séquences d'ADN sont constituées uniquement de SNPs.

Les SNPs ne sont cependant pas les seuls polymorphismes[23] que l'on retrouve. Toute variation provoquée par le processus de mutation génétique, présentée à la section 1.1.1.1, constitue un polymorphisme.

Il existe plusieurs avantages à l'utilisation des SNPs comme données de polymorphismes pour étudier les effets de la sélection naturelle dans le génome humain. Premièrement, les méthodes de génotypage actuelles (section 1.3.1.2) permettent d'obtenir des jeux de données de grande taille très rapidement. De plus, les modèles expliquant l'apparition et la transmission des polymorphismes bi-alléliques ont suscité l'intérêt de recherches mathématiques très poussées (section 1.2). Finalement, pour un grand nombre d'analyses, l'état ancestral du variant doit être connu, ce qui nécessite l'alignement de séquences humaines avec leurs homologues prove-

[23] Variations entre individus dans leur séquence d'ADN.

nant d'une espèce d'un groupe externe, comme le chimpanzé. Pour les séquences d'ADN et les SNPs, cette tâche est, de nos jours, relativement facile à accomplir [Altschul et al., 1990].

1.3.1.2 Séquençage, génotypage et haplotypage

Les données de SNPs sont générées par de nombreux protocoles. La méthode la plus évidente passe par le séquençage complet d'une région génomique chez plusieurs individus. De telles données peuvent être analysées en utilisant des méthodes applicables à des séquences d'ADN complètes comme l'alignement de séquences [Edgar, 2004] des différents individus pour retrouver les sites polymorphes.

Étant donné les coûts (en temps et en argent) que peuvent générer les protocoles de séquençage, la plupart des données de SNPs ne sont pas générées directement par le séquençage de tous les individus de l'échantillon. Les SNPs peuvent être identifiés de deux façons différentes :

- par un balayage des bases de données de fragments génomiques ou de EST[24] à la recherche de sites polymorphes;
- par le séquençage d'une petite proportion des individus de l'échantillon, pour y déterminer les sites polymorphes présents.

Seuls les SNPs identifiés seront alors génotypés chez tous les individus de l'échantillon. Cela permet de générer de grands échantillons aux moindres coûts.

Pour chaque SNP génotypé, nous connaissons alors les deux allèles que chacun des individus portent. Cependant, nous ne savons pas sur quel brin d'ADN ces allèles se situent et à quels autres allèles des sites voisins ils sont liés : les haplotypes exacts nous sont inconnus. La détermination des haplotypes, après génotypage, par les méthodes de laboratoire est une tâche terriblement coûteuse, en argent et en temps. C'est donc à partir des données recueillies que les haplotypes sont inférés *in silico*. Le programme `PHASE` [Stephens et al., 2001] implémente une méthode statistique

[24]Expressed Sequence Tag : séquences de nucléotides transcrites, utilisées pour trouver des gènes dans une séquence ADN.

bayésienne pour reconstruire les haplotypes. `PHASE` est actuellement la méthode la plus précise, mais d'autres, plus rapides, existent [Marchini et al., 2006]. Le fait d'avoir dans l'échantillon des trios familiaux[25] permet d'inférer les haplotypes de façon plus fiable.

La production de ce type de jeux de données peut être faite à petite échelle dans certains laboratoires de recherche, ou à grande échelle dans le cadre de projets regroupant un grand nombre de chercheurs.

1.3.1.3 Le projet international HapMap

Grâce au projet HapMap, il existe des données de SNPs pour tout le génome humain, disponibles publiquement et prêtes à être analysées.

En bref, HapMap se veut un catalogue des variations génétiques les plus fréquentes chez l'humain. On y trouve des informations sur la nature des variants, leur emplacement dans le génome humain et une estimation de leur distribution au sein des différentes populations humaines. Le projet vient en aide aux chercheurs en recherche fondamentale et dans le secteur biomédical, notamment pour leurs travaux visant à comprendre le génome humain et à découvrir de nouveaux gènes, ou variants, impliqués dans des maladies complexes et dans la réponse pharmacologique aux médicaments.

Les populations étudiées sont d'origine africaine, asiatique et européenne et les échantillons d'ADN ont été recueillis auprès de 270 personnes :

- la population africaine est représentée par 30 trios familaux (60 parents et 30 enfants d'âge adulte) provenant du Nigeria : il s'agit de la population des Yoruba d'Ibadan.
- la population européenne est également représentée par 30 trios familiaux. Ces échantillons ont été recueillis auprès de résidents des États-Unis originaires de l'Europe du Nord et de l'Ouest par le Centre d'Étude du Polymorphisme Humain (CEPH).

[25] Les haplotypes de deux parents et de leur enfant d'âge adulte

– Au Japon et en Chine, 45 individus sans lien de parenté de la région de Tokyo et 45 individus sans lien de parenté de la région de Beijing ont été échantillonnés.

Le processus de génotypage (voir section 1.3.1.2) s'est déroulé dans dix centres, en utilisant cinq technologies différentes, mais la qualité des résultats est vérifiée de très près. Toutes les informations relatives au projet international HapMap [Consortium, 2005, Consortium, 2004] peuvent être trouvées sur le site internet : `www.hapmap.org`.

1.3.1.4 Simulations

Les modèles sont grandement utilisés en génétique des populations, afin de comparer les mesures empiriques de paramètres populationnels (taux de mutation, taux de recombinaison, déséquilibre de liaison, coefficient de sélection) aux prédictions faites sous une distribution neutre. Cette distribution doit donc être connue, mais elle est souvent difficile à obtenir analytiquement. L'idée est donc de simuler des échantillons de SNPs en précisant, entre autres, la longueur de la séquence à simuler, l'effectif efficace de la population, le nombre d'individus dans l'échantillon, le taux de mutation, etc. Plusieurs groupes de recherche composent eux mêmes leurs méthodes de simulations de données, mais deux programmes disponibles publiquement ont été grandement exploités afin de réaliser le travail décrit au chapitre 3 : les programmes `ms` et `Selsim`.

`ms` **[Hudson, 2002] :** ce programme permet de générer des échantillons de données de diversité génétique, sous une grande variété de scénarios neutres (sans sélection). Il se base sur un modèle de Wright-Fisher (section 1.2.1) et utilise une approche par coalescence (section 1.2.3.2) pour générer l'histoire généalogique aléatoire d'un échantillon. Il est possible de simuler différents phénomènes évolutifs, tels que les scénarios démographiques, la recombinaison, et la conversion génique[26].

[26] Remplacement d'une séquence d'ADN par une autre par recombinaison.

`Selsim` [**Spencer and Coop, 2004**] **:** cet outil permet de simuler des données de diversité génétique dans lesquelles un des sites mutés est soumis à la sélection naturelle. Il permet de simuler des scénarios de sélection positive et balancée. On peut spécifier le coefficient de sélection et la fréquence de l'allèle sous sélection. Pour la sélection positive, celle si peut être égale à 1 si l'on veut simuler un événement de sélection où l'allèle sélectionné est fixé et il est possible de définir le temps écoulé depuis la fixation.

1.3.2 Méthodes

De façon globale, les méthodes ont pour but de tenter de déterminer, par l'analyse des données de SNPs disponibles, si celles-ci ont été, au moins partiellement, profilées par des événements de sélection adaptative. En pratique, les approches en génétique des populations tentent de répondre à la question suivante : si l'on postule le modèle neutre de la théorie neutraliste de l'évolution [Kimura, 1983], les patrons de variation génétique observés dans les données sont-ils improbables? Si la réponse est affirmative, alors nous sommes éventuellement en présence de séquences cibles de la sélection naturelle.

1.3.2.1 Signatures de sélection adaptative

La sélection positive peut causer plusieurs types de variations génétiques spécifiques dans les séquences d'ADN. On identifie actuellement cinq signatures moléculaires [Sabeti et al., 2006] de sélection adaptative. Chaque signature moléculaire nous permet de détecter des événements de sélection survenus à des temps évolutifs différents.

Une importante proportion de mutations altérant la fonction des protéines Les mutations dans les séquences d'ADN qui altèrent la fonction d'une protéine encodée sont généralement délétères et soumises à la sélection négative. Une sélection positive ancienne peut également augmenter le taux de mutations al-

térant la fonction protéique. On mesure cet effet par la comparaison des séquences d'ADN entre différentes espèces. Cette signature permet d'identifier des événements de sélection survenus il y a plusieurs millions d'années.

Une réduction de la diversité génétique Lors d'un *selective sweep* complet (voir Fig.1.3), l'allèle sélectionné atteint la fixation, entraînant avec lui les allèles qui lui sont génétiquement liés, ce qui élimine la diversité génétique dans le voisinage proche du site sous sélection positive. Ceci laisse une empreinte caractérisée par une faible diversité génétique, avec un excès d'allèles rares (singletons, doubletons, et tripletons par exemple - voir note 20). Cette signature permet d'identifier des événements de sélection entamés il y a moins de 250 000 ans.

La présence d'allèles dérivés à hautes fréquences Après l'apparition de mutations, les allèles dérivés ont normalement des fréquences plus basses que les allèles ancestraux. Dans un *selective sweep* partiel, les allèles dérivés liés génétiquement à l'allèle avantageux sont plus fréquents qu'attendus sous neutralité. Cette signature permet d'identifier des événements de sélection entamés il y a moins de 80 000 ans.

Les différences de variation génétique entre populations Lorsque les populations sont géographiquement séparées et soumises à des environnements différents, la sélection va affecter les patrons de variation dans une seule population. Les différences de variation génétique entre les populations d'une même espèce est donc un signal d'une sélection positive potentielle. Cette signature permet d'identifier des événements de sélection entamés il y a entre 50 000 et 75 000 ans.

De longs haplotypes fréquents À la section 1.1.2.3, la notion de liaison génétique à été introduite. Autour d'un site sélectionné, la recombinaison n'a pas le temps de rompre l'association entre les allèles. La région sous sélection naturelle présente donc un haut taux de déséquilibre de liaison. Ceci va créer un haplotype de

grande taille, fréquent dans la population. Cette signature permet d'identifier des événements de sélection récents, entamés il y a moins de 30 000 ans.

1.3.2.2 Tests de sélection

Sous un modèle neutre (sans sélection), il est possible de faire des prédictions sur certaines relations entre le taux de mutation μ et des paramètres propres de la population étudiée.

Les tests statistiques permettant de détecter les signatures de la sélection adaptative dans les données tirent avantage de ces prédictions, qui sont comparées aux paramètres inférés directement depuis les données empiriques. Il s'agit de tests de neutralité : ils testent un large spectre d'hypothèses pour déterminer si la dérive seule explique la diversité génétique des données. Ces hypothèses caractérisent une population à l'équilibre, se sont les hypothèses de départ de la loi de Hardy-Weinberg (section 1.2.1). L'action de la sélection adaptative dans une population va provoquer une déviation significative du modèle neutre, les fréquences génotypiques observées différeront des attendus théoriques. Il est cependant à noter que, pour plusieurs tests de neutralité, la violation de n'importe quelle hypothèses de départ de la loi de Hardy-Weinberg entraîne le rejet du modèle neutre, même en absence de sélection.

On peut classer les tests de détection de sélection dans les données moléculaires en quatre catégories, reportées ci-dessous [Harris and Meyer, 2006, Nielsen, 2005]. Le tableau 1.2 présente les tests de sélection les plus influents de chacune des catégories.

Tests basés sur les spectres de fréquences des polymorphismes. La sélection positive affecte la distribution des fréquences alléliques et haplotypiques dans une population (Fig. 1.9). Les tests les plus utilisés qui exploitent ce principe se basent sur des statistiques sommaires résumant l'information contenue dans les spectres de fréquences par site et par haplotypes (voir section 1.2.2.2). Ces tests permettent de mettre en évidence les signatures moléculaires causées par la réduction de la diversité génétique et la présence d'allèles dérivés à hautes fréquences

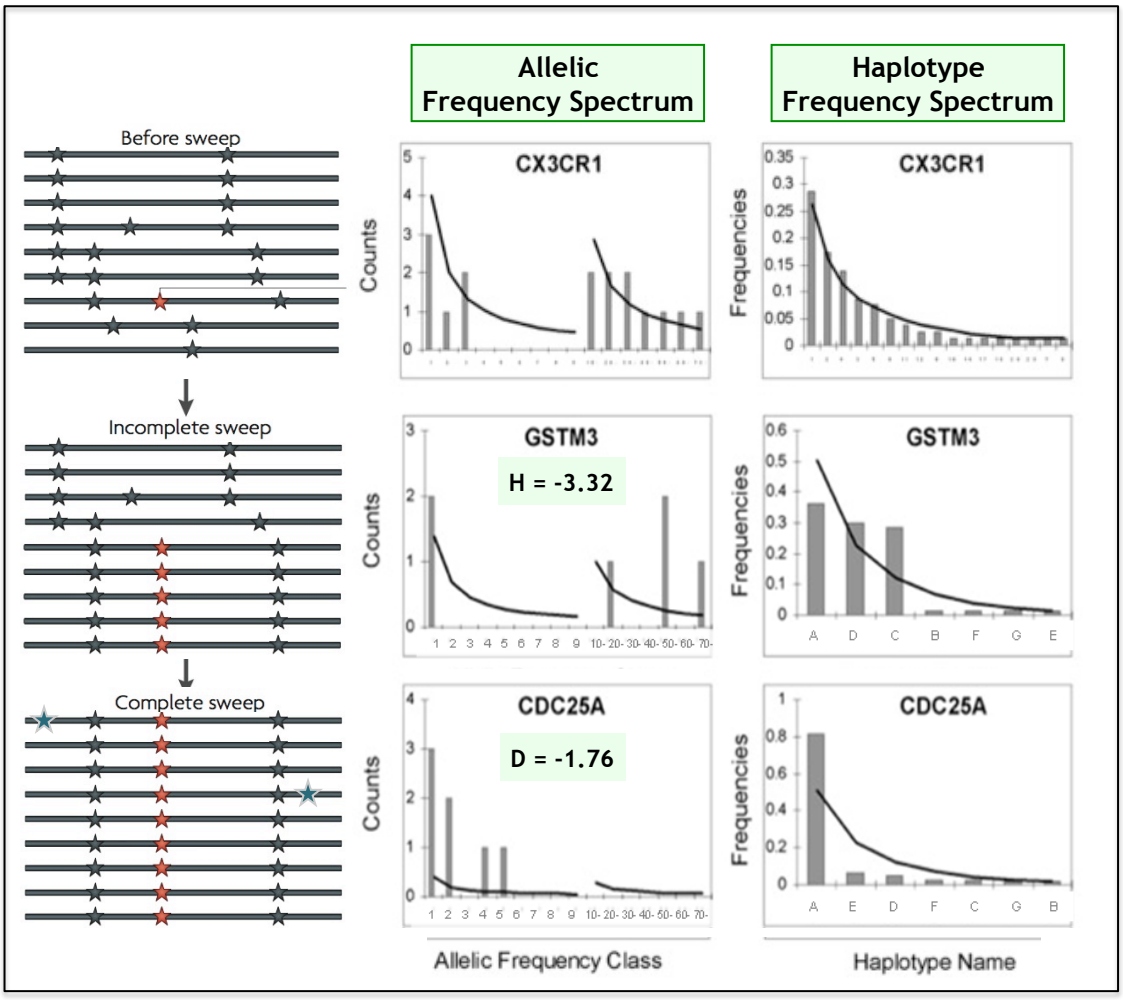

FIG. 1.9 – Selection sweep et spectres de fréquences

Spectres de fréquences alléliques et haplotypiques pour trois segments en amont de sites de transcription de trois gènes. Les spectres pour CX3CR1 sont les spectres typiques attendus sous neutralité, alors que ceux pour les locus GSTM3 et CDC25A présentent une déviation notable du scénario neutre, causée par un *selective sweep* partiel et complet, respectivement. GSTM3 présente une valeur du H de Fay et Wu [2000] hautement négative, due à l'excès d'allèles dérivés à hautes fréquences dans la région étudiée. CDC25A présente une valeur du D de Tajima [1989] hautement négative due à un excès d'allèles rares et d'une réduction de la diversité génétique. Les histogrammes représentent les valeurs observées; les lignes représentent les distributions théoriques attendues sous neutralité. (source : [Labuda et al., 2007]).

(section 1.3.2.1). L'exemple le plus célèbre est le test du D de Tajima [1989]. Ce test compare deux mesures de diversité : le nombre moyen de différences entre deux séquences d'un échantillon et le nombre de sites polymorphes dans cet échantillon. Si la différence de ces deux mesures est plus importante que celle attendue sous neutralité, le modèle neutre est rejeté. Ce test est détaillé à la section 1.3.2.3. Fu et Li [1993] développèrent un test similaire pour tenir compte de l'état des allèles (dérivé ou ancestral). Fu [1997] chercha ensuite à prendre en compte également le nombre d'haplotypes différents dans l'échantillon : le test Fs est basé sur la probabilité de n'avoir pas moins de a_O haplotypes dans l'échantillon étant donné le nombre moyen de différences entre deux séquences de l'échantillon. Le test H de Fay et Wu [2000] cherchent à identifier les signatures laissées par un *selective sweep* partiel et est présenté précisément à la section 1.3.2.3. Ce type de tests sont les tests de neutralité les plus utilisés aujourd'hui.

Tests basés sur les différentes classes de variations entre les espèces et populations. D'après le modèle neutre, il est possible de prédire le niveau de polymorphisme attendu entre populations d'une même espèce et le niveau de divergence[27] attendu entre différentes espèces. En fait, sous neutralité, ces deux paramètres sont proportionnels au taux de mutation. Le test HKA [Hudson et al., 1987] se base sur le calcul du ratio polymorphisme/divergence pour plusieurs régions génomiques. Si les valeurs de ce ratio, pour les différentes régions génomiques, varient plus qu'attendu sous le modèle neutre, alors la neutralité est rejetée. En général, on compare un gène candidat à d'autres locus, considérés comme neutres. Si un allèle avantageux a été fixé dans la population par l'effet de sélection positive, le niveau de polymorphisme du gène porteur est bas, par rapport au niveau de divergence. Ainsi, le gène candidat montrera un excès de divergence par rapport aux locus neutres. Ce test permet de mettre en évidence la signature moléculaire causée par la réduction de la diversité génétique (section 1.3.2.1).

[27]Ici, divergence désigne les différences entre les pools génétiques de deux espèces

McDonald and Kreitman [1991] proposent un test de neutralité basé sur la comparaison de la proportion de substitutions synonymes et non-synonymes. Sous neutralité, que l'on calcule ces proportions pour des individus de la même espèce ou pour des individus d'espèces différentes, les proportions pour ces deux types de substitutions seront les mêmes. Dans une région sous sélection positive, on s'attend à ce que une plus grande proportion de substitutions ait une importance fonctionnelle (soit de type non-synonymes). Cet effet sera plus marqué entre les individus d'espèces différentes. Le test basé sur le ratio dN/dS [Hughes et al., 1988] s'intéresse aussi à la distinction faite entre les mutations synonymes et non-synonymes. Cependant, il tient compte du fait que certains sites ne peuvent engendrer que des substitutions synonymes, et que d'autres n'engendrent que des substitutions non-synonymes. dN est le nombre de substitutions non-synonymes survenues par site non synonyme et dS est le nombre de substitutions synonymes survenues par site synonyme (notons qu'un site peut-être à la fois synonyme et non synonyme). Sous neutralité, dN/dS = 1. Sous sélection positive, dN/dS > 1, car on s'attend à ce que les substitutions non-synonymes soient plus fréquentes. Les tests de McDonald and Kreitman [1991] et dN/dS permettent de mettre en évidence la signature moléculaire qui consiste en une importante proportion de mutations altérant la fonction des protéines (section 1.3.2.1)

Tests basés sur la différenciation entre les populations. La sélection peut, dans un grand nombre de cas, augmenter le degré de différenciation entre les populations. Si un variant apparu par le processus de mutation est sous sélection positive dans une population, sa fréquence n'augmentera que dans cette population particulière. En effet, ce variant ne se retrouvera pas à haute fréquence dans d'autres populations où la pression sélective est absente. Cet effet va augmenter la différenciation entre les populations, signature moléculaire de sélection positive (section 1.3.2.1). Lorsque l'on détecte un niveau particulièrement haut de différenciation à un locus isolé, il est très probable que ce locus soit sous sélection positive. Le tout pre-

TAB. 1.2 – Revue partielle de tests de sélection positive précédemment publiés

Test	Données	Signature	Sél. positive	Référence
Fréquence des polymorphismes				
D	intra-populationelles	(ii)	D très négatif	[Tajima, 1989]
F, F^*D, D^*	intra-populationelles	(ii)	Excès de singletons	[Fu and Li, 1993]
F_s	intra-populationelles	(ii)	Excès de singletons	[Fu, 1997]
H	intra-populationelles et séquence ancestrale	(iii)	H très négatif	[Fay and Wu, 2000]
Différentes classes de variations génétiques				
HKA	intra-populationelles et inter-espèces	(ii)	Excès de divergence au locus candidat	[Hudson et al., 1987]
dN/ds	inter-populationelles et/ou inter-espèces	(i)	$dN/ds>1$	[Hughes et al., 1988]
$McDonald\text{-}Kreitman$	intra-populationelles et inter-espèces	(i)	Excès de divergence non-synonymes	[McDonald and Kreitman, 1991]
différenciation entre les populations				
F_{ST}	inter-populationelles	(iv)	F_{ST} extrème	[Lewontin and Krakauer, 1973]
$rMHH$ et rHH	inter-populationelles	(iv)	$rMHH<0.05$ et $rHH<0.3$	[Kimura et al., 2007]
Diversité des haplotypes et le niveau de déséquilibre de liaison				
EHH	intra-populationelles et séquence ancestrale	(v)	Longs haplotypes avec peu de diversité	[Sabeti et al., 2002]
iHS	intra-populationelles et séquence ancestrale	(v)	$\|iHS\|$ très positif	[Voight et al., 2006]

Signatures : (i) Proportion importante de mutations altérant la fonction (ii) Réduction de la diversité génétique (iii) Présence d'allèles dérivés à hautes fréquences (iv) Différences entre populations (v) Longueur des haplotypes.

mier test de neutralité proposé se base sur cette idée [Lewontin and Krakauer, 1973]. A travers la statistique F_{ST}, la mesure de différenciation entre les populations la plus commune, ce principe est encore grandement exploité. Récemment, Kimura et collaborateurs [2007] ont développé une méthode, reprenant cette idée, pour identifier des *selective sweeps* fixés. La différenciation des haplotypes est mesurée en comparant l'homozygotie entre deux populations, au lieu d'utiliser F_{ST}.

Tests basés sur la diversité des haplotypes et le niveau de déséquilibre de liaison (LD). Quand une mutation est sous sélection positive, sa fréquence augmente très rapidement dans la population, plus rapidement que l'apparition des recombinaisons. De ce fait, les allèles des SNPs autour de l'allèle sélectionné vont également voir leur fréquence augmenter, créant ainsi de long haplotypes très fréquents (section 1.3.2.1). Cela va se traduire par une région étendue où le taux de LD sera élevé (voir section 1.1.2.3). Particulièrement, un *selective sweep* partiel laisse une marque très caractéristique sur la structure haplotypique. Cela a mené au développement de plusieurs méthodes statistiques pour détecter la sélection positive basée sur le LD. La statistique EHH permet de quantifier l'étendue de la diversité d'un haplotype potentiellement porteur d'une mutation sous sélection. Différentes méthodes sont basées sur cette statistique, comme le test LRH (pour *Long-Range-Haplotype*) [Sabeti et al., 2002] et la statistique iHS (pour *integrated Haplotype Score*) [Voight et al., 2006]. Cette dernière statistique est détaillée à la section 1.3.2.3.

1.3.2.3 Détection des signatures moléculaires

Dans ce qui suit, trois des tests présentés au tableau 1.2 sont détaillés. Ces tests sont ceux qui ont inspirés le développement de notre méthode de détection de la sélection.

Les tests basés sur le spectre de fréquences par site prennent avantage des dis-

tinctions entre les différents types d'estimateurs de θ. Chaque estimateur capte un type d'information particulier que contiennent les données. Le D de Tajima [1989] permet de détecter un *selective sweep* après fixation, alors que, grâce au H de Fay et Wu [2000], des *selective sweeps* en cours sont détectés. Un *selective sweep* en cours va également déformer le spectre de fréquences par haplotypes. La variation génétique étant réduite dans l'entourage du site sous sélection, il y création de longs haplotypes présentant une faible diversité entre les individus. C'est cet effet que le test iHS [Voight et al., 2006] va tenter de capter.

1.3.2.3.a Le D de Tajima [Tajima, 1989]

Fumio Tajima a proposé une statistique basée sur la différence normalisée de θ_W et θ_π (voir section 1.2.2.3) pour résumer l'effet d'un *selective sweep* complet sur le spectre de fréquences :

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{\mathbb{V}(\theta_\pi - \theta_W)}} \tag{1.14}$$

Sous neutralité, les estimateurs θ_W et θ_π présentent essentiellement la même valeur, correspondant au taux de mutation θ unique, on s'attend donc à ce que D = 0. On peut obtenir la distribution de D sous l'hypothèse nulle de neutralité en simulant des jeux de données par ordinateur. Cette distribution neutre est utilisée pour déterminer si une valeur de D observée dans les données est significativement différente des valeurs attendues.

Le D de Tajima, permet d'évaluer si il y a plus d'allèles rares qu'attendu par rapport au nombre de sites polymorphes. θ_π sera plus faible que le taux de mutation réel, car les variants à basses fréquences (i.e les allèles rares) vont très peu contribuer au nombre moyen de différences entre les paires de séquences de l'échantillon. Cependant, θ_W n'est pas sensible aux variation de fréquences des SNPs, il sera ainsi supérieur à θ_π. On obtiendra une valeur de D inférieure à 0, nous indiquant un événement de sélection positive.

Notons tout de même que, dans certains cas, des événements de sélection balancée

vont nous amener à une valeur positive du D de Tajima. Il y a un maintien des deux allèles à un même site sélectionné : le nombre moyen de différences entre séquences sera proportionnellement supérieur à la mesure de diversité basée sur le nombre de site polymorphes ($\theta_\pi - \theta_W > 0$).

1.3.2.3.b Le H de Fay et Wu [Fay and Wu, 2000]

Le principal problème du D de Tajima lorsque l'on cherche à détecter des événements de sélection positive est que, sous sélection négative, la valeur de D sera aussi inférieure à 0. En effet, les nouveaux allèles sous sélection négative réduisent le *fitness* des individus porteurs et la fréquence de ces allèles n'augmentera quasiment pas, créant ainsi un excès de variants à basses fréquences.

Pour détecter des événements de sélection positive, il convient donc de se baser sur une autre signature moléculaire : l'excès d'allèles dérivés à hautes fréquences, avant la fixation de l'allèle sélectionné. L'idée est de comparer la mesure de diversité basée sur le nombre moyen de différences entre séquences (θ_π) avec une mesure sensible aux allèles dérivés à hautes fréquences (θ_H).

Fay et Wu ont proposé une statistique basée sur la différence entre θ_W et θ_H pour résumer l'effet d'un *selective sweep* partiel sur le spectre de fréquences :

$$H = \theta_\pi - \theta_H \tag{1.15}$$

Pour normaliser cette statistique comme cela a été fait pour le D de Tajima, la variance est requise. Cependant, la variance de $\theta_\pi - \theta_H$ n'est pas facile à obtenir. Zeng et collaborateurs [2006] montrent que $\theta_H = 2\theta_L - \theta_\pi$ et ainsi, H = $2(\theta_\pi - \theta_L)$. La variance de $\theta_\pi - \theta_L$ peut-être calculée plus facilement, nous permettant d'utiliser la version normalisée de H :

$$H_{norm} = \frac{\theta_\pi - \theta_L}{\sqrt{\mathbb{V}(\theta_\pi - \theta_L)}} \tag{1.16}$$

Comme pour le D de Tajima, sous neutralité, on s'attend à ce que H = 0. Quand un *selective sweep* se produit, l'excès d'allèles dérivés à hautes fréquences donne lieu

à un H très négatif (ce qui n'est pas le cas sous sélection négative). Ainsi, les tests de Tajima et de Fay et Wu se complètent. Un test conjoint, le test DH, à d'ailleurs été proposé [Zeng et al., 2006] et n'est sensible qu'à la sélection positive.

1.3.2.3.c Le test iHS [Voight et al., 2006]

Le test iHS est basé sur la statistique EHH (pour *Extended Haplotype Homozigosity*) [Sabeti et al., 2002]. Cette statistique a été développée pour détecter si un haplotype particulier est sous sélection positive. De part et d'autre de cet haplotype central, un ensemble de SNPs est considéré. EHH permet de calculer le degré de LD (section 1.1.2.3) entre l'haplotype central et les SNPs situés à différentes distances physiques. EHH est la probabilité que deux chromosomes, choisis au hasard et qui portent l'haplotype central, soient identiques, de l'haplotype central jusqu'à un SNP donné. Plus ce SNP sera éloigné de l'haplotype, plus les valeurs de EHH vont décroître. Ces valeurs seront plus élevées autour d'un haplotype sous sélection positive qu'autour d'un haplotype neutre.

Le test iHS (pour *integrated Haplotype Score*) [Voight et al., 2006] a été développée pour utiliser EHH sur de larges jeux de données. L'haplotype central est réduit à un seul SNP, qu'on appelle le site central. On mesure séparément la décroissance de EHH le long des haplotypes porteurs de l'allèle dérivé et de l'allèle ancestral au site central. Si les variants ancestraux et dérivés sont présents sur des haplotypes de taille similaire, le ratio de ces mesures sera égal à 1. Le ratio s'éloigne de 1 lorsque l'un des allèles présente un EHH qui décroît plus rapidement que celui de l'autre allèle. Sous sélection positive récente, l'allèle dérivé se trouve sur un haplotype plus long que l'allèle ancestral: des valeurs extrêmes de iHS vont être obtenues.

L'approche utilisant iHS ne consiste pas à rejeter ou à accepter une hypothèse nulle, en fonction des données. Le but de cette approche est de détecter les régions présentant des valeurs atypiques, qualifiées d'*outliers*, en balayant les jeux de données. Ce type de balayage consiste à faire coulisser une fenêtre d'une taille fixe et de considéré toujours un site central différent.

1.3.2.4 Difficultés

Malgré les progrès faits pour détecter les signatures de sélection positive dans les données génomiques, plusieurs difficultés se posent. Les biais peuvent être causés par des paramètres intrinsèques à la population étudiée ou par les méthodes expérimentales de génération de données.

1.3.2.4.a L'histoire démographique des populations

Sous le modèle neutre, la taille de la population est considérée comme constante. En fait, il est important que la population soit de taille constante depuis un assez grand laps de temps, pour que les événements démographiques du passé ne laissent plus d'empreinte sur les données.

Les événements démographiques compliquent la détection de signaux de sélection positive à partir des données : en effet, certains événements de l'histoire démographique d'une espèce peuvent créer des patrons dans les données, que les méthodes ne différencient pas des signatures de sélection. Par exemple, une réduction soudaine de la taille d'une population, un *bottleneck*, cause la perte des allèles à basses fréquences, produisant ainsi un excès d'allèles dérivés à hautes fréquences, comme peut le faire un événement de sélection positive en cours. Une expansion démographique provoque une augmentation de la proportion des allèles à basses fréquences, comme lors d'un *selective sweep* complet.

Les tests basés sur le spectre de fréquences sont grandement affecté par ce problème. Les déviations observées par rapport à la neutralité peuvent être attribuées soit à des événements de sélection, soit à des changements dans la taille de la population. De plus, puisque la signature de sélection elle même est sensible à certains événements démographiques, ceux-ci annulent les signatures moléculaires de la sélection, et sont donc responsables d'un nombre important de faux négatifs, pour lesquels des régions sous sélection sont perçues comme neutres.

Une première idée est de prendre en compte l'histoire démographique des populations dans le modèle neutre et ainsi, de ne pas supposer que la population se trouve

à l'équilibre. Pour le D de Tajima [1989] et le H de Fay et Wu [2000], les distributions « neutres » peuvent être simulées en introduisant des scénarios démographiques spécifiques à la population étudiée. Cependant, cela nécessite des informations précises sur les événements démographiques de la population en question, ce qui n'est pas toujours évident.

Une autre approche consiste à comparer les locus entre eux. On s'attend à ce que les scénarios démographiques affectent l'entièreté des variations génétiques du génome, alors que la sélection naturelle n'affecte que la région génomique concernée. Pour cette raison, les jeux de données de génomes complets (par exemple, les données du projet HapMap, section 1.3.1.3) vont nous permettre de discriminer les signaux « démographie » des signaux « sélection ».

1.3.2.4.b Biais expérimentaux

Comme mentionné précédemment et principalement pour des raisons financières et méthodologiques, la génération de jeux de données de SNPs est rarement faite via un re-séquençage complet de tous les individus (voir 1.3.1.2). Cependant, à cause des faibles tailles d'échantillons, la probabilité d'identifier un SNP par balayage de base de données est proportionnelle aux fréquences alléliques. Ceci implique que les SNPs présentant un allèle rare ont une faible probabilité d'être découverts, par rapport aux SNPs dits communs. Les SNPs présentant un allèle à très basse fréquence seront sous-représentés dans l'échantillon. Cela cause, entre autre, une déformation importante du spectre de fréquences. De ce fait, tous les tests basés directement sur le spectre de fréquences (i.e. le D de Tajima, le H de Fay et Wu, F_{ST}, ...) seront biaisés par le biais de recrutement (ou *ascertainment bias*) [Nielsen et al., 2004, Clark et al., 2005]. Les méthodes d'inférence des haplotypes (comme `PHASE` [Stephens et al., 2001]) peuvent amener des biais d'haplotypage. En effet, il n'y a pas de méthodes bio-informatique entièrement fiable pour inférer les haplotypes depuis des génotypes diploïdes [Andrés et al., 2007]. Un large taux d'erreur

est en effet constaté, même avec les algorithmes les plus performants, particulièrement lorsque l'on dispose de longues séquences (en pb) avec peu de SNPs, génotypés dans un petit nombre d'individus. De plus, la présence de SNPs présentant un allèle rare (dont la fréquence est inférieure à 0,1 par exemple) est un facteur affectant grandement la reconstruction des haplotypes. Les scénarios démographiques, mais aussi l'action de la sélection peuvent affecter négativement la fiabilité d'inférence des haplotypes. Comme expliqué précédemment (section 1.3.2.1), un *selective sweep* complet, comme une expansion démographique, laisse une empreinte caractérisée par une faible diversité génétique, avec un excès d'allèles rares.

1.4 L'intolérance au lactose, un trait sous sélection positive

Le fait que la sélection positive laisse des empreintes caractéristiques sur nos chromosomes est aujourd'hui une évidence pour les généticiens. L'une des empreintes les plus claires trouvées à ce jour se situe sur le gène qui exprime l'enzyme de lactase. Ce locus présente, au niveau moléculaire, un signal de sélection positive considéré par les chercheurs comme « l'un des signaux de sélection positive récente les plus forts documenté à ce jour dans le génome humain » [Bersaglieri et al., 2004].

1.4.1 Consommation de lait chez l'humain

La lactase-phlorizin hydrolase est une enzyme digestive intestinale. Il s'agit de la protéine responsable de la bonne digestion du lactose, le glucide contenu dans le lait (2 à 8% des éléments constituant le lait). Son rôle est d'hydrolyser le lactose en glucose et en galactose (Figure 1.10), molécules qui peuvent ensuite être absorbées par l'organisme au niveau de l'intestin grêle.

Chez les mammifères, le lait est la principale source d'alimentation des nouveaux nés. Cependant, en général, l'activité de la lactase diminue une fois la phase d'allaitement terminée : les individus perdent alors peu à peu leur capacité à digérer

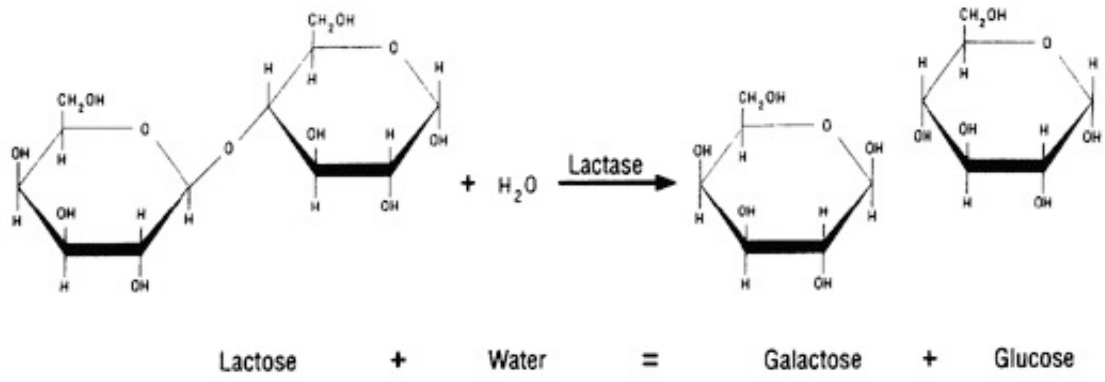

FIG. 1.10 – Mode d'action de la lactase. source : [Keeton and Gould, 1996]
L'enzyme de lactase va rompre le disaccharide en deux monosaccharides : le galactose et le glucose

le lactose. Ce déclin d'activité est aussi une caractéristique propre à la plupart des humains. On fait communément référence à ce phénotype comme l'alactasie ou l'intolérance au lactose. L'alactasie peut être partielle ou totale. Cela se produit lorsque le lactose passe dans l'intestin sans avoir été digéré. Sa présence provoque des troubles gastro-intestinaux : ballonnements, diarrhées, crampes, douleurs abdominales, etc. Chez d'autres individus, l'enzyme de lactase reste active après le sevrage : ces individus sont dits lactase-persitants. Les individus lactase-persitants peuvent consommer de grande quantité de lait et produits laitiers sans complication.

La fréquence de l'alactasie varie considérablement entre les populations humaines : la figure 1.11 montre la distribution géographique approximative des humains lactase-persitants à travers le monde. Dans les détails les patrons de distribution de ce trait sont complexes et font encore l'objet d'études épidémiologiques. Dans les populations du nord de l'Europe, les individus lactase-persitants sont très fréquents (au delà de 90% des individus dans certaines populations scandinaves). Plus l'on se déplace vers le sud et l'est, plus ces fréquences baissent. Le phénotype lactase-persitant est généralement absent des populations asiatiques et africaines, bien que certaines études récentes montrent que le trait est présent à des fréquences inter-

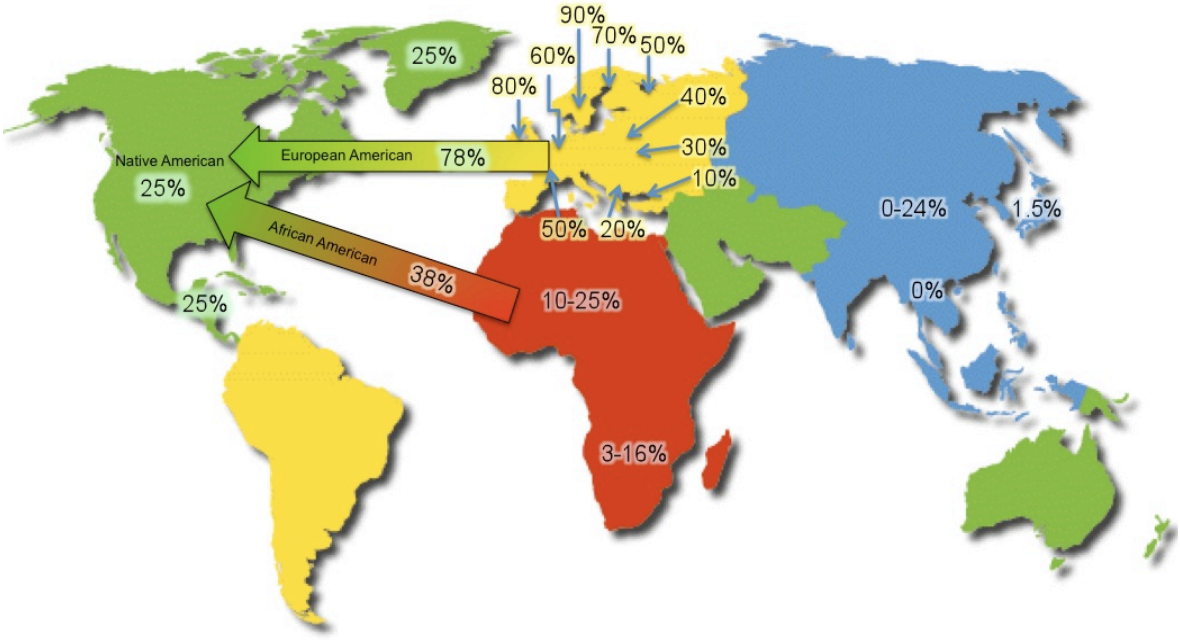

FIG. 1.11 – Distribution de la tolérance au lactose dans les populations humaines. Dans les régions de haute tolérance au lactose (comme dans le nord de l'Europe), les populations consomment en général beaucoup de lait et produits laitiers. Dans les régions de basse tolérance au lactose (en Asie, par exemple), les gens ne consomment pas de produits laitiers à l'âge adulte. source : [Harris and Meyer, 2006, Rozin and Pelchat, 1988]

médiaires dans quelques populations du nord de l'Afrique et du Moyen-Orient. La culture alimentaire de certains peuples africains est basée sur les produits laitiers, comme pour les Peuls du Cameroon, les Wolofs du Sénégal et les Tutsi du Rwanda, du Congo et d'Uganda. Dans ces populations, la proportion d'individus lactase-persistants est plus élevée que celle d'autres populations habitant les mêmes pays [Mulcare et al., 2004, Harris and Meyer, 2006].

On sait actuellement que le lait a eu une importance capitale dans la survie des humains, lors de leur arrivée dans certaines régions du monde. La distribution atypique du trait ainsi que la relation entre cette distribution, les habitudes culturelles des populations et la distribution géographique des élevages laitiers suggèrent que la sélection positive est responsable de la persistance de l'enzyme de lactase. Différentes hypothèses sélectives ont été proposées [Holden and Mace, 1997, Mace et al., 2003]. Une des théories stipule que les individus pouvant boire du lait à l'âge adulte bénéficieraient d'un avantage nutritif important pour leur survie dans certaines régions du monde.

1.4.2 Génétique de l'intolérance au lactose

Aujourd'hui, nous savons que l'alactasie a une base génétique et qu'il ne s'agit pas d'un trait résultant de l'influence d'un substrat[28]. La persistance de l'enzyme de lactase à l'âge adulte est un trait héréditaire autosomal dû à l'existence d'un allèle dominant LCT*P. On trouve des individus homozygotes lactase-persistants (LCT*P/LCT*P), homozygotes récessifs souffrant d'alactasie (LCT*R/LCT*R) et hétérozygotes (LCT*P /LCT*R) qui ont des niveaux intermédiaires d'activité enzymatique de la lactase.

L'enzyme de lactase est constituée d'un peu moins de 2000 acides aminés et est encodée sur un gène (LCT) de 17 exons sur 50 Kb, situé sur le chromosome 2 [Boll et al., 1991, Swallow, 2003]. Quatre haplotypes communs de 11 SNPs ont été

[28]Une molécule utilisée comme produit de départ dans une réaction chimique catalysée par une enzyme.

trouvés dans le monde entier [Hollox et al., 2001] et il a été montré que la persistance de la lactase est fortement associée à un seul de ces haplotypes [Swallow, 2003]. Des études familiales de déséquilibre de liaison et d'association génétique proposent qu'un variant C → T [Enattah et al., 2002], 13910 nucléotides en amont du gène de lactase, pourrait être le polymorphisme responsable de la persistance de la lactase dans les populations européennes. Des études *in vitro* montrent que ce variant T fonctionnent tel un *cis*-élément[29]. Par un effet régulateur positif sur la région promotrice du gène codant pour la lactase, l'allèle en augmente la transcription [Lewinsky et al., 2005]. D'autre part, dans des populations finlandaise, un variant G → A, 22018 nucléotides en amont du gène LCT, montre aussi un haut niveau d'association avec le trait de persistance de la lactase. Les variants A-22018 et T-13910 sont situés dans le minichromosome de maintenance 6 (MCM6) dans l'intron 9 et 13, respectivement.

Ces variants sont cependant absents ou extrêmement rares dans la plupart des populations africaines, y compris celles où les individus lactase-persistants sont nombreux. Dans ces groupes, la persistance de la lactase est très probablement causée par un autre variant. Cette idée a été récemment confirmée par deux études de populations africaines et moyen-orientales [Tishkoff et al., 2007, Enattah et al., 2008]. Quatre variants, tous situés dans MCM6, démontrent une association avec le phénotype de persistance de la lactase :

- C → G (-13907, intron 13 de MCM6) dans la population Beja du nord du Soudan.
- T → G (-13915, intron 13 de MCM6) dans les populations du Kenya et d'Arabie Saoudite
- T → G (-14010, intron 13 de MCM6) dans les populations du Kenya et de Tanzanie
- T → G (-3712, intron 17 de MCM6) dans la population d'Arabie Saoudite.

Cela signifie que la persistance de la lactase est apparue indépendamment dans les populations européennes et africaines.

[29]Région d'ADN ou ARN qui régule l'expression de gènes localisés sur le même brin

1.4.3 Lactase et sélection positive

Deux signatures moléculaires de sélection positive ont été trouvées dans les populations européennes : un haut degré de différenciation entre populations et une taille anormalement longue de l'haplotype associé (voir section 1.3.2.1 pour la description de ces signatures). Une étude a comparé 101 SNPs de la région du gène LCT, provenant de 3 populations (européenne, afro-américaine et asiatique de l'est) [Bersaglieri et al., 2004]. L'analyse des données grâce à la statistique F_{ST} a montré que ce locus est sous sélection positive dans la population européenne. Une autre étude [Voight et al., 2006] confirme ce résultat grâce à l'analyse génomique des données HapMap en utilisant le test de sélection iHS. Finalement, une autre analyse de données [Harris and Meyer, 2006] utilise la statistique EHH [Sabeti et al., 2002] et apporte une troisième preuve de l'existence d'un *selective sweep* en cours dans la population européenne pour cette région génomique. Aucun test basé sur les spectres de fréquences à conclut positivement à la présence d'un *selective sweep* dans cette région.

L'intolérance au lactose est l'un des traits sous sélection naturelle les plus étudiés de nos jours. Ceci a motivé l'utilisation de ce locus comme système modèle pour valider notre nouvelle méthode de détection de la sélection positive récente, présentée dans la suite de ce mémoire. Retrouver d'autres locus avec les mêmes propriétés que ce système modèle nous permettrait de mettre en évidence d'autres gènes candidats ayant permis à l'Homme moderne de s'adapter à son environnement et d'être ce qu'il est actuellement.

Chapitre 2

Méthodologie

La méthodologie utilisée est statistique et bio-informatique. Ainsi toutes les analyses ont été faites *in silico*. Nous avons développé un nouveau test de neutralité pour détecter la sélection positive. Nous avons programmé des outils liés à ce nouveau test, nécessaires à l'analyse de données biologiques informatisées et disponibles dans les bases de données publiques. Pour une étude complète de la nouvelle statistique Svd, de son développement à sa validation, deux approches méthodologiques ont été utilisées : une approche par simulation et une approche empirique.

Tous les outils présentés sont codés en JAVA. Les outils sont disponibles sur `www.iro.umontreal.ca/~hussinju/Svdtools.html`)

2.1 Développement du test statistique

Nous avons choisi de focaliser notre attention, dans ce projet, uniquement sur la détection de la sélection positive récente et en cours le long de séquence d'ADN, où l'allèle sous sélection est dominant dans la population mais pas encore fixé (sa fréquence est supérieure à 0.5 et inférieur à 1). Les empreintes moléculaires laissées par ce type de sélection sont causées par un *selective sweep* partiel. Ces signatures se traduisent, au niveau du locus sous sélection, par la présence d'allèles dérivés à hautes fréquences et de longs haplotypes fréquents dans la population (voir section

1.3.2.1 pour les détails sur ces signatures).

Lorsque de telles pressions sélectives poussent un allèle avantageux à se retrouver à haute fréquence dans une population, la variation réduite aux alentours de ce site modifie significativement les spectres de fréquences par haplotype et par site, par rapport à ce qui est attendu sous neutralité. Chaque spectre capte un certain type d'information issue des données : le spectre de fréquences par site considère les données site par site, alors que le spectre de fréquences par haplotype les considère séquence par séquence. Nous pensons que le fait de combiner ces deux types d'information augmenteraient la détection de *selective sweeps* partiels dans les données génomiques.

2.1.1 Les Classes Alléliques d'Haplotypes (HAC)

Les classes alléliques d'haplotypes (ou HAC pour *Haplotype Allelic Classes*) constituent un nouvelle manière de considérer les données de polymorphismes en génétique des populations. Elles ont été introduites par Labuda et collaborateurs [2007] pour étudier les patrons de variation génétique présents dans des segments génomiques situés en amont de sites de transcription. Les HAC constituent le point de départ méthodologique du projet reporté dans ce mémoire.

Comme le nom l'indique, les classes alléliques d'haplotypes consistent à classer les différentes séquences d'un échantillon en différentes classes d'haplotypes. On définit d'abord un haplotype de référence avec lequel seront comparées les séquences de l'échantillon. La classe allélique k regroupe toutes les séquences qui sont a une distance k de l'haplotype de référence. Dans ce contexte, la distance correspond au nombre de différences entre une séquence et l'haplotype de référence. Notons que l'haplotype de référence n'existe pas nécessairement dans l'échantillon. La figure 2.1 présente un exemple de construction graphique des HAC à partir d'un jeu de données fictif, avec deux haplotypes de référence distincts : l'haplotype porteur de tous les allèles ancestraux et l'haplotype porteur de tous les allèles majeurs

Dans l'analyse de diversité faite par Labuda et collaborateurs [2007], l'haplotype

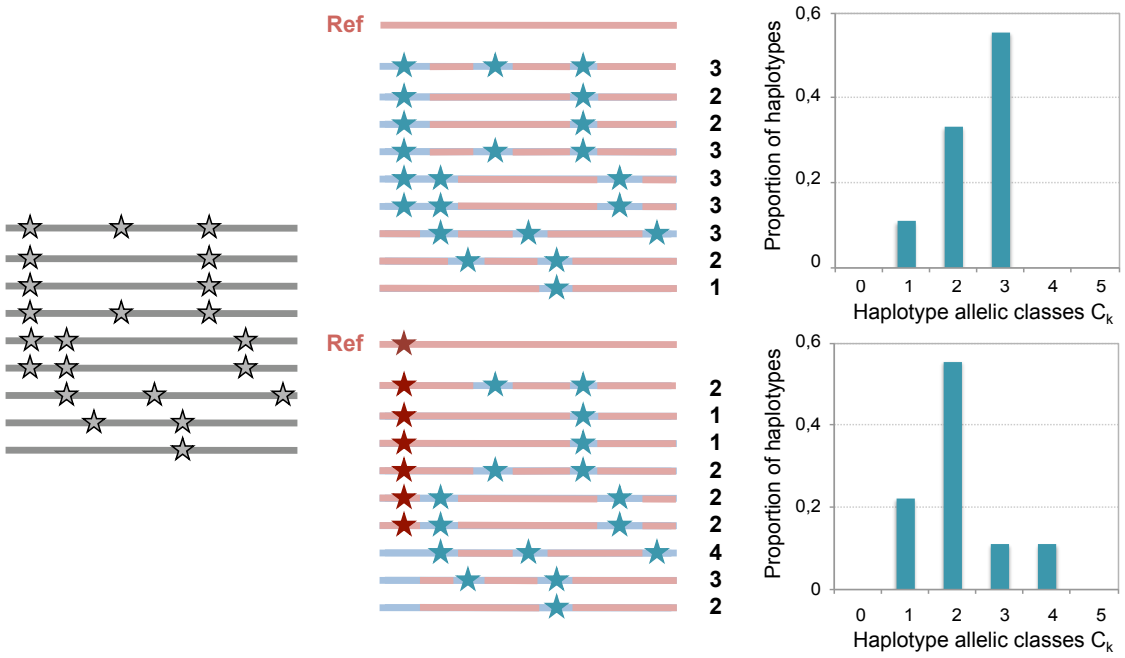

FIG. 2.1 – Calcul des HAC avec deux types d'haplotype de référence

Sur le jeu de données fictif à gauche, les étoiles représentent les mutations (allèles dérivés). Les HAC dépendent de l'haplotype de référence que l'on choisit. Pour un même jeu de données, nous présentons deux représentations graphiques des HAC, construite avec deux haplotypes de référence distincts : l'haplotype composé uniquement des états ancestraux et l'haplotype composé uniquement des états majeurs. Au centre de la figure, on présente en bleu les différences avec chaque haplotype de référence, qui sont comptabilisées à droite de chaque séquence. À droite, les HAC sont représentées à l'aide d'un histogramme : on compte le nombre de séquences qui ont 0, 1, 2, etc. différences avec l'haplotype de référence.

de référence formé des allèles ancestraux a été utilisé pour calculer les HAC. Cependant, lorsque le but est de détecter des régions génomiques sous sélection positive, l'haplotype formé par les allèles majeurs est mieux adapté. Lorsque la variabilité génétique d'un locus est causée par un *selective sweep* partiel, l'haplotype formé de tous les allèles majeurs est probablement très proche de l'haplotype porteur de la mutation sous sélection.

2.1.2 La statistique Svd

Les statistiques comme le D de Tajima [1989] ou le H de Fay et Wu [2000] sont des statistiques sommaires qui résument l'information contenue dans les spectres de fréquences. Nous cherchons à réaliser, de la même façon, une statistique sommaire qui résumeraient les effets captés par les HAC. Dans un premier temps, nous avons testé plusieurs statistiques susceptibles de résumer l'informations des HAC.

La première statistique sur laquelle nous avons travaillé s'appelle Sv2 :

$$Sv2 = \frac{1}{n}\sum_{k=0}^{S} k^2 C_k \tag{2.1}$$

avec n le nombre de séquences dans l'échantillon, S le nombre de SNPs par séquence, k la distance à l'haplotype de référence et C_k le nombre de séquences à une distance k de l'haplotype de référence (i.e le nombre de séquences dans la classe allélique k). On peut remarquer que Sv2 est la moyenne empirique au carré du nombre de différences entre les séquences de l'échantillon et l'haplotype de référence. Cette statistique pourrait donc être utilisée comme estimateur de l'espérance du nombre de différences. Cependant, des tests préliminaires nous ont permis de conclure que, pour la détection de *selective sweeps* partiels, cette statistique ne s'avère pas très puissante.

La stratégie choisie a donc été de s'appuyer sur l'idée de la statistique iHS : évaluer séparément la variabilité génétique autours des deux allèles d'un SNP. La diversité n'est réduite que pour les haplotypes portant l'allèle sélectionné, les autres

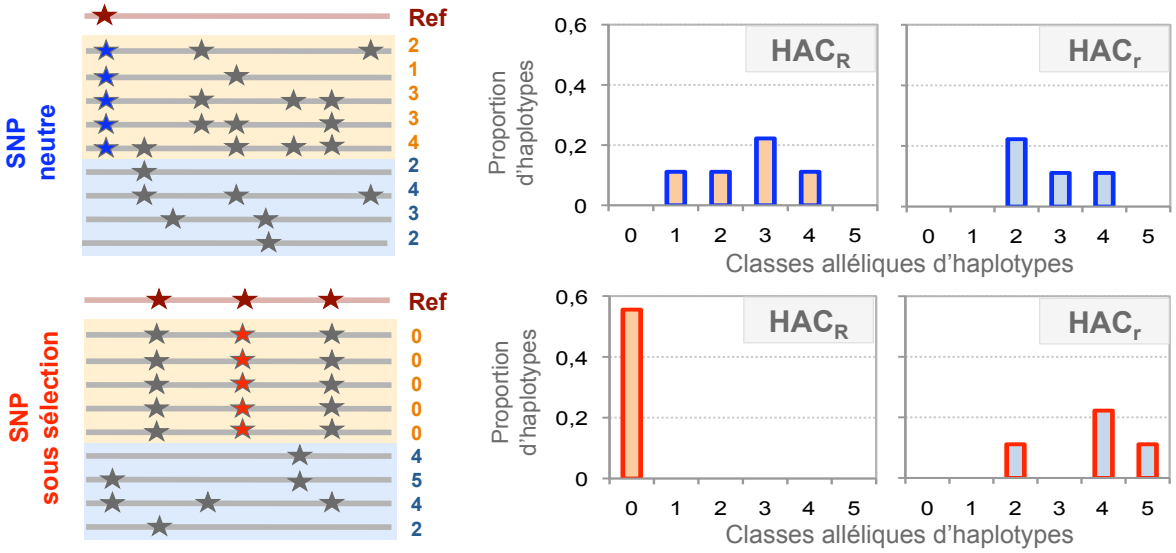

FIG. 2.2 – Séparation des données pour calculer les distributions des HAC.
Le SNP coloré en bleu est un SNP neutre, alors que le SNP en rouge est sous sélection positive. Le nombre de différences entre la séquence et l'haplotype de référence est indiqué à droite de chaque séquence des jeux de données fictifs. Les jeux de données sont ensuite séparés en deux sous-échantillons, selon les deux états des SNPs colorés : l'état R en jaune pour l'état (majeur) présent sur l'haplotype de référence et l'état r en bleu pour l'autre état (mineur). Pour chaque jeu de données, les distributions des HAC pour les sous-échantillons sont compilées séparément et représentées par deux histogrammes.

présentant une diversité semblable à celle attendue sous neutralité. On va donc regarder séparément les distributions des HAC pour deux sous-échantillons : celui contenant les haplotypes porteurs de l'allèle majeur (celui sur l'haplotype de référence) et celui contenant les haplotypes porteurs de l'allèle mineur.

En probabilité et statistiques, la variance est une mesure permettant de caractériser la dispersion d'une distribution. En effet, la variance est zéro si toutes les données sont identiques et augmente lorsque les données se diversifient. Même si les distributions des HAC pour deux allèles d'un SNP neutre seront différentes à cause de la variabilité neutre dans l'échantillon, la dispersion, et donc la variance, des distributions seront similaires. Pour un SNP sous sélection positive, la variance des HAC pour les haplotypes porteurs de l'allèle mineur sera comparable aux variances calculées sous neutralité. Cependant, la distribution des HAC pour les haplotypes

porteurs de l'allèle sélectionné majeur sera plus étroite et la variance sera plus petite, car les haplotypes seront moins variables.

Basée sur cette séparation, la statistique Svd a été développée. L'idée est d'évaluer la différence entre les variances empiriques v des HAC pour les deux sous-échantillons. Une définiton formelle de la statistique et des explications détaillées se trouvent au Chapitre 3 (section 3.2). Svd, à chaque SNP d'un jeu de données, se calcule de la façon suivante :

$$Svd_i = \frac{v(HAC_{r,i}) - v(HAC_{R,i})) \cdot f_{d,i}}{S} \tag{2.2}$$

avec

$$v(HAC) = \frac{1}{n}(\sum_{k=0}^{S} k^2 C_k - (\sum_{k=0}^{S} kC_k)^2). \tag{2.3}$$

$f_{d,i}$ est la fréquence de l'allèle dérivé au SNP i, $HAC_{R,i}$ est la distribution des HAC pour les haplotypes porteurs de l'allèle sur l'haplotype de référence (l'allèle majeur) au SNP i et $HAC_{r,i}$ est la distribution des HAC pour les haplotypes porteurs de l'autre allèle (l'allèle mineur) au SNP i.

La statistique Svd peut être utilisée comme variable de décision d'un test statistique de neutralité avec :

- une hypothèse nulle H_0 : le site est sous neutralité sélective
- une hypothèse alternative H_1 : le site est sous sélection positive et les patrons de diversité génétique qui l'entourent sont causés par un *selective sweep* partiel.

Sous neutralité, on s'attend à ce que la différence des variances théoriques soit proche de zéro, ou négative, et qu'elles soient très positives sous sélection positive. Ces différences seront estimées grâce aux variances empiriques des HAC (voir eq. 2.2 et 2.3). Lorsque les deux sous échantillons seront de tailles semblable, elle se rapprochera de zéro. Sinon, $v(HAC_{r,i})$ sera plus petite que $v(HAC_{R,i})$ car elle est calculée pour moins d'haplotypes, produisant une valeur de Svd négative. Pour un

SNP sous sélection positive et les sites génétiquement liés, la distribution des HAC étant étroite, $v(HAC_{R,i})$ sera petite et les valeurs de Svd seront très positives.

Nous allons donc pouvoir définir une valeur critique de la statistique Svd, de sorte que, lorsque Svd>c, H_0 soit rejetée. Il s'agit d'un test d'hypothèse unilatéral, car la région de rejet est entièrement située à l'extrémité positive de la distribution d'échantillonnage.

2.2 Approche par simulation

Les simulations sont utilisées dans un premier temps pour valider l'utilisation des HAC dans l'étude de la sélection puis pour le développement de la statistique Svd. Elles vont aussi nous permettre de calculer les valeurs critiques du test pour la détection d'un *selective sweep* partiel basé sur Svd.

Pour simuler des échantillons de données en génétique des populations, différents programmes existent. Deux d'entre eux ont été décrits à la section 1.3.1.4 et sont utilisés dans ce travail car nous n'avons pas programmer de générateur de jeux de données. Les outils développés permettent d'analyser les jeux de données simulés dans le cadre précis de l'analyse des indices de sélection, des HAC et de Svd.

L'approche par simulation s'est faite en deux phases. Premièrement, la phase de génération des données grâce aux programmes `ms` et `Selsim` et outils supplémentaires et deuxièmement, la phase d'analyse des données simulées.

2.2.1 Génération des données simulées

2.2.1.1 Simulation des données

Trois groupes d'ensembles de jeux de données ont été simulés. Chaque ensemble est une suite de R jeux de données simulés, appelés réplicats. Chaque réplicat est composé de $2n$ lignes de S caractères, où $2n$ est le nombre de séquences simulés (n étant le nombre d'individus haploïdes) et S est le nombre de SNPs par séquence. Chaque SNP présente deux allèles : le caractère 0 représente l'état ancestral et le

caractère 1 représente l'état dérivé, après mutation. Tous les réplicats d'un même ensemble sont simulés suivant les mêmes paramètres.

Tous les ensembles contiennent R=1000 réplicats de $2n$=50 séquences de 100 Kb, sur lesquels les mutations sont distribués aléatoirement selon un taux de mutation $\theta/Kb = 2.23$, pour avoir, en moyenne 1000 SNPs par séquence sous neutralité. En effet, d'après l'estimateur du taux de mutation θ_W,

$$\theta \simeq \frac{S}{\sum_{i=1}^{n-1} 1/i} = \frac{1000}{\sum_{i=1}^{50-1} 1/i} \simeq 2.23/Kb.$$

Le premier groupe de données simulées est constitué de 6 ensembles de réplicats, présenté au tableau 2.1. Ils nous permettront d'évaluer l'impact de facteurs démographiques, de la recombinaison et de la sélection sur les HAC, sur Svd et sur d'autres tests de sélection, présentés à la section 1.3.2.3. Dans le scénario « Neutralité », il n'y a pas de recombinaison, pas de sélection et la taille effective de la population est constante. A cause de limitations techniques intrinsèques aux logiciels de simulations, nous avons fixé l'effectif efficace de la population à 1000, bien quelle soit plutôt de 3000, chez les populations européens et asiatiques, à 7500 chez la population africaine [Tenesa et al., 2007]. Le bottleneck simulé correspond à une réduction sévère de 95% de la population, à une époque correspondant au Néolithique, le début de l'agriculture [Excoffier and Schneider, 1999]. Une expansion récente de la population, qui la fait doubler de taille, est également simulée. Les simulations des scénarios démographiques ont été faites par Philippe Nadeau dans le cadre de son projet de maîtrise en cours. Des réplicats ont été simulés avec un taux de recombinaison de $\rho/\text{Kb} = 1$, pour avoir $\rho \simeq \theta/2$, ce qui est le cas en moyenne chez l'humain. Deux scénarios sous sélection positive de force modérée ($s = 0.15$) ont été simulés. Le premier reproduit les effets d'un *selective sweep* partiel, où l'allèle sous sélection a une fréquence de $f = 0.75$. Le second reproduit les effets d'un *selective sweep* complet, ce qui signifie que l'allèle sous sélection est fixé et donc $f = 1$.

Le deuxième groupe de données simulées est constitué de 14 ensembles de répli-

TAB. 2.1 – Paramètres de simulations sous différents scénarios

Les paramètres présentés dans cette table sont ceux qui varient entre les scénarios.

Les paramètres communs à tous les scénarios sont précisés dans le texte.

Scenario	**Effectif efficace N_e**	ρ/Kb	**Paramètres de sélection**		**Outil**
			coefficent s	freq. du site f	
Neutralité	Constante sur 4000 gen. : 1000	0	Pas de sélection		`ms`
Bottleneck	Gen. 0 à 260 : 1000 Gen. 260 à 340 : 50 Gen. 340 à 4000 : 1000	0	Pas de sélection		`ms`
Expansion	Gen. 0 à 3700 : 500 Gen. 3700 à 4000 : 1000	0	Pas de sélection		`ms`
Recombinaison	Constante sur 4000 gen. : 1000	1	Pas de sélection		`ms`
Selective Sweep partiel	Constante sur 4000 gen. : 1000	0	0.15	0.75	`Selsim`
Selective Sweep complet	Constante sur 4000 gen. : 1000	0	0.15	1	`Selsim`

TAB. 2.2 – Paramètres de simulations sous des scénarios de *selective sweep* partiel
Les paramètres présentés dans cette table sont ceux qui varient entre les scénarios.
Les paramètres communs à tous les scénarios sont précisés dans le texte.

Scenario	**Effectif efficace N_e**	**Taux ρ/Kb**	**ρ/Kb dans un Hotspot**
Défaut	1000	0	sans Hotspot
Petite N_e	500	0	sans Hotspot
Grande N_e	2000	0	sans Hotspot
Faible taux de recombinaison	1000	1	sans Hotspot
Haut taux de recombinaison	1000	5	sans Hotspot
Hotspot x10	1000	1	10
Hotspot x100	1000	1	100

cats, dont 7 ensembles sont simulés sous neutralité ($s = 0$) et 7 autres, identiques, sont simulés sous sélection modérée ($s = 0.15$). Les 7 différents scénarios sont présentés au tableau 2.2. Ils nous permettront d'évaluer l'impact de l'effectif efficace et de la recombinaison sur le pouvoir de détection du test basé sur Svd (voir section 2.2.5). L'effet des recombinaisons est étudié avec des taux constants de recombinaison plus ou moins élevés et des hotspots de recombinaison de différentes intensités. Dans tous les cas, le site central de chaque séquence a une fréquence de $f = 0.75$. Ce site est celui sous sélection dans les ensembles de réplicats où $s = 0.15$.

Le dernier groupe de données simulées est constitué de 20 ensembles de réplicats, conçus pour évaluer l'impact des deux paramètres du modèle de *selective sweep* partiel : la fréquence f du site sélectionné et le coefficient de sélection s. Pour chacune des cinq valeurs prises par f (f=[0.6, 0.7, 0.75, 0.8, 0.9]), on simule quatre réplicats avec des valeurs de s différentes (s =[0, 0.05, 0.15, 0.5]). Les autres paramètres utilisés pour ces simulations sont ceux du scénario de sélection par défaut, présenté à la première ligne du tableau 2.2.

2.2.1.2 Modifications des données simulées

Avant d'être utilisés par les outils d'analyse, certains ensembles de données simulées présentés à la section précédentes sont modifiés. Ils vont être transformés afin de reproduire des biais expérimentaux que l'on retrouve dans les données expérimentales provenant de séquence d'ADN humaines, tel que le biais de recrutement et le biais causé par l'haplotypage par `PHASE` (biais présentés à la section 1.3.2.4.b).

Les deux modules suivants implémentent les méthodes pour recréer le biais de recrutement :

`msAscSim`. Ce module implémente la procédure décrite par Nielsen et ses collaborateurs [Nielsen et al., 2004] pour recréer le biais de recrutement. Pour chaque réplicat, nous choisissons au hasard un sous-ensemble des 50 séquences simulés. Nous ne retenons, dans les 50 séquences, que les SNPs qui sont polymorphes dans ce sous-ensemble de séquences. Les nouveaux jeux de données ne contiendront donc qu'un sous-ensemble des SNPs présents dans les jeux originaux.

`classFreq`. Ce module nous permet de retirer les sites de certaines classes de fréquences du spectre de fréquences par site (voir section 1.2.2.2). Lors du reséquençage en laboratoire, des erreurs peuvent survenir de temps à autre et certains laboratoires enlèvent de leurs données les sites dont le MAF (pour fréquence de l'allèle mineur) est inférieure à 0.05. Comme pour `msAscSim`, les nouveaux jeux de données ne contiendront donc qu'un sous-ensemble des SNPs présents dans les jeux originaux.

Pour recréer le biais d'haplotypage instauré par des logiciels comme `PHASE`, la procédure est la suivante :

- les haplotypes sont regroupés aléatoirement deux à deux afin de constituer des individus diploïdes ;
- les deux haplotypes d'un individu sont re-transformés en génotypes : pour chaque SNP, on connaît les deux allèles que porte l'individu mais on ne sait plus sur lequel des chromosomes ils se trouvent et à quels allèles des sites

voisins ils sont liés ;

– à l'aide de `fastPHASE`, on résout les génotypes en haplotypes.

`fastPHASE` est plus rapide que `PHASE`, ce qui est pratique lorsque l'on effectue cette procédure sur des ensembles de 1000 réplicats. La précision de cette méthode est similaire à celle de `PHASE 2.0` [Scheet and Stephens, 2006].

Finalement, pour uniformiser la longueur des haplotypes considérés dans nos analyses, nous utilisons le module `msCut`. Il permet d'obtenir des ensembles de données simulées dans lesquels le nombre de SNPs est le même pour tous les réplicats. Il convient de choisir un site sur lequel seront centrées toutes les séquences de chaque réplicat. `msCut` s'applique à des fichiers directement issus de `ms` ou `Selsim`, ou sur des fichiers modifiés par `msAscSim`, par `classFreq` ou par la procédure d'haplotypage par `fastPHASE`. L'utilisation de ce module va nous permettre de tester l'impact de la longueur de l'haplotype considéré lors du calcul des statistiques sur le pouvoir de détection de celles ci (voir section 2.2.5). Dans nos analyses, `msCut` a été utilisé pour produire des haplotypes de 25, 50 et 200 SNPs.

2.2.2 Analyse des données simulées

Les données simulées vont nous permettre de valider le potentiel des HAC dans l'étude de la sélection et de comparer notre statistique Svd à d'autre tests de sélection. On va également se servir de données simulées pour évaluer le pouvoir de détection des différents tests de sélection et particulièrement de Svd.

2.2.3 Les outils

Voici quelques outils qui nous permettront de réaliser ces tâches:

`msCalSel.` Ce module permet de traiter les réplicats simulés par `ms` ou `Selsim` et implémente les méthodes de calcul :

– des spectres de fréquences par site et par haplotypes;

– des HAC avec les deux haplotypes de référence décrits à la section 2.1.1 (l'haplotype porteur des allèles ancestraux ou des allèles majeurs);

– des valeurs de θ_W, θ_π, θ_H, θ_L;
– du D de Tajima [1989], du H de Fay et Wu [2000], de la version normalisée de H [Zeng et al., 2006].

`msToiHS.` Ce module permet de traiter les réplicats simulés par `ms` ou `Selsim` et implémente les méthodes de calcul de la statistique iHS et iHS *unstandardized*. Les méthodes utilisées nous ont été acheminées par Rachel Myers par le biais de Philip Awadalla et ont été codées par Jon Keebler.

`SelDiff.` Ce module permet de calculer Svd sur les réplicats simulés par `ms` ou `Selsim`. Dans un même fichier, tous les réplicats doivent présenter le même nombre de SNPs, ils auront donc été préalablement traiter par `msCut`. À chaque SNP, le module sépare chaque réplicat en deux sous-échantillons pour calculer Svd, comme expliqué à la section 2.1.2. Pour chaque réplicat, on obtient autant de valeurs qu'il y a de sites.

`Compile.` Ce module permet de calculer la distribution de chacune des statistiques lorsque l'on a les valeurs observées pour chaque réplicat (calculées par `msCalSel`, `msToiHS` ou `SelDiff`). Une fonction en escalier permet une bonne estimation des distributions d'échantillonage. Le module nous permettra ainsi de construire un histogramme dont la taille des classes est constante et spécifiée et de classer les valeurs observées selon leur rang, en ordre croissant.

2.2.4 Distributions sous différents scénarios

2.2.4.1 Distribution des HAC

La distribution moyenne des HAC sous différents scénarios est obtenue grâce aux données simulées analysées par `msCalSel`. La figure 2.3 montre ces distributions moyennes lorsque l'haplotype de référence est celui porteur des allèles majeurs. Les distributions pour les deux scénarios de sélection positive différent grandement de la distribution attendue sous neutralité. Particulièrement, la distribution obtenue sous un *selective sweep* partiel est beaucoup plus étroite et proche de zéro que

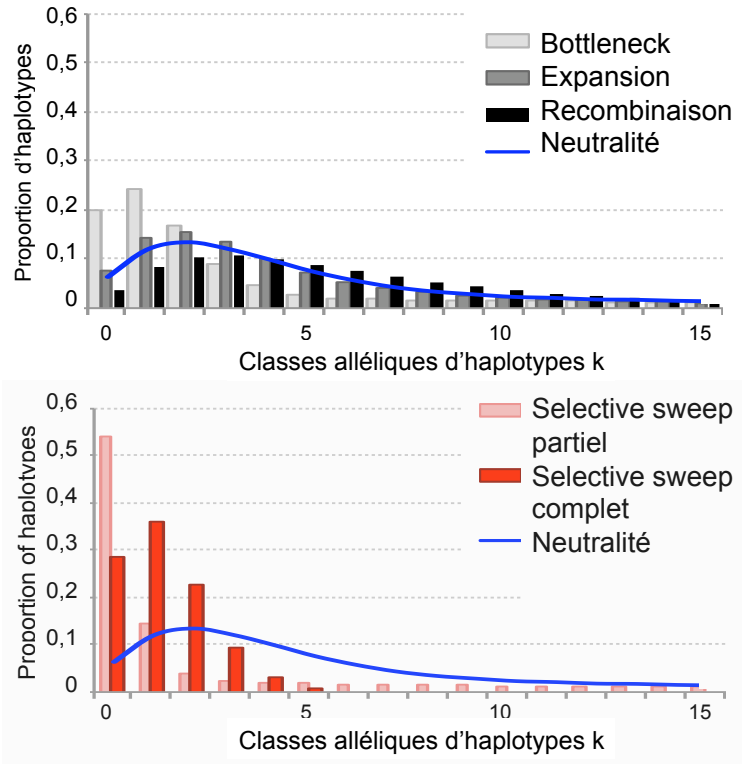

FIG. 2.3 – Effets de différents scenarios sur les HAC

Les distributions pour les différents scénarios sont calculées à partir des données simulées du premier groupe dont les paramètres sont décrits au tableau 2.1 et dans la section 2.2.1.1.

celle attendue sous neutralité. Cela traduit la capacité des HAC à capter ce type d'événement de sélection et a motivé le développement d'un test pour les détecter.

2.2.4.2 Distribution des statistiques de sélection

Les statistiques sont calculées par `msCalSel` pour D et H, `msToiHS` pour iHS et `Seldiff` pour Svd. Pour Svd et iHS, seules les valeurs pour le SNP central de chaque réplicat sont retenues. Les distributions sont obtenues grâce à l'outil `Compile`. Ces distributions sont présentées aux figures 3.1 et 3.7.

2.2.5 Pouvoir de détection

Pour vérifier que la statistique présente des valeurs significativement différentes lorsqu'elle est calculée sur des réplicats simulés sous le modèle neutre (coefficient

de sélection $s = 0$) et sous un modèle de sélection positive (coefficient de sélection $s > 0$), nous calculons le pouvoir de détection de la statistique. Le pouvoir de détection est la sensibilité du test, c'est à dire, la probabilité de rejeter H_0 lorsque H_0 est fausse. Ici, il s'agit de la probabilité que $Svd > c$ quand le réplicat est simulé sous l'effet d'un *selective sweep* partiel, avec c la valeur critique du test.

Comment déterminer la valeur critique c ?

On va définir c tel que P(Svd>c sous le modèle neutre) = 0.05. Ceci signifie que l'on obtiendra le pouvoir de détection de Svd à p=0.05. Cette procédure peut-être représentée graphiquement, comme ce qui est montré à la figure 2.4. On compare l'ordre des valeurs obtenues pour les réplicats simulés sous sélection avec l'ordre de celles obtenues à partir de réplicats simulés avec des paramètres identiques mais sous neutralité (coefficient de sélection $s = 0$). Dans l'exemple de la figure 2.4, 95% des réplicats simulés sous neutralité donnent une valeur de Svd inférieure à c=0.44, alors que seulement 19% des réplicats simulés sous sélection donnent une valeur de Svd inférieure à c. Donc, en acceptant 5% de faux positifs, on a un pouvoir de détection 81%. Nous avons procéder ainsi pour obtenir les résultats présentés au tableau 3.1 aux figures 3.2 et 3.3.

Cette même approche est aussi utilisée pour assigner des p-valeurs aux valeurs observées. La p-valeur est la probabilité d'obtenir au hasard un résultat au moins aussi extrême que la valeur réellement observée, sous l'hypothèse nulle. Dans l'exemple de la figure 2.4, on peut supposer que les valeurs utilisées pour construire la courbe rouge sont des valeurs observées auxquelles l'on veut attribué des p-valeurs grâce à la distribution nulle, représentée en bleu. À la valeur de 0.44, on attribuerait donc la p-valeur de 0.05.

2.3 Approche empirique

2.3.1 Traitement des données des bases de données publiques

L'information sur les données utilisées pour réaliser ce projet est détaillée au paragraphe **SNPs data information** de la section 3.5. Il convient simplement de préciser que les modules suivants ont été programmés pour préparer les analyses des

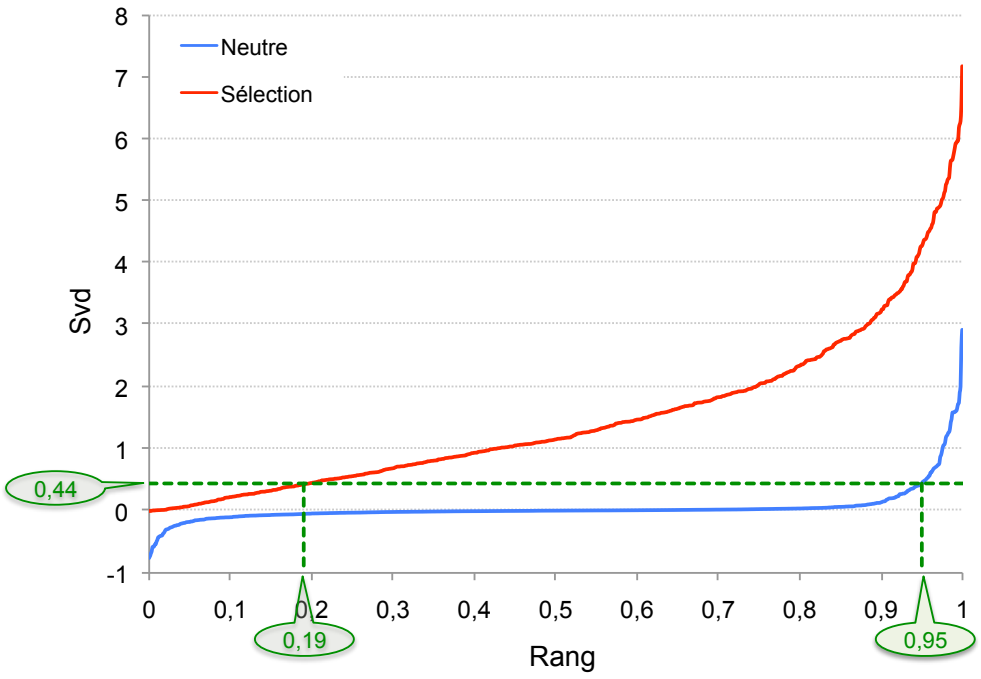

FIG. 2.4 – Détermination du pouvoir de détection de la sélection de Svd

Deux ensembles de 1000 réplicats de 50 séquences ont été générés, l'un sous neutralité et l'autre sous l'effet d'un *selective sweep* partiel. La fréquence de l'allèle dérivé du site central est de 0.75 pour tous les réplicats. Pour les réplicats neutres, le coefficient de sélection est de $s = 0$ alors qu'il est de $s = 0.15$ pour les réplicats simulés sous sélection. Les valeurs de Svd pour le site central de chaque réplicat ont été calculées, en utilisant les haplotypes formés par 25 sites de part et d'autre du site sélectionné (haplotypes de 51 SNPs en tout). Ces valeurs sont classées en ordre croissant (rangs reporté entre 0 et 1). (Paramètres de simulations : $\theta/Kb = 2.23$, effectif efficace $N_e = 1000$, taux de recombinaison $\rho/Kb = 0$).

données extraites des bases de données publiques.

`BiomartToPhaseOut` Ce module nettoie les données HapMap de génotypage issue de Biomart (`http://hapmart.hapmap.org/BioMart/martview`). Ce module va s'assurer que les données sont « propres », c'est-à-dire qu'il n'y a pas de site non polymorphe et pas de SNP avec plus de deux allèles. Un site qui ne serait pas polymorphes sera retiré des données. Un site présentant plus de deux allèles sera transformé en un site bi-allélique : seul les deux états les plus fréquents seront gardés. Les autres états passeront à l'état de l'allèle le moins fréquent des deux allèles conservés.

`SetAncestral` Ce module nous permet d'obtenir les allèles ancestraux pour chaque position du génome humain présentée dans les données HapMap de la base de données Biomart. Cela est fait à partir des allèles que présentent les génomes du chimpanzé ou du macaque, issues de la banque de données génomique de UCSC (`http://genome.ucsc.edu/cgi-bin/hgTables`). Si aucun allèle n'est connu ni pour le chimpanzé ni pour le macaque, ou si ceux-ci ne correspondent à aucun des deux allèles humains, c'est l'allèle humain majeur qui est définit comme ancestral. Notons que, dans Biomart, un allèle de référence pour chaque position est disponible, mais cet allèle ne correspond pas à l'allèle ancestral. En effet, l'allèle de référence correspond à l'allèle trouvé lors du séquençage initial du génome humain [Lander, 2001].

`GenomeToMs` Ce module permet de transformer les données issues des bases de données en réplicats, dans le même format que les données simulées. En définissant un nombre S de SNPs par réplicat, on obtient les données génomiques sous forme d'une série de réplicats de S SNPs. En choisissant des valeurs min et max, on obtiendra que les réplicats de S SNPs consécutifs qui s'étendent sur plus de min Kb et moins de max Kb.

`Portion` Ce module permet de sélectionner dans les données une portion de SNPs consécutifs. On définit la position de départ et la position de fin de la portion à isoler et on obtient uniquement les données pour les SNPs inclus entre ces positions.

2.3.2 Analyses des données par scan génomique

En utilisant le module `GenomeToMs`, on peut transformer le jeu de données expérimentales en plusieurs réplicats. Ainsi, on peut utiliser l'outil `msCalSel` (voir section 2.2.3) pour calculer les spectres de fréquences et les HACs de notre région génomique. Avec l'aide du module `Compile` (voir section 2.2.3), on peut également calculer les distributions empiriques des estimateurs de θ, des statistiques D, H.

L'outil `SvdSlide` nous permet d'obtenir la distribution empirique de Svd par une approche par fenêtres coulissantes (*sliding window*). On choisit la taille fixe, en nombre de SNPs, de la fenêtre qui coulissera, un SNP à la fois, le long des données. Chaque SNP est placé au centre d'une fenêtre et la valeur de Svd qui lui correspond est calculée en tenant compte des SNPs de part et d'autre du SNP central. Par exemple, si l'on fixe la taille de fenêtre à 200, la valeur de Svd pour un SNP est calculée sur les haplotypes contenant les 100 SNPs qui le précèdent et les 100 SNPs qui le suivent. Dans cet exemple, les valeurs de Svd pour les 100 premiers et derniers SNPs de la région génomique sont calculées en se basant sur les haplotypes formés par les 200 premiers et derniers SNPs de la région, respectivement.

Nous avons tenter de prendre en compte le taux de recombinaison pour déterminer la taille de la fenêtre. Cette option a été intégrée à l'outil `SvdSlide`, mais n'a pas été utilisée dans l'analyse présentée au chapitre 3. L'idée est de faire varier la taille en SNPs de la fenêtre selon le taux de recombinaison local. Soit R le taux de recombinaison maximum observé dans toute la région à scanner. Soit S_{min} et S_{max} la taille de fenêtre minimale et maximale et $\sigma = \frac{S_{min}}{S_{max}}$. Soit r_i le taux de recombinaison maximum observé dans une fenêtre de taille S_{max} centrée sur le SNP i. La nouvelle taille de fenêtre $N(i)$ utilisée pour calculer Svd au SNP i est :

$$N(i) = S_{max} * \left(\frac{(\sigma - 1) * r_i}{R} + 1 \right) \tag{2.4}$$

Lorsque r_i est proche de R, la taille de fenêtre est grandement réduite, alors qu'elle reste proche de S_{max} quand r_i est proche de 0. Cet algorithme nous permet de prendre en compte les variations du taux de recombinaison et plus particulièrement la présence de hotspots de recombinaison dans les données. Cependant un problème

se pose : dans certains cas, le hotspot de recombinaison considéré pour réduire la taille de fenêtre se retrouve hors de la nouvelle fenêtre. L'haplotype considéré dans un tel cas est artificiellement réduit. Pour ne pas réduire arbitrairement ces fenêtres, nous utilisons l'algorithme suivant :

1. On calcule $N(i)$
2. Soit d la distance entre le SNP i et le hotspot présentant un taux de recombinaison r_i. Si r_i a été trouvé hors de la nouvelle fenêtre de taille $N(i)$ centré sur le SNP i, on calcule $N'(i)$:

$$N'(i) = S'_{max} * \left(\frac{(\sigma' - 1) * r_i}{R} + 1 \right) \tag{2.5}$$

   avec $S'_{max} = 2(d-1)$ et $\sigma' = \frac{S_{min}}{S'_{max}}$.

3. $N(i)_{Final} = max\{N(i), N'(i)\}$

La fonction proposée pour calculer $N(i)$ est une fonction linéaire la plus simple, mais d'autres fonctions peuvent être utilisées.

Nous utilisons `SvdSlide`, avec taille de fenêtre fixe ou variable, pour trouver des signaux exceptionnels dans les données, ce qui permet d'identifier des régions génomique potentiellement sous sélection positive récente et en cours. Ces régions candidates doivent ensuite être analysées plus finement, par des approches de biologie moléculaire ou par l'utilisation d'autres tests de sélection. Les résultats sont présentés au tableau 3.2 aux figures 3.4 et 3.5.

2.3.3 Analyse d'un locus candidat

Une fois un locus candidat identifié à la suite du scan génomique, on peut utiliser le module `Portion` pour isoler la région génomique d'intérêt et faire une analyse plus fine de ces SNPs. L'outil `SelDiff` nous permet de calculer les valeurs de Svd à chaque SNP de la région isolée, pour les haplotypes constitués par les SNPs compris dans la région, puis de leur attribuée une p-valeur en spécifiant la distribution neutre des valeurs. Nous avons procéder ainsi pour obtenir les résultats présentés à la figure 3.6, pour les 26 SNPs du locus de l'intolérance au lactose (Fig. 3.3).

L'algorithme effectué par `SelDiff` sur un locus est le même que celui réalisé sur chaque réplicat généré par `ms` pour séparer les données en deux sous-échantillons à chaque SNP (voir section 2.2.3). On obtient une valeur pour chacun des sites du jeu de données.

La statistique Svd va donc nous être utile à la fois pour scanner le génome à la recherche de régions candidates, puis pour faire une première validation de ces régions. Évidemment, il conviendra d'appliquer d'autres tests de sélection, d'utiliser d'autres données et de valider la cible en laboratoire avant de conclure à un événement de sélection positive causé par un *selective sweep* partiel.

Chapitre 3

Article

3.1 Introduction

Evidence accumulates about the role of positive selection in the evolution of modern humans and in their local adaptations. The study of how recent positive selection affects patterns of variation has become a field of intense research in order to understand evolutionary factors shaping genetic diversity of extant humans. Numerous methodological advances have come to forth, however it is not always easy, according to the relatively low rate of overlap between results [Nielsen et al., 2007], to determine exactly which type of selective events are captured by the available methods. Recent and incomplete selective sweeps, where the selected allele is dominant but not fixed, are particularly interesting selective events when we aim to learn more about the biological basis of the evolution of modern humans and loci under such evolutionary forces are very likely to be of functional importance, otherwise selection would not be operating.

To detect ongoing selective sweeps affecting current population genetic variation, approaches based on linkage disequilibrium (LD) try to assess the extent to which genetic variation in a genomic region surrounding a locus is indicative of selection [Sabeti et al., 2002, Zhang et al., 2006, Voight et al., 2006, Kim and Nielsen, 2004]. These haplotype diversity-based approaches are based on the idea that genealogical changes caused by positive selection distort the haplotype frequency spectrum [Ewens, 1972]. Other tests use the site frequency spectrum, skewed by the action

of selection, to reject neutrality [Tajima, 1989, Fay and Wu, 2000, Fu and Li, 1993, Zeng et al., 2006]. Fu [1997] proposed the Fs test which combines site frequencies and haplotype informations. Because an incomplete selective sweep tends to skew both haplotype and site frequency spectra, this idea appears as a efficient strategy to detect recent and ongoing positive selection events.

To combine allelic and haplotypic informations in a single plot, a new way of looking at polymorphism data was previously proposed : the haplotype allelic classes (HAC) [Labuda et al., 2007]. The idea is to order the sampled haplotypes by their distance to a predefined reference haplotype. This distance, also called allelic class, is calculated in terms of allele differences between haplotypes. We can detect the imprint of incomplete selective sweeps, by looking at the distribution of these classes, using the haplotype composed of all major alleles as the reference haplotype. This reference haplotype has the highest probability of being the closest haplotype to a hypothetical selected haplotype. Note that the reference haplotype does not necessarily exist in the sample.

A selective sweep drastically narrows the HAC neutral distribution, by increasing considerably the proportion of haplotypes similar to the reference haplotype. The more intense a selective sweep is, the more quickly it will raise the selected allele and linked allele to major allele frequency, since recombination will not have time to break the LD. Therefore, HAC derived statistics are expected to help us identify selection events in population genetics data. We present a summary statistic based on the HAC to detect ongoing selective sweeps, less sensitive than other tests to confounding factors (changes in population size, variation in the fine-scale recombination rate). We further apply the new method to a widely studied trait, showing very strong evidence for recent positive selection : the lactase persistence locus. We show that our approach succeeds to identify the lactase-persistence locus as being a strong target of positive selection by a genome-scanning approach as well as by a candidate approach.

3.2 The Svd statistic

The statistic we developed based on the HAC is called Svd. The idea for Svd is similar to the one on which the iHS statistic [Voight et al., 2006] is based : we evaluate separately the two different alleles of a SNP. Diversity is reduced on haplotypes carrying advantageous alleles favored by positive selection and these haplotypes are in sharp contrast with more variable haplotypes which do not carry alleles under selection. In order to highlight this contrast, we compare, at each SNP, the HAC for the haplotypes carrying one allele to the distribution for the haplotypes carrying the other allele. For clarity purposes, we call the major allele of a SNP the reference allele, since it is on the reference haplotype. For a neutral SNP, the spread of the distributions, and therefore the HAC variances, are expected to be similar. For positively selected SNPs and the SNPs highly linked to them, the allele under selection is likely to present a tighter distribution and a lower HAC variance than the other allele.

Svd : Selection statistic based on HAC Variance Difference. Formally, let D be the number of pairwise differences between a haplotype and the reference haplotype. In a sample containing $2n$ chromosomes and S segregating sites, the empirical variance v is :

$$v(D) = (\sum_{k=0}^{S} k^2 * \mathbb{P}(D = k)) - (\sum_{k=0}^{S} k * \mathbb{P}(D = k))^2 \tag{3.1}$$

$\mathbb{P}(D = k)$ is well approximated by $\frac{C_k}{2n}$. For each SNP, the sample is separated in two sub-samples: the sub-sample R containing the haplotypes carrying the reference allele and sub-sample r containing the remaining haplotypes. The D values, called $D_{R,i}$ and $D_{r,i}$ at SNP i for the haplotypes in the sub-sample R and r are calculated. We then compute :

$$vd_i = v(D_{r,i}) - v(D_{R,i}) \tag{3.2}$$

When the alleles of a SNP i are not linked to a selected allele, vd_i is expected to be negative or close to zero since, because of haplotype sampling, the variance of

$D_{R,i}$ is likely to be the highest. For the selected SNP, the variance of $D_{R,i}$ is likely to be particularly small because the distribution is tighter. Hence, vd_i is expected to be positive when computed for a selected polymorphism and its linked sites.

Even under neutrality, SNPs having a major allele with a particularly high frequency tend to give higher vd values than SNPs having a major allele at intermediate frequency. Hence, ancestral alleles at high frequencies can lead to positive vd values, as well as high frequency derived alleles which are under positive selection. In order to distinguish these two cases, we weight the difference of variances by the derived allele frequency. Because, when sites are independents, vd values should not depend on haplotypes length in number of SNPs, we divide the weighted difference of variances by S (see Material ans Methods). The resulting standardized statistic is :

$$Svd_i = \frac{(v(D_{r,i}) - v(D_{R,i})) * f_{d,i}}{S} \tag{3.3}$$

where $f_{d,i}$ is the frequency of the derived allele at site i. By weighting the difference of variances by $f_{d,i}$, the majority of significantly positive values are expected to be computed for derived alleles linked to a selected allele. The significance of Svd values is assessed by comparing the results with simulated or empirically obtained critical Svd value.

Svd under different population scenarios. Using computer simulations, the distribution of Svd under demographic (population bottleneck and exponential growth), recombination and ongoing selection scenarios were computed (Fig. 3.1). The ongoing selection scenario is the only distribution that obviously differs from the distribution expected under neutrality, contrary to other statistics previously developed to detect selection events in genomic data (Supplementary Fig. 3.7.A). This result suggests that Svd is not sensitive to bottlenecks, population expansions and recombination, which makes it quite robust.

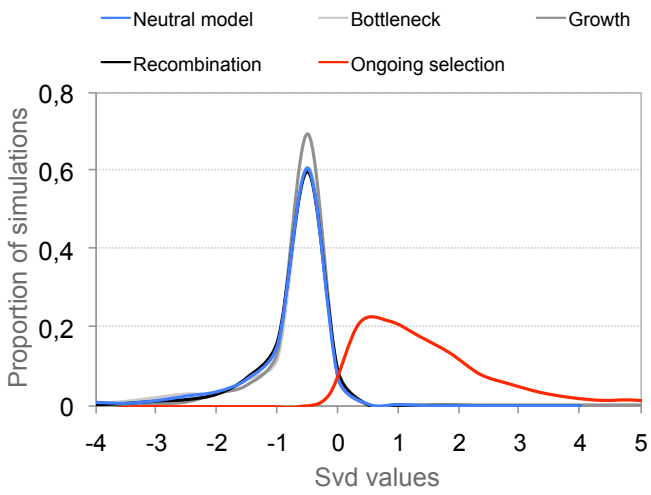

FIG. 3.1 – Distribution of Svd under simulations for five population scenarios

Shift in the Svd values distribution for the ongoing selection model. See also supplementary Fig. 3.7 for the distribution of other selection statistics under the same population scenarios. The distribution of Svd are computed from 1000 simulated replicates of 50 chromosomes (see Material and Methods for details on simulations) for each scenario with $\theta/Kb = 2.23$ in all cases. Under neutrality the effective population size $N_e = 1000$ does not change over the 4000 generations since the most recent common ancestor. For the bottleneck scenario, the population size experiences a 95% reduction during 80 generations (between generations 260 and 340) and in the population growth scenario, the initial population size is $\frac{N_e}{2} = 500$ and doubles in the last 300 generations. For the recombination scenario, $N_e = 1000$, recombination rate $\rho/Kb = 1$. For the ongoing selection scenario, the central SNP has a frequency of $f = 0.75$ and is under moderate positive selection (selection coefficient $s = 0.15$).

3.3 Power to detect ongoing positive selection

We used the Svd statistic as a neutrality test. Neutrality is rejected when Svd>c, where c is the critical value of the test. For all subsequent analyses, the critical value c is defined such as $P(\text{Svd}>c \mid \text{neutrality}) = \text{p}$, with p = 0.05. We call detection power the sensitivity of the test, which is the probability of having Svd>c when a selective sweep is in progress.

A model system (see Material and Methods) was designed using computer simulations to measure the power of Svd to detect an incomplete selective sweep with a dominant selected allele. In a locus with 50 SNPs, without recombination, the detection power of an ongoing selective sweep, at $p = 0.05$, is 81% (Table 3.1).

Comparison with other statistics. Detection power of Svd to detect ongoing selection events and three other statistics was measured on simulated data at different false discovery rates (FDR) (Supplementary Fig. 3.7.B). For FDR over 2.5%, Svd outperformed the site frequency spectrum-based statistics. The specificity of the iHS statistic is the highest, since it has a better detection power than the other statistics at low FDR. However, Svd presents the highest sensitivity to detect actual selection events (a sensitivity of 0.95 for a specificity of 0.9).

Influence of population parameters. Population parameters and locus characteristics are likely to affect the detection power of the test (Table 3.1). When the effective population size N_e doubles, the number of false positives is expected to be reduced, because the impact of genetic drift is weaker : the detection power of the test is indeed better with $N_e = 2000$ than with $N_e = 1000$ or 500. When the test is performed on a locus of 25 SNPs instead of 50, the detection power drops to 74%, whereas it increases when computed on a larger locus, meaning that Svd values are more accurate when a larger number of SNPs is considered in haplotypes. To evaluate the impact of recombinations, we simulated sequences with constant recombination rates and with hotspots of recombination. A constant recombination rate of $\rho/Kb = 5$ causes the strongest drop in the detection power, although constant recombination rate of $\rho/Kb = 1$ also reduces considerably the power of the statistic to detect ongoing selective sweeps. Two types of recombination hotspots of 1Kb are

TAB. 3.1 – Detection power of Svd under various population parameters

Detection power values are computed from 1000 simulated replicates of 50 chromosomes (see Material and Methods for details on simulations) for each scenario with $\theta/Kb = 2.23$ in all cases. The selected site is the central site of the locus, its frequency is $f = 0.75$ and the selection coefficient is $s = 0.15$ in all cases. The 1Kb hotspots are located 1Kb downstream the selected site. Default parameters are presented in bold. Detection power values are measured at p = 0.05, critical values were obtained using identical simulations with $s = 0$.

Haplotypes length		**Simulation parameters**		**Detection Power**
(in number of SNPs)	N_e	Background ρ/Kb	ρ/Kb in hotspot	
50	**1000**	**0**	**no hotspot**	0.81
50	500	0	no hotspot	0.81
50	2000	0	no hotspot	0.91
25	1000	0	no hotspot	0.74
200	1000	0	no hotspot	0.85
50	1000	1	no hotspot	0.68
50	1000	5	no hotspot	0.63
50	1000	1	10	0.67
50	1000	1	100	0.65

simulated, in which the crossover rate is 10 and 100 times higher than the background rate of $\rho/Kb = 1$. With respect to the constant recombination rate scenario, the smallest hotspot does not affect the detection power by more than 1% and the highest hotspot causes a loss in power not greater than 3%.

We investigated the capacity of Svd to differenciate between derived alleles at high frequency due to genetic drift and the same frequencies caused by ongoing positive selection, for several selection coefficients s (Fig. 3.2). Svd performs the best when the selected polymorphism is at high frequency and when the strength of selection is high ($s = 0.5$). Furthermore, the gap between detection powers measured

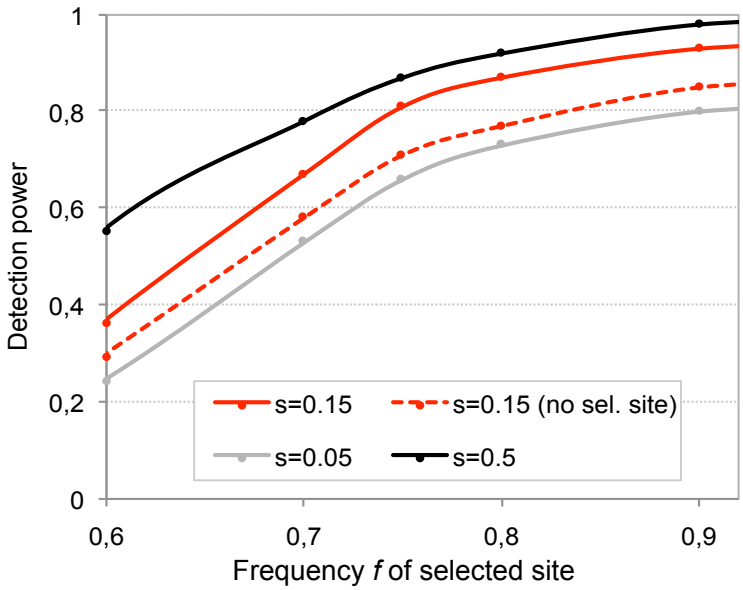

FIG. 3.2 – Power to detect sweeps-in-progress with Svd

Each detection power values is computed from 1000 simulated replicates of 50 chromosomes (see Material and Methods for details on simulations) without recombination with $\theta/Kb = 2.23$ and $N_e = 1000$ in all cases. The selected site is the central site of the locus. Power is evaluated for different frequencies of the selected site (f), for different selection coefficient ($s > 0$) and is measured at p = 0.05, critical values were obtained using identical simulations with $s = 0$. The dashed line indicates detection power when the selected SNP is excluded.

at a site under strong ($s = 0.5$) and moderate ($s = 0.15$) intensity of selection is smaller for high frequency selected SNPs than for SNPs at intermediate frequencies. The power to detect selective sweeps depends on the two selection parameters in conjunction : a selected allele having a frequency of 0.75 because of strong selective forces is detected with the same power than an allele driven to a frequency of 0.8 by a selective sweep of moderate strength. These results show that, with high confidence, Svd distinguishes an allele with increased frequency due to a selective sweep from an allele on its way to fixation by genetic drift. Finally, due to the surrounding SNPs, when the site under selection is removed from the simulated replicates, the selective sweep can still be detected, even though the power is decreased as if the selective sweep of moderate strength had a lower selection coefficient.

Influence of experimental bias. SNP data available in databases present experimental biases such as the ascertainment bias and errors in haplotype phasing (Fig. 3.3). To test the impact of ascertainment bias, SNPs are taken off from the replicates simulated with default parameters (see Table 3.1). Only the SNPs found in a panel of m individuals are retained in all n individuals (see Material and Methods). When we genotyped the SNPs in no more than 5 individuals out of 25, ascertainment bias negatively influences the detection power, which is below the default power value of 81%. When more than 10 out of 25 individuals are genotyped, the detection power remains unchanged. Between these bounds, the power increases, getting over 81%. By considering SNPs only present in these proportions of individuals, rare SNPs that create noise are more likely to go undiscovered, which raises the sensitivity of the test because it amplifies the molecular signature of positive ongoing selection left by high-frequency derived alleles and captured by Svd. To validate this hypothesis, singletons and doubletons are taken off from the replicates simulated with default parameters with the 50 chromosomes are genotyped. This procedure increases the detection power to 88%, meaning that the SNPs from these removed frequency classes are probably the one producing noisy signals, lowering the detection power of Svd.

The determination of haplotypes by experimental techniques is considerably expensive and time-consuming. In general, the estimation of haplotype phase is done accurately by statistical methods, such as the PHASE algorithm [Stephens et al., 2001]. Because this algorithm tends to cluster together the sampled chromosomes into groups of similar haplotypes, the method narrows the HAC distribution. Under selection, $\mathbb{V}(D_{r,i})$ tends to be artificially smaller, reducing Svd values and causing a important drop in the detection power of our method to 56%, when computed on 50 SNPs long haplotypes, after phasing the simulated datasets (see Material and Methods for details on simulations with haplotype phasing bias). However, increasing haplotype length increases the detection power up to 80% when 800 SNPs are considered to compute Svd values.

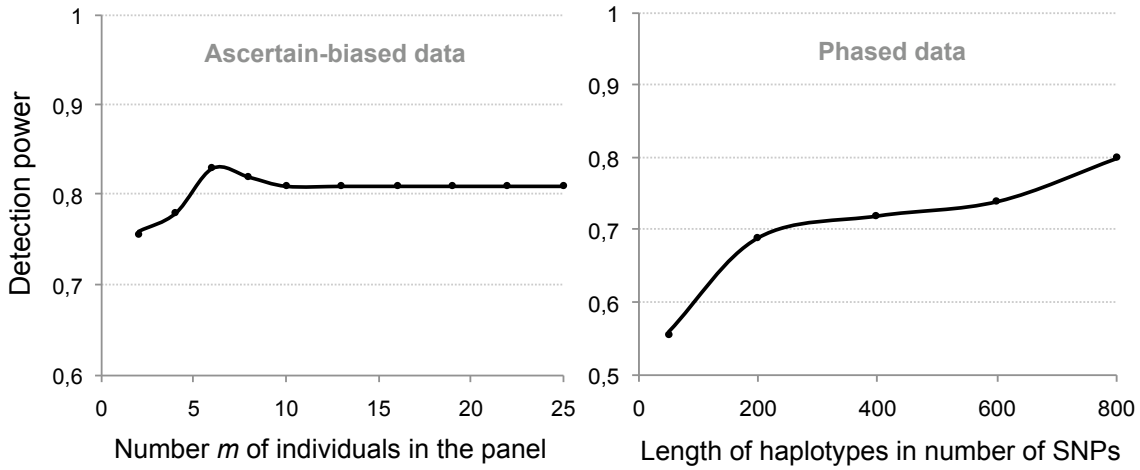

FIG. 3.3 – Impact of experimental bias on the detection power of Svd

Biaised data sets are composed by 1000 simulated replicates of 50 chromosomes without recombination with $\theta/Kb = 2.23$ and $N_e = 1000$, the selected site is the central site of the locus, its frequency is $f = 0.75$ and the selection coefficient is $s = 0.15$. Detection power calculated for ascertained biased data is shown in the left-hand side graph. For each replicate, the SNPs typed in a panel of $2m$ randomly chosen chromosomes are retained in all chromosomes, creating new replicates with a smaller SNP density (see Material and Methods for details on simulations with ascertainment bias). Detection power calculated for phased biased data, on haplotypes of different lengths, is shown in the right-hand side graph (see Material and Methods for details on phased simulations). Power is measured at p = 0.05, critical values are obtained using identical simulations with $s = 0$ and the same procedures.

3.4 Application to data

Genome-scan approach. Our approach can be used for genome-wide studies of natural selection by using a sliding window approach to calculate Svd at all SNPs. Each SNP considered is placed at the center of a haplotype of fixed length, in number of SNPs. We identify signals of selection by detecting regions that behave as outliers, i.e. clusters of SNPs displaying remarkably high values of Svd. In figure 3.4, we plotted the top 1% values of positive Svd computed for SNPs of chromosome 2, between 50 and 200 Mb, from HapMap data in asian population.

Clusters of signals for six loci previously identified as under selection in recent studies are detected [Tang et al., 2007, Williamson et al., 2007, Carlson et al., 2005], where many SNPs show evidence for selection. The intensity and clarity of the signals vary depending on the length of haplotypes chosen to compute Svd. Large haplotypes allow a better detection of an ongoing selective sweeps, but are more likely to present a varying fine scale recombination rate and recombination hotspots, which lower the detection power. When Svd is calculated for haplotypes of 50 SNPs, signals are noisy and unclear. Noise is reduced when larger haplotypes are considered, but signals may disappear as the length of haplotypes increases. For example, around 177Mb, a clear signal is detected at small haplotype lengths, but disappears when Svd is computed for haplotypes of 600 SNPs and more. The fading of this signal can be explained by the presence of a hotspot of high intensity located in that region. At this location in chromosome 2, an expressed gene, LOC375295, is encoded, but no in vivo function has yet been reported for this gene. Signals visible at all haplotype lengths are in loci previously unidentified as target of selection. One such region, at 124-125 Mb, contains CNTNAP5, a gene that belongs to the neurexin family, involved in cell contacts and communication in the nervous system.

Comparing signals between populations can help us to validate targets of selection. In figure 3.5, a strong signal of positive selection is found in a 1 Mb region including the lactase (LCT) and MCM6 genes in the HapMap european-derived

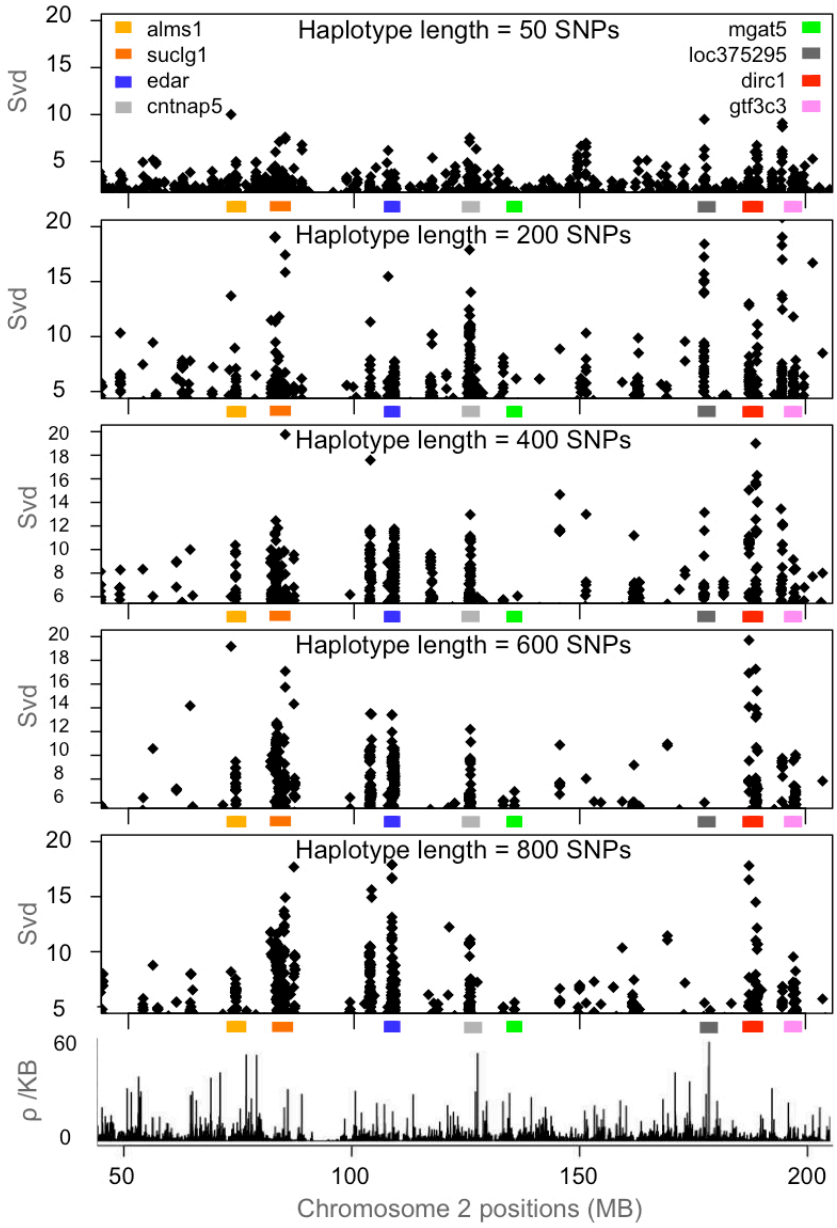

FIG. 3.4 – Comparison of Svd clustered signals for different length of haplotypes. Top 1% positive Svd values computed with 5 different haplotype lengths, for the HapMap asian population in the region between 50 and 200 Mb in chromosome 2. The bottom plot shows an estimation of the fine scale recombination rates in chromosome 2, calculated using InfRec [Lefebvre and Labuda, 2008]. The colored boxes below each plot indicate the location of five loci previously identified as being targets of selection in recent studies [Tang et al., 2007], [Williamson et al., 2007]. The gray shading boxes indicate the location of two loci not previously identified as target of selection for which clustered signals were found with Svd.

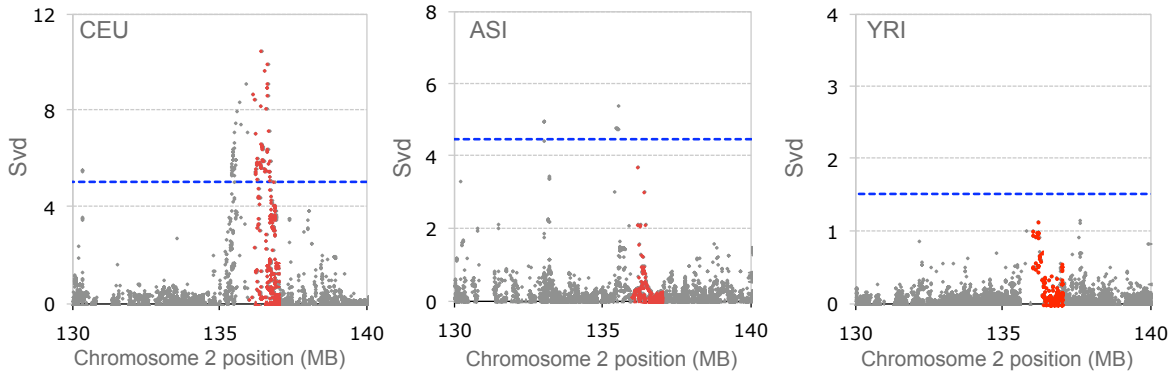

FIG. 3.5 – Positive Svd values in a 10 Mb region in chromosome 2

Plots of positive Svd values for the three HapMap populations with a window size of 800 SNPs. Svd values plotted above the dashed blue lines are in the top 1% of positive values of chromosome 2 in each population. A 1Mb locus, containing the LCT and MCM6 genes, is plotted in red. A strong and clear signal of positive selection is found in this locus in the european-derived population, whereas no signal is detected in the two other populations.

population only, where domestication of cattle appeared 10.000 yrs ago and where cultural habits of consuming milk could have been advantageous for individuals in Europe (nutritional benefit, improved calcium absorption [Mace et al., 2003]). This signal is absent from the two other populations, where no *Svd* value is in the top 1% of positive Svd values in the 1Mb region. It was recently reported, however, that positive selection could have acted on the lactase persistence locus in some african and middle-eastern populations [Tishkoff et al., 2007, Enattah et al., 2008], but these populations are not included in the HapMap Phase II data.

Other strong signals of ongoing positive selection for several loci in chromosome 2, previously identified in other studies, are found by the Svd statistic and are presented in Table 3.2).

Candidate approach. Signals found when performing a genome scan depend on a large amount of information, especially when haplotypes of 800 SNPs are considered. The Svd test can also be used to examine the site that are present in a particular locus in order to validate the signal and to detect the SNP which triggered the

TAB. 3.2 – Validation of nine candidate regions on chromosome 2

Preliminary identification (✓) of some selection signals for ten candidate regions where strong signals of positive selection has been reported in recent genome scan for the three populations. Light blue colouring represents the populations in which the corresponding studies found evidence of selection.

	Genes (chr2)	CEU	ASI	YRI	References
Signal in two populations	SLC3A1	✓		✓	[Tang et al., 2007]
	MGAT5	✓	✓		[Tang et al., 2007]
	GTF3C3	✓	✓		[Tang et al., 2007]
	SUCLG1	✓	✓		[Carlson et al., 2005]
Signal in one population	LRP1B*			✓	[Williamson et al., 2007]
	LCT/ MCM6	✓			[Voight et al., 2006]
	ALMS1		✓		[Tang et al., 2007]
	EDAR		✓		[Tang et al., 2007, Williamson et al., 2007, Carlson et al., 2005]
	DIRC1		✓		[Williamson et al., 2007]
	NCOA1/ ADCY3*			✓	[Voight et al., 2006]

* The signal appears in the upstream region of the gene

selective sweep in the locus. The idea of the candidate approach is to accurately assign p-values to each Svd value computed, which can be used to measure confidence in inference of selection in a single genomic region.

We choose the lactase persistence locus as a genetic system to show the potential of Svd in analyzing candidate targets of ongoing positive selection. This locus has already been identified as a target of natural selection in numerous studies. In the genome-scan analysis of the chromosome 2, the lactase persistence locus exhibits one of the strongest clustered signal. Over the past few years, the haplotype associated with lactase persistence and the variant responsible for the trait were found for european-derived populations. The mutation proposed to be the causal factor for the persistence phenotype is located 13910 base pairs upstream of the LCT initiation codon in the MCM6 gene [Enattah et al., 2002]. Haplotype diversity-based tests showed formal evidence for positive selection at this locus in european populations [Bersaglieri et al., 2004, Voight et al., 2006], but previous methods based on the site frequency spectrum failed to validate this result in the region.

We focus the candidate analysis on the 26 SNPs available in MCM6 for the european-derived HapMap population. Svd values at each SNP are calculated on haplotypes formed by these 26 SNPs (Fig. 3.6) and $p-$values are assigned. The C→T -13910 variant is present on a haplotype carrying 18 ancestral and 8 derived alleles, which happens to be the reference haplotype. Here, 7 of the 8 derived alleles on the reference haplotype are above have $p-$values above 0.1 which is very unlikely in a haplotype of 26 SNPs, according to our simulated data. The remaining derived allele (rs2236783) is a very old mutation, it has been subjected to genetic drift for a long time and was already on its way to fixation when the sweep began. This allele is present on both selected and unselected background.

The value computed for the C→T -13910 variant is significantly higher than the others and its $p-$value is smaller than 0.05, a result suggesting that this SNP may be the one which is under positive selection, a suggestion that happens to be supported by undeniable evidences. The significance of the results increases when additional

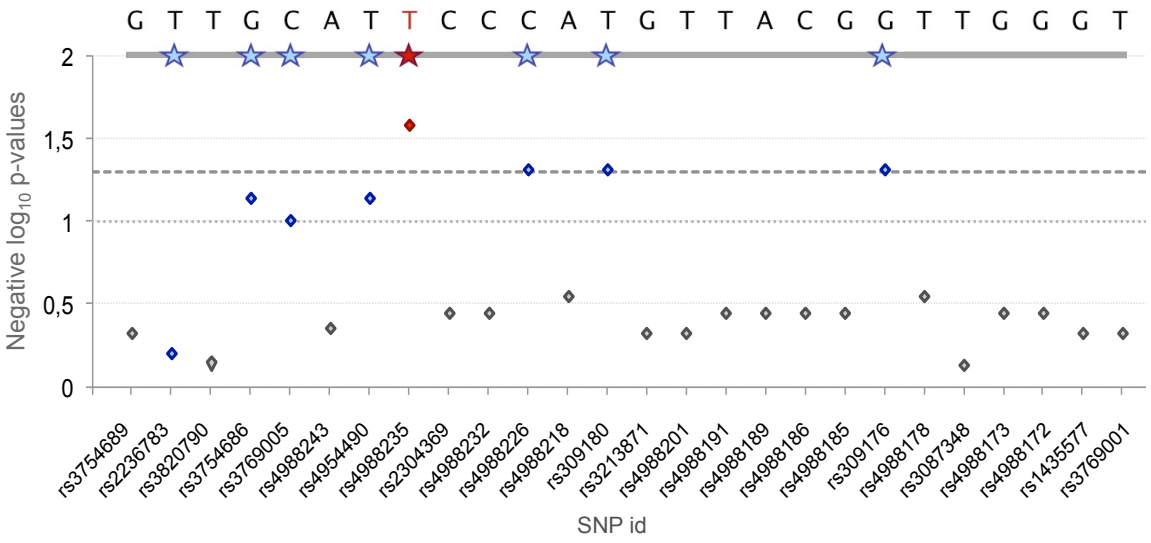

FIG. 3.6 – Svd values for the 26 SNPs at the MCM6 locus

The reference haplotype used to compute the HAC in the sample is presented on top of the plot. Stars represent the derived alleles present on the haplotype. The red star corresponds to the T -13910 variant (rs4988235), found to be responsible for the trait in europeans-derived populations [Enattah et al., 2002]. We generated 1000 replicates of a locus with the same characteristics than the lactase persistence locus (120 ascertainment-biased chromosomes with 26 SNPs, see Material and Methods for details), simulated under neutrality. $P-$values were estimated by comparing the Svd values from the real data to the distribution observed in the simulated data sets. The dotted line and dashed line represent the 10% and 5% cutoff, respectively.

SNPs are considered in haplotypes.

3.5 Material and Methods

Simulations Data was simulated under a variety of population neutral and selection scenarios using coalescent programs. Each set of simulations contains 1000 replicates with 50 chromosomes of 100 Kb, and mutations appears at a rate $\theta/Kb = 2.23$ to have an average of 1000 SNPs by chromosome under neutrality (because $S \simeq \theta_W \cdot \sum_{i=1}^{n-1} 1/i$).

The program `ms` [Hudson, 2002] is used to simulate set of replicates under the neutral scenario, under demographic scenarios (bottleneck and exponential growth) and with constant recombination rate set to $\rho \simeq \theta/2$. See Fig. 3.1 for parameter details for each scenario. The program `Selsim` [Spencer and Coop, 2004] is used to simulate set of replicates under ongoing positive selection scenarios. In all cases, the central SNP, at position 50000, is under positive selection. The default selection scenario has the following parameters : effective population size $N_e = 1000$, frequency of the selected SNP $f = 0.75$, selection coefficient $s = 0.15$, constant recombination rate $\rho = 0$, with no recombination hotspot. Different ongoing positive selection scenarios are generated by changing one parameter at a time as specified in figure captions.

Ascertainment biased replicates. The ascertainment scheme we recreated is the one involving a preliminary identification of SNPs in a panel of m individuals, or $2m$ chromosomes, and typing them in a larger sample of size $2n$, containing all the panel chromosomes. To evaluate the impact of panel size, different values of m were considered ($m = [2, 4, 6, 8, 10, 13, 16, 19, 22, 25]$). To generate ascertainment biased sets, the set of replicates simulated under the default selection scenario is used. For each of the 1000 replicates, $2m$ out of 50 chromosomes are randomly chosen and only SNPs identified in this panel are kept in the replicate.

In some ascertainment protocols, SNPs are reported only if they have some minimum frequency in the sample. Because a large proportion of sites with a minor allele frequency (MAF) below 0.05 have a large probability of being caused by sequencing or genotyping errors, these sites are often removed from the analyses. To generate such simulated set of replicates of 50 chromosomes, singletons and doubletons (which have a MAF below 0.05) are removed from the replicates simulated under default selection scenario.

Haplotype phasing biased replicates. The effect of haplotype phasing on the detection power of Svd was assessed by randomly assigning, for each replicate, the 50 simulated chromosomes to 25 individuals, resolving the resulting genotypes back to the haplotypes using the *fastphase* program [Scheet and Stephens, 2006] and recomputing Svd. The accuracy of *fastphase* is nearly the same as PHASE 2.0 [Stephens et al., 2001], used to phase the HapMap data.

Svd standardization. When sites are independent, *vd* values should not depend on haplotypes length in number of SNPs S. Let X_i be the random variable which is 1 when the state at site i in a haplotype is different from the state at site i in the reference haplotype, and 0 when the states are identical. $X_i = 1$ with probability p_i which is the frequency of the minor allele, and $X_i = 0$ with probability $1 - p_i$. The distance D can be defined as the sum of S Bernoulli random variables X_i, with i from 1 to S. Hence, the variance of D is in $O(S)$, and $\frac{\mathbb{V}(D)}{S}$ is independent of haplotypes length.

Detection of selection in simulated data. Four statistics are used to detect ongoing selection sweeps on set of replicates simulated under ongoing positive selection scenarios : Svd, the unstandardized version of iHS [Voight et al., 2006], Tajima's D [Tajima, 1989] and the normalized version of Fay and Wu's H [Zeng et al., 2006]. The statistics are computed on haplotypes with a fixed number of SNPs with the selected site always located at the center of the haplotypes. The default haplotype

length is 50 SNPs, however Svd is also computed on haplotypes of 25 and 200 SNPs, to evaluate the impact of haplotype length on the results (see Table 3.1). Svd and iHS are computed at the central site of the haplotypes whereas Tajima's D and Fay and Wu's H summarize the site frequency spectrum using all the SNPs in the haplotypes.

The detection power of iHS, Tajima's D and Fay and Wu's H is computed for the default selection scenario. The detection power of Svd is computed for a variety of selection scenarios. For each scenario, critical values at $p = 0.05$ are obtained by computing the statistics on identical simulations, with identical ascertainment or re-phasing procedures, but with $s = 0$. When the selected SNP (central site) is excluded, the maximum Svd value in the replicate is used to assess the detection power.

SNPs data information. The analysis of experimental data is based on data from the HapMap project release 21a. The japanese and chinese samples were merged together in order to create an asian combined population with 89 individuals denoted ASI. The european derived population (CEU) and the Yoruba population from Nigeria (YRI) samples contains trio data : we analyzed only the 60 unrelated individuals. The data was browsed from the biomart HapMap browser (`http://hapmart.hapmap.org/BioMart/martview`). We used the phased haplotype data available (phase reconstruction done by PHASE 2.0 [Stephens et al., 2001]).

For the lactase persistence locus analysis, we selected the 26 polymorphic sites in the CEU individuals available in HapMap data for the MCM6 gene, in the genomic region Chr2:136424478..136459810. A list of the 26 marker ids (rs numbers) is provided (Supplementary Table 3.3). The entire chromosome 2 was analyzed for the three populations CEU, ASI and YRI : the number of SNPs analyzed were 221 956, 206 665 and 252 249, respectively. To compute Svd, we needed information about the ancestral state of each SNP. The chimp allele, or the macaque allele when the chimp allele was unavailable, was considered as ancestral, and was found

through the UCSC table browser (`http://genome.ucsc.edu/cgi-bin/hgTables`). When the chimp or macaque was unavailable in UCSC database, the major allele was considered as ancestral.

Replicates matching MCM6 locus. To assign $p-$values to observed data, we simulated a set of 1000 replicates, with 120 chromosomes, the mutation rate is $\theta/Kb = 2.23$ and effective population size $N_e = 1000$. The central SNP of all replicates is under positive selection with $f = 0.78$ (frequency of the selected T variant in MCM6 in HapMap european-derived population) and $s = 0.15$. To model SNP ascertainment, we used a rejection sampling to modify the simulated frequency spectrum to match the site frequency spectrum observed in chromosome 2 from HapMap data, for the european-derived population (CEU). The rejection sampling function we used is the one described in [Voight et al., 2006]. To match the MCM6 locus in the CEU population, haplotypes of 26 SNPs were chosen in such a way that the 8^{th} SNP of each replicate is the one under positive selection. Uncertainty in haplotype phase was ignored in simulated data because, in europeans-derived HapMap data, trio data is used to reconstruct the haplotypes accurately.

3.6 Discussion

Many large-scale studies have aimed at detecting positive selection and point out that a substantial number of regions in the human genome have possibly experienced selective sweeps. Classical models recognize genetic drift as the main force shaping the genetic variation. However, selected loci can cause changes in the frequency of genetically linked sites remarkably similar to fluctuations caused by genetic drift, as Gillespie's model of genetic draft suggests [Gillespie, 2000]. This means that if there are in fact many genes experimenting partial selective sweeps in the human genome, genetic variation might be shaped by selective forces acting on adaptive mutations and not mainly by genetic drift. To test whether variation should be interpreted in

the light of models of draft instead of models of drift, we think that developing a more specific statistical test to detect incomplete selective sweeps is a good strategy.

Because the Svd statistic is based on variability present in the allelic composition of haplotypes, the method tends to be less sensitive to demography than three statistics, widely used to detect selection based on allelic frequencies or LD patterns. In human populations, where the effective size is estimated to be larger than N_e =2000 [Tenesa et al., 2007], simulation results suggest that the Svd test have a good detection power and will perform well on a variety of population models.

Furthermore, Svd summarizes the overall haplotype diversity and can therefore accurately be applied to ascertainment-biased data. The removal of rare SNPs increases the detection power, suggesting that the Svd test on data with common SNPs, genotyped from previously ascertained variants, will perform slightly better than on datasets containing rare SNPs, generated by resequencing studies for example. Yet, considering the overall diversity of haplotypes to detect selective events makes the method quite sensitive to haplotypes reconstruction errors. This is especially true when Svd values are computed on haplotypes that contain a small number of SNP. A good way to by-pass this problem is then to use haplotypes of hundreds of SNPs to compute Svd values, because reconstruction errors are diluted by larger patterns in the data. However, considering large haplotypes can cause an increase of the false negative rate, because of the action of recombination which breaks linkage between SNPs. Our results on simulations confirm that recombination hotspots lower the detection power of Svd when they are close to selected loci, suggesting that recombination rates and the number of SNPs in haplotypes are parameters that have to be carefully taken into account when comparing Svd values at different loci in the same population, in approaches like genome scans.

This study will be followed by a complete scan of the human genome using Svd to find regions that have potentially experimented partial selective sweep. A detailed comparison between the list of loci under selection according to Svd and sets of candidates found in previous genome scans will be released. In this paper, we wanted to

demonstrate the potential of such an approach and some of the preliminary results from HapMap chromosome 2 analysis are presented. To assign statistical significance, we first used the empirical approach of identifying outliers to find candidate loci. Each locus is then investigated individually, by accurately assigning p-values to observed Svd values, based on prior information we may have for the region and population in which the signal was found (recombination rates, demography, SNP-ascertainement protocol). As an example, the case of the lactase persitence locus was closely reviewed and analyzed. In this locus, we already knew which SNP was the one associated with the selected trait in european populations, but our analysis demonstrate the great potential of the method in detecting new candidate polymorphisms for association studies.

Even though interesting candidate genes are standing out in the first-glance analysis of the top 1% values in chromosome 2, different factors need to be considered to obtain a collection of outliers that properly represent the set of potential targets of recent and ongoing selection. We already mentioned the importance of recombination rates and haplotypes length when computing Svd values, but questions about clustering strategies and cutoff specifications also need to be addressed. The impact of ancestral misidentification on the method must be assessed [Hernandez et al., 2007], especially since we weight the standardized difference of variances by the frequency of the presumed derived allele.

3.7 Conclusion

The main objective of this research is to show the usefulness of the HAC in population genetic analysis. As a first example, we developed the summary statistic Svd along with an intuitive and computationally efficient statistical test designed to find incomplete selective sweeps in genomic data. Several elements presented in this study can however be modified to fit more accurately properties of the data and research questions. For example, other reference haplotypes can be considered :

a previous study defined the reference haplotype as being composed of all ancestral alleles [Labuda et al., 2007]. Furthermore, because the HAC distribution is also sensitive to a complete selective sweep, an approach similar to the one proposed by [Kimura et al., 2007] to identify fixed loci under positive selection could be developed using the HAC instead of haplotype homozygosity.

3.8 Supplementary material

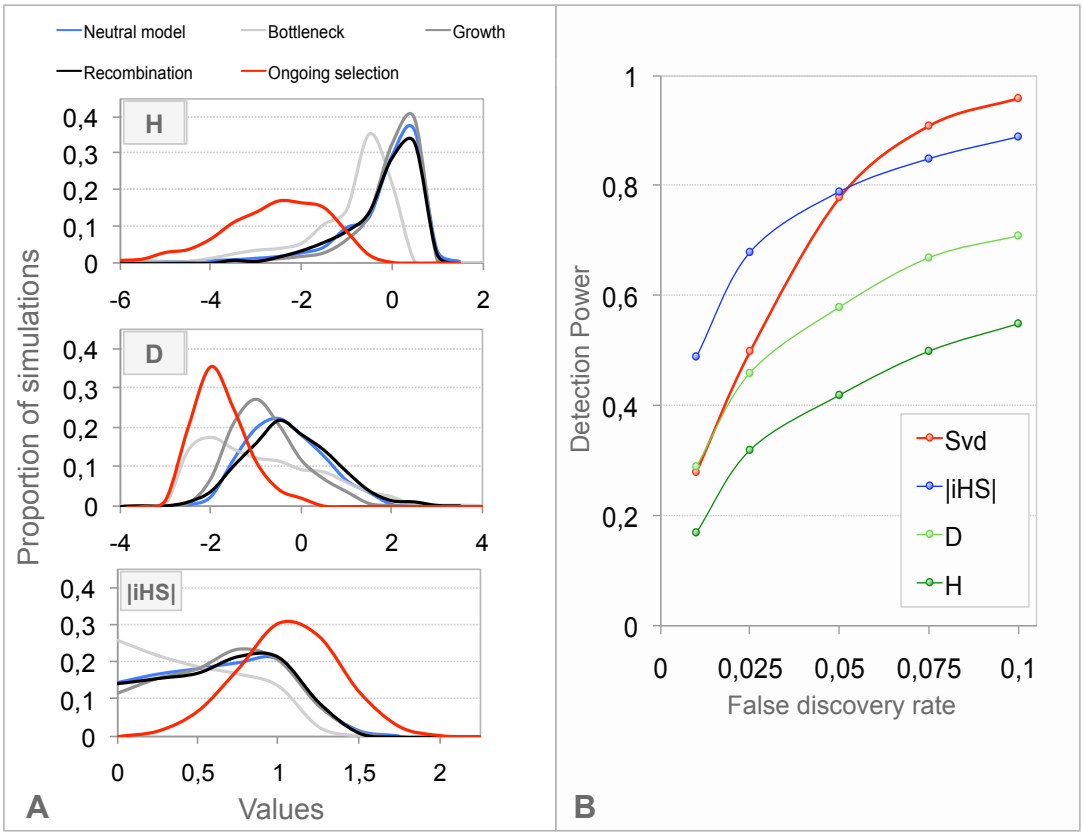

FIG. 3.7 – Comparing Svd with three widely used statistics to detect selection (A) Distribution of the values for three statistics used to detect selection : two site frequency spectrum-based statistics, the normalized version of Fay and Wu's H [Zeng et al., 2006] and Tajima's D [Tajima, 1989] and the haplotype diversity-based statistic iHS [Voight et al., 2006]. The distributions were computed under demographic (population bottleneck and exponential growth), recombination and ongoing selection models. The empirical distributions are computed from simulated data for the five population scenarios, which are the same that were used to compute the HAC distributions presented at Fig. 3.1. (B) Power to detect an ongoing sweep for the four statistics on simulated data at different false discovery rates.

TAB. 3.3 – List of marker ids for the 26 SNPs used for the lactase persistence analysis
The allele considered as ancestral because of its presence in the homologous chimpanzee sequence is presented in **blue**.

Chr.2 position	Marker id	Alleles	Strand
136424478	rs3754689	A/**G**	-
136427890	rs2236783	**C**/T	-
136432949	rs3820790	**A**/T	-
136437008	rs3754686	**A**/G	-
136437098	rs3769005	**C**/G	-
136441435	rs4988243	**A**/G	-
136441963	rs4954490	A/**G**	+
136442378	rs4988235	**C**/T	-
136443052	rs2304369	A/**G**	+
136443403	rs4988232	**C**/T	-
136444330	rs4988226	C/**T**	-
136447193	rs4988218	**A**/G	-
136447987	rs309180	A/**G**	+
136448954	rs3213871	**C**/T	+
136452239	rs4988201	C/**T**	-
136453476	rs4988191	G/**T**	-
136454689	rs4988189	**A**/G	-
136455590	rs4988186	A/**C**	-
136455673	rs4988185	A/**G**	-
136455948	rs309176	C/**T**	+
136456679	rs4988178	C/**T**	-
136458046	rs3087348	**G**/T	-
136458114	rs4988173	A/**G**	-
136458418	rs4988172	C/**G**	-
136458679	rs1435577	**C**/G	+
136459810	rs3769001	C/**T**	-

Chapitre 4

Synthèse

La question initiatrice de la recherche présentée dans ce mémoire se formulait ainsi : de quelle façon peut-on utiliser les HAC, sensibles aux empreintes laissées par la sélection naturelle dans les séquences biologiques, pour retrouver des locus fonctionnellement importants dans le génome humain. La relation entre sélection naturelle et effet phénotypique est aujourd'hui bien établie, particulièrement dans le cas des facteurs causaux de maladies héréditaires. Une recherche bibliographique de littérature générale sur ce sujet vaste qu'est la selection naturelle nous apprend qu'il en existe plusieurs types, et que même si toutes ces formes ont en commun la définition qu'en fait Darwin dans *Origin of species* [Darwin, 1859] (i.e. la préservation des variations individuelles favorables et la destruction des variations nuisibles), il ne s'agit pas des mêmes réalisations moléculaires. Les phénomènes de sélection positive, négative et balancée diffèrent, en effet, dans leur mode d'action. Dés le début du projet, il a donc fallut précisément définir le type d'événement de sélection que nous allions rechercher.

La sélection positive est le type de sélection qui laisse les empreintes les plus claires sur les séquences d'ADN. Ces dernières années, elle a fait l'objet de nombreuses recherches et revues de littérature détaillée [Nielsen, 2005, Sabeti et al., 2006, Harris and Meyer, 2006]. Il s'agit du type de sélection le plus recherché dans les données de polymorphismes. Intuitivement, il est assez clair que la sélection négative

soit associée à des régions fonctionnelles de l'ADN, mais en général, un locus sous sélection naturelle positive se doit également d'avoir une importance fonctionnelle, sinon, la sélection ne pourrait agir. De ce fait, trouver des régions génomiques sous sélection positive peut nous permettre de trouver des variations fonctionnellement importantes reliées à des phénotypes particuliers. Il est intéressant d'identifier des régions qui sont actuellement sous sélection positive, c'est à dire de détecter des *selective sweeps* récents et en cours [Nielsen et al., 2007]. Les gènes qui sont cibles de la sélection positive à cet instant de l'évolution de l'Homme, sont très probablement associées à certaines maladies génétiques humaines : même un effet favorable limité peut, sur une grande échelle de temps, laisser des patrons de variations importants et distincts sur les chromosomes qui nous permettront d'identifier des facteurs génétiques responsables de maladies complexes.

Le cas de la lactase est un très bon exemple pour illustrer ce qui vient d'être dit. La mutation, dans la région régulatrice du gène codant pour la lactase permet la persistance de l'action de l'enzyme à l'âge adulte chez l'Homme. Ce qui est considéré chez les européens comme une maladie, l'intolérance au lactose, est en fait l'état ancestral d'un trait actuellement sous forte sélection positive. De façon simplifiée, on peut dire que les « non-malades » sont les « mutants ». La nature trouve donc un moyen de pallier à certains problèmes que peuvent avoir les individus d'une espèce grâce à la sélection naturelle agissant sur les porteurs de mutations favorables dans un environnement. Ainsi, l'identification de ces cibles de sélection nous permet d'élucider les mécanismes que la nature a installé pour rendre certains individus plus résistants et s'en servir dans des traitements thérapeutiques adaptés. D'autres raisons nous incitent à penser que des gènes impliqués dans les maladies génétiques sont des cibles de sélection positive. Par exemple, durant un *selective sweep*, il arrive que des allèles modérément délétères soient entraînés à de hautes fréquences par un effet d'auto-stop génétique (ou *hitchhiking*).

Nous avons ainsi focalisé nos efforts sur la recherche d'événements de sélection laissant des signatures moléculaires sur les séquences d'ADN comme celles, par

exemple, laissées par la sélection positive agissant sur le lactase. Svd est la statistique sommaire de l'information des HAC qui nous permet d'atteindre cet objectif.

Nous avons développé un test basé sur cette statistique pour détecter la sélection positive par une technique de balayage génomique, communément appelé scan génomique, et pour l'analyse de locus candidats (voir les sections 2.3.2 et 2.3.3). Plusieurs études ont récemment utilisé des approches par scan génomique pour rechercher des cibles potentielles de sélection dans le génome humain [Carlson et al., 2005, Voight et al., 2006, Wang et al., 2006, Zhang et al., 2006, Williamson et al., 2007, Tang et al., 2007, Kimura et al., 2007]. Un problème important, entravant l'analyse des résultats, est qu'il est difficile de savoir exactement quels événements de sélection les nombreuses méthodes développées retrouvent. Elles sont toutes conçues pour retrouver des locus sous sélection positive, mais les signaux observés sont-ils causés par des *selective sweeps* complets, partiels, ou les deux? Des régions ayant subies d'autres types de sélection (balancée, négative) sont-elles aussi retrouvées? Ce sont des questions dont les réponses restent difficiles à déterminer. En effet, Nielsen et collaborateurs [2007] soulèvent ce problème et en démontrent l'effet principal : une faible concordance entre les résultats des différentes études. Nous pensons que le développement d'un test spécifique à la détection de *selective sweeps* partiels dominants nous permettrait de réduire l'univers des possibles pour rendre nos résultats plus facile à interpréter. Notre test présente en effet un très faible pouvoir de détection sur des données simulées sous des modèles de *selective sweeps* complets ou de sélection balancée.

Particularités de Svd

Il paraît important de souligner deux principaux avantages de Svd : la robustesse de la statistique aux scénarios démographiques en l'absence de sélection et la possibilité de l'utiliser sur des données de génotypage, présentant un biais de recrutement ou *ascertainment bias*. Ce sont les deux premiers points soulevés par Nielsen

et collaborateurs [2007] dans leur revue bibliographique sur la sélection positive récente, lorsqu'ils mentionnent les problèmes que posent l'identification de la sélection positive dans les données de polymorphismes.

En principe, les tests de neutralité supposent que la taille de la population étudiée est constante et que la population n'est pas subdivisée, ce qui les rend particulièrement sensibles aux événements démographiques qu'auraient subi la population étudiée. Par exemple, dans le cas d'un *bottleneck*, le H de Fay et Wu [2000] aura tendance à rejeter la neutralité en absence de sélection positive. La comparaison de multiples locus le long du génome est historiquement la façon la plus utilisée pour contourner le problème des effets négatifs de l'histoire démographique sur les tests de neutralité et s'est aujourd'hui étendue à l'approche par scan génomique. La sélection naturelle positive n'agit que sur certaines régions, fonctionnellement importantes, alors que les effets des facteurs démographiques sont présents dans tout le génome, sauf en ce qui concerne la subdivision des populations, qui elle, peut affecter la variation entre les locus [Nielsen, 2001]. Nous n'avons pas testé l'impact de modèles de populations subdivisées. Cependant, nous estimons que l'échantillonnage des individus lors de la collecte des données est fait de manière à ce que la probabilité qu'il y ait subdivision dans la population étudiée est faible. En basant notre méthode sur la composition allélique des haplotypes, nous détenons une statistique plus robuste que d'autres tests couramment utilisés, se basant principalement sur la fréquence des allèles. Il ne va sans dire, cependant, que les effets démographiques doivent quand même être considérés avec attention dans l'analyse des résultats, car aucune méthode ne peut se proclamer complètement robuste à ces effets, et cela parce que la signature de sélection positive elle même peut-être sensible aux événements démographiques. Le programme `Selsim` ne nous permet pas de simuler de la sélection couplée à des scénarios démographiques comme le *bottleneck* ou l'expansion démographique. Par contre, l'utilisation de la méthode de simulations développée par Teshima et collaborateurs [2006] pourrait nous être utile pour évaluer plus précisément les effets de la démographie sur la détection par Svd des empreintes génétiques de sélection. De

plus, certains phénomènes, comme ce qu'on appelle les *surfing mutations* récemment identifiées, peuvent confondre notre méthode. Il s'agit de nouvelles mutations qui apparaîtraient lors d'une expansion géographique et se trouveraient rapidement emportées comme par une vague à de hautes fréquences dans certaines régions géographiques uniquement [Klopfstein et al., 2006, Edmonds et al., 2004]. Les patrons de variations qui en découlent ressemblent grandement aux effets d'un *selective sweep*. La manière dont les deux phénomènes pourront être différenciés n'est pas claire car la sélection naturelle peut justement favoriser les *surfing mutations*, bien que ces mutations peuvent également être complètement neutres.

Les projets de génotypage ayant mené à l'apparition de grands jeux de données des variations génétiques dans le génome humain n'ont pas, à la base, été développés pour la recherche d'événements de sélection. Ce sont cependant des jeux de données contenant énormément d'informations, très utiles pour ce genre de tâche. Il va donc de soi que, malgré les biais présents dans les données pour le choix des marqueurs génétiques étudiés, de nombreux efforts sont faits pour les utiliser dans le contexte des méthodes de détection de la sélection naturelle. L'*ascertainment bias* des SNPs présent dans ces données affecte particulièrement les patrons de fréquences alléliques. En général, les tests basés sur le spectre de fréquences par site, très sensibles à l'*ascertainment bias* ne peuvent être utilisés sur de telles données et nécessitent un re-séquençage complet des individus pour être performants. L'*ascertainment bias* affecte également les patrons de LD [Nielsen and Signorovitch, 2003]. Les tests de sélection qui prennent en compte ce type d'information seront donc également biaisés. En prenant comme mesure le nombre d'allèles mineurs sur chaque séquence, nous avons développé une statistique qui est principalement sensible à la diversité globale des haplotypes et ce type de test est moins sensible à l'*ascertainment bias* [Sabeti et al., 2006]. Nous avons montré que, sur des données simulées, l'absence de SNPs rares (effet causé par l'*ascertainment bias*) affecte positivement notre statistique, augmentant le pouvoir de détection de notre test. Malgré ces résultats encourageants, l'*ascertainment bias* reste un problème à considérer dans certains jeux de

données tel que celui du projet HapMap où les protocoles de recrutement de SNPs diffèrent selon la région génomique étudiée. Dans certaines régions, les panels d'individus dans lequel les SNPs ont été découverts sont plus grands que dans d'autres. Par exemple, les SNPs des données HapMap dans des régions de 500Kb ciblées par le projet ENCODE sont issus d'un re-séquençage d'environ 25% des individus, plus important que dans la plupart des autres régions. Ces régions présenteront, théoriquement, une proportion plus importante de SNPs rares et la diversité haplotypique sera augmentée. Il est difficile d'évaluer actuellement quels sont les effets de cette hétérogénéité dans les données HapMap sur Svd, mais les problèmes causés par l'*ascertainment bias* continuent d'entraver les analyses. Finalement, notons qu'il existe des jeux de données, tel que les données Perlegen [Hinds, 2005], dont le processus de recrutement des SNPs est connu et ne varient pas le long des séquences.

Deux facteurs affectent particulièrement la détection des événements de sélection par notre statistique : le biais d'haplotypage et les taux de recombinaison le long des séquences. La première est, comme l'*ascertainment bias*, un biais introduit lors de la génération de données génomiques. L'inférence des haplotypes à partir des génotypes est une étape importante, cependant, le pouvoir de détection de notre approche est sensible aux erreurs de reconstruction des haplotypes. Cela est dû au fait que notre approche se base sur la diversité globale des haplotypes pour détecter la sélection. Nous avons montré sur des données simulées que considérer de larges haplotypes pour calculer les valeurs de Svd nous permet d'augmenter considérablement le pouvoir de détection d'un *selective sweep* partiel. Il se peut que, lorsque les haplotypes considérés contiennent peu de SNPs une erreur de phasage affecte significativement la diversité globale des haplotypes, mais que ces erreurs soient diluées dans la quantité importante d'information présente dans des haplotypes de plusieurs centaines de SNPs. Des analyses plus poussées méritent d'être menées pour nous permettre de mieux comprendre les raisons précises d'une telle amélioration du pouvoir de détection.

Considérer de longs haplotypes lors de l'analyse de données génomiques va clari-

fier les signaux de sélection positive dans des données phasées, mais ceux-ci risquent d'être fortement fragmentés par la recombinaison. En effet, un *selective sweep* partiel laissera toujours une marque plus distincte dans une région où le taux de recombinaison est faible. C'est d'ailleurs ce qui se passe dans la région du gène de lactase chez les européens : ce gène se trouve dans une région où le taux de recombinaison est très bas et où aucun hotspot ne fragmente les haplotypes. Cela explique pourquoi le signal de sélection positive ressort aussi clairement, dans un grand nombre d'études. Dans une région ayant subit un *selective sweep* récent, il n'est pas étonnant que les mesures du taux de recombinaison présentent de faibles valeurs : par rapport à d'autres régions sous neutralité où elle agit sans cesse depuis un grand laps de temps, la recombinaison n'a pas encore eu le temps d'agir pour « casser » les haplotypes. Malgré tout, bien que sous neutralité la recombinaison n'affecte pas la distribution des valeurs de Svd (voir Fig. 3.1), nos résultats montrent que les hotspots de recombinaison diminuent le pouvoir de détection de la statistique. Plusieurs manières de traiter ce problème nous apparaissent valables. La première est d'utiliser des valeurs critiques de rejet de l'hypothèse nulle qui seraient fonction du taux de recombinaison. Cela se fait communément lors de l'analyse d'une région candidate, dont les caractéristiques sont connues, mais s'applique difficilement à la détection de signaux de sélection par scan génomique. Une variante de cette idée, pour analyser les résultats de scans génomique en tenant compte du taux de recombinaison consiste à ne comparer entre eux que les locus présentant des taux de recombinaison similaires. Cependant, puisque les recombinaisons surviennent principalement dans les hotspots, il paraît difficile d'établir un critère fiable pour estimer la similarité des taux de recombinaison locaux dans de longues régions génomiques. Parce que nous pensons que ce sont principalement les hotspots qui vont fragmenter le signal de sélection dans les haplotypes, nous avons testé une méthode qui prend l'intensité des hotspots en considération pour déterminer la longueur adéquate, en nombre de SNPs, des haplotypes à considérer lors du scan génomique (cette méthode est décrite à la section 2.3.2). Il est important de tester précisément, sur plusieurs jeux de

données, ces différentes approches afin d'opter pour celle qui est la plus puissante avec notre statistique.

Le calcul de la statistique Svd nécessite l'identification de l'allèle ancestral à chaque SNP car un des termes de l'expression de Svd, à un site donné, est la fréquence de l'état dérivé à ce site (la fréquence de la mutation). Le signal ne ressortira que lorsque l'allèle dérivé est arrivé à une fréquence intermédiaire ou haute. Les simulations nous ont permis de vérifier que ce terme améliore grandement le pouvoir de détection de la statistique, particulièrement dans les cas, pas si rares, où les haplotypes porteurs d'un allèle majeur et ancestral présentent par hasard une faible variance des HAC. À cause de la pondération de la différence des variances par la fréquence de la mutation, les SNPs dont les allèles ancestraux sont liés à la mutation sélectionnée ne donneront pas de signal. Pour analyser les résultats d'un scan génomique, une approche prometteuse, utilisée par Voight et collaborateurs [2006] est d'évaluer des regroupements de signaux. Nous étudions actuellement différentes possibilités pour déterminer la meilleure façon de faire ressortir les regroupements de SNPs exhibant de hautes valeurs de Svd mais l'absence de signal lorsque les allèles ancestraux sont majeurs devra être considérée. La nécessité d'identifier l'état ancestral pose un autre problème, récemment soulevé par Hernandez et collaborateurs [2007]. Ils montrent que, dans les données de genotypage, on observe un excès global d'allèles dérivés à haute fréquence, et ceci même dans les régions non-codantes, ou la sélection positive agit rarement. Cela peut être causé par la mauvaise identification de l'état ancestral. En effet la méthode d'identification de l'ancestral qui consiste à comparer les séquences génomiques humaines avec des séquences orthologues provenant une espèce comme le chimpanzé se base sur le modèle ISM. Ce modèle suppose que chaque nouvelle mutation se produit à un nouveau site, jamais touché par le processus de mutation, et qu'il n'y a pas de *back mutations*. Si l'on postule que ces *back mutations* peuvent survenir, alors l'allèle présent chez le chimpanzé ne sera pas nécessairement l'allèle ancestral, ce qui va provoqué l'identification erronée de l'état ancestral. Une proportion conséquente d'allèles dérivés rares seront alors pris pour

des allèles dérivés très fréquents et le patron de variation résultant sera similaire à l'empreinte laissée par un *selective sweep* partiel, ce qui va confondre les tests basés sur le spectre de fréquences. L'impact de cette identification erronée de l'allèle ancestral sur le pouvoir de détection de Svd n'est pas clair et devra être testé sur des données simulées.

Analyse des résultats

Lorsque les points mentionnés ci dessus seront pris en considération de manière adéquate (effets démographiques, *surfing mutations, ascertainment bias*, erreurs de phasage, taux de recombinaison, regroupement de signaux et identification de l'ancestral) il conviendra d'utiliser Svd pour réaliser un scan génomique complet du génome humain, sur des données génomiques qui conviendront (données HapMap, Perlegen, Projet des 1000 génomes [G., 2008], ou autres).

Il sera alors important de déterminer quels résultats sont statistiquement significatifs. Pour réaliser cette tâche, des méthodologie, basées sur deux écoles de pensée, sont présentées dans la littérature : les approches basée sur des modèles d'une part et les approches empiriques d'autre part. Sabeti et collaborateurs [2006] soutiennent que, bien qu'il soit nécessaire d'avoir de bons modèles théoriques pour interpréter les résultats des tests, il est crucial d'utiliser la distribution empirique génomique des valeurs pour déterminer si les résultats pour un locus sont significatifs. En effet, dans un scan génomique, l'information préalable propre à chaque locus ne peut être intégrée. L'idée est d'éviter de spécifier un modèle statistique pour identifier les cibles potentielles de sélection, ceux-ci pouvant engendrer des erreurs d'interprétation (voir par exemple le commentaire à ce sujet, à propos du gène ASPM chez l'humain, publié par Yu et collaborateurs [2007]). Il convient d'utiliser plutôt une approche strictement empirique pour rechercher les locus présentant des signaux atypiques, qualifiés d'*outliers*. On va déterminer la valeur critique c du test basé sur Svd (voir section 2.2.5) en utilisant les valeurs produites par le scan génomique

au lieu de les déduire d'un modèle. D'un autre côté, Nielsen et collaborateurs [2007] soutiennent que, même si les approches empiriques vont mener à l'identification d'un nombre important de candidats intéressants, cet ensemble de gènes ne consiste pas nécessairement en une représentation fiable des cibles de la sélection positive à travers le génome humain, biaisé par les hypothèses spécifiques utilisées pour définir les *outliers*. Ces conclusions se basent sur l'étude faite par Teshima et collaborateurs [2006], qui ont testé la fiabilité des approches empiriques en utilisant des statistiques sommaires des spectres de fréquences par site (estimateurs de θ, D de Tajima [1989], D de Fu et Li [1993], H de Fay et Wu [2000]) et par haplotype (calcul de l'homozygotie haplotypique). Il serait intéressant de tester la fiabilité d'une approche empirique utilisant Svd avec la méthode et les données utilisées par Teshima et collaborateurs [2006] afin de tester si les résultats se maintiennent avec notre statistique. Une alternative possible pour déterminer la signification statistique des résultats est évoquée par ces auteurs. L'idée est d'utiliser une distribution nulle qui ne soit pas basée sur un modèle théorique défini mais sur les sites connus pour être neutres dans le génome (tels que les sites des régions non-codantes ou les sites synonymes) et pas sur un modèle neutre.

Compte tenu des éléments que nous possédions, nous pensons que déterminer la signification statistique par l'approche empirique d'identification d'*outliers* est l'approche la mieux adaptée pour retrouver des locus candidats. Une bonne méthode d'identification de regroupement des signaux et des limites adéquatement déterminée devront cependant être spécifiée afin d'obtenir un ensemble d*outliers* donnant une bonne vue d'ensemble des cibles de sélection. Lorsqu'un candidat est identifié, cependant, il est préférable d'attribuer des p-valeurs aux valeurs observées afin de valider l'*outlier*. Ces p-valeurs sont estimées en comparant les valeurs de Svd du jeux de données génomiques à la distribution issue de jeux de données simulées grâce aux informations préalables que nous pourrions avoir sur le locus particulier dans la population et les données en question (taux de recombinaison, démographie, *ascertainment bias* dans les données étudiées, etc.).

Améliorations et développements

Certaines améliorations et ajouts à apporter à notre étude ont été proposées et discutées lors de la rédaction de l'article, de ce mémoire et de la présentation des résultats en colloque.

La première concerne la statistique elle-même. Nous avons construit une statistique qui permet de comparer les distributions des HAC pour les haplotypes porteurs de l'allèle sous sélection et les haplotypes non porteurs en utilisant la différence des variances. Cela a été motivé par le fait que la distribution des HAC est plus étroite lorsque la variation a été affectée par un *selective sweep*. Or, cette distribution est non seulement plus étroite, mais également plus proche de zéro. En effet, lors d'un *selective sweep* partiel, les allèles liées à l'allèle avantagé par sélection positive voient leur fréquence augmentée, et deviennent des allèles majeurs, présents sur l'haplotype de référence. De ce fait, les haplotypes porteurs de l'allèle sélectionné ne portent que très peu d'allèles mineurs, aucun dans certains cas, comme pour le cas du lactase chez les européens. La classe allélique d'haplotypes de $k = 0$ (où k est le nombre d'allèles mineurs) contient alors un nombre d'haplotypes beaucoup plus important que ce qui est attendu sous neutralité (voir Fig. 2.3). Il serait donc intéressant de non seulement considérer l'étendue de la distribution mais aussi le déplacement de la distribution vers la gauche. Une idée proposée serait de calculer également la médiane empirique m des HAC dans les deux sous-échantillons pour obtenir la variable de décision suivante :

$$SvdBIS_i = \frac{v(HAC_{r,i}) \cdot g(m(HAC_{r,i})) - v(HAC_{R,i}) \cdot g(m(HAC_{r,i}))}{S} \cdot f_{d,i} \quad (4.1)$$

$$\text{avec la variance } v(HAC) = \frac{1}{n}(\sum_{k=0}^{S} k^2 C_k - (\sum_{k=0}^{S} k C_k)^2)$$

et g une fonction de la médiane $m(HAC)$. La médiane, qui divise en deux l'échan-

tillon des haplotypes, semble plus adéquate pour quantifier le déplacement de la distribution que la moyenne arithmétique, car elle permet d'atténuer l'influence perturbatrice de valeurs extrêmes obtenues dans des circonstances exceptionnelles.

Une deuxième proposition intéressante est de développer une méthode statistique, basée sur les HAC, pour retrouver un *selective sweep* complet en utilisant le même haplotype de référence. En effet, la figure 2.3 montre qu'un *selective sweep* complet déforme aussi, de façon importante, la distribution des HAC par rapport à celle attendue sous neutralité. Lorsque la mutation sous sélection positive se fixe dans une population, les allèles qui lui sont génétiquement liés vont aussi atteindre une fréquence de 1 et les autres mutations seront perdues (voir Fig. 1.3). La distribution des HAC sera plus étroite, à cause de cette perte de variabilité génétique. Le processus de mutation continue d'agir après le *sweep* et restaure lentement la diversité, ce qui va entraîner la présence d'un excès d'allèles dérivés rares sur les haplotypes et tous les allèles ancestraux sont ainsi majeurs. Parce que la plupart des séquences ne présentent pas de mutation du tout, le nombre d'haplotypes dans la classe allélique d'haplotypes de $k = 0$ (où k est le nombre d'allèles mineurs, tous dérivés) est plus important que celui attendu sous neutralité. La plupart des mutations sont apparues sur des haplotypes qui ne portaient pas d'allèle mineur, ce qui explique que la classe allélique d'haplotypes de $k = 1$ contienne aussi plus d'haplotypes qu'attendus sous neutralité. L'analyse de données simulées sous un modèle de *selective sweep* complet suggèrent que N_e générations après la fixation de l'allèle (avec N_e l'effectif efficace de la population), la distribution des HAC s'apparente à nouveau à la distribution attendue sous neutralité. Ce résultat concorde avec l'étude faite par Przeworski [2002], qui prédit sous des hypothèses simplificatrices, qu'on s'attend à ce qu'un allèle sous sélection positive déforme les patrons de variations durant approximativement N_e générations.

Ces résultats suggèrent que les HAC pourraient également être utile pour détecter des *selective sweeps* complets chez l'Homme moderne. Svd n'est pas une statistique adaptée pour cette tâche, puisque son calcul nécessite, pour détecter les événements

de sélection, la présence de deux allèles pour la comparaison des haplotypes dans une population. Une nouvelle statistique sommaire résumant l'impact d'un *selective sweep* complet sur les HAC est donc requise. Pour retrouver de tels événements de sélection, deux populations sont requises : une population test, où l'on recherche un signal, et une population de référence. Kimura et collaborateurs [2007] présentent une approche utilisant deux populations pour retrouver des allèles fixés par sélection positive. Ils définissent deux indices : l'homozygotie haplotypique qui est la probabilité que deux haplotypes tirés au hasard dans la population soient identiques et l'homozygotie haplotypique de l'haplotype le plus fréquent qui est la probabilité que deux haplotypes tirés au hasard soient identiques à l'haplotype le plus fréquent. En se basant sur ces deux indices, deux statistiques sont développées pour détecter des changements dans les fréquences haplotypiques entre les populations et/ou une baisse de la diversité haplotypique dans la population test.

Une idée serait d'utiliser une approche similaire à celle développée par Kimura et collaborateurs [2007], en utilisant les HAC comme mesure de diversité haplotypique, au lieu de l'homozygotie haplotypique.

Études fonctionnelles

L'étude de la sélection naturelle dans les génomes des êtres vivants et particulièrement de l'Homme, est une problématique de biologie fondamentale d'une part et de biologie appliquée d'autre part.

Biologie fondamentale, car cette problématique provient d'une curiosité naturelle de comprendre le passé évolutif de l'Homme et les mécanismes qui ont menés, par évolution moléculaire, à la grande diversité d'espèces que nous côtoyons dans le monde.

Biologie appliquée, car les traits sélectionnés en eux-mêmes sont d'un grand intérêt en génétique médicale par exemple. Les instances particulières de gènes, ou régions génomiques, identifiées comme des cibles potentielles de la sélection naturelle

doivent donc être validées, puis caractérisées et finalement décodées.

Ces tâches requièrent la collaboration de plusieurs approches méthodologiques pour parvenir à « disséquer » les locus candidats, telle que l'expérimentation biochimique, la génomique comparative, des études d'association génétique dans les populations, etc. Pour un seul locus, cela peut demander beaucoup d'effort et de temps. Par exemple, comprendre la relation entre les propriétés biochimiques des drépanocytes et la malaria a constitué un travail laborieux [Kwiatkowski, 2005]. La compréhension des traits sous sélection n'est pas une tâche évidente à réaliser pour plusieurs raisons. D'abord la sélection peut agir sans être phénotypiquement détectable. En effet, elle peut agir sous certaines conditions uniquement, de sorte que les effets sur un phénotype peuvent être difficilement mesurables. Ensuite, alors qu'un locus doit avoir une importance fonctionnelle pour être sous sélection, il est faux de penser que les différences fonctionnelles que l'on peut observer démontre une action passée ou présente de la sélection [Gould and Lewontin, 1979]. Les arguments de Gould et Lewontin [1979], qui démontrent cela sont oubliés ou peu connus par la nouvelle génération de biologiste de l'évolution, essentiellement formée en biologie moléculaire, génomique et bio-infomatique [Nielsen et al., 2007].

Conclusion

L'importance attribuée à la sélection naturelle comme facteur d'évolution est discutée depuis l'époque de Darwin. Bien qu'il ait convaincu la communauté scientifique que les espèces évoluent dans le temps, le concept de sélection naturelle n'était pas très populaire à son époque. Il a fallu attendre les travaux des années 1930-40 pour que ce concept soit vu comme une force majeure de l'évolution. Aujourd'hui encore, les opinions sont partagées : certains pensent que cette force explique presque qu'entièrement la diversité complexe du monde vivant alors que d'autres pensent qu'elle n'a que très peu d'importance. Le problème vient du fait que les réalisations de la sélection naturelle, réelles ou supposées, agissent sur des échelles de temps très larges et ne sont pas accessibles à nos sens d'humains : nous traitons donc un problème « invisible » [Macbeth, 1971, Bergman, 1992].

Un des points important à retenir de ce mémoire est que, de nos jours, la sélection naturelle ne constitue plus un problème invisible. La bio-informatique, en faisant le lien entre les théories mathématiques et les données réelles disponibles grâce aux nouvelles ressources en biologie moléculaire, nous permet maintenant d'étudier cette question de façon plus structurée pour obtenir des réponses de plus en plus indiscutables. Il ne reste plus qu'à trouver les bonnes façons de le faire.

Pour participer à cet effort, le travail décrit dans ce mémoire propose une nouvelle façon de regarder la diversité génétique et d'analyser les données, basée sur la fusion de deux modèles de mutation classiques, les modèles ISM et IAM. J'ai présenté ici un exemple d'application de l'approche à la détection de sélection naturelle dans les données génomiques, premier témoin du potentiel de ce type d'analyse dans les études en génétique des populations.

Nous avons d'abord construit une statistique sommaire, Svd, susceptible de pouvoir détecter des *selective sweeps* récents et en cours dans les données de polymor-

phismes, et nous avons démontré qu'elle en était capable sur des données simulées. Cette première validation nous a permis de développer un test statistique qui utilise Svd comme variable de décision, et nous l'avons appliqué à des données génomiques. La méthode développée, à partir des classes allélique d'haplotypes, est une approche statistique simple, intuitive et facile à implémentée.

Nous nous sommes servis du système génétique, aujourd'hui célèbre chez les généticiens, de l'intolérance au lactose pour démontrer le potentiel de notre méthode sur des données génomiques. Il s'agit là de la meilleure illustration de l'étude d'un trait qui, initialement basée sur aucune information génétique, a bénéficié des données génomiques. Notre méthode retrouve avec succès le signal remarquable de sélection positive à ce locus, prédit par les recherches précédentes dans la population européenne, et nous indique même la mutation sélectionnée dans l'haplotype sous sélection.

L'importance du rôle de la sélection naturelle, façonnant les patrons de variations dans le génome humain, est de mieux en mieux mis en évidence. Le scan génomique qui suivra cette étude va mener à l'identification de locus qui s'ajouteront à ceux de nombreuses études. Cette explosion de résultats suggèrent que la sélection positive est plus commune que ce qui n'était supposé auparavant. Si les preuves s'accumulent pour dire que la sélection positive domine les patrons de variabilité génétique, c'est la théorie neutraliste de l'évolution qui sera remise en cause, au profit de modèles modernes comme celui de Gillepsie [2000]. Celui-ci est basé sur le fait que les sites sous sélection positive causent des fluctuations de fréquences chez les sites liés, comme le décrit le modèle de *selective sweep*. De ce fait, les changements de fréquences dans les régions génomiques ne seraient pas principalement le résultat de la dérive génétique, mais résulteraient également des effets de la sélection de locus voisins. La question maintenant est de savoir si ce type de modèle est plus approprié pour expliquer les niveaux de variation génétique chez l'humain, et il semble qu'étudier les *selective sweeps* partiels, en action actuellement, sera une stratégie de choix dans les prochaines années.

Bibliographie

[Altschul et al., 1990] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biology*, 215:403–410.

[Andolfatto et al., 1999] Andolfatto, P., Wall, J., and Kreitman, M. (1999). Unusual haplotype structure at the proximal breakpoint of in(2l)t in a natural population of drosophila melanogaster. *Genetics*, 153:1297– 1311.

[Andrés et al., 2007] Andrés, A., Clark, A., Shimmin, L., Boerwinkle, E., and Sing, C. e. a. (2007). Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.*, 31:659–671.

[Ardlie et al., 2002] Ardlie, K., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3:299–309.

[Balding et al., 2001] Balding, D., Bishop, M., and Cannings, C. (2001). *Handbook of Statistical Genetics*. John Wiley and Sons, Ltd.

[Bergman, 1992] Bergman, J. (1992). Some biological problems with the natural selection theory. available at `http://www.rae.org/natsel.html`.

[Bersaglieri et al., 2004] Bersaglieri, T., Sabeti, P., Patterson, N., Vanderploeg, T., Schaffner, S., Drake, J., Rhodes, M., Reich, D., and Hirschhorn, J. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, 74:1111–1120.

[Boll et al., 1991] Boll, W., Wagner, P., and Mantei, N. (1991). Structure of the chromosomal gene and cdnas coding for lactase-phlorizin hydrolase in humans

with adult-type hypolactasia or persistence of lactase. *Am. J. Hum. Genet.*, 48:889–902.

[Carlson et al., 2005] Carlson, C., Thomas, D., Eberle, M., Swanson, J., Livingston, R.J. Rieder, M., and Nickerson, D. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res*, 15:1553–65.

[Clark et al., 2005] Clark, A., Hubisz, M., Bustamante, C., Williamson, S., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, 15:1496–1502.

[Consortium, 2004] Consortium, I. H. (2004). Integrating ethics and science in the international hapmap project. *Nat Rev Genet*, 5:467–475.

[Consortium, 2005] Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, 437:1299–1320.

[Darwin, 1859] Darwin, C. (1859). *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray.

[Durrett, 1999] Durrett, R. (1999). *Essentials of Stochastic Processes*. Springer.

[Edgar, 2004] Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, 32:1792–97.

[Edmonds et al., 2004] Edmonds, C., Lillie, A., and Cavalli-Sforza, L. (2004). Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci.*, 101:975–9.

[Enattah et al., 2008] Enattah, N., Jensen, T., Nielsen, M., and Lewinski, R. e. a. (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.*, 82:57–72.

[Enattah et al., 2002] Enattah, N., Sahi, T., Savilahti, E., Terwilliger, J., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30:233–237.

[Ewens, 1972] Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112.

[Excoffier and Schneider, 1999] Excoffier, L. and Schneider, S. (1999). Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc. Natl. Acad. Sci.*, 96:10597–10602.

[Fay and Wu, 2000] Fay, J. and Wu, C. (2000). Hitchhiking under positive darwinian selection. *Genetics*, 155:1405–1413.

[Fearnhead and Donnelly, 2002] Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *JRSS Series B*, 64:657–680.

[Fisher, 1930] Fisher, R. (1930). *The genetical theory of natural selection.* Clarendon.

[Freeman, 2006] Freeman, J. e. a. (2006). Copy number variation: New insights in genome diversity. *Genome Res.*, 16:949–961.

[Fu, 1995] Fu, Y. (1995). Statistical properties of segregating sites. *Theor Popul Biol*, 48:172–197.

[Fu, 1997] Fu, Y. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147:915–925.

[Fu and Li, 1993] Fu, Y. and Li, W. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133:693–709.

[G., 2008] G., S. (2008). International consortium announces the 1000 genomes project. available at `http://www.1000genomes.org/files/1000Genomes-NewsRelease.pdf`.

[Gillespie, 2000] Gillespie, J. (2000). Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, 155:909–919.

[Gould and Lewontin, 1979] Gould, S. J. and Lewontin, R. C. (1979). Spandrels of san-marco and the panglossian paradigm — a critique of the adaptationist program. *Proc. R. Soc. London Series B Biol. Sci.*, 205:581–598.

[Haldane, 1932] Haldane, J. B. S. (1932). *The causes of evolution*. Longman, New York.

[Hardy, 1908] Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28:49–50.

[Harris and Meyer, 2006] Harris, E. and Meyer, D. (2006). The molecular signature of selection underlying human adaptations. *American Journal of Physical Anthropology*, 131:89–130.

[Hernandez et al., 2007] Hernandez, R., Williamson, S., and Bustamante, C. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol.*, 24:1792–800.

[Hinds, 2005] Hinds, D. e. a. (2005). Whole-genome patterns of common dna variation in three human populations. *Science*, 307:1072–1079.

[Holden and Mace, 1997] Holden, C. and Mace, R. (1997). A phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol*, 69:605–628.

[Hollox et al., 2001] Hollox, E., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A., and Swallow, D. (2001). Lactase haplotype diversity in the old world. *Am J Hum Genet.*, 68:160–172.

[Hudson, 2001] Hudson, R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159:1805–1817.

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

[Hudson and Kaplan, 1988] Hudson, R. and Kaplan, N. (1988). The coalescent process in models with selection and recombination. *Genetics*, 120:831–840.

[Hudson et al., 1987] Hudson, R., Kreitman, M., and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116:153–159.

[Hughes et al., 1988] Hughes, A., Nei, M., and Aguade, M. (1988). Pattern of nucleotide substitution at mhc class i loci reveals overdominant selection. *Nature*, 335:167–170.

[Karlin and Taylor, 1981] Karlin, S. and Taylor, H. (1981). *A Second Course in Stochastic Processes*. Academic Press, San Diego, CA.

[Keeton and Gould, 1996] Keeton, W. T. and Gould, W. W. (1996). *Biological Science, 6th ed.* W. W. Norton and Co., New York, N.Y.

[Kim and Nielsen, 2004] Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167:1513–1524.

[Kimura, 1968] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.

[Kimura, 1983] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. The Neutral Theory of Molecular Evolution. Cambridge University Press.

[Kimura and Crow, 1964] Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738.

[Kimura et al., 2007] Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS one*, 2:e286.

[Kingman, 1982] Kingman, J. (1982). The coalescent. *Stochastic processes and their application*, 13:235–248.

[Klopfstein et al., 2006] Klopfstein, S., Currat, M., and Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol.*, 23:482–90.

[Kuhner et al., 2000] Kuhner, M., Yamato, J., and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–1401.

[Kwiatkowski, 2005] Kwiatkowski, P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet.*, 77:171–192.

[Labuda et al., 2007] Labuda, D., Labbé, C., Langlois, S., Lefebvre, J.-F., and Freytag, V. e. a. (2007). Patterns of variation in dna segments upstream of transcription start sites. *Human mutation*, 0:1–10.

[Lander, 2001] Lander, E. e. a. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

[Lefebvre and Labuda, 2008] Lefebvre, J.-F. and Labuda, D. (2008). Fraction of informative recombinations: a heuristic approach to analyze recombination rates. *Genetics*, 178:2069–2079.

[Lewinsky et al., 2005] Lewinsky, R., Jensen, T., Moller, J., Stensballe, A., Olsen, J., and Troelsen, J. (2005). T-13910 dna variant associated with lactase persistence interacts with oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet*, 14:3945–3953.

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

[Li and Stephens, 2003] Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233.

[Macbeth, 1971] Macbeth, N. (1971). *Darwin retried; an appeal to reason.* Gambit. Boston.

[Mace et al., 2003] Mace, R., Jordan, F., and Holden, C. (2003). Testing evolutionary hypotheses about human biological adaptation using cross-cultural comparison. *Hum Biol*, 69:605–628.

[Manica et al., 2007] Manica, A., Amos, W., Balloux, F., and Hanihara, T. (2007). The effect of ancient population bottlenecks on human phenotypic variation. *Nature*, 448:346–348.

[Marchini et al., 2006] Marchini, J., Cutler, D., Patterson, N., Stephens, M., and Eskin, E. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, 78:437–450.

[McDonald and Kreitman, 1991] McDonald, J. and Kreitman, M. (1991). Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351:652–654.

[McVean et al., 2002] McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–1241.

[Mulcare et al., 2004] Mulcare, C., Weale, M., Jones, A., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow, D., Bradman, N., and Thomas, M. (2004). The t allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (lct) (c-13.9kbt) does not predict or cause the lactase-persistence phenotype in africans. *Am. J. Hum. Genet.*, 74:1102–1110.

[Myers et al., 2005] Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324.

[Nachman and Crowell, 2000] Nachman, M. and Crowell, S. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304.

[Neuhauser and Krone, 1997] Neuhauser, C. and Krone, S. (1997). The genealogy of samples in models with selection. *Genetics*, 145:519–534.

[Nielsen, 2001] Nielsen, R. (2001). Statistical test of selective neutrality in the age of genomics. *Heredity*, 86:641–647.

[Nielsen, 2005] Nielsen, R. (2005). Molecular signature of natural selection. *Ann. Rev. Genet.*, 39:197–218.

[Nielsen et al., 2007] Nielsen, R., Hellman, I., Hubisz, M., Bustamente, C., and Clark, A. (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet*, 8:857–868.

[Nielsen et al., 2004] Nielsen, R., Hubisz, M., and Clark, A. (2004). Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, 168:2373–2382.

[Nielsen and Signorovitch, 2003] Nielsen, R. and Signorovitch, J. (2003). Correcting for ascertainment biases when analyzing snp data: applications to the estimation of linkage disequilibrium. *Theor Popul Biol.*, 63:245–55.

[Pritchard and Przeworski, 2001] Pritchard, J. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.*, 69:1–14.

[Przeworski, 2002] Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, 160:1179–1189.

[Rozin and Pelchat, 1988] Rozin, P. and Pelchat, M. (1988). Memories of mammaries: adaptations to weaning from milk. epstein a. n. morrison a. r. eds. *Progress in Psychobiology and Physiological Psychology*, 1988:1–29.

[Sabeti et al., 2002] Sabeti, P., Reich, D., Higgins, J., and Levine, H. e. a. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837.

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

[Scheet and Stephens, 2006] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78:629–644.

[Spencer and Coop, 2004] Spencer, C. and Coop, G. (2004). Selsim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 20:3673–3675.

[Stephens et al., 2001] Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68:978–989.

[Swallow, 2003] Swallow, D. (2003). Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet*, 37:197–219.

[Tajima, 1983] Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105:437–460.

[Tajima, 1989] Tajima, F. (1989). Statistical methods to test for nucleotide mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595.

[Tang et al., 2007] Tang, K., Thornton, K., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *Plos Biology*, 5:e171.

[Tenesa et al., 2007] Tenesa, A., Navarro, P., Hayes, B., Duffy, D., Clarke, G., Goddard, M., and Visscher, P. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17:520–6.

[Teshima et al., 2006] Teshima, K., Coop, G., and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Res.*, 16:702–712.

[Tishkoff et al., 2007] Tishkoff, S., Reed, F., Ranciaro, A., Voight, B., and Babbitt, C. (2007). Convergent adaptation of human lactase persistence in africa and europe. *Nature Genetics*, 39:31–40.

[Voight et al., 2006] Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS Biol*, 4:e72.

[Wang et al., 2006] Wang, E., Kodama, G., Baldi, P., and Moyzis, R. (2006). Global landscape of recent inferred darwinian selection for homo sapiens. *Proc Natl Acad Sci.*, 1:135–40.

[Watterson, 1975] Watterson, G. (1975). On the number of segregation sites. *Theor Popul Biol*, 7:256–276.

[Weinberg, 1908] Weinberg, W. (1908). Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64:368–382.

[Williamson et al., 2007] Williamson, S., Hubisz, M., Clark, A., Payseur, B., Bustamante, C., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *Plos Genetics*, 3:e90.

[Wright, 1931] Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

[Yu et al., 2007] Yu, F., Hill, R., Schaffner, S., Sabeti, P., Wang, E., Mignault, A., R.J., F., Moyzis, R., Walsh, C., and Reich, D. (2007). Comment on « ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens ». *Science*, 316:370.

[Zeng et al., 2006] Zeng, K., Fu, Y., Shi, S., and Wu, C. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 155:1405–1413.

[Zhang et al., 2006] Zhang, C., Dione, K., Awad, T., and Guoying, L. e. a. (2006). A whole genome long-range haplotype (wglrh) test for detecting imprints of positive selection in human populations. *Bioinformatics*, 22:2122–2128.