Analyses bioinformatiques dans le cadre de la génomique du SIDA

Cédric Coulonges

le cnam

ÉCOLE DOCTORALE ARTS et MÉTIERS

Laboratoire Génomique Bioinformatique et Applications (GBA)

THÈSE présentée par :

Cédric COULONGES

soutenue le : **16 décembre 2011**

pour obtenir le grade de : Docteur du Conservatoire National des Arts et Métiers

Discipline/ Spécialité : **Bioinformatique**

Analyses bioinformatiques dans le cadre de la génomique du SIDA

THÈSE dirigée par :

M. ZAGURY Jean-François	Professeur, Conservatoire National des Arts et Métiers

RAPPORTEURS :

Mme GUINOT Christiane	Docteur, C.E.R.I.E.S, Université de Tours
Mme TIRET Laurence	Docteur, Inserm U525

JURY :

M. GUEDJ Mickaël	Docteur, Pharnext
M. LATOUCHE Aurélien	Professeur, Conservatoire National des Arts et Métiers

Remerciements

J’exprime toute ma reconnaissance au Pr Jean-François Zagury, mon directeur de thèse, mon patron et mon ami qui m'a accepté dans son équipe, m'a fait découvrir le goût de la science et de la recherche.

Je suis très reconnaissant envers le Dr Christiane Guinot et le Dr Laurence Tiret pour avoir accepté de juger mon travail de thèse en tant que rapporteur.

Je remercie également le Dr Mickaël Guedj et le Pr Aurélien Latouche de m'avoir fait l’honneur de participer au jury de cette thèse.

Je remercie le Conservatoire National des Arts et Métiers (CNAM) et tout son personnel pour m'avoir accueilli ainsi que le département IMATH pour leur disponibilité et leur écoute. Je pense en particulier à Christiane, ma secrétaire préférée.

Je tiens à remercier chaleureusement toute mon équipe dont certains sont devenus mes amis. A mes compagnons de bureaux de ces dernières années, Lieng le « R-addicted », Olivier l'haplotypeur, Taoufik l'expert BDD et Vincent le sélectionneur. A Sigrid la généticienne en chef pour son aide et Sophie pour son expertise. Aux «drug designers», Matthieu, Hélène, Nathalie et Nesrine. Aux « cytokines », Hadley, Rojo et Gabriel. Aux stagiaires que j'ai pu encadrer ou côtoyer. Enfin à Hervé l'inclassable pour tous ces moments de bonne humeur.

Enfin j'ai une pensée toute particulière pour ma « Lu » que j'aime, merci pour ton soutien et pour m'avoir aidé à trouver la motivation.

Merci à mes grands parents, mes parents, mes frères et mes amis et tout ceux qui sont toujours restés à mes cotés.

Résumé

Après 25 ans de recherche intensive, il n'y a pas de vaccin contre le SIDA ni de traitement définitif face à la maladie. Les technologies génomiques permettent aujourd'hui d'explorer la quasi-totalité du génome humain, ceci afin d'identifier les facteurs génétiques influençant le développement de la maladie. Elles peuvent ainsi permettre de mieux comprendre les mécanismes moléculaires de l'infection par VIH-1, ce qui devrait favoriser le développement rationnel de nouvelles stratégies thérapeutiques. La bioinformatique joue un rôle central dans le traitement des données génomiques à l'interface de l'informatique, des statistiques, et de la biologie.

Ma thèse a porté sur l'exploitation des données issues de puces de génotypage dans le cadre du SIDA, et en particulier de la cohorte GRIV (Génétique de la Résistance au VIH-1) de patients à profil extrême du SIDA composée d'une part de patients non-progresseurs à long terme (NP, n=275) et pour une autre part de patients progresseurs rapides (PR, n=86).

L'analyse simple marqueur « classique » des NP de la cohorte GRIV a révélé le rôle majeur de la région HLA dans la non progression à long terme, associée à un contrôle fort de la charge virale. Nous avons donc recherché si des associations génétiques particulières pouvaient expliquer le sous-groupe des NP dits "non elite"(i.e. avec une charge virale >100 copies/mL) et un polymorphisme du gène CXCR6 a été mis en évidence. Cette association avec l'infection VIH-1 est le seul résultat répliqué issu d'études d'association « génome entier », en dehors du locus HLA.

Il a aussi fallu que je me familiarise avec des concepts de bioinformatique allant de l'haplotypage, ce qui m'a conduit à développer le logiciel SUBHAP, jusqu'à l'imputation à grande échelle qui est devenue aujourd'hui une étape obligée. Récemment la mise en place du projet international IHAC (International HIV-1 Acquisition Consortium) a nécessité la conception d'un « pipeline » d'analyse complexe mais efficace et des résultats encore préliminaires pointent vers un nouveau gène, KIF26B, potentiellement associé à l'acquisition du VIH-1.

Il reste encore beaucoup de progrès à faire dans l'exploitation bioinformatique des données génomiques. Les avancées de biologie moléculaire actuelles permettant le séquençage massivement parallèle (Next Generation Sequencing) des génomes vont soulever de nouveaux défis probablement aussi difficiles à gagner que notre espoir d'arriver au développement de nouvelles stratégies diagnostiques et thérapeutiques.

Mots clés : étude d'association « génome entier », VIH-1, SNP, haplotypes, imputation, méta-analyse

Résumé en anglais

After 25 years of intensive research, there is no vaccine and no definitive cure against AIDS. The genomic technologies allow today to explore most of the human genome and identify the genetic factors impacting disease progression. They could thus lead to the understanding the molecular mechanisms of HIV-1 infection, which should contribute to the rational development of new therapeutic strategies. Bioinformatics plays a central role in the treatment of genomic data at the interface of computer science, statistics, and biology.

My PhD thesis has dealt with the exploitation of data from genotyping chips in AIDS, in particular in the GRIV (Genetics of Resistance against Infection by HIV-1) cohort comprising patients with extreme profiles of progression to AIDS, on the one hand long term non progressors (NP, n=275) and on the other hand, rapid progressors (RP, n=86).

The traditional ‘individual marker' analysis revealed the major role of the HLA region in long term non progression, correlated with the strong control of viral load. We have thus looked for genetic associations that could explain the subgroup of NP called “non elite“ (*i.e.* NP with a viral load > 100 copies/mL) and a polymorphism in the CXCR6 gene was found. This new association with HIV-1 infection is the sole replicated result stemming from a “genome wide“ association study outside of the HLA locus.

I have become familiar with various bioinformatics concepts, ranging from haplotyping, that led me to develop the software SUBHAP, to large-scale imputation which is now a mandatory tool. Recently, the setting-up of the IHAC international project (International HIV-1 Acquisition Consortium) has required the design of a complex but efficient analysis « pipeline », and still preliminary results point to the possible association of the KIF26B gene with HIV-1 acquisition.

Much progress remains to be done for the bioinformatics exploitation of genomic data. The current technological breakthroughs allowing the large-scale parallel sequencing (Next generation sequencing) of genomes will raise new challenges likely as difficult to overcome as our hope to develop new diagnostic and therapeutic strategies.

Keywords : GWAS, HIV-1, SNP, haplotypes, imputation, meta-analysis

Table des matières

Liste des figures

Liste des abréviations

ACP	Analyse en Composantes Principales
ACS	Amsterdam Cohort Study
ADN	Acide Désoxyribo-Nucléique
ARN	Acide Ribo-Nucléique
ARNm	Acide Ribo-Nucléique messager
AZT	AZidoThymidine
BAT1	Spliceosome RNA helicase BAT1
CCR	C-C chemokine receptor
CD	Cluster de Différenciation
CDC	Center for Diseases Control
CEPH	Centre d'Etude du Polymorphisme Humain
CNG	Centre National de Génotypage
CNV	Copy Number Variation
CTL	Cytotoxic T Lymphocyte
CXCR	C-X-C récepteur
ddl	degré de liberté
Genevar	GENe Expression VARiation
GRIV	Génomique de la Résistance face à l'Infection par le VIH-1
GWAS	Genome-Wide Association Study
HAART	Highly Active Anti-Retroviral Therapy
HCP5	HLA Complex P5
HEPS	Highly Exposed Persistently Seronegative
HERV	Human Endogenous RetroVirus
HLA	Human Leukocyte Antigen, ou CMH Complexe Majeur d'Histocompatibilité
IFN	Interféron
IL	Interleukine

kb	kilobase
KIF26B	KInesin-like protein Family
LD	Déséquilibre de Liaison
LTNP	Long-Term Non-Progressor
MAF	Minor allele frequency
Mb	mégabase
MCMC	Chaines de Markov de Monte Carlo
Nb	nombre
NIH	National Institutes of Health
NNRTI	Non Nucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
pb	paire de bases
PR	Progresseur Rapide
Q-Q plot	Quantile-Quantile plot
RFLP	Restriction Fragment Length Polymorphism
SHAPEIT	Segmented HAPlotype Estimation and Imputation Tool
SIDA	Syndrome d'Immuno-Déficience Acquise
SIV	Simian Immunodeficiency Virus
SNP	Single Nucleotide Polymorphism
TNF	Tumor Necrosis Factor
VIH	Virus de l'Immuno-déficience Humaine

Introduction

I. Introduction à la génétique épidémiologique

Depuis longtemps, l'homme s'intéresse au fonctionnement de la nature et de son corps. Dans un but d'auto-préservation, nous cherchons à contrôler et comprendre les maladies afin d'en atténuer les conséquences. Depuis la découverte des premiers médicaments jusqu'à la compréhension des mécanismes biologiques sous-jacents, nous n'avons cessé de progresser dans notre connaissance des pathologies. La démarche historique pour découvrir des médicaments et des vaccins - encore utilisée dans l'industrie pharmaceutique aujourd'hui - est d'évaluer expérimentalement les effets de milliers de composés pour choisir les meilleurs candidats à tester par des essais cliniques. En 1953 Watson et Crick allaient mettre à jour une nouvelle molécule, l'ADN, base de notre hérédité qui allait permettre l'avènement de la génétique médicale. Aujourd'hui on sait que toutes les maladies ont une composante génétique du fait des réponses inégales des individus. Comprendre cette composante génétique c'est commencer à comprendre les mécanismes pathogènes de ces maladies et cela devrait permettre ainsi d'envisager rationnellement de nouvelles stratégies thérapeutiques ou diagnostiques.

I.1 Notions de génétique

La génétique est la science qui étudie les gènes et leur transmission. L'ensemble du matériel génétique - appelé génome - est contenu dans chacune des cellules d'un individu sous la forme de chromosomes. Leur nombre est de 23 paires chez l'homme, pour chaque paire un chromosome venant de la mère et l'autre du père. Il y a 22 paires d'autosomes (chromosomes non sexuels) numérotés de 1 à 22 par taille décroissante, et une paire de chromosomes sexuels (XX pour la femme, XY pour l'homme). Chaque chromosome est constitué d'une molécule d'ADN sur laquelle se succèdent 4 nucléotides différents : adénine, cytosine, guanine et thymine aussi notées A,C,G,T. On peut considérer chaque chromosome comme un texte composé de 4 lettres. Certaines séquences d'ADN permettent de coder pour la synthèse de protéines – constituants de base du vivant – au niveau des gènes.

Depuis Mendel au 19e siècle, pionnier de la génétique, jusqu'au séquençage complet des 3 milliards de bases du génome humain en 2003, la génétique n'a cessé de prendre de l'importance dans la compréhension des maladies et une branche de celle ci, la génétique humaine, s'est particulièrement intéressée aux différences qui peuvent exister entre les individus.

I.1.1 Polymorphismes et diversité génétique

Le polymorphisme génétique est défini comme l'existence, au sein des individus d'une même espèce, de formes distinctes (appelées allèles) au niveau d'un locus chromosomique. Chaque polymorphisme est le produit d'un événement de mutation provoquant une modification soudaine et transmissible d'un fragment de l'ADN lorsque cet événement se produit dans une cellule germinale. Les polymorphismes moléculaires dans un gène donné peuvent être simplement neutres ou bien affecter la fonction du gène selon trois modalités : perte de fonction, maintien partiel de la fonction avec interférences, gain de fonction. Ces variations sont transmissibles d'une génération à l'autre. On définit comme haplotypes la combinaison de plusieurs allèles situés sur des locus différents d'un même chromosome. Chaque individu possédant 22 paires d'autosomes, on trouvera à un locus donné une combinaison de deux allèles appelée génotype.

a) Les polymorphismes chromosomiques

Les polymorphismes chromosomiques sont des variations structurales liées ou non à des anomalies phénotypiques. Ces variations sont le résultat d'événements de translocation, inversion, fusion, ou fission de fragments chromosomiques. On peut aussi observer des anomalies portant sur le nombre de chromosomes (par exemple figure 1 avec la trisomie 21).

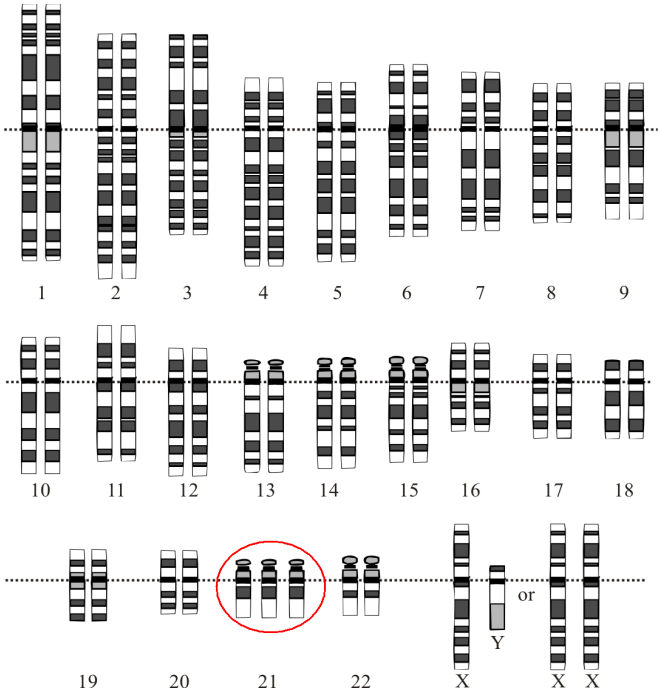


Figure 1 : Représentation schématique des 23 paires de chromosomes chez l'homme, avec une anomalie du nombre d'exemplaires au niveau du chromosome 21

b) Les séquences répétées en tandem

Les séquences répétées en tandem (ou VNTR pour Variable Number of Tandem Repeat), sont de taille variable et constituées de répétitions en tandem d'un motif unitaire de taille également variable. Selon la taille du motif et de la répétition on distingue les satellites, les mini-satellites et les micro-satellites : les micro-satellites sont des motifs de 1 à 5 nucléotides répétés 2 à 50 fois (taille totale < 300 pb). Les mini-satellites sont des motifs de 15 à 100 nucléotides répétés 15 à 50 fois (taille totale entre 1 et 5kb) ; les satellites sont des grands motifs (α : 171, β : 168, et γ : 220 nucléotides respectivement) répétés les uns à la suite des autres.

c) Les indels

Les indels correspondent à une insertion ou une délétion de nucléotides dans la séquence d'ADN (figure 2).

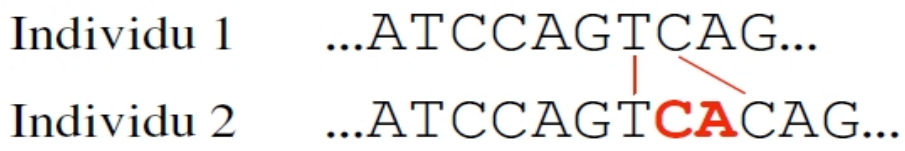


Figure 2 : Représentation d'une insertion de deux nucléotides chez un individu

d) Les SNPs

Les SNPs (Single Nucleotide Polymorphism) sont des variations ponctuelles d'un seul nucléotide (figure 3). Ils sont répartis sur l'ensemble du génome humain et constituent la forme la plus abondante de variation génétique. Selon les dernières données connues du projet 1000 Génomes [1], leur nombre est estimé à environ 40 millions et ils représentent ainsi plus de 90% des différences entre individus. Du fait de leurs propriétés, ils permettent de dresser des cartes génétiques très denses (e.g. dbSNP) et représentent le marqueur génétique actuel de prédilection. Ils peuvent servir de marqueurs génétiques pour identifier un individu [2]. Pour la suite de cette thèse, je vais me focaliser exclusivement sur ce type de marqueur.

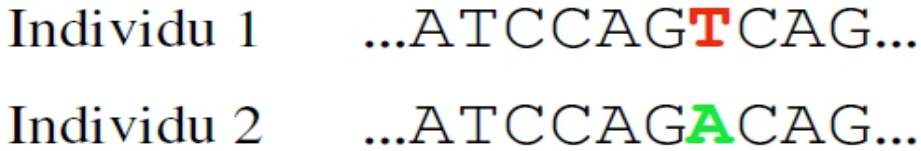


Figure 3 : Représentation du polymorphisme d'un nucléotide entre 2 individus

e) **Les CNV**

Les CNV (Copy Number Variation) sont définis comme des segments d'ADN d'une longueur supérieure à 1 kb et présents dans le génome un nombre de fois variable en comparaison avec un génome de référence. Plus de 12% du génome humain serait concerné par le polymorphisme de nombre de copies géniques. Cette forme de polymorphisme découverte récemment connaît un intérêt croissant et ouvre de nouvelles perspectives dans la recherche de prédisposition ou de susceptibilité à des maladies.

I.1.2 Équilibre de Hardy-Weinberg

L'équilibre de Hardy-Weinberg est un principe fondamental en génétique des populations qui établit que les fréquences alléliques dans une population restent constantes au fil des générations, sous certaines conditions :

1. Population de grande taille
2. Générations non chevauchantes
3. Panmixie, c'est-à-dire dans une population où le croisement des individus est pris au hasard.
4. Non sélection, non mutation et non migration des populations

Dans le cas d'un locus bi-allélique (A, a)

avec les fréquences f_A et $f_a = 1 - f_A$, on a l'équilibre:

$$\begin{cases} f_{AA} = f_A^2 \\ f_{Aa} = 2 f_A f_a \\ f_{aa} = f_a^2 \end{cases}$$

Lorsque l'équilibre de Hardy-Weinberg n'est pas respecté et qu'on observe une déviation significative, il convient de chercher les causes qui ont pu engendrer ce déséquilibre. Celui-ci peut être dû à des problèmes de génétique des populations (dérive, sélection, consanguinité...).

I.1.3 Déséquilibre de liaison

Il faut introduire la notion de recombinaison qui est un phénomène qui se produit par enjambement des chromosomes homologues durant le processus de formation des gamètes qu'on appelle méiose. La probabilité qu'un événement de recombinaison se produise entre deux loci

chromosomiques augmente avec la distance qui les sépare. Le déséquilibre de liaison est l'association non aléatoire des allèles de deux ou plusieurs loci polymorphes sur le même chromosome. Lors de la formation des gamètes, les loci d'un chromosome peuvent être indépendants du fait de la recombinaison et ces loci peuvent donc être transmis de manière indépendante. Cependant, plus les loci sont proches plus la recombinaison est faible. Si nous considérons des loci polymorphes indépendants, les fréquences des combinaisons d'allèles possibles sur un chromosome dans une population, correspondent alors au produit des fréquences de ces allèles, c'est l'équilibre de liaison. On peut formaliser cet équilibre entre de 2 loci bi-alléliques:

Locus A d'allèles A et a de fréquences respectives f_A et f_a

Locus B d'allèles B et b de fréquences respectives f_B et f_b

Il existe 4 combinaisons possibles entre les allèles, les fréquences de ces combinaisons sont les suivantes:

$$\begin{cases} f_{AB} = f_A f_B \\ f_{Ab} = f_A f_b \\ f_{aB} = f_a f_B \\ f_{ab} = f_a f_b \end{cases}$$

Si nous considérons des loci polymorphes qui ne sont pas indépendants, les combinaisons entre les allèles de ces loci ne se font plus au hasard. Les fréquences des combinaisons d'allèles possibles sont alors différentes du produit des fréquences alléliques, c'est le déséquilibre de liaison. Le déséquilibre de liaison peut se mesurer par la valeur du coefficient D de déviation entre les fréquences des combinaisons observées et celles attendues sous l'hypothèse d'indépendance entre les loci. On peut formaliser ce déséquilibre entre 2 loci bi-alléliques:

$$\begin{cases} f_{AB} = f_A f_B - D \\ f_{Ab} = f_A f_b + D \\ f_{aB} = f_a f_B + D \\ f_{ab} = f_a f_b - D \end{cases}$$

Le déséquilibre de liaison se mesure également par une valeur de D normalisée : D' variant entre -1 et 1 ; et par un coefficient r^2 de corrélation variant entre 0 et 1 :

$$D' = \begin{cases} \dfrac{D}{min(f_{AB}, f_{ab})} \; pour \; D \geq 0 \\ \dfrac{D}{min(f_{Ab}, f_{Ba})} \; pour \; D \leq 0 \end{cases} \qquad r^2 = \frac{D^2}{f_A f_B f_a f_b}$$

Pour résumer, $r^2 = 0$ traduit une indépendance entre les allèles des deux SNPs, alors que lorsque $r^2 = 1$, les allèles des deux SNPs sont parfaitement corrélés et systématiquement co-transmis : c'est le déséquilibre de liaison total (*e.g.* A et B sont systématiquement co-transmis -et parallèlement a et b-, et seules les combinaisons AB et ab existent).

I.1.4 Notion d'haplotype

Un haplotype correspond à la combinaison d'allèles de 2 ou plusieurs loci polymorphes sur le même chromosome. Les haplotypes ont été crées au cours de l'évolution au gré des mutations. Ces combinaisons sont le résultat dans la plupart des cas de l'émergence de polymorphismes au sein d'une population et de la recombinaison entre ces loci polymorphes. Le maintien des haplotypes peut aussi être influencé par la sélection naturelle, la dérive génétique et la migration (figure 4).

L'étude des haplotypes est biologiquement pertinente, puisqu'ils constituent un reflet de l'évolution et correspondent à une information plus complexe que les polymorphismes individuels. La problématique des haplotypes réside dans leur reconstruction car les données expérimentales ne permettent pas toujours de déterminer la phase entre allèles des polymorphismes, en d'autres termes déterminer la combinaison d'allèles au sein d'un chromosome.

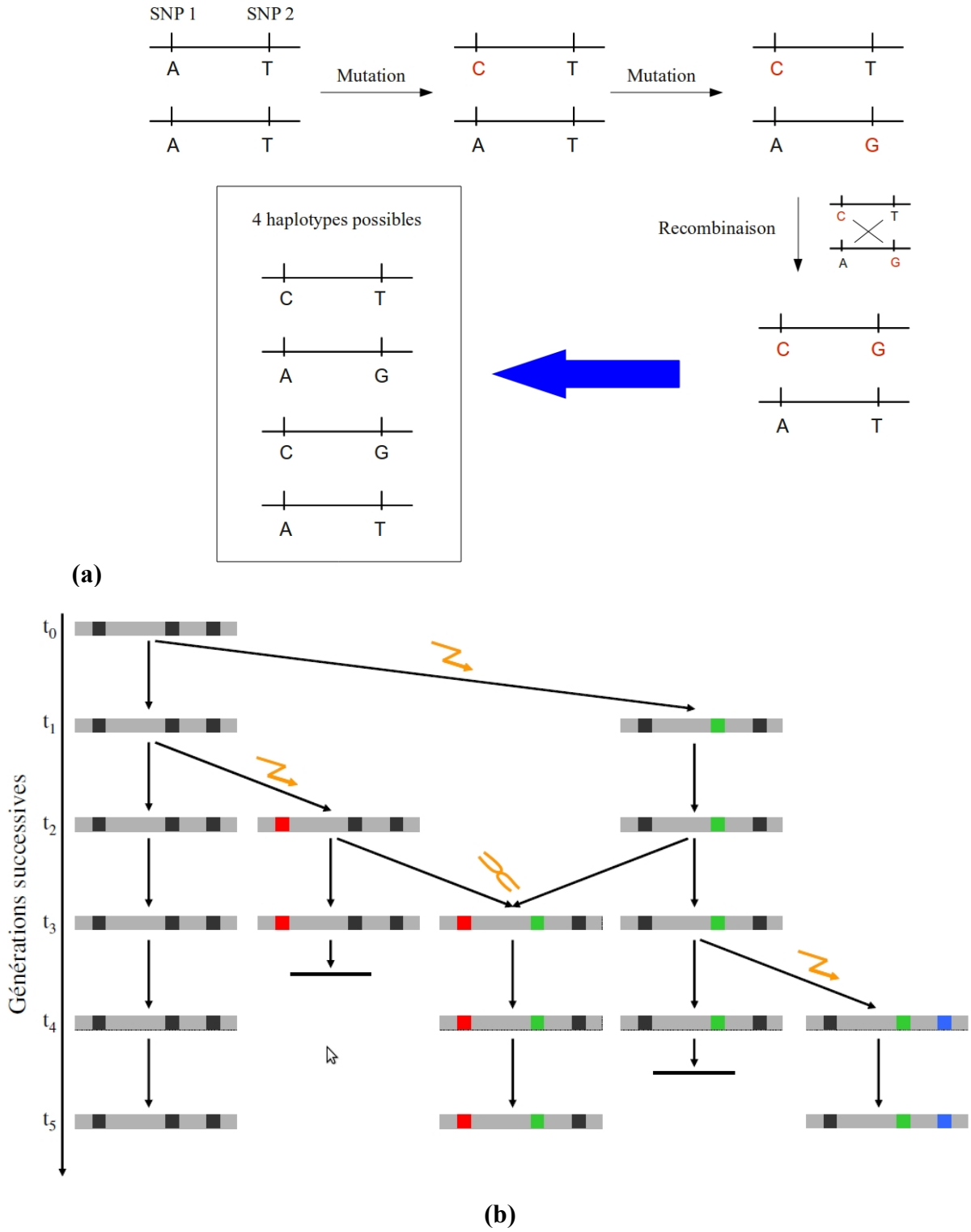


Figure 4 : Représentation schématique de la genèse des haplotypes de deux SNPs **(a)** et de 3 SNPs **(b)** au cours de l'évolution prise à 6 temps précis (t_0-t_5). Les événements successifs de mutation et de recombinaison permettent la création de nouveaux SNPs et de nouveaux haplotypes. Les phénomènes de pression de sélection ou de dérive génétique sont à l'origine de la disparition de deux haplotypes au cours des générations entre les temps t_3 et t_4, et t_4 et t_5

Dans certaines régions du génome, on a pu mettre en évidence des blocs de marqueurs en fort déséquilibre de liaison appelés blocs d'haplotypes [3, 4]. En d'autres termes on a découvert que la recombinaison se faisait dans des régions préférentielles. Cependant ce découpage des chromosomes n'est pas absolu, plusieurs méthodes basées sur le déséquilibre de liaison proposent la définition des limites de ces blocs mais ne sont pas toujours concordantes entre-elles. En effet il a été montré que les limites des blocs étaient très sensibles aux paramètres de la méthode utilisée (par exemple le seuil du déséquilibre de liaison) et des fréquences alléliques des SNPs aux limites des blocs.

I.2 Structure des populations

La structure des populations (ou stratification) est la présence de différences de fréquences alléliques entre des sous populations dont l'origine ancestrale diffère. Les causes de la structure des populations reposent sur la séparation physique de groupes d'individus sur plusieurs générations. S'ensuit alors la dérive génétique de certains marqueurs conduisant à la variation des fréquences alléliques dans chacun de ces groupes jusqu'à fixation. Dans les populations humaines, les trois grands groupes ancestraux sont les Européens, les Africains et les Asiatiques.

Dans des populations homogènes, comme en Europe, on peut modéliser la structure des populations en fonction de leur éloignement géographique (figure 5), les fréquences alléliques ayant tendance à varier en fonction de la distance séparant les groupes d'individus.

Récemment sont apparus des mélanges de populations dont les individus ont différentes origines ancestrales (comme les Afro-américains par exemple). La représentation des individus en fonction de leurs origines géographiques devient impossible.

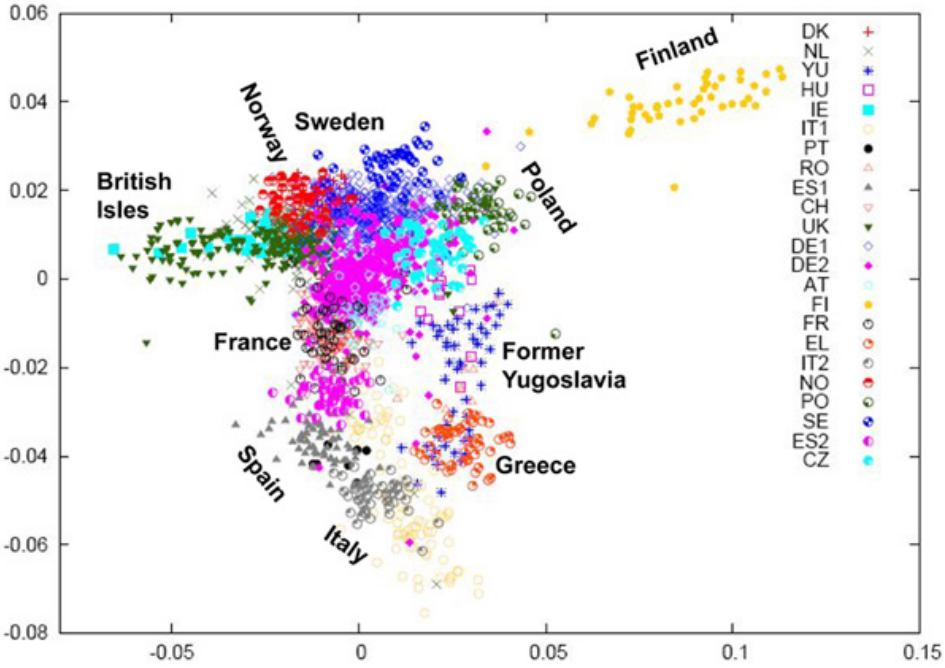


Figure 5 : Représentation par Analyse en Composante Principales (ACP) des différences alléliques entre les individus de plusieurs pays européens. On peut observer la représentation géographique des populations en fonction de leurs fréquences alléliques. *Extrait de anthropology.net*

I.3 Épidémiologie génétique

L'épidémiologie génétique concerne l'étude de la diffusion des gènes dans les populations et l'impact des facteurs génétiques sur les phénotypes. L'étude des facteurs génétiques impliqués dans les maladies est rapidement devenu l'enjeu majeur de l'épidémiologie génétique et amène de nombreux défis. Un facteur n'apporte qu'une prédisposition ou une vulnérabilité pour développer la pathologie. Ainsi être porteur d'une mutation de susceptibilité ne signifie pas nécessairement que le patient développera la pathologie. De plus, l'intervention de facteurs environnementaux va jouer un rôle prépondérant dans le déclenchement de ces maladies. Ainsi deux vrais jumeaux n'auront pas nécessairement les mêmes maladies même s'ils disposent des mêmes facteurs de prédisposition. La transmission de ces maladies est non mendélienne, la pénétrance est incomplète et nécessite donc une analyse à grande échelle portant sur un très grand nombre de cas afin de statistiquement reproduire plusieurs fois le même effet.

Aujourd'hui, les connaissances concernant les gènes prédisposant ne sont que parcellaires, on ne connaît pas encore précisément le nombre de gènes impliqués, ni leur degré d'implication dans la plupart des maladies actuellement étudiées. De plus lorsqu'un facteur est identifié, on ne sait souvent pas encore expliquer le rôle fonctionnel véritable. Ainsi concernant par exemple le VIH, le

variant génétique CCR2-64I a été clairement identifié comme impliqué dans la progression de la maladie [5] sans pour autant savoir quel en est le mécanisme biologique sous-jacent.

Un des enjeux actuels de la recherche médicale est donc d'identifier des gènes impliqués dans les maladies multi-factorielles, d'en comprendre les mécanismes avec pour objectif d'aboutir enfin à l'élaboration de traitements ou tests diagnostics permettant d'intervenir dans ces mécanismes. Une des branches de la génétique épidémiologique s'intéresse au rôle de la composante génétique dans les maladies. Des marqueurs génétiques spécifiques font l'objet d'études d'association afin de savoir s'ils sont impliqués ou non dans la maladie étudiée. Le risque associé est ensuite estimé en prenant en compte éventuellement les interactions avec d'autres facteurs génétiques (GxG) ou des facteurs environnementaux (GxE).

Historiquement, les premiers facteurs génétiques causaux identifiés portaient sur des maladies rares à transmission mendélienne, dont une mutation présente sur un seul gène était l'unique responsable de la maladie. L'étude de ces maladies dites monogéniques telles que la Myopathie de Duchenne ou la chorée de Huntington a ouvert la voie de l'analyse génétique en médecine. Cependant la plupart des maladies sont dites multi-factorielles car elles sont dues à l'interaction de multiples événements environnementaux et génétiques. Ces maladies, parmi les plus courantes, regroupent aussi bien le diabète, l'asthme, la polyarthrite rhumatoïde, les maladies cardiovasculaires ou encore des maladies infectieuses comme le SIDA.

II. Exploration des génomes

II.1 Historique

1865	Mendel pose les bases de la génétique moderne (notions d'allèles, de dominance, d'hétérozygote...)
1910	Morgan découvre la méiose et la recombinaison (base de la théorie chromosomique de l'hérédité)
1913	Première carte génétique
1953	Watson et Crick utilisent la diffraction pour découvrir la molécule d'ADN
1965	Monod, Jacob et Lwoff découvrent les ARNs et la régulation de l'expression génique
1972	Le premier séquençage d'une séquence d'ARN d'un bactériophage est réalisé.
1977	Séquençage d'un organisme entier
1980	Découverte des marqueurs polymorphiques
1983	Découverte de la PCR
1990	Lancement du projet génome humain
1992-1996	Publication des premières cartes du génome
2000	Fischer publie les premiers résultats de thérapie génique sur l'homme
2004	Assemblage complet du premier génome humain

Figure 6 : Repères historiques dans l'exploration des génomes

Depuis les années 1970 et jusqu'à aujourd'hui, la cartographie du génome, afin de localiser les loci ayant un impact sur les maladies, constitue un enjeu colossal (figure 6). La découverte durant cette période des enzymes de restriction, capables de découper précisément l'ADN, donne aux chercheurs un outil précieux dans l'exploration des génomes. Couplé à la technologie de l'ADN recombinant, il va devenir possible de pratiquer la transgenèse permettant l'insertion d'une portion

d'ADN dans un autre ADN et en étudier leurs fonctions.

Plus tard, durant les années 1980, la découverte des marqueurs polymorphiques, des techniques de RFLP (Restriction Fragment Length Polymorphism), des micro-satellites et les grands progrès du génie génétique commencent à permettre une cartographie intensive du génome humain. Le séquençage de l'ADN devient également possible grâce notamment aux travaux de Sanger au Royaume-Uni et de Gilbert aux États-Unis aboutissant en 1977 au séquençage du premier organisme vivant, le bactériophage φX174. Ils obtiendront le prix Nobel de chimie pour ces travaux.

L'amplification en chaîne par polymérase (PCR) constituera au début des années 1990 une avancée majeure pour le clonage et le séquençage des ADNs par la méthode de Sanger. En effet cela a permis de multiplier la quantité d'ADN disponible et d'en obtenir une quantité suffisante pour le clonage ou le séquençage direct.

En 1992, le Généthon publie dans Cell la première carte physique couvrant la moitié du génome humain [6] et la première carte d'un chromosome [7]. Entre 1993 et 1996, les premières cartes physiques du génome humain sont dévoilées par le Généthon et le Centre d'Étude du Polymorphisme Humain (CEPH) dans la revue Nature [8-10] qui en fera même une édition complète. Parallèlement le projet génome humain est lancé dans le monde entier et en 2001 la séquence brute du génome est dévoilée [11] . Finalement en 2004 le consortium international public publie la séquence complète du génome d'un humain [12].

A partir des années 2000, la recherche des polymorphismes s'intensifie, la recherche des variations génétiques par le génotypage haut-débit aboutit à l'élaboration de bases de données publiques internationales. Le dernier saut technologique actuel repose sur la possibilité, pour un prix de plus en plus faible et dans un temps de plus en plus court, du séquençage complet d'un individu. Cela ouvre la voie à la découverte toujours plus exhaustive des polymorphismes humains existants dans les populations humaines.

De par leurs propriétés, les SNPs sont devenus les marqueurs les plus utilisés dans le développement des études d'association. En effet, comme nous l'avons vu, les SNPs sont répartis sur l'ensemble du génome, ce qui constitue un avantage comparé aux autres polymorphismes. De plus, ils sont très nombreux et représentent directement ou indirectement une grande partie de la variabilité du génome humain, enfin, ils sont très faciles à analyser expérimentalement. Dans la suite de mon travail de thèse, je vais ainsi m'intéresser exclusivement à ce type de marqueurs.

II.2 Ressources bioinformatiques en génomique

II.2.1 dbSNP

La base de données « SNP database » est une base de données publique qui répertorie les variations génétiques découvertes dans différentes espèces animales dont évidemment l'Homo sapiens. Malgré son nom, cette base de donnée répertorie non seulement les SNPs mais aussi d'autres variants génétiques comme les indels ou les micro-satellites. Les données sont composées aujourd'hui pour l'homme de près de 20 millions de polymorphismes validés. (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi).

Cette base de données qui permet à l'ensemble de la communauté scientifique d'avoir accès à tous les variants découverts à travers le monde est un outil très important. DbSNP regroupe des applications en pharmacogénomique, en génomique fonctionnelle, dans les études d'évolution et enfin dans les études d'association à grande échelle qui vont m'intéresser tout particulièrement.

II.2.2 HapMap

En 2002, le projet HapMap [13] a été mis en place, il vise à référencer les variants génétiques communs que sont les SNPs (http://hapmap.ncbi.nlm.nih.gov/). Ce projet s'est déroulé en trois phases :

Lors de la **phase I** du projet, environ 1 million de SNPs chez 270 individus issus de 4 populations d'origines africaine, asiatique et européenne ont été génotypés :

- 90 nigériens (Yoruba d'Ibadan) composés de 30 trios,
- 45 japonais (région de Tokyo) non apparentés,
- 45 chinois (région de Beijing) non apparentés,
- 90 résidents des États-Unis originaires d'Europe du Nord et de l'Ouest recrutés par le CEPH, composés de 30 trios.

Lors de la **phase II**, la densité de la carte génétique a été augmentée en génotypant environ 5 800 000 de SNPs supplémentaires chez les mêmes individus [14].

Lors de la **phase III**, la densité de la carte génétique a encore été augmentée, de nouveaux individus issus de nouvelles populations ont été inclus [15] :

- Pour les populations originelles du projet HapMap, le nombre de participants a été augmenté

à 180 individus pour les populations ancestrales africaine et européenne, et à 90 individus pour les populations asiatiques. Le nombre de SNPs génotypés sur l'ensemble de ces sujets est de ~4 millions.

- Pour les 7 populations additionnelles (afro-américains, chinois des USA, indiens Gujarati des USA, kenyans Luhya, kenyans Masaï, mexicains des USA, et toscans italiens), les participants sont au nombre de 90-100 (excepté pour les kenyans Masaï au nombre de 180) et le nombre de SNPs génotypés dans chacune de ces populations est de ~1 400 000.

A l'heure actuelle, ce projet a permis de constituer un catalogue des SNPs communs (MAF> 5%) qui décrit la nature des SNPs avec leur localisation dans le génome, ainsi que la manière dont ils sont distribués dans les populations et entre les populations dans différentes parties du monde. L'autre objectif du projet HapMap était de décrire les relations de dépendance entre les SNPs en utilisant l'information du déséquilibre de liaison et ainsi répertorier les haplotypes communs (>5%). En effet dans le génome humain, il existe des régions (ou blocs) riches en déséquilibre de liaison et caractérisées par une diversité haplotypique limitée, ainsi, de nombreux SNPs du génome se retrouvent corrélés au sein de ces blocs d'haplotypes. Cette propriété permet de cerner la diversité du génome humain en se limitant à un nombre de SNPs suffisants pour différencier tous les haplotypes communs, appelés SNPs marqueurs ou « **tagSNPs »** (figure 7) [16, 17] . Il s'agit d'une forme de compression de l'information consistant à extraire un sous-ensemble de SNPs qui capture l'essentiel de l'information génétique. De cette façon, on estime que seuls 300 000 à 600 000 tagSNPs sont suffisants pour couvrir les ~10 millions de SNPs représentant l'essentiel de la diversité du génome humain.

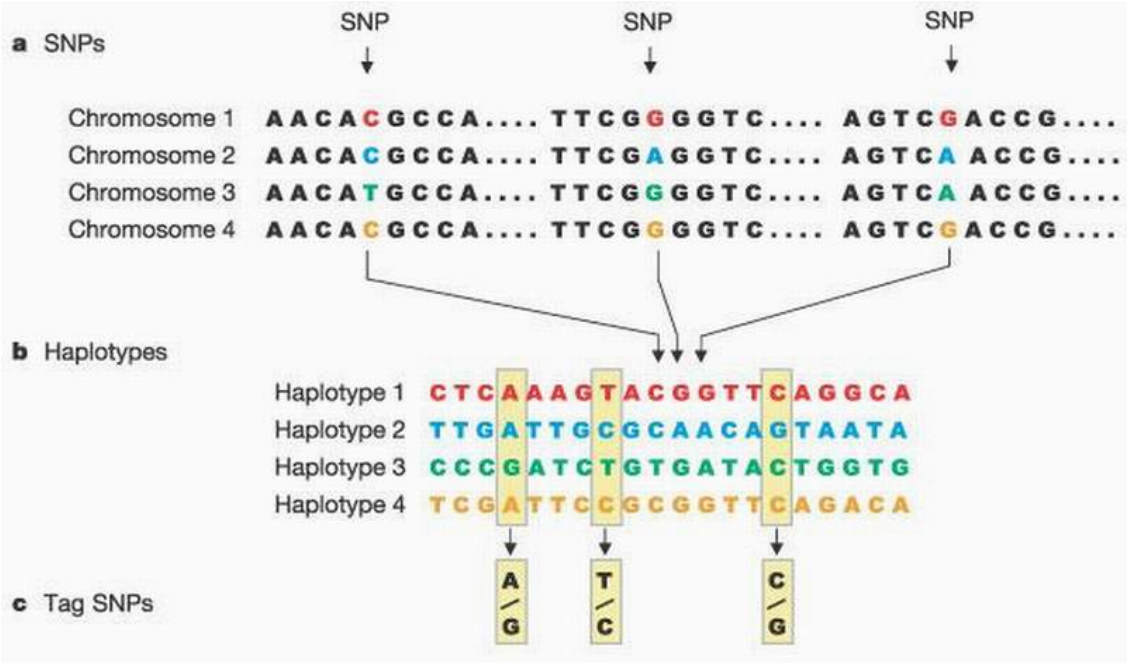


Figure 7 : Illustration de la notion de « tagSNP ». (**a**) Portion de chromosome séquencée dans 4 individus différents. Les SNPs sont représentés en couleur. (**b**) A partir des SNPs, détermination des haplotypes de la région composée ici de 20 SNPs, dont les 3 SNPs du panel a. (**c**) Le génotypage des 3 tagSNPs, ou SNPs marqueurs, est suffisant pour identifier de façon unique les 4 haplotypes de la population. Par exemple, un profil A-T-C pour ces 3 tagSNPs correspond à l'haplotype 1.
Extrait de hapmap.org

II.2.3 1000 genomes

Dans le but de faciliter l'analyse et la recherche des SNPs de faible fréquence, un nouveau projet a vu le jour : le projet 1000genomes (http://www.1000genome.org) [1] avec pour objectif de séquencer la totalité du génome de 2500 individus originaires de 28 populations différentes.

Ces données sont en cours de production et devraient permettre bientôt d'avoir une cartographie plus complète des variants humains, de leurs haplotypes et des déséquilibres de liaison qui existent entre eux. Le degré de définition des polymorphismes a pour but de déterminer tous les variants dont la MAF est inférieure à 1% dans le génome et de 0,1% à 0,5% dans les gènes (figure 8). Le projet HapMap ayant été axé sur les SNPs avec des MAF > 5%, l'implication des SNPs de faible fréquence dans les maladies complexes est aujourd'hui fortement soutenue [18]. De ce fait de nouvelles puces de génotypage, plus denses, basées sur les données du projet HapMap et sur une partie des données du projet 1000genomes ont été développées. La cartographie plus fine des polymorphismes existants dans les populations et leur exploitation dans l'analyse des maladies est en marche.

Cette couverture plus fine de la diversité génétique apporte son lot de défis à relever. L'analyse des données de séquençage tout d'abord. Aujourd'hui, la couverture du projet 1000 génomes n'est que de 4X (un locus est lu en moyenne quatre fois), les génotypes des individus ne sont que sous la forme d'une probabilité. La multiplication des marqueurs, qui plus est de faible fréquence, pose des problèmes statistiques qui pourront être insolubles dans certains cas (par exemple un marqueur indépendant très rare impactant une maladie sera très difficile à identifier).

Les versions successives publiques de ce projet sont les suivantes :

- 1000 génomes Pilot en juin 2010,
- 1000 génomes 2010 interim en décembre 2010 avec 13 Millions de SNPs,
- 1000 génomes phase I interim en juin 2011 avec 37 millions de SNPs.

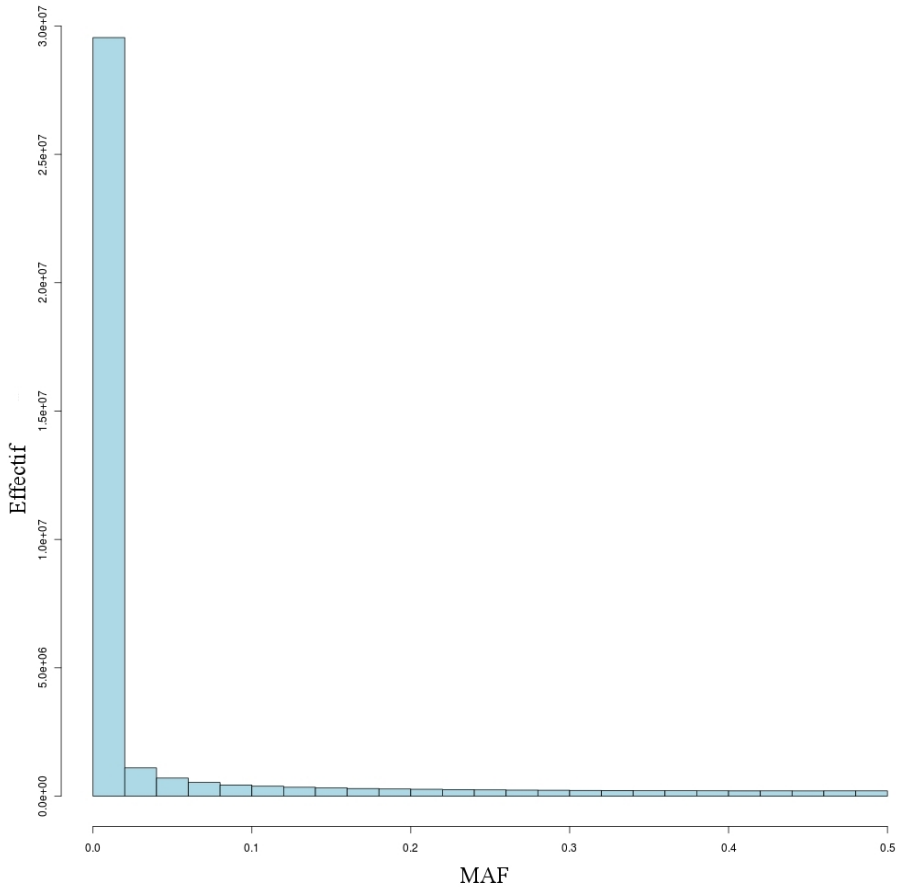


Figure 8 : Histogramme de la distribution des SNPs suivant leurs fréquences alléliques dans 1000genomes phase 1 interim (37M SNPs). On voit que la majorité des SNPs ont des MAFs très faibles

II.3 Puces de génotypage

Le développement de HapMap a permis aux généticiens de tirer profit de l'organisation des SNPs au sein des chromosomes. Depuis lors les chercheurs peuvent examiner l'ensemble du génome à partir d'un nombre restreint de tagSNPs [19, 20] au lieu d'étudier les 10 millions de SNPs qu'il contient.

Dans cette optique, les entreprises Illumina et Affymetrix ont développé des puces de génotypage permettant l'analyse de 300 000 à 1 000 000 tagSNPs basées sur les premières données du projet HapMap. Ces technologies permettent ainsi l'étude simultanée d'une partie importante de la diversité du génome humain à un coût raisonnable. La grande différence entre les puces Illumina et Affymetrix repose sur la sélection des tagSNPs, le but étant de minimiser les coûts tout en capturant le maximum d'information. C'est pourquoi ces deux entreprises ont mis sur le marché des puces contenant des panels de SNPs différents. Dans les deux cas, ces technologies ont permis l'essor des études d'association « génome entier », avec près de 1000 publications en seulement quelques années (figure 9). Ainsi, des gènes jusqu'à présent insoupçonnés ont pu être associés à des maladies complexes comme par exemple la dégénérescence maculaire liée à l'âge [21], la sclérose en plaques [22], ou l'obésité [23].

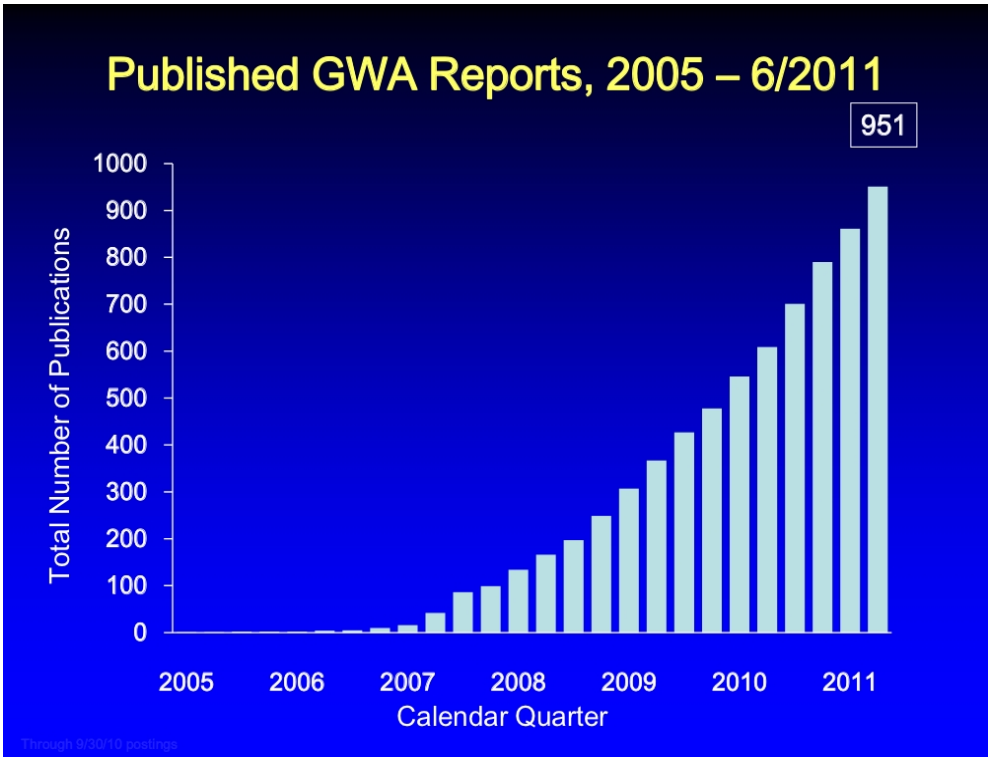


Figure 9: Evolution jusqu'en juin 2011 du nombre de publications pourtant sur les analyses d'association « génome entiers » (GWAS). *Extrait de www.genome.gov/gwastudies/*

Les puces de génotypage de l'entreprise Illumina sont de tailles différentes intégrant des SNPs et parfois des CNVs:

- *HumanOmni2.5-Quad* avec environ 2.5 millions de marqueurs
- *HumanOmni1S-8* avec environ 1.25 millions de marqueurs
- *HumanOmni1-Quad* et *Human1M-Duo* avec environ 1 million de marqueurs
- *Human660W-Quad* avec environ 658,000 marqueurs
- *HumanCytoSNP-12* avec environ 300,000 marqueurs

La stratégie d'Illumina est essentiellement basée sur la sélection des tagSNPs à partir de Hapmap2, avec une composante gène centrée afin de capter au maximum l'information génétique des individus et assurer ainsi une bonne couverture (figure 10).

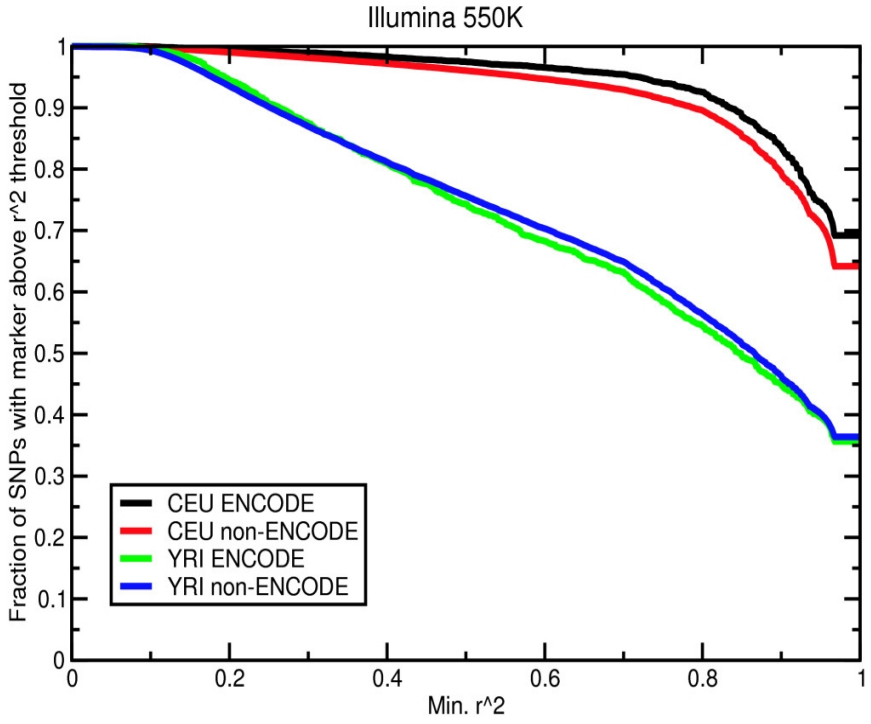


Figure 10 : Représentation de la couverture du génome (HapMap2 > 2,3 millions de SNPs) en fonction du seuil de déséquilibre de liaison choisi dans les 3 grandes populations ancestrales (CEU : Européens, CHB+JPT : Asiatiques, YRI : Africains) pour une puce Human660W-Quad. On voit par exemple qu'en prenant un $r^2 > 0,8$, on capte avec cette puce 90% de l'information génétique en moyenne. *Extrait de Illumina.com*

III. Études génétiques des maladies

Il existe deux grandes stratégies d'identification de facteurs de risque génétiques : les analyses de liaison sur des familles et les analyses d'association impliquant des individus sans lien de parenté.

III.1 Analyses de liaison

Les analyses de liaison génétique ont pour but de localiser une région chromosomique présentant une co-ségrégation avec le phénotype étudié. En d'autres termes, ce type d'analyse consiste à rechercher des loci polymorphes dont la transmission au sein des familles n'est pas indépendante de la transmission de la maladie. Ces analyses de liaison se sont révélées particulièrement efficaces dans le cas de l'étude de maladies monogéniques telles que la mucoviscidose ou la Chorée de Huntington. Ce type d'approche s'est aussi avéré fructueux dans le cas de maladies infectieuses comme la lèpre [24]. En revanche, ces études sont plus limitées pour la détection des facteurs génétiques impliqués dans les maladies multi-factorielles où des effets familiaux sont moins évidents :

- chaque facteur n'explique qu'une fraction du phénotype
- ces pathologies ne touchent pas nécessairement plusieurs membres d'une même famille (par exemple dans le SIDA).

III.2 Analyses d'association

Les études d'association cherchent à déceler une association entre un variant génétique et la maladie, au niveau d'une population et non plus seulement d'une famille. On cherche donc à identifier les polymorphismes dont la distribution d'allèles est associée avec la maladie étudiée. Classiquement ce type d'étude se réalise à l'aide d'un groupe de cas comparé à un groupe de contrôles dans une étude dite « cas-contrôles ». Les études cas-témoins sont des études dichotomiques pour lesquelles on compare deux groupes de sujets qui ont été sélectionnés en fonction d'un phénotype précis. Ce phénotype, qui les différencie, doit être très précis et la connaissance des différences externes à ce phénotype doit être la plus exhaustive possible. Ce type d'étude est particulièrement adapté pour comparer un phénotype dit « cas » relativement rare avec le reste de la population dite « contrôle » [25].

Les principaux risques inhérents à ce type d'étude sont la mauvaise définition des groupes, soit par le phénotype observé soit de la population d'où sont tirés les échantillons composant ces groupes. De prime abord, le recrutement de cas et de contrôles peut paraître plus facile que celui de familles, du fait de la contrainte imposée dans ce dernier cas par l'obtention des génotypes d'individus apparentés. Les études d'association étaient initialement utilisées sur des gènes candidats. Depuis quelques années, les avancées techniques de la génomique permettent de réaliser les études d'association sur le génome entier [26, 27] (Genome Wide Association Study).

III.2.1 Approche « gène candidat »

Les approches « gènes candidats » consistent à sélectionner un ensemble de gènes qui sont pertinents pour intervenir dans l'étiologie de la maladie à étudier et à évaluer leurs polymorphismes directement par association. Le choix des gènes peut être guidé par des a priori biologiques tels que la fonction ou l'appartenance à une voie métabolique associée à une maladie, ou encore sur la base de la localisation dans une région chromosomique d'intérêt, suggérée par une précédente étude de liaison ou d'association. Ce type d'approche repose donc sur des a priori et n'est pas adaptée pour explorer de manière exhaustive les causes génétiques d'une maladie.

Du fait des limitations du séquençage et du génotypage, ce fut longtemps la seule source d'exploration génétique des maladies. Elle reposait donc sur une connaissance préalable de la maladie et une connaissance approfondie des mécanismes moléculaires associés. Par exemple, mon laboratoire s'intéressait à la génomique de l'hote du VIH-1 dans le SIDA. L'approche initiale a été de génotyper des gènes de l'immunité, des gènes connus ou suspectés d'avoir un rôle dans la pathogenèse du VIH-1.

III.2.2 Approche « génome entier »

Une étude d'association « génome entier » (ou systématique) explore une grande partie du génome sans aucun a priori sur l'identité des loci génétiques impliqués. Cette approche représente une stratégie ouverte et assez complète, pouvant être mise en place en l'absence d'indices sur la fonction ou la position des loci de susceptibilité. Elle a d'abord été appliquée pour des études de liaison à l'aide des micro-satellites et a permis de mettre en évidence la plupart des gènes responsables des maladies monogéniques connues. Malheureusement, cette approche a eu des difficultés à s'étendre aux maladies multi-factorielles, l'excès de transmission chez des apparentés

atteints étant plus faible pour des effets modérés. Les études d'association « génome entier » sont donc apparues comme une alternative de choix et devaient constituer dès 1996 pour Risch et Merikangas [28] l'avenir des études génétiques des maladies complexes.

Cependant il n'est pas interdit de réintroduire une notion de «candidat» [29] dans un second temps en ne sélectionnant que certains SNPs en y ajoutant un a priori biologique [30-32]. Pour cela il existe une multitude d'angles d'approches:

- On peut par exemple ne s'intéresser qu'aux gènes impliqués dans la maladie, on s'approche alors d'une approche gène candidat globale.
- On peut s'intéresser à plusieurs types de SNPs en fonction de leur position chromosomique (exoniques, régions promotrices...)
- On peut aussi croiser les informations des bases de données. On a à disposition des bases de données cherchant à identifier des régions eQTL (expression Quantitative Trait Loci), c'est à dire des SNPs qui sont associés à une expression d'ARNm différentielle d'un ou plusieurs gènes [33, 34]. Par exemple, la base de données GENEVAR met à disposition pour tous les gènes du génome humain la valeur d'expression de ceux-ci pour chaque patient de Hapmap2 et Hapmap3.
- On peut croiser l'information basée sur les siRNA ou ARN interférents [35-37], qui cherchent à établir une liste de gènes candidats dans la maladie étudiée.
- Des informations sont également disponibles sur des données d'épissage, de polyadénylation et de sites de fixation de facteurs de transcription.
- On peut s'intéresser à l'enrichissement des voies de signalisation (« pathways ») [38].

Durant ma thèse, je n'ai pas eu l'occasion de travailler sur des données familiales et je vais me concentrer dans la suite de ma rédaction sur l’étude génétique de sujets sans lien de parenté connu.

IV. Méthodes bioinformatiques d'analyse de données génomiques

IV.1 Contrôle qualité

Le contrôle qualité dans les études d'association « génome entier » est fondamental car il peut éviter de mauvaises interprétations et de faux résultats [39].

IV.1.1 Données manquantes

Lorsqu'un SNP présente un fort taux de données manquantes, on peut penser que la fiabilité de la caractérisation des génotypes (« SNP calling ») sur la plate forme de génotypage a été défectueuse. Cette notion est très largement empirique mais reste néanmoins un filtre assez fiable de contrôle qualité.

De la même façon, lorsque les données d'un patient présente beaucoup de SNPs non renseignés, cela peut laisser penser que l'ADN extrait est de mauvaise qualité et que ce patient doit être enlevé de l'étude.

IV.1.2 Déviation de l'équilibre de Hardy-Weinberg

En dehors des déviations de l'équilibre dues à la consanguinité, la sélection ou d'autres concepts de génétique des populations, cela peut être le symptôme d'une erreur de génotypage ou d'une contamination. En effet dans certains cas, un génotype peut être détecté préférentiellement à un autre, ce qui induit une rupture de l'équilibre de Hardy-Weinberg sur un site neutre. Lorsque cela dévie avec une p-value $< 10^{-3}$ ou 10^{-4} on élimine ce locus de l'analyse.

Il faut bien garder en tête cependant que la déviation de Hardy-Weinberg, surtout quand elle s'affiche uniquement dans les cas, peut être le témoin d'une association avec la maladie. Le test le plus simple est un χ^2, cependant pour les faibles fréquences il convient plutôt d'utiliser un test exact de Fisher [40, 41].

IV.1.3 Liaison entre les patients

Lors de l'inclusion dans des études, il peut arriver que le même patient soit inclus deux fois, que des tubes aient été inversés par erreur pendant la collecte, qu'il y ait eu une contamination ou

que deux individus soient apparentés. Pour vérifier que deux individus issus d'une même population ne soient pas trop proches par rapport à ce qu'on pourrait attendre s'ils avaient été tirés aléatoirement dans la population, on réalise un test « Identical By State » (IBS) entre chaque couple d'individus pour évaluer le nombre de génotypes qu'ils ont en commun. A condition de prendre un nombre important de SNP, généralement plusieurs milliers, on peut détecter les éventuelles contaminations d'échantillons ou des liaisons familiales inconnues (dans le cas le plus extrême deux échantillons qui se révèlent venir du même patient).

IV.2 Localisation des sites de susceptibilité: le problème de la puissance statistique

La puissance du test, c'est-a-dire la capacité à détecter l'association lorsque celle-ci est vraie va dépendre de la localisation exacte du site de susceptibilité :

- Soit le marqueur testé est directement responsable de l'association auquel cas la puissance sera maximale.
- Soit le marqueur est en déséquilibre de liaison avec le site de susceptibilité auquel cas la puissance sera d'autant moins forte que le déséquilibre de liaison avec celui-ci sera faible.

Dans tous les cas, la capacité à détecter une association dépendra :

- De la pénétrance de la maladie au site de susceptibilité, observée par l'odd-ratio. Plus celui-ci sera faible et plus il sera difficile de détecter le signal.
- De la fréquence allélique du marqueur testé. Plus celle-ci sera faible et plus il sera difficile de déterminer l'association.

Pour un cas donné la puissance va dépendre du nombre de patients testés, d'où la nécessité avant toute étude d'évaluer le nombre de patients nécessaires – en fonction par exemple des odd-ratio attendus – pour pouvoir obtenir des signaux significatifs.

IV.3 Stratification et le rôle des variables externes

Lorsque les contrôles ne sont pas issus de la même population, ce qui est rarement le cas en pratique, on risquera d'interpréter faussement une association entre la maladie et le polymorphisme. On peut citer l'exemple de Lander et Schork [42] : un généticien qui voudrait étudier l'habilité de

manger avec des baguettes dans la population de San Fransisco trouvera certainement une association avec le *HLA-A1*, non pas parce-que des déterminants immunologiques jouent un rôle dans la dextérité à se servir des baguettes mais parce-que l'allèle *HLA-A1* est plus fréquente chez les asiatiques que chez les caucasiens. Il convient donc de prendre des précautions concernant le mélange sous-jacent de populations dans les études sur génome entier. Outre la sélection préalable des individus, il existe différentes méthodes de correction de la stratification.

Méthode du « genomic control »

Il s'agit d'une méthode statistique qui utilise l'information du Q-Q plot et l'éventuelle déviation de la courbe théorique des données observées (figure 11) [43]. On utilise un facteur d'inflation λ calculé par rapport à l'ensemble des valeurs du χ^2 entre tous les tests de marqueurs non associés à la maladie. Ce facteur λ peut être estimé par: λ = médiane(x_1,x_2,...,x_n)/0,456 quand x suit une loi de distribution du χ^2. On utilise alors ce facteur pour corriger la statistique obtenue. La faiblesse de cette approche réside dans le fait que tous les marqueurs sont corrigés avec un poids équivalent alors que la stratification est représentée par un nombre limité de SNPs. Lorsque λ est supérieur à 1,10 on considère que la stratification est trop importante, il convient de vérifier les données, et le cas échéant la stratification ne peut être ignorée.

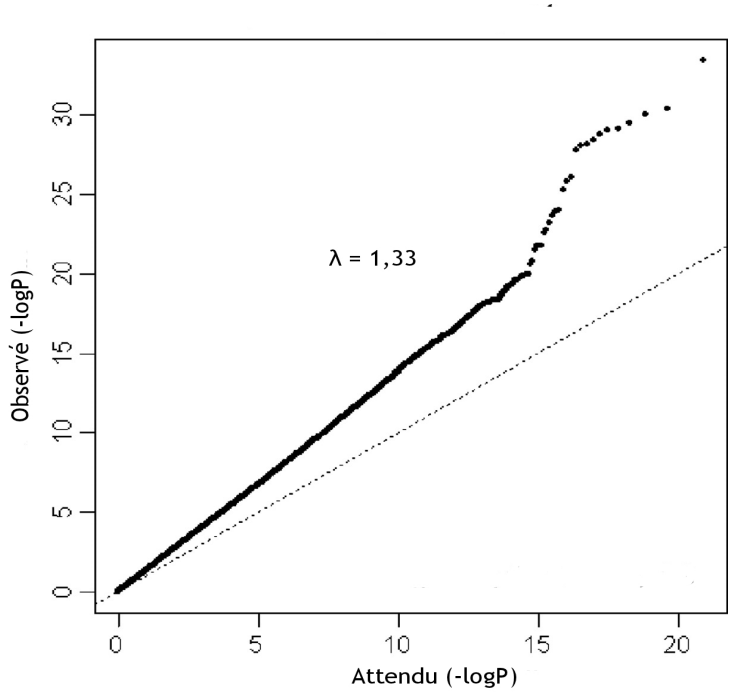


Figure 11 : Exemple de Q-Q plot où le facteur de dispersion λ = 1.33 ce qui représente une stratification forte

Méthode « STRUCTURE »

Cette méthode permet de détecter des groupes de patients appartenant à la même population (figure 12).

Ce logiciel est basé sur la méthodes des moyennes mobiles (K-means) et sur le fait que la stratification fait dévier de l'équilibre de Hardy-Weinberg plusieurs populations mélangées. Cela va permettre soit de prendre en compte dans l'étude ces «clusters» soit d'éliminer les patients marginaux n'appartenant pas à la population générale étudiée.

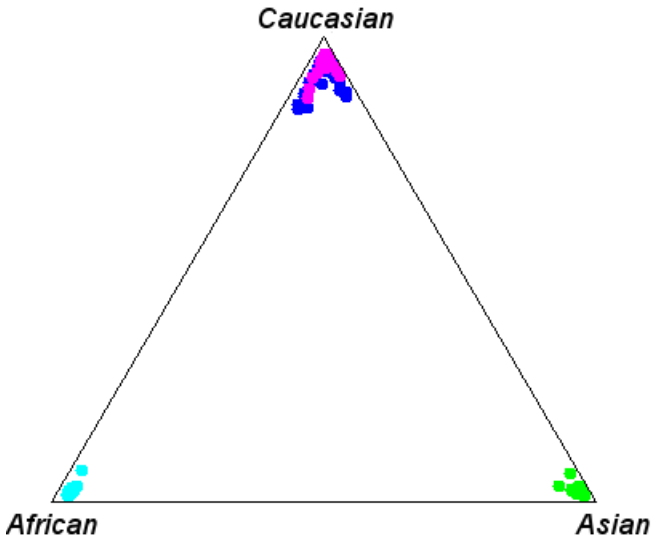


Figure 12 : Représentation des groupes d'individus issu de Hapmap2 grâce au logiciel STRUCTURE [44]. Quelques SNPs hautement informatifs (environ 500) permettent de décomposer chaque point (individu) pour le placer dans son groupe ethnique ancestral

Analyse en Composantes Principales (ACP)

Cette méthode permet d'identifier des axes de variations génétiques pour chaque individu.

1. On fait une ACP pour identifier les composantes principales et réduire ainsi l'information. Ainsi on «résume» les différences génétiques entre les individus sur plusieurs axes (exemple figure 13).
2. Les individus sont placés les uns par rapport aux autres grâce à ces axes de différentiation. On pourra également repérer les SNPs qui différencient préférentiellement les groupes de populations en contribuant plus ou moins à l'élaboration des axes.
3. Les patients qui sont trop différents des autres sont écartés de l'étude
4. Enfin on pourra ensuite utiliser les composantes principales comme covariables pour ajuster les statistiques d'association et prendre en compte la stratification résiduelle.

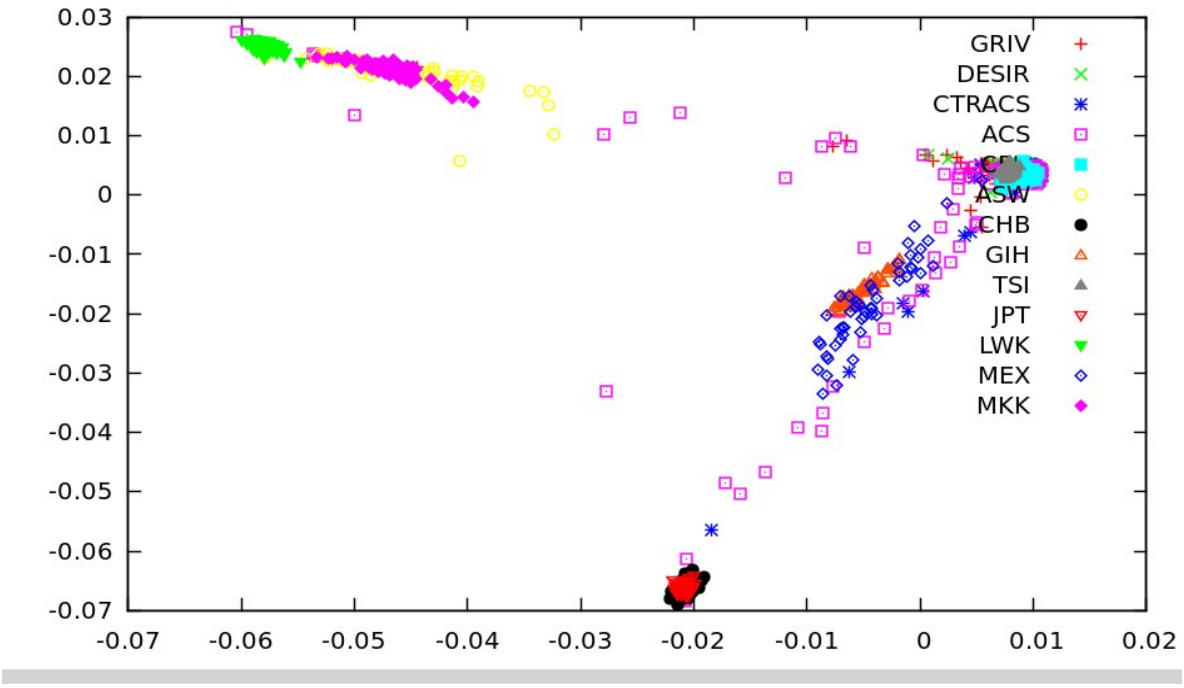


Figure 13 : Représentation des deux premiers axes de l'analyse en composantes principales réalisée à partir de plusieurs dizaines de milliers de SNPs grâce au logiciel EIGENSTRAT. On peut observer sur les données hapmap3 les différents groupes ethniques et leur degré de rapprochement.

IV.4 Tests multimarqueurs

L'analyse individuel de chaque SNP n'est pas le seul angle d'analyse possible. En effet il convient de s'intéresser à des approches globales qui permettent de prendre en compte plusieurs marqueurs et qui ont souvent, en plus de l'approche statistique, une réalité biologique.

IV.4.1 Haplotypes: le problème de l'haplotypage

A partir des données actuelles issues du génotypage ou du séquençage, l'information de la phase (*i.e.* combinaison allélique sur un même chromosome) est perdue. Les logiciels d'haplotypage ont pour objectif de reconstruire cette information à partir de l'ensemble des génotypes d'une population pour pouvoir attribuer une paire d'haplotypes à chaque individu (figure 15).

L'haplotype étant défini comme la combinaison d'allèles sur un même chromosome à différents loci, l'haplotypage va consister à passer de l'information des génotypes à celle de l'haplotype correspondant.

Concernant n loci bialléliques :

- On a en théorie 2^n haplotypes possibles. En pratique, à cause des déséquilibres de liaison, on en observe généralement moins.
- Quand n= 2 (2 SNP d'allèles Aa et Bb) :
  - $r^2 = 1$ signifie que le déséquilibre de liaison est total on a alors que 2 haplotypes possibles.
  - D' = 1 signifie qu'il n'y a pas de recombinaison entre les marqueurs, il y a donc au maximum 3 haplotypes possibles

	SNP_2			
SNP_1		AA	Aa	aa
	BB	AB/AB	AB/aB	AB/aB
	Bb	AB/Ab	AB/ab ou Ab/aB	AB/ab
	bb	Ab/Ab	Ab/ab	Ab/ab

Figure 14 : Répartition des haplotypes théoriquement possibles pour 2 SNPs bi-alléliques

On voit (figure 14) que l'ambiguïté se situe sur les doubles hétérozygotes, il va falloir utiliser des méthodes statistiques basées sur les déséquilibres de liaison pour déterminer la probabilités des

haplotypes dans ces cas.

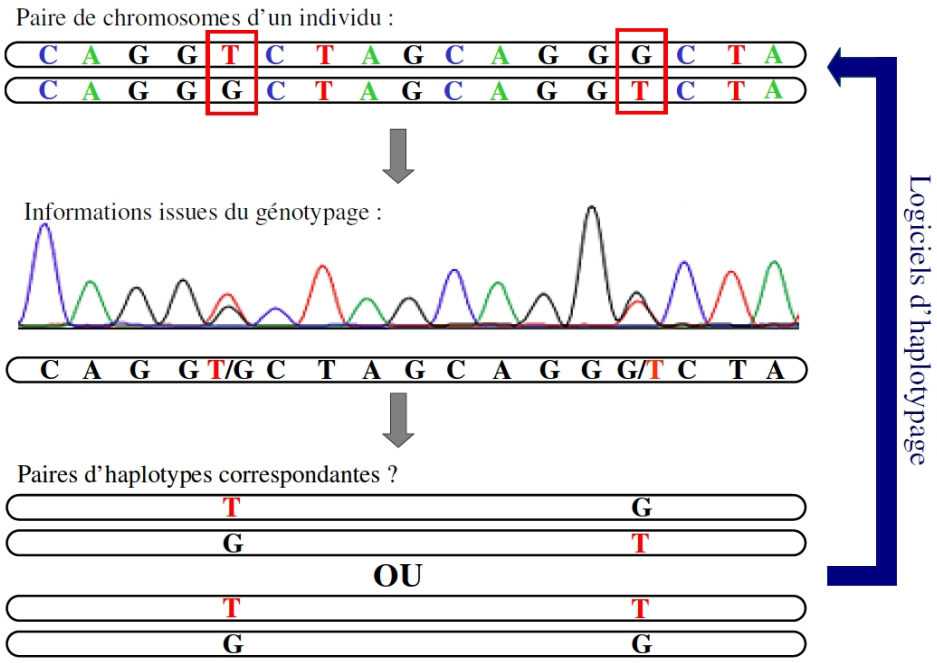


Figure 15 : Problématique de l'haplotypage

Il existe différentes méthodes pour reconstruire les haplotypes.

1) Les **méthodes combinatoires** qui ont pour principe de tester toutes les combinaisons possibles d'haplotypes pour ensuite fixer un critère de choix parmi toutes ces combinaisons. Ce critère peut être un critère de parcimonie [45] ou un critère de phylogénie [46].

2) Les **méthodes d'inférence statistiques**

   1. Méthode basée sur la vraisemblance

   L'algorithme d'Espérance-Maximisation (EM) est un algorithme itératif qui se décompose ainsi:

   Soit k haplotypes possibles dans la population, de fréquences respectives f_k

   - On part avec des fréquences aléatoires $f_k(0)$
   - On pratique une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en calculant pour chaque individu la fréquence des paires d'haplotypes.
   - Ensuite vient une étape de maximisation (M), où l'on estime le maximum de vraisemblance en se basant sur l'équilibre de Hardy-Weinberg.

- On utilise ensuite les paramètres trouvés comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi jusqu'à convergence et jusqu'à ce que les fréquences $f_k(t+1) \approx f_k(t)$

La première méthode d'inférence statistique des haplotypes fondée sur la vraisemblance a été introduite par Excoffier et Slatkin en 1995 [47] puis développée ensuite par Fallin et al. [48].

2. Méthodes basées sur des algorithmes bayésiens ou pseudo-bayésiennes (MCMC)

Ces méthodes se servent à la fois de l'information a priori concernant les fréquences haplotypiques et l'information des génotypes pour calculer la distribution a posteriori des haplotypes sachant les génotypes observés. Le modèle de distribution des haplotypes utilise le modèle de coalescence et prend en compte la recombinaison. Un échantillonneur de Gibbs (algorithme MCMC) est utilisé pour approximer cette distribution à partir des génotypes observés.

Des auteurs ont essayé de mesurer la différence de performance des différents algorithmes [49]. Il semblerait que cette dernière méthode, appliquée dans différents logiciels dont PHASE [50] et SHAPEIT [51, 52] , soit la plus précise dans la reconstruction des haplotypes [53].

Une revue détaillée des méthodes d'haplotypage a été publiée cette année [54].

IV.4.2 Test haplotypiques

Intérêt

Les haplotypes se révèlent être un axe d'étude de la variabilité génétique de plus en plus important dans les études d'association. Le premier intérêt des haplotypes tient à leur capacité à capturer les structures de déséquilibre de liaison sur des vastes régions chromosomiques et de calculer les coefficients de corrélation (D, D' et r^2) robustement. D'autre part, ceci permet aussi de résumer l'essentiel de la diversité génétique humaine en seulement quelques centaines de milliers de SNP, par l'intermédiaire des tagSNPs.

Le second intérêt des haplotypes tient à leur utilité en tant que marqueurs multi-alléliques sur plusieurs SNPs. En regroupant l'information portée par plusieurs SNP et en modélisant de façon implicite les corrélations qui existent entre les SNP qu'ils regroupent, ils permettent de tester simultanément leur association avec la maladie. En effet, dans l'exemple d'une région codante,

plusieurs polymorphismes peuvent conjointement modifier la conformation d'une protéine et donc être responsables du phénotype observé. Cependant, l'avantage de cette approche par rapport aux méthodes SNP par SNP est un sujet très débattu dans la communauté [55]. En effet, certaines études montrent que les tests haplotypiques sont plus puissants que les tests « simple marqueur », alors que d'autres affirment le contraire.

Le troisième intérêt des haplotypes concerne cette fois-ci leur signification biologique. Ils correspondent à des combinaisons d'allèles transmises de génération en génération, ils peuvent donc être corrélés à un facteur héréditaire de risque, composé d'une combinaison de plusieurs allèles, impliquant une expression (allèles dans le promoteur) et/ou une structure (allèles dans les introns) particulière d'une protéine donnée. De plus, leur complexité les rend certainement plus adaptés pour se corréler et décrire la diversité des mécanismes biologiques [56].

Plusieurs exemples d'haplotypes ayant un effet chez l'homme ont été décrits [57-62]. Cependant, pour illustrer toutes ses notions sommairement décrites, on peut reprendre l'exemple dans GRIV du rôle du variant de *CCR5* dans la progression de l'infection par le VIH-1 [60]. En analyse simple SNP, la mutation dans *CCR5* ressort avec une valeur de significativité p-value=10^{-3} alors qu'au sein d'un haplotype couvrant 3 SNP de la région, elle ressort à $4x10^{-5}$. De plus, un autre haplotype défini sur ces trois mêmes SNP révèle une combinaison de mutations sur *CCR5* qui favorise cette fois-ci l'infection, alors qu'en analyse simple SNP cet effet reste totalement muet.

Méthodes

Lorsque les marqueurs sont indépendants, une idée consiste à tester l'homogénéité de la distribution des haplotypes entre les cas et les contrôles. La méthode la plus simple consiste à tester l'homogénéité en utilisant un test de $\chi 2$ dont le nombre de degrés de liberté sera égal à $k - 1$.

On peut également tester les haplotypes individuellement pour observer l'effet spécifique d'un haplotype avec la maladie. On va alors tester cet haplotype versus tous les autres allèles d'haplotypes.

Les données issues des logiciels d'haplotypage se présentent sous la forme de données probabilisées. Pour traiter ces données, on peut prendre en compte pour chaque individu sa paire d'haplotypes la plus probable. Le désavantage est que cela ne prend pas en compte la distribution de probabilité introduisant un biais statistique et pouvant conduire à des conclusions erronées.

Heureusement il est possible de prendre en compte cette incertitude dans des régressions basées sur les données probabilisées pour chaque patient.

Choix des haplotypes

Le nombre potentiel d'allèles d'haplotype augmente exponentiellement en fonction du nombre de SNPs (2^{n-1} en théorie), et le nombre d'haplotypes à tester augmente en fonction du sous-groupe choisi.

Par exemple pour une région composée de 200 SNPs, il faudrait tester $C_2^{200} = 2 \times 10^4$ combinaisons d'haplotypes. Pour un chromosome entier de 20000 SNPs, il faudrait tester $C_2^{20000} \approx 2 \times 10^8$ combinaisons.

Il est donc nécessaire de faire un choix sur le nombre de SNPs composant les haplotypes et sur les régions ciblées afin de limiter le nombre de tests.

Une première idée largement répandue est d'utiliser une fenêtre coulissante d'une taille définie et d'ensuite la faire glisser SNP par SNP le long du chromosome.

La deuxième idée est de se servir des blocs d'haplotypes comme défini précédemment ou des zones d'intérêt sur lesquelles on ne gardera que les haplotypes « tagSNP » (htSNP).

Enfin, on peut tester plus exhaustivement des régions données en testant toutes les combinaisons possibles de 2 ou 3 SNPs (en tester plus risquerait d'augmenter de manière exponentielle le nombre de combinaisons possibles).

Dans tous les cas, il convient de prendre des précautions avec la redondance d'information représentée par le déséquilibre de liaison.

IV.4.3 Épistasies

Méthode de régression logistique

Les méthodes de régression sont également utilisées de manière très classique pour estimer les interactions entre plusieurs variables explicatives.

La méthode de la régression logistique permet de tester la relation entre une variable par exemple dichotomique (malade – non malade) et des variables x_i , qui peuvent être qualitatives ou quantitatives. On va ensuite pouvoir calculer la probabilité P(x) d'être malade en fonction des variables étudiées.

$$logit\, P(X) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

avec x_1 génotype pour un SNP_1, x_2 le génotype pour un SNP2.

Approche « Multiple Dimensionary Reduction »

Il s'agit d'une méthode combinatoire. En pratique, les méthodes combinatoires cherchent à déterminer toutes les combinaisons possibles de génotypes pour expliquer la maladie. Il existe donc des méthodes dont la plus connue est le « Multiple Dimensionary Reduction » (MDR) qui permet de réduire le nombre de combinaisons à explorer [63]. Elle a été mise en application dans l'étude d'interaction entre des gènes du métabolisme des estrogènes dans le cancer du sein [63]. Cependant cette technique présente le désavantage d'être très lente et non applicable en « génome entier ».

IV.5 Méta-analyses

La multiplicité des études à grande échelle « génome entier » sur la même maladie permet souvent de pouvoir en confronter les résultats. Le but principal est d'augmenter la puissance statistique de découverte d'un locus impliqué dans la maladie ou de valider un résultat déjà obtenu.

Le but principal d'une méta-analyse est d'estimer l'effet combiné d'un signal trouvé dans plusieurs études. Si toutes les études sont d'égale importance, on peut simplement calculer la moyenne de l'effet. Si au contraire certaines études sont plus importantes que d'autres, il faut faire une moyenne pondérée des effets en assignant plus de poids à une étude et moins à une autre. La question est de savoir comment ces poids sont assignés et qu'est ce que signifie un « effet

combiné ». Il y a deux modèles utilisés en méta-analyse, le modèle fixe et le modèle aléatoire.

Il convient de vérifier que les résultats sont homogènes entre les études. On peut mesurer ces différences à l'aide du test de Cochran :

$$Q = \sum_i w_i (\theta_i - \theta)^2$$

Un résultat statistiquement significatif signifie que l'écart entre les études est plus important que leur effet.

L'indicateur I^2 est une alternative au test: il mesure la proportion d'hétérogénéité dans les études qui ne peut pas être expliquée par le hasard seul mais en tenant compte du nombre d'études analysées. Les valeurs de I^2 égales à 25%, 50%, et 75% représentent respectivement une hétérogénéité basse, modérée et forte.

Une approche historique est d'utiliser les p-values des différentes études, les seules valeurs souvent disponibles. On peut pour cela utiliser la méthode de Fisher :

$$\chi^2 = -2 \sum_{i=1}^{k} \ln(p_i) \quad \text{pour ddl} = 2k$$

Le principal défaut de cette méthode est qu'elle ne prend pas en compte le sens de l'effet dans les différentes études. De plus chaque étude a la même importance que les autres.

Lorsque plus d'informations sont disponibles, on peut également utiliser la méthode des z-score qui peut prendre en compte le nombre d'individus et le sens de l'association - c'est-à-dire l'allèle associé - dans chaque étude.

Pour cela nous devons convertir les p-values P_i de chaque étude en z-score Z_i puis

$$Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

Il existe enfin des méthodes dites « inverse de la variance » qui prennent en compte l'écart type dans chaque étude. Par exemple l'écart-type du beta de la régression.

Il convient enfin de mesurer l'hétérogénéité entre les résultats à l'aide des tests de Cochran Q. Une très forte hétérogénéité dénote souvent une différence significative dans la constitution des cohortes et doit avertir l'expérimentateur d'un potentiel problème.

Lorsque les études présente une bonne homogénéité d'effet on va privilégier le modèle « fixe » qui prend comme hypothèse que toutes les études, ont de manière sous-jacente, la même force de l'effet tandis que le modèle « aléatoire » va permettre de considérer que l'effet peut varier d'une étude à l'autre. Par exemple, si un groupe d'étude est plus vieux ou ne présente pas exactement le même phénotype, l'effet mesuré ne sera pas exactement le même.

Cette étape de méta-analyse est indispensable pour confirmer les résultats obtenus et peut souvent permettre de découvrir de nouveaux signaux associés à la maladie. Pour la publication d'une étude « génome entier » aujourd'hui, il est devenu absolument nécessaire de passer par cette étape [64]. Généralement pour pouvoir publier un résultat il faut impérativement que la p-value issue de la méta-analyse soit inférieure à 5.10^{-8} [65].

Malheureusement, la réplication d'un résultat est souvent difficile à cause de la diversité des différentes études (population d'origines différentes, différences liées à l'expérimentation, complexité de la maladie ou encore les différences d'analyse).

IV.6 Imputation

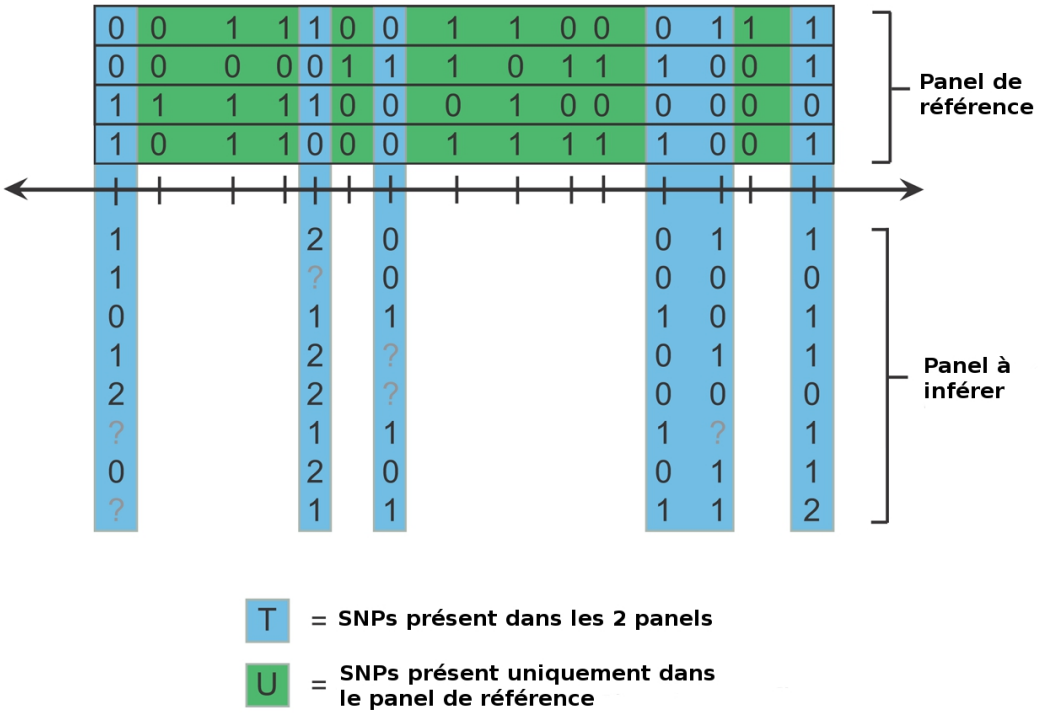


Figure 16 : Représentation schématique de l'imputation à partir des haplotypes d'un large panel de référence. L'information du panel à inférer va servir de base pour reconstruire les données manquantes.

Les puces de génotypage ont une couverture du génome bien définie suivant la technologie utilisée (de 300K à 1M de SNPs). L'imputation permet de déterminer les génotypes des marqueurs non typés en s'appuyant sur la structure du déséquilibre de liaison et sur l'information des taux de recombinaisons des haplotypes du panel de référence contenant un très grand nombre de SNPs phasés (figure 16).

Ce processus d'imputation permet notamment [66-71]:

- D'augmenter la puissance statistique.
- D'augmenter la couverture du génome des données de puces de génotypage déjà existantes et de permettre de caractériser des variations non génotypées initialement.
- De permettre les méta-analyses entre plusieurs jeux de données issus de différentes plate-formes de génotypage.
- De compléter les éventuelles données manquantes.

Il existe de nombreux logiciels permettant d'inférer les génotypes et d'estimer la distribution de probabilité des SNPs non-observés à partir d'haplotypes déjà connus [67, 72, 73]. Ces approches sont assez similaires aux méthodes d'haplotypage et tentent d'estimer la distribution a posteriori des génotypes en utilisant un modèle de Markov caché (MMC).

IV.6.1 Panels de référence

On peut théoriquement prendre n'importe quelles données haplotypées comme panel de référence présentant une couverture supérieure à sa puce de génotypage. Ces données sont généralement issues des données publiques de 1000genomes et de Hapmap. L'amélioration des techniques de séquençage et de génotypage ont augmenté rapidement la taille des panels de référence.

Nom	NCBI build	Date de publication	Nombre de SNPs	Nombre d'individus
1000 Genomes Phase I	b37	Juin 2011	37M	1094
1000 Genomes Pilot	b36	Juin 2010	11,9M	179
HapMap 3	b36	Février 2009	1,4M	165
HapMap 2	b36	Octobre 2008	3.1M	90

IV.6.2 Pré haplotypage

Plus le panel de référence est grand, plus la qualité de l'imputation sera bonne. En général la complexité de l'imputation augmente quadratiquement en fonction des données, rendant le calcul difficile sur les grappes de calculs courantes, existant dans la plupart des laboratoires.

Il peut être avantageux de passer par une étape dite de pré-haplotypage (figure 17). L'idée est que le processus d'imputation sur des données déjà phasées est considérablement plus rapide que sur des données non phasées. L'autre avantage réside dans le fait que ces données phasées pourront être utilisées avec des panels de référence différents à mesure que ceux-ci sont publiés et disponibles. Cette approche a été décrite comme ne faisant perdre que très peu de précision dans l'imputation et donnant des résultats très similaires à l'approche standard [74].

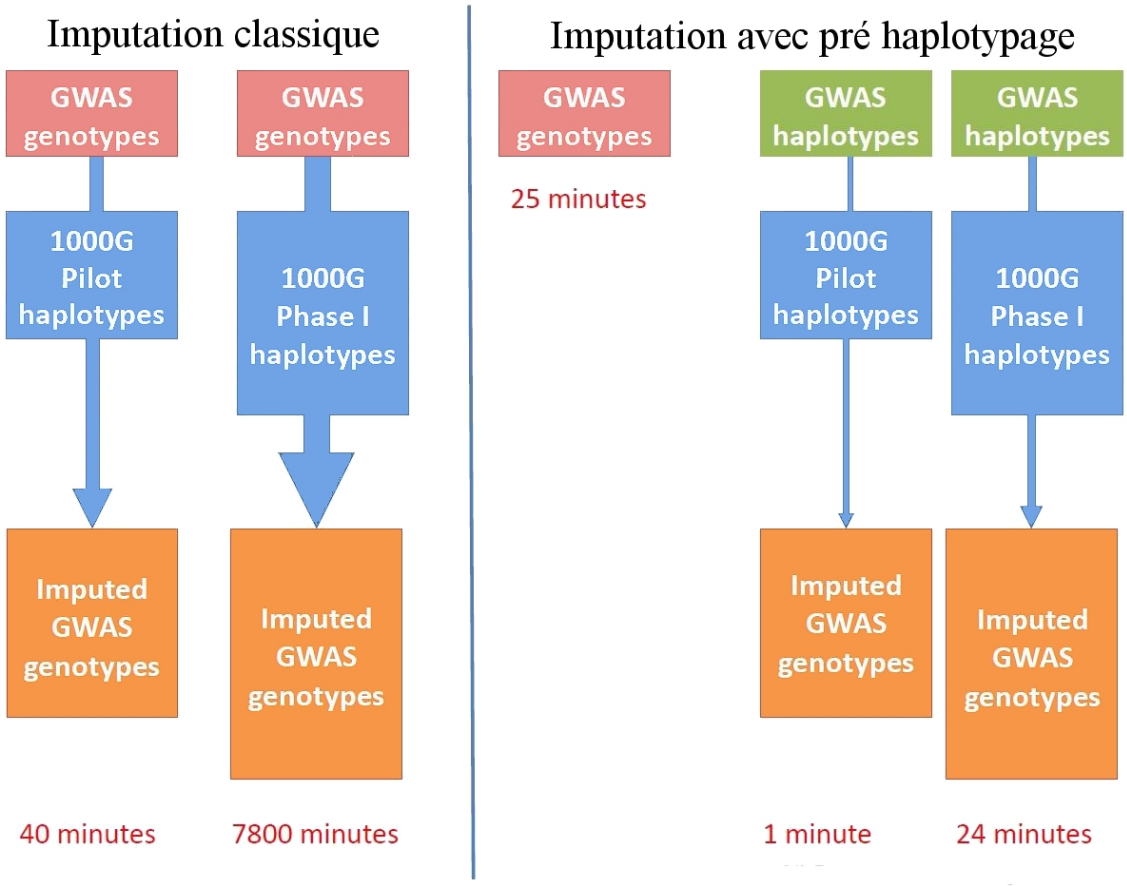


Figure 17 : Ordre de grandeur du temps de calcul pour imputer un génome sur une machine standard avec la méthode classique et avec la méthode de pré-haplotypage

V. Projet GRIV et génomique du SIDA

V.1 La maladie

Le premier cas de Syndrome d'Immuno-Déficience Acquise (SIDA) fut découvert aux États-Unis en 1981. Le Docteur Michael Gottlieb décrit la découverte de symptômes rares chez quatre jeunes homosexuels de Los Angeles : pneumocystoses (pneumonies à *Pneumocystis carinii*) et candidoses buccales [75]. Ces symptômes sont associés à une perturbation du système immunitaire (amaigrissement, fièvre...). Dès 1982, les examens biologiques révélèrent chez tous les malades une immunodépression caractérisée par une chute significative du taux de lymphocytes T $CD4^+$. Cette caractéristique donna naissance à l'appellation « syndrome d'immuno-déficience acquise » ou SIDA.

Le virus du SIDA fut isolé pour la première fois en 1983 par l'équipe du Pr Montagnier à partir de lymphocytes T provenant d'un ganglion de patient atteint de lymphadénopathie généralisée [76]. Ce virus a été initialement dénommé LAV (LymphAdenopathy Virus), le terme VIH-1 (Virus de l'Immuno-déficience 1 Humaine) sera adopté par la communauté scientifique en 1986. Le test de dépistage sérologique mis au point par l'équipe du Pr Gallo en 1984 a permis d'établir que le VIH-1 était l'agent étiologique du SIDA, en démontrant que tous les patients atteints du SIDA étaient porteurs d'anticorps dirigés contre ce virus [77]. Ce test a permis d'éviter la contamination des banques de sang, sauvant ainsi des millions de vies.

Un virus proche du VIH-1 fut découvert en 1986 par l'équipe du Pr Clavel. Ce virus, appelé VIH-2 [78], existe essentiellement à l'état endémique en Afrique de l'ouest, mais il existe aussi d'autres sites de présence sporadique ailleurs dans le monde, notamment les pays entretenant des liens commerciaux avec les pays d'Afrique de l'ouest. Très apparenté sur le plan morphologique au VIH-1, le VIH-2 est cependant décrit comme moins pathogène que le VIH-1 : un temps de latence plus long avant l'apparition du syndrome d'immuno-déficience, une charge virale plus faible, un taux de progression vers le SIDA plus faible, et des risques de transmission (notamment mère-enfant) plus faibles [79, 80].

V.1.1 Le SIDA, épidémie mondiale

Depuis le début de la pandémie, plus de 25 millions de personnes sont décédées des suites d'une infection par le VIH. Elle représente un véritable fléau qui touche plus de 33,4 millions de personnes dans le monde (figure 18). En 2008, on estime à 2,7 millions le nombre de personnes nouvellement contaminées et 2 millions le nombre de personnes décédées du SIDA (estimation datant du dernier rapport de l'ONU-SIDA en 2009). Les données épidémiologiques montrent clairement l'importance des progrès dans la prévention de nouvelles infections et la diminution du nombre annuel de morts. Cependant, l'accès au traitement reste coûteux, limitant de ce fait leur accès aux personnes les plus pauvres qui sont les personnes les plus touchées par l'infection. Il est donc urgent de rechercher de nouvelles pistes thérapeutiques efficaces et accessibles à tous.

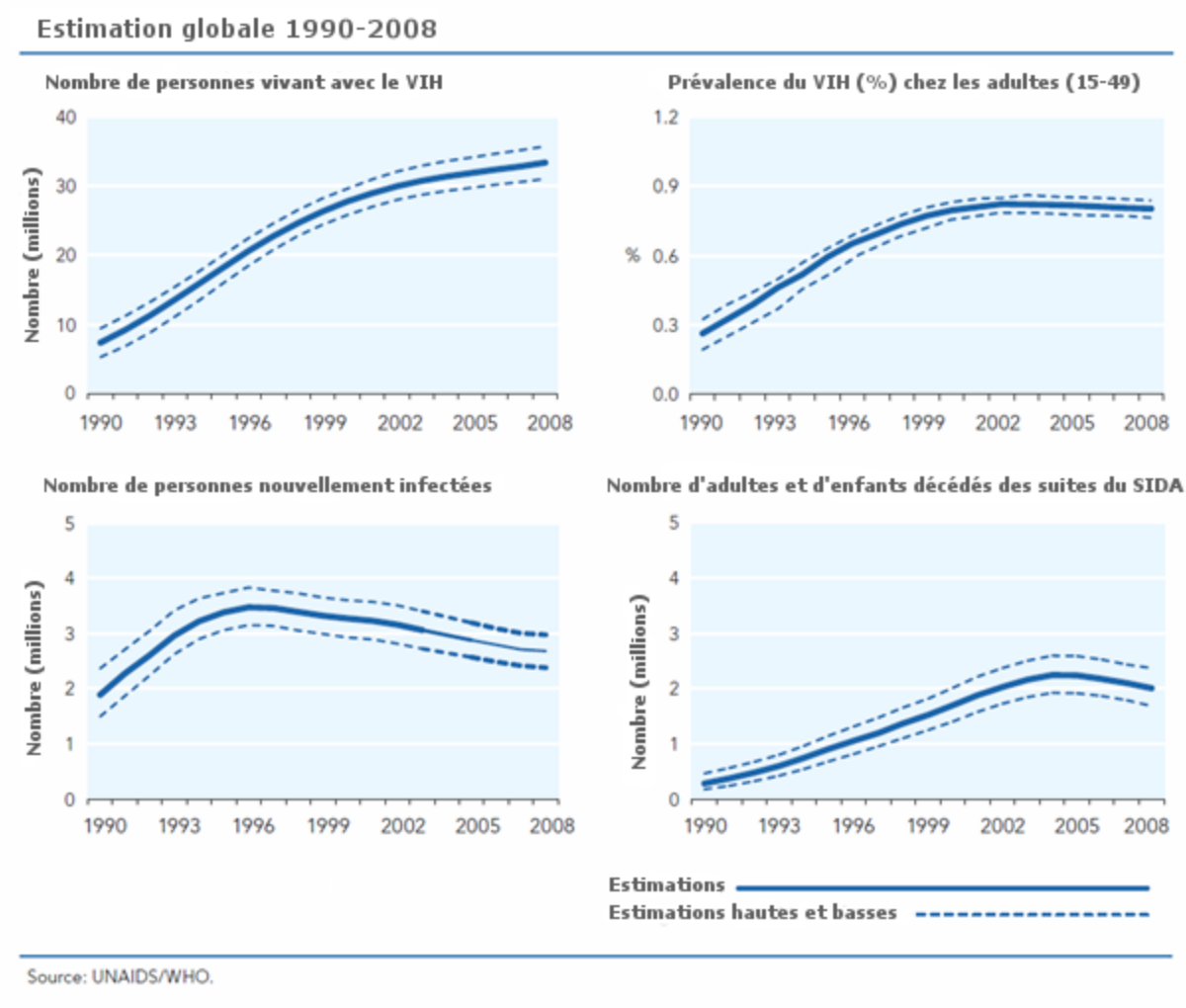


Figure 18 : Estimation globale entre 1990 et 2008 du nombre de personnes vivant avec le VIH, de la prévalence du VIH chez les adultes, du nombre de personnes nouvellement infectées par le VIH et du nombre d'enfants et d'adultes décédés des suites du SIDA

V.1.2 Virus de l'immuno-déficience humaine 1 : VIH-1

Le VIH-1 est un rétrovirus appartenant à la sous-famille des lentivirus, caractérisée par une longue période d'incubation. Les rétrovirus se distinguent par la présence de la transcriptase inverse, enzyme virale responsable de la rétrotranscription (ou transcription inverse) de leur génome d'ARN en ADN. Ainsi, l'ADN viral est intégré sous forme de provirus dans le génome de la cellule hôte. Le provirus est stable et se réplique grâce à la machinerie cellulaire en même temps que l'ADN de l'hôte.

Sur le plan évolutif, les rétrovirus sont générateurs de diversité, du fait des fréquentes erreurs commises lors de la rétrotranscription [81] et de leur intégration dans le génome de l'hôte [82-84]. Cette dernière étape peut conduire :

- à l'incorporation de portions du génome de l'hôte lors de la génération de nouveaux virus
- à l'altération de l'activation/inactivation des gènes situés à proximité du site d'intégration (fort pouvoir oncogène des rétrovirus)
- à leur cooptation complète ou partielle dans les cellules germinales hôtes (HERV, Human Endogenous Retrovirus).

Le cycle de réplication virale [85, 86]

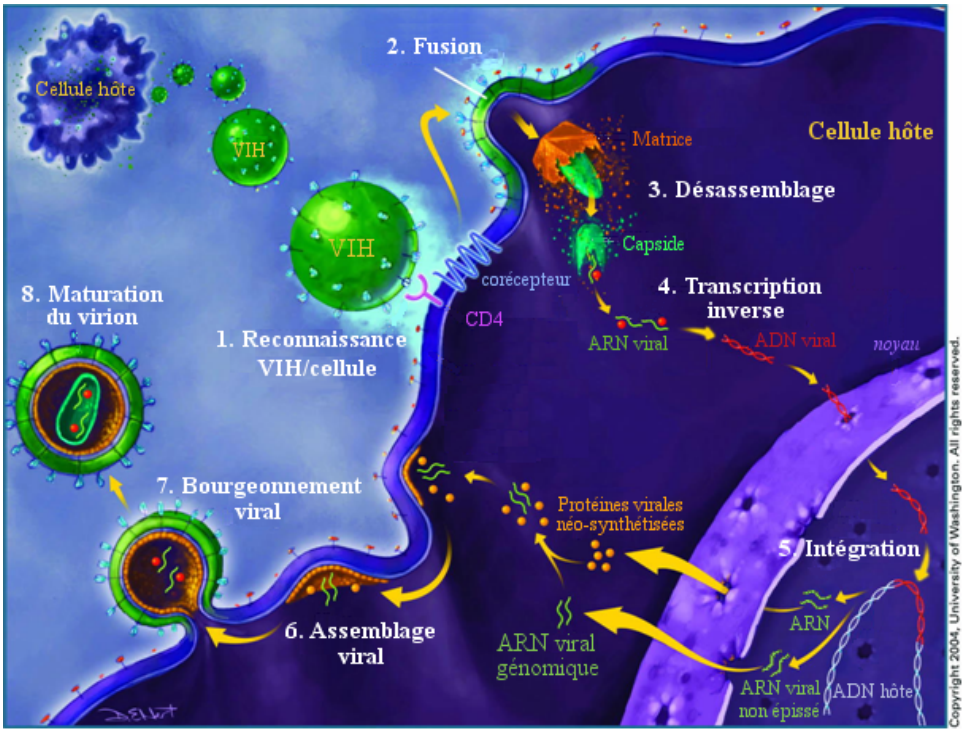


Figure 19 : Cycle de réplication du VIH-1. *Extrait de HIVwebstudy.org*

Les cellules sensibles à l'infection VIH sont essentiellement celles exprimant à leur surface les molécules CD4, CCR5 et CXCR4, nécessaires à l'entrée du virus. Ces cellules incluent notamment :

- la sous-population CD4 (dont la déplétion progressive au cours du temps est caractéristique de l'infection VIH)
- les cellules de la lignée monocytaire-macrophagique (réservoir principal du virus)
- les cellules dendritiques
- les cellules de Langerhans de l'épiderme, du sang et des muqueuses.

De même les lymphocytes B, T $CD8^+$ et les cellules NK peuvent être infectés par le VIH, mais la démonstration *in vivo* reste à confirmer [87]. Cependant, l'expression en surface des molécules nécessaires à l'entrée du VIH n'implique pas forcément une infection, les adipocytes par exemple, expriment CD4, CCR5 et CXCR4 et ne sont pas sensibles à l'infection *in vivo* [88, 89].

Des cellules telles que certains précurseurs hématopoïétiques, les fibroblastes et certaines cellules intestinales et nerveuses, ne présentent pas les récepteurs nécessaires à l'entrée du VIH à leur surface et s'avèrent pourtant sensibles à l'infection. Il existe donc des mécanismes de pénétration du virus dans la cellule différents de ceux dont nous avons parlé jusqu'ici, mais ils sont mal connus.

V.1.3 Evolution clinique

La progression de l'infection par le VIH-1 en l'absence de traitement anti-retroviraux peut être suivie à l'aide de deux indicateurs biologiques évoluant en sens opposé : la charge virale et le nombre de lymphocytes T $CD4^+$ dans le sang. Elle se découpe en trois phases successives : la primo-infection, la phase de latence clinique et la phase symptomatique ou phase SIDA (figure 20).

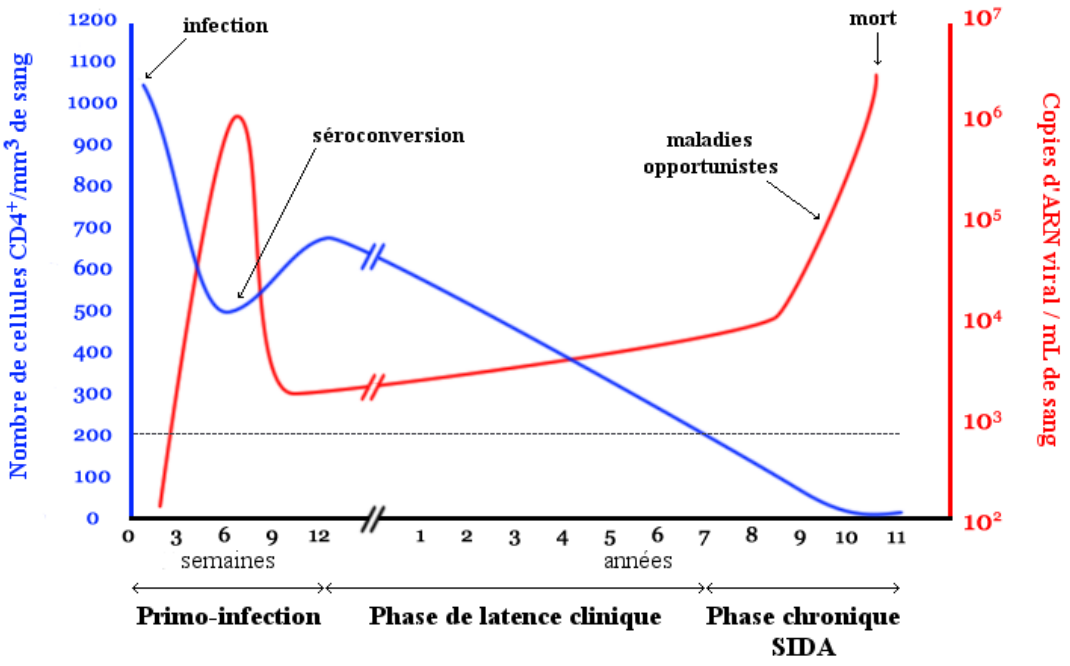


Figure 20 : Profil d'évolution de l'infection par le VIH-1

1) Phase de primo-infection

Elle peut être asymptomatique ou symptomatique. Lorsqu'elle est symptomatique (50% à 70% des cas), elle se manifeste par des signes généraux peu spécifiques : fièvre, pharyngite, fatigue, myalgies, courbatures, éruptions cutanées... Dans certains cas, des symptômes plus évocateurs de l'infection VIH, tels qu'une lymphodénopathie généralisée ou éruption fébrile, peuvent être constatés. Les premiers signes de primo-infection apparaissent en moyenne 20 jours après la contamination. La primo-infection peut durer 3 à 6 semaines, cette phase correspond à une multiplication intense du virus provoquant alors une chute brutale du nombre de cellules $CD4^+$. Par la suite le nombre de $CD4^+$ augmente et la charge virale diminue. Cet événement reflète l'activation du système immunitaire et des réponses humorales et cellulaires, caractérisées par l'apparition dans le sang d'anticorps dirigés contre les protéines du virus et de lymphocytes T cytotoxiques (CTL, Cytotoxic T Lymphocyte) : le sujet est alors séropositif pour le VIH.

2) Phase de latence clinique

Une période de latence clinique suit la primo-infection, cette phase est caractérisée par un équilibre entre la production virale et son élimination. En conséquence, le sujet infecté ne présente pas de signe clinique lié à l'infection [90]. La durée de cette phase est variable, pouvant aller de quelques mois à plusieurs dizaines d'années, avec une moyenne entre 7 et 10 ans sans traitement. La période précédant la phase SIDA est définie par une augmentation de la virémie et une détérioration progressive du système immunitaire attestée par la chute du taux de lymphocytes T $CD4^+$. Selon la classification CDC1993, lorsque le seuil de 200 cellules CD4+/mm3 de sang est atteint, le sujet entre dans le stade symptomatique.

3) Phase symptomatique

La phase symptomatique, qui correspond à la phase SIDA à proprement parler, se manifeste principalement par des infections opportunistes sévères. La moyenne de survie des patients parvenus au stade SIDA est brève, de l'ordre de deux ans.

V.1.4 Les traitements actuels

Depuis l'introduction de l'AZT en 1987 [91], de nombreux progrès ont été réalisés dans la thérapie anti-rétrovirale notamment avec l'utilisation de la trithérapie (HAART).

Lors du développement de médicaments antiviraux, il est important d'interrompre le cycle viral sans tuer la cellule hôte. Ainsi, le premier médicament anti-VIH visait la transcriptase virale, la transcription inverse étant un processus absent des cellules eucaryotes. Les inhibiteurs de transcriptases virales sont séparés en deux classes : NRTI (Nucleoside Reverse Transcriptase Inhibitor) et NNRTI (Non Nucleoside Reverse Transcriptase Inhibitor); les NRTI sont des inhibiteurs compétitifs de la transcriptase inverse, ils se fixent au niveau du site actif et sont reconnus comme des nucléotides. Ils sont donc incorporés dans l'ADN et provoquent l'arrêt de la synthèse provirale. L'AZT (Zidovudine®) et l'Abacavir (Ziagen®) font partis de cette famille d'inhibiteurs NRTI. Les NNRTI sont des inhibiteurs non compétitifs de la transcriptase virale c'est à dire qu'ils ne ciblent pas le site actif.

Par la suite, sont apparus les inhibiteurs de la protéase virale. Ces molécules anti-VIH, telles que le Rinatovir (Norvir®) ou l'Indivanir (Crixivan®), sont des analogues mimant un peptide de liaison au niveau du site actif de la protéase.

Pour ces deux classes d'inhibiteurs, il a été rapidement constaté une pression de sélection, avec l'émergence de virus mutants résistants à ces molécules antivirales (entre quelques semaines et quelques mois) aboutissant à l'échec des monothérapies. Cette observation a mené au développement de thérapies combinées dès le début des années 1990 avec les trithérapies. Les trithérapies actuelles sont composées de deux NRTI combinés à un NNRTI ou à un inhibiteur de protéase.

Plus récemment, de nouvelles classes d'anti-rétroviraux sont apparues. Le Maraviraoc (Celsentri®) est une petite molécule bloquant l'interaction entre le corécepteur de chimiokine CCR5 [92] et la protéine virale gp120, qui inhibe ainsi l'infection des cellules par les virus R5. Le T20 (Fuezon®) bloque les changements conformationnels de la protéine gp41 nécessaire à la réalisation de l'étape de fusion entre le virus et la cellule cible. Enfin, des anti-intégrases ont vu le jour tels que l'Insentress (Raltégravir®) (80 Riordan 2009).

Tous ces traitements ont permis d'améliorer la qualité et l'espérance de vie des patients infectés par le VIH. Cependant, comme évoqué précédemment, ils ne permettent pas l'éradication du virus. De plus, ces traitements sont à l'origine de nombreux effets secondaires indésirables, de virus résistants, et enfin, ces traitements sont très onéreux. Toutes ces considérations encouragent le développement de nouvelles thérapies anti-VIH, le graal étant ici bien entendu la mise en place d'un vaccin prophylactique ou curatif.

V.1.5 Profils d'évolution particuliers

1) Les non-progresseurs à long terme

Le recul au niveau de l'épidémie du SIDA a permis d'observer différents profils de progression vers la phase SIDA. En effet, certains individus infectés par le VIH-1 depuis de nombreuses années ne présentent aucun signe de progression. Ces sujets appelés « non-progresseurs à long terme » (LTNP) , ils sont asymptomatiques et présentent un taux de lymphocytes T $CD4^+$ stable sans prise de traitement anti-rétroviral. Ils représentent moins de 1% des patients séropositifs [93]

Il existe un sous-groupe particulier de LTNP, les « elite controllers » [94, 95] , qui ne présentent aucune charge virale (ou une charge virale très faible selon les définitions). Bien qu'ayant un phénotype précis, leurs profils immunologique et génétique sont hétérogènes [96].

2) **Les exposés non infectés**

Il existe un autre profil particulier d'individus face à l'infection par le VIH-1: en effet, certains sujets, après des expositions répétées au VIH-1, ne sont pas infectés : les HEPS [97] (Highly Exposed Persistently Seronegative). Jusqu'ici ce type de résistance a pu être observé chez les prostituées, les usagers de drogue par voie intraveineuse, les enfants nés de mères infectées... L'étude de ce type de résistance face à l'infection présente un grand intérêt car elle pourrait permettre de développer de nouvelles voies dans l'élaboration d'une thérapie ou d'un vaccin.

Un facteur génétique important permettant d'expliquer cette résistance est la délétion de 32 bases au niveau du co-récepteur CCR5 [98-100] (mutation CCR5-Δ32). Il y a trois ans, un patient sidéen atteint d'une leucémie a reçu une greffe de moelle osseuse provenant d'un sujet porteur de cette mutation homozygote, et depuis sa charge virale est devenue indétectable [101]. Ce résultat a justifié la mise en œuvre de traitements visant à bloquer le récepteur CCR5-Δ32 dans le SIDA. De tels traitements sont expérimentés aujourd'hui, soit sous forme de petites molécules anti-CCR5 [102], soit sous forme de thérapie génique où les cellules de moelle osseuse sont traitées de façon à détruire le gène CCR5 *ex-vivo* et sont ensuite réinjectées au patient séropositif [103].

V.2 Approches génétiques du SIDA par gènes-candidats.

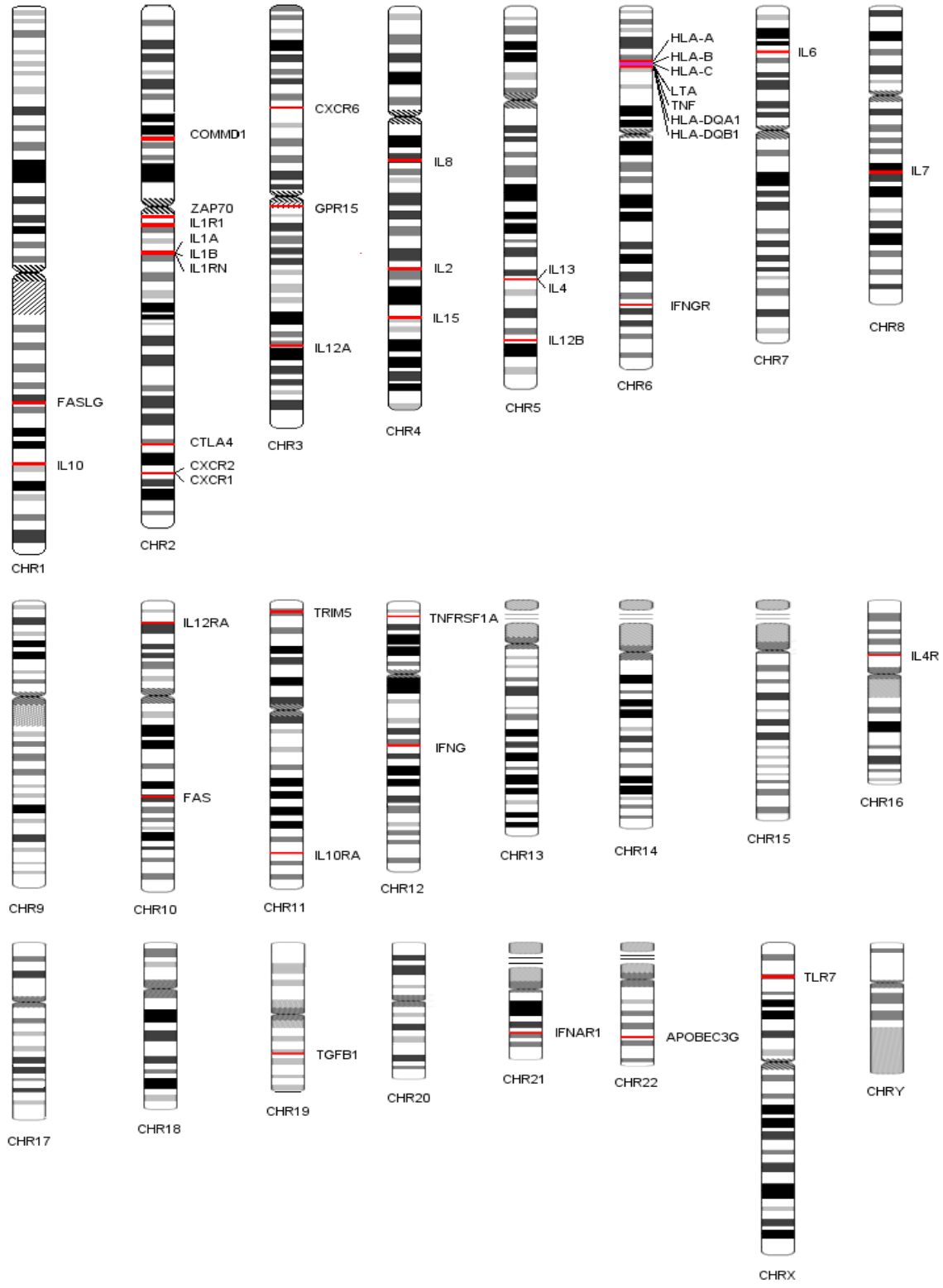


Figure 21 : Listes des gènes candidats testés dans le cadre de la cohorte GRIV avant 2007

Au début de ma thèse en 2006, plusieurs études d'association génétique avaient été réalisées en France et à l'étranger au travers de l'approche gène-candidat (figure 21).

Les principaux résultats d'association portaient sur les variants du locus des corécepteurs du VIH-1 avec les variants *CCR5-Δ32* [104-106], *CCR5-P1* [107, 108] et *CCR2-64I* [5], ainsi que sur les variants du HLA avec notamment *HLA-B35* [61, 109] et *HLA-B57* [110, 111].

De nombreuses autres études ont appréhendé les gènes candidats ciblant notamment des cytokines. J'ai par ailleurs eu la chance d'être en charge de l'analyse statistique de plusieurs publications portant sur des gènes candidats dans le SIDA [112-117].

On peut noter l'excellente revue de An et al. qui répertorie tous les variants génétiques étudiés [118].

V.3 Projet Génétique de la Résistance à l'Infection par le VIH-1 (GRIV)

L'observation de profils d'évolutions particuliers est à la base du projet GRIV (Génétique de la Résistance face à l'Infection par le VIH) [119] qui a pour but de déterminer quels sont les facteurs génétiques de l'hôte qui entrent en jeu durant l'infection. Ce projet est soutenu par Sidaction depuis 1995 et permet une compréhension plus approfondie des mécanismes impliqués dans la pathogenèse du virus.

La cohorte est constituée de deux populations à profil extrême de progression :

- Les **non-progresseurs** composées d'environ 300 personnes qui ont la particularité de résister longtemps à l'infection par le virus (voir la partie matériels et méthodes pour leur description plus complète).
- Les **progresseurs rapides** composées d'environ 100 personnes qui ont la particularité d'aller très rapidement dans la phase SIDA.

La cohorte GRIV constitue ainsi la plus grande cohorte d'individus VIH^+ à profil extrême de progression au monde. L'étude des extrêmes confère une grande puissance statistique [119, 120], et représente un outil de travail exceptionnel pour l'identification de facteurs génétiques associés à la progression de l'infection VIH-1.

Le projet GRIV se place dans un contexte d'étude génétique « cas-contrôles » : les individus extrêmes PR et LTNP sont comparés à des sujets séronégatifs de même origine ethnique (cohortes contrôles **SU.VI.MAX** et **DESIR)**.

L'analyse de ces patients a commencé par l'étude de gènes candidats [119] depuis 1996 puis ensuite sur des puces de génotypage Illumina 317K.

V.4 Autres études « génome entier » dans le SIDA

V.4.1 Étude Euro-CHAVI

Cette cohorte a été la première à avoir été étudiée par approche « génome entier » dans le SIDA, en 2007 [121], et réunit 486 patients séropositifs d'origine européenne suivis depuis leur séroconversion (consortium de 9 cohortes provenant d'Angleterre, Australie, Danemark, Suisse, Espagne et Italie). Les génotypes obtenus sur ces patients ont été étudiés selon deux phénotypes, la charge virale plasmatique stable au cours de la phase asymptomatique et le phénotype de progression défini par la durée avant l'initiation d'un traitement ou par la durée avant la chute du taux de cellules T $CD4^+$ sous 350/µL de sang.

Cette étude a permis de révéler deux signaux passant le seuil de significativité de Bonferroni, les SNPs **rs2395029** du locus *HCP5/HLA-B*57* et rs9264942 du gène *HLA-C* associé au contrôle de la charge virale ainsi que l'association du locus *ZNRD1/RNF39* avec la progression de l'infection.

V.4.2 Étude PRIMO

Parallèlement à notre analyse « génome entier » sur la non progression à long terme, une autre étude d'association de ce type a été réalisée sur la cohorte française PRIMO de patients séroconvertis composée de 605 patients [122].

L'analyse sur la cohorte PRIMO est basée sur deux phénotypes distincts : le niveau d'ARN viral plasmatique et le niveau d'ADN viral dans les PBMC (Peripheral Blood Mononuclear Cells). Pour les SNPs significatifs selon le FDR≤25%, la fréquence allélique au sein de la cohorte PRIMO a été comparée avec celle observée dans une population de *controllers* du VIH (charge virale <400 copies/mL après 10 ans d'infection, n=45).

Au niveau de l'ARN viral, le seul SNP passant le seuil de Bonferroni est le **rs10484554** ($3{,}58 \times 10^{-9}$) localisé dans une région intergénique proche des gènes *HLA-C* et *HLA-B*. Sur 15 SNPs

appartenant au locus 6p21 et passant le seuil statistique FDR≤25%, 4 SNPs (rs2395029, rs13199524, rs12198173 et rs3093662) présentent des différences significatives au sein de la cohorte de *controllers*. Dans les régions en dehors du chromosome 6, seul le polymorphisme rs11725412 du chromosome 4 (p=5,97x10^{-6}) est répliqué dans la cohorte de *controllers* du VIH : p=6,58x10^{-4}. Ce polymorphisme est intergénique et les gènes les plus proches sont *TBC1D1* et *KLF3*. *TBC1D1* pourrait être impliqué dans la régulation de la prolifération et la différenciation cellulaire [123] alors que *KLF3* coderait pour un facteur de transcription [124].

Au niveau de l'ADN viral, le meilleur résultat concerne le SNP **rs2395029** (p=6,72x10^{-7}) localisé dans le gène *HCP5*. Ce signal ne passe pas le seuil de significativité classique, mais il est tout de même supporté par l'approche statistique du FDR avec un taux d'erreur de seulement 1,4%.

Cette étude « génome entier » a souligné l'importance de la région *HLA* pour le contrôle de la charge virale ARN plasmatique et ADN cellulaire et a également soulevé de nouvelles associations potentielles pour le contrôle des charges virales.

V.4.3 Étude MACS156

En 2009 une GWAS a été conduite sur un sous-groupe de la cohorte américaine d'origine européenne MACS [125], enrichi en patients à profil extrême de progression (n=156) : 51 progresseurs rapides, 57 progresseurs modérés et 48 non-progresseurs à long terme.

Cette étude a mis en évidence la région du gène *PROX1* (chromosome 1) . *PROX1* est impliqué dans des fonctions biologiques liées à l'infection VIH, en particulier en tant que régulateur négatif de l'expression *IFNγ* par les cellules T [126] . L'*IFNγ* joue un rôle important dans la progression vers le SIDA via ses activités d'immunorégulation et d'activation de l'inflammation.

Cette étude « génome entier » a ainsi dévoilé un nouveau candidat potentiel pour les mécanismes moléculaires de pathogenèse de l'infection VIH. Il est important de préciser que le SNP rs2395029 n'est pas ciblé par les puces Affymetrix utilisées dans cette étude.

V.4.4 Étude sur une cohorte afro-américaine

Toujours en 2009 est parue la première étude « génome entier » sur une cohorte d'afro-américains composée de 515 individus [127]. Les génotypes obtenus sur ces patients ont été étudiés selon un phénotype : charge virale plasmatique stable au cours de la phase asymptomatique.

Aucun signal n'a atteint le seuil de significativité « génome entier ». L'association la plus

forte obtenue dans la région *HLA* concerne le SNP **rs2523608** localisé dans le gène *HLA-B* ($p=2,3x10^{-6}$). L'analyse fine des allèles du *HLA-B* au sein de 285 participants a permis de déceler un fort déséquilibre de liaison entre ce polymorphisme et l'allèle HLA-B*5703 ($r^2=0,075$ et D'=1). Individuellement, cet allèle *HLA-B*5703* exprime une très forte association avec la charge virale, respectant même le seuil de significativité « génome entier » ($p=5,6x10^{-10}$), et explique ~10% de la variabilité de la charge virale au sein de cette cohorte afro-américaine. De façon similaire à ce qui a été observé dans les populations d'origine européenne, un allèle *HLA-B*57* a été identifié dans les individus afro-américains comme l'association majeure impliquée dans le contrôle de la charge virale suggérant un mécanisme commun.

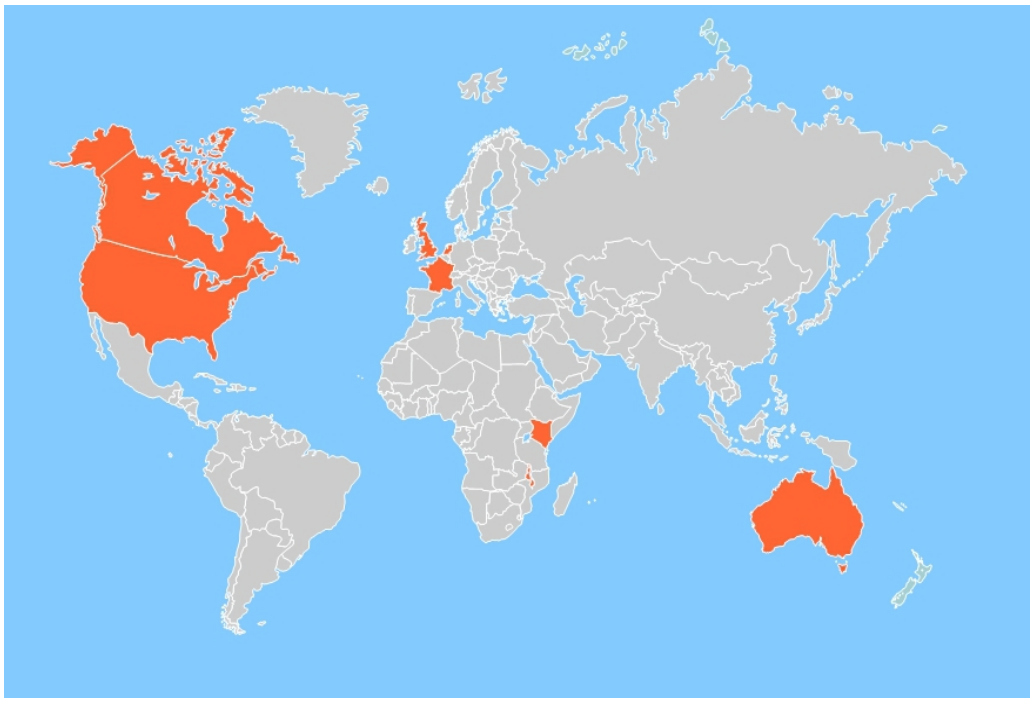
De faibles associations sont observées pour les polymorphismes rs2395029 (très rare dans les populations d'origine africaine -0,8% ici- ; p=0,030) et rs9264942 (p=0,018), ne suggérant donc pas de rôle majeur de ces SNPs pour le contrôle de la charge virale dans la population afro-américaine, contrairement à ce qui est observé dans les populations d'origine européenne.

De précédentes études avaient déterminées que l'allèle *HLA-B*5703* pouvait être impliqué dans le contrôle de la charge virale au sein de populations d'origine africaine [128] mais cette étude « génome entier » a mis en avant cet allèle comme l'association la plus significative affectant le contrôle viral dans cette population.

V.4.5 Étude génome entier à partir de phénotypes cellulaires

Des études « génome entier » ont également été réalisées pour caractériser de nouveaux gènes comme *DYRK1A* localisé sur le chromosome 21 codant pour une kinase qui serait impliquée dans la réplication de VIH-1 *in vitro* [129].

V.5 Projet International HIV Acquisition Consortium (IHAC)



En octobre 2009, sous l'égide du NIH américain, une réunion a eu lieu entre différents groupes internationaux travaillant sur la génétique du SIDA et ayant utilisé des études à grandes échelles (GWAS) sur des patients séropositifs. Ceux-ci se sont réunis pour mettre en commun les données de leurs patients et étudier le phénotype d'acquisition du VIH-1. Une collaboration internationale (International HIV Acquisition Consortium) a débuté donnant l'accord de plusieurs groupes à travers le monde pour collecter les données "genome entier" de SNPs de patients seropositifs. En corrolaire il fut également recolté un grand nombre de patients contrôles, c'est-à-dire non infectés par le virus du VIH. Le but de la collaboration est de comparer les génotypes des patients pour tenter de détecter une éventuelle association génétique avec la contamination par le virus.

L'ensemble de ces équipes est décrit dans le tableau ci-dessous et détaillé dans les Matériels et Méthodes. Cette collaboration a abouti à la constitution de six cohortes de patients caucasiens cas-contrôles regroupant 13785 individus. Ces six cohortes ont été construites en fonction de la plate-forme de génotypage utilisée à l'origine (Illumina ou Affymetrix) et de l'homogénéité génétique des patients afin d'éviter au maximum les biais de stratification.

Le phénotype étudié est celui de l'infection par le virus (« HIV-1 acquisition »), les seules informations disponibles étant l'infection ou non des patients par le virus VIH-1, leur sexe et leur

origine ethnique (Européenne, Africaine ou Afro-américaine). Ce projet est constitué d'une première étape importante au cours de laquelle les équipes de ce consortium vont apprendre à travailler ensemble sur le phénotype d'acquisition du virus.

Ensuite, une deuxième étape de mise en commun des données cliniques aura lieu afin d'étudier les facteurs génétiques de progression vers la maladie (charge virale ou clinique).

VI. Objectifs de thèse

Je suis arrivé dans l'équipe de recherche du Pr Zagury comme stagiaire en 2005 pour développer des outils bioinformatiques d'exploitation des polymorphismes identifiés au niveau des gènes candidats de la cohorte GRIV.

Avec l'avènement des données génomiques haut-débit des puces en 2007, l'Agence Nationale de Recherches sur le SIDA et les Hépatites (ANRS) a décidé de lancer un projet de génomique sur deux cohortes françaises, la cohorte PRIMO et la cohorte GRIV.

J'ai décidé de m'engager sur une thèse pour l'exploitation de ce nouveau type de données haut-débit issues de puces Illumina dans la cohorte GRIV.

Au cours du temps, mon projet de thèse s'est dessiné sur 3 niveaux:

1. Mettre en place des outils pour exploiter l'information issue des haplotypes, ce qui a conduit au logiciel SUBHAP.
2. Mettre en place les outils bioinformatiques et statistiques d'exploitation des données de puces de génotypage. Ce travail a conduit à plusieurs publications dont notamment celle portant sur le phénotype non-progresseur non « elite », qui a conduit à l'identification d'un nouveau locus associé à ce phénotype..
3. Utiliser le savoir-faire précédemment acquis sur la cohorte GRIV pour servir de centre d'analyse européen de référence pour le projet de méta-analyse IHAC.

Matériels et Méthodes

I. Description des cohortes

I.1 Populations françaises

I.1.1 Étude GRIV

La cohorte **GRIV** [119] est composée de patients présentant un profil extrême dans la résistance aux VIH-1 et est constituée de 85 prog**resseurs rapides** (PR) et 275 **non-progresseurs à long terme** (LTNP). Les PR sont définis par une chute du taux de cellules T $CD4^+$ à moins de $300/mm^3$ moins de 3 ans après le dernier test séronégatif. Les LTNP sont des individus séropositifs et asymptomatiques depuis plus de 8 ans, et présentant un taux de cellules T $CD4^+$ supérieur à $500/mm^3$ en absence de tout traitement antirétroviral.

Les groupes PR et LTNP étudiés dans le cadre de ce travail de thèse sont respectivement composés de 73 hommes et 12 femmes âgés à l'inclusion entre 21 et 55 ans (médiane=32), et de 201 hommes et 74 femmes âgés à l'inclusion entre 19 et 62 ans (médiane=35). A l'inclusion, la médiane du taux de cellules T $CD4^+$ était de $230/mm^3$ pour la population PR (min.-max.=20-297), et de $706/mm^3$ pour la population LTNP (min.-max.=501-2298). Divers autres paramètres phénotypiques à l'inclusion ont été mesurés comme la charge virale plasmatique.

Le projet GRIV se déroule en 4 phases successives :

1. Le prélèvement d'un échantillon de sang sur chaque sujet satisfaisant aux critères d'inclusion sus-mentionnés (taux de CD4, etc). Les patients retenus proviennent d'hôpitaux de toute la France et sont tous de type Caucasien afin de limiter la diversité génétique.
2. Le génotypage des sujets inclus afin de déterminer leurs variations génétiques. Jusqu'en 2006 par la technique du séquençage, puis en 2007 par la méthode SNPlex permettant de sélectionner une région précise d'intérêt (la région 21q22). Enfin, plus récemment, par le biais de puces de génotypage Illumina de 317K SNPs.
3. L'analyse et l'exploitation des données ainsi récoltées par des outils bioinformatiques et biostatistiques.
4. L'interprétation biologique des résultats obtenus.

I.1.2 Population contrôle SU.VI.MAX

L'étude SU.VI.MAX [130] (Supplémentation en Vitamines et Minéraux Antioxydants) a été initialement développée pour évaluer l'effet d'une supplémentation nutritionnelle quotidienne en vitamines et minéraux antioxydants sur la réduction de problèmes de santé publique, tels que les cancers et maladies cardio-vasculaires. Le groupe contrôle SU.VI.MAX dont nous avons disposé pour réaliser nos études « génome entier » est composé de 1352 individus, tous d'origine européenne vivant en France et séronégatifs pour le VIH-1. Cette population regroupe 525 hommes et 827 femmes, âgés en moyenne respectivement de 53,1 et 48,5 ans.

I.1.3 Population contrôle DESIR

L'étude DESIR [131] (Data from an Epidemiological Study on Insulin Resistance syndrome) consiste en un suivi de 9 ans du développement du syndrome d'insulino-résistance. Le groupe contrôle utilisé dans nos études « génome entier » est composé de 697 participants à ce programme, tous non obèses, normo-glycémiques, d'origine européenne vivant en France et séronégatifs pour le VIH-1. Cette cohorte regroupe 281 hommes et 416 femmes, âgés entre 30 et 64 ans.

I.2 Autres populations étudiées

I.2.1 Cohorte hollandaise ACS

L'étude ACS [132, 133](Amsterdam Cohort Study) regroupe 335 hommes homosexuels et toxicomanes infectés par le virus VIH-1. La moyenne d'age des participants au moment de leur contamination est estimée à 31,5 ans (entre 19 et 55 ans) . Les phénotypes étudiés sont des données de survie parfois censurées concernant le temps d'apparition du SIDA suivant la définition du CDC 1993 (Centers for Disease Control) et la durée de survie avant le décès du patient. Divers autres paramètres sont longitudinalement mesurés comme la charge virale plasmatique.

I.2.2 Cohortes américaines MACS156

L'étude MACS (Multicenter AIDS Cohort Study) est constituée de 156 patients homosexuels américains blancs infectés par le virus VIH-1 [125]. C'est un groupe de patients issus d'une cohorte plus importante choisi pour leurs profils dans la progression vers le SIDA, parmi ceux-la 51 sont des progresseurs rapides et 59 sont des non-progresseurs suivant la définition de GRIV. Il s'agit là encore d'une étude longitudinale observant la durée entre la séroconversion et

l'apparition du SIDA et le délai de survie avant décès principalement.

I.2.3 Cohortes du projet IHAC

Les cohortes inclues dans le projet IHAC sont diverses et regroupent en partie les cohortes déja décrites précédemment. Elles incluent des patients contaminés par usage de drogue, par voie sexuelle ou par transfusion sanguine ainsi que des cohortes avec des profils particuliers comme dans le cas de GRIV. Les phénotypes associés à chacun des patients en dehors du sexe ne sont pas pour l'heure communiqués au consortium.

Après un contrôle qualité effectué pour chaque cohorte (IBS pour détecter les doublons et les individus apparentés) il résulte 34 jeux de données distincts issus de 8 plate-formes de génotypage différentes (voir figure 24). Au total 9978 patients cas VIH+ et 9504 patients contrôles VIH- furent collectés (avec une grande proportion de patients ayant une origine européenne, figure 22).

Origine	Phénotype	Nombre
Européens	Cas	6556
	Contrôles	7253
	Total	**13809**
Afro-Americains	Cas	2542
	Contrôles	1416
	Total	**3958**
Africains	Cas	880
	Contrôles	835
	Total	**1715**

Figure 22 : Nombre total de patients inclus dans le projet IHAC intial

Le nombre de patients infectés par le virus VIH-1 étudiés dans ce projet est sans précédant. Néanmoins, l'hétérogénéité des cohortes rend impossible le regroupement de tous les patients dans une seule analyse. Il fallut donc regrouper les patients par plate-forme et par origine ethnique pour diminuer les biais de stratification.

La stratégie d'analyse choisie par le consortium fut d'abord d'étudier les cohortes dont les patients sont d'origine européenne puis d'éventuellement confirmer les résultats à travers les autres origines ethniques (Afro-américaine et Africaine).

Deux pôles d'analyse ont été choisis :

Description des cohortes

- Le Broad Institute à Boston aux États-Unis
- Le Conservatoire National des Arts et Métiers (CNAM)

Les analyses sont réalisées en parallèle afin de s'assurer de la concordance des résultats et d'évaluer les stratégies d'analyse (utilisation de logiciels différents, approches différentes ou complémentaires). A l'issue d'un travail de contrôle qualité et d'homogénéisation des populations dans l'analyse cas-contrôles, un consensus s'est dégagé en Juillet 2011 pour analyser six cohortes distinctes d'origine européenne (figure 23).

Nom	Phénotype	Nombre	SNPs	Plate-Forme
DUTCH	Cas	401	297 364	Illumina
	Contrôles	998		
	Total	**1399**		
FRENCH	Cas	850	282 612	Illumina
	Contrôles	672		
	Total	**1522**		
ILLUMINA 1	Cas	2759	393 414	Illumina
	Contrôles	2759		
	Total	**5518**		
ILLUMINA 2	Cas	986	281 865	Illumina
	Contrôles	513		
	Total	**1499**		
USA 1	Cas	1185	383 698	Affymetrix
	Contrôles	1190		
	Total	**2375**		
USA 2	Cas	357	393 401	Affymetrix
	Contrôles	1115		
	Total	**1472**		

Figure 23 : Détail des six cohortes cas-contrôles obtenues après correction de la stratification étudiées dans le projet IHAC

Description des cohortes

Nom de la cohorte	Origine	Phenotype	Puces	Nb SNPs	Nb Individus
407_lgd_cel_calls	Européens	Cas	Affy_6.0	713803	22
LGD_eur_gwas_ihac2	Européens	Cas	Affy_6.0	651432	839
550_calls_095conf.EUR	Européens	Cas	Affy_6.0	681432	283
920_calls_095conf.EUR	Européens	Cas	Affy_6.0	691531	421
IHCS.phase2.eur	Européens	Cas	Illumina_1M	877698	578
IHCS.phase3.eur	Européens	Cas	Illumina_1M	819058	580
MACS_Duke_1204whites_1M	Européens	Cas	Illumina_1M	792119	768
MACS_Duke_1204whites_550	Européens	Cas	Illumina_550	487406	422
IHCS.phase1.eur	Européens	Cas	Illumina_650	526384	505
GRIV_Cas	Européens	Cas	Illumina_300	300916	346
PRIMO	Européens	Cas	Illumina_300	306828	581
ACS_Cas	Européens	Cas	Illumina_370	305191	408
Other_EuroCHAVI2	Européens	Cas	Illumina_550	484825	436
SHCS_EuroCHAVI2	Européens	Cas	Illumina_550	489007	829
Additional_SHCS	Européens	Cas	Illumina_650	555981	242
MIGen	Européens	Contrôles	Affy_6.0	690614	5802
iControlDB_Europeans	Européens	Contrôles	Illumina_550	494344	3050
GRIV_CTR_GRP1	Européens	Contrôles	Illumina_1M	798283	518
GRIV_CTR_GRP2	Européens	Contrôles	Illumina_300	299208	678
ACS_CTR	Européens	Contrôles	Illumina_370	307304	1011
GWA-RS3-v1	Européens	Contrôles	Illumina_610	546038	1604
LGD_aa	Afro Americains	Cas	Affy_6.0	762441	775
920_calls_095conf.AfrAm	Afro Americains	Cas	Affy_6.0	659308	44
550_calls_095conf.AfrAm	Afro Americains	Cas	Affy_6.0	648647	32
Duke_DOD_AfrAmer	Afro Americains	Cas	Illumina_1M	924938	464
IHCS.phase2.aa	Afro Americains	Cas	Illumina_1M	951928	379
IHCS.phase3.aa	Afro Americains	Cas	Illumina_1M	885158	412
MACS_Duke_144AA_1M	Afro Americains	Cas	Illumina_1M	875640	123
MACS_Duke_144AA_550	Afro Americains	Cas	Illumina_550	480003	9
IHCS.phase1.aa	Afro Americains	Cas	Illumina_650	593194	380
iControlDB_Afro Americains	Afro Americains	Contrôles	Illumina_550	510199	1476
Pumwani.HIV.infected	Africain	Cas	Affy_5.0	389339	358
Duke_Africains_4	Africain	Cas	Illumina_1M	841068	522
MALAWI_neg_4	Africain	Contrôles	Illumina_1M	842171	835

Figure 24 : Détail des différentes cohortes utilisées dans le projet IHAC

II. Puces de génotypage

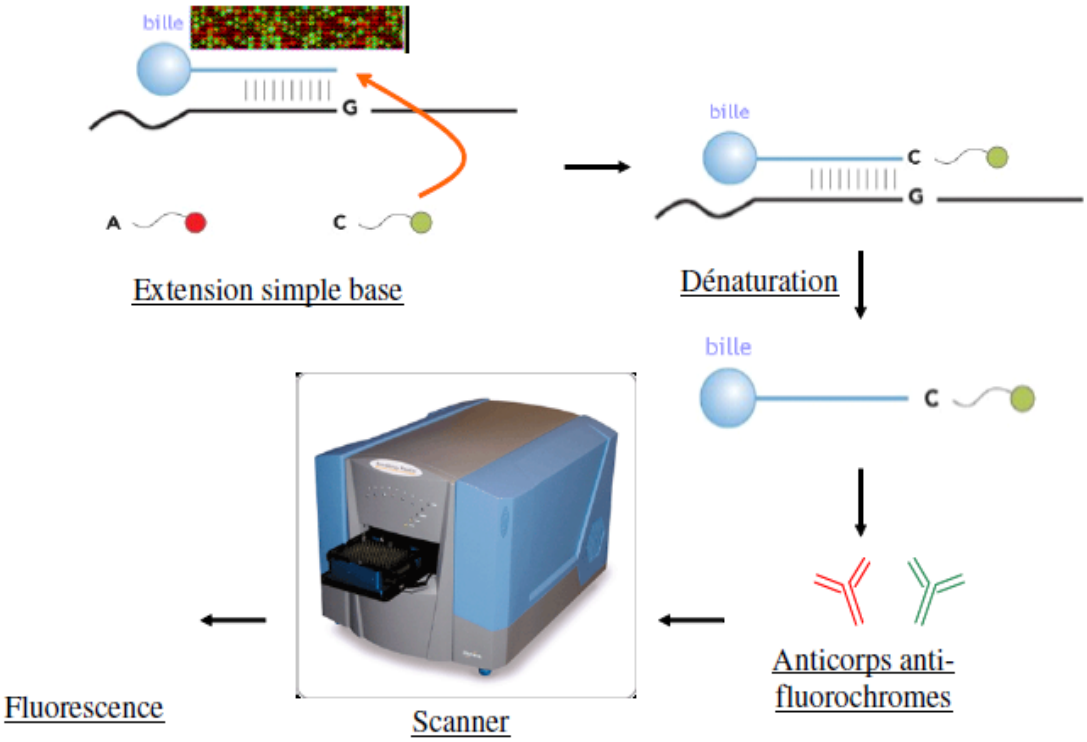
Les puces Illumina HumanHap300 permettent de génotyper les individus sur 317 000 SNPs. Elles ont été élaborées d'après la Phase I du projet HapMap. Il a été estimé que l'ensemble des SNPs fréquents de la Phase I du projet HapMap pouvait être capturé par ~294 000 tagSNPs avec un seuil $r^2 > 0,8$. Dans cette optique, les puces Illumina HumanHap300 ciblent surtout des tagSNPs fréquents (>5%) de la Phase I de HapMap avec un seuil de $r^2 > 0,8$ pour les régions génétiques situées dans les gènes à ±10kb et pour les régions génétiques conservées au cours de l'évolution et avec un seuil de $r^2 > 0,7$ pour les autres régions génétiques. Ces puces ont également été enrichies avec ~8 000 SNPs exoniques non synonymes et avec ~1 500 SNPs de la région génétique complexe, mais importante, du HLA.

La technologie des puces de génotypage offre de nombreux avantages:

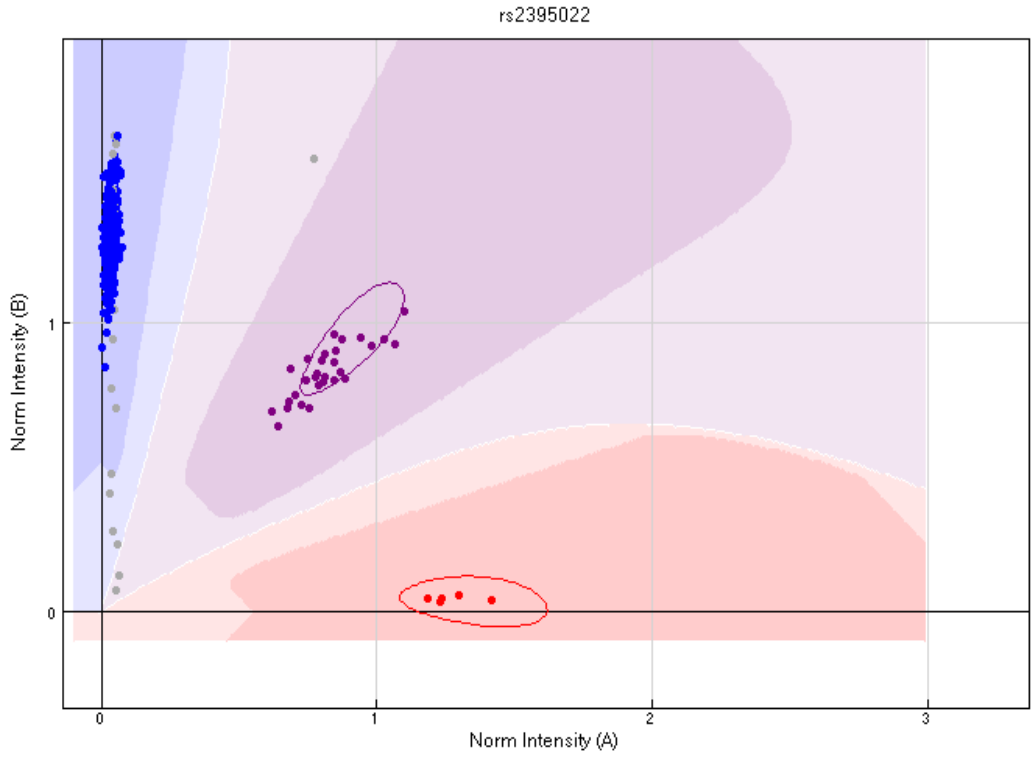
- l'ensemble du protocole est standardisé et automatisé (utilisation de kits et de robots), ce qui offre une grande robustesse et reproductibilité des résultats.
- la consommation d'ADN est faible (750ng) pour la quantité de SNPs génotypés.
- les SNPs ciblés sont répartis sur l'ensemble du génome de façon gène-centrée, ce qui facilite la découverte et l'interprétation de nouvelles associations génétiques dans le cadre de pathologies.

Les puces de génotypage Illumina HumanHap300 se présentent sous la forme de lames de verre, sur lesquelles sont fixées des micro-billes de verre de 3µm de diamètre, qui permettent le génotypage simultané de deux individus grâce à une séparation centrale hermétique (figure 25). Chaque micro-bille est recouverte d'oligonucléotides (50mères) spécifiques d'un SNP et est présente en une vingtaine d'exemplaires sur chaque puce afin d'assurer la réplication et la robustesse des résultats.

N.B. : Dans le procédé HumanHap300, seuls deux fluorochromes sont utilisés : les bases A et T sont couplées au dinitrophényl (DNP) qui est reconnu par un anticorps anti-DNP fluorescent (Cy-5 rouge), et les bases C et G sont couplées à la biotine qui est reconnue par la streptavidine fluorescente (Cy-3 verte). De fait, dans cette technologie, les SNPs A/T et C/G ne peuvent pas être ciblés directement sur les puces. Cette considération est prise en compte par l'utilisation de tagSNPs de HapMap.



(a)



(b)

Figure 25 : (a) Schéma des étapes successives nécessaires au génotypage sur plate-forme Illumina Infinium. **(b)** Sortie de fluorescence du logiciel Beadstudio. On voit clairement les 3 groupes correspondant aux 3 génotypes sur un SNP.

III. Traitement des données

III.1 Contrôle qualité

III.1.1 Données manquantes

Les données brutes issues du génotypage ont été analysées à l'aide du logiciel Illumina BeadStudio v3.1. Les filtres sur les données de fluorescence sont:

- Le « call rate » (pourcentage de SNPs génotypés par individu). S'il est inférieur à 95%, les individus sont éliminés.
- Le « call frequency » (pourcentage d'individus génotypés par SNP). S'il est inférieur à 98%, les SNPs sont supprimés ce qui correspond à tous les SNPs dont le taux de données manquantes est supérieur à 2%.

L'ensemble de ces étapes assure des données de génotypage fiables avec peu de données manquantes.

III.1.2 Déviation de l'équilibre de Hardy-Weinberg

On vérifie pour chaque SNP que l'équilibre d'Hardy-Weinberg est bien respecté, en partant de l'hypothèse qu'un écart important de cet équilibre résultera plutôt d'une erreur de génotypage que d'un réel déséquilibre dû à un concept de génétique des populations comme la sélection, la mutation ou la migration.

En général une erreur de génotypage résulte d'une erreur sur un génotype particulier et l'équilibre de Hardy-Weinberg ne sera pas du tout respecté. Il n'est donc pas nécessaire d'être fortement stringent dans ce contrôle pour éliminer les SNPs mal génotypés.

Lorsque les cas-contrôles ont été génotypés sur la même plate-forme il est préférable de n'éliminer que les SNPs échouant pour ce test uniquement dans la population contrôle, une variation de cet équilibre chez les cas pouvant résulter d'une association avec le phénotype étudié.

On réalise un test exact [41] testant la déviation de cet équilibre. Une p-value inférieure à 10^{-4} ou 10^{-5} est généralement utilisée comme seuil.

III.1.3 Faible fréquences

L'élimination des SNPs de faible fréquence est une étape classique du contrôle de qualité assurant la fiabilité des données de génotypage et facilitant l'analyse statistique postérieure. Les SNPs dont la fréquence de l'allèle mineur est inférieure à 1% dans la population globale sont donc éliminés pour des cohortes de faible taille.

IV. Correction de la stratification

Dans le cadre des études « génome entier », un critère de jugement essentiel sur l'homogénéité des cohortes repose sur l'analyse du Q-Q plot permettant de calculer un **facteur d'inflation** λ. La technique dite du « genomic control » peut utiliser ce facteur lambda pour ajuster les p-values en prenant en compte la stratification sous-jacente [43]. Cependant cette méthode, même si elle peut dans certains cas s'avérer efficace, a plutôt tendance à sur estimer la stratification en corrigeant trop fortement les p-values résultants d'une baisse significative de puissance. Tandis que lorsque peu de marqueurs sont porteurs de la stratification, cette technique aura tendance à ne pas corriger la stratification [134].

Pour corriger l'éventuelle stratification de nos populations au niveau intercontinental, les génotypes de tous les individus (cas et contrôles) ont été analysés en utilisant le logiciel **STRUCTURE**.

Pour cela, un jeu de 328 SNPs informatifs de l'origine ancestrale (index de fixation FST >0,2) d'après les données Perlegen et distants de plus de 5Mb (afin d'éviter le déséquilibre de liaison) a été sélectionné. Les génotypes des individus non apparentés issus des populations du projet HapMap ont également été inclus dans notre analyse, afin de séparer au mieux les individus cas et contrôles selon leur origine continentale, et d'exclure ceux d'origine non européenne.

Enfin la méthode par **Analyse en Composantes Principales** (ACP) en utilisant le logiciel Eigenstrat permet également de détecter et corriger la stratification d'une population en modélisant sur des axes continus de variations les différences génétiques entre les individus.

Par cette méthode on peut :

- Déterminer les patients étant génétiquement trop éloignés des autres qui seront par la suite retirés de l'analyse.
- Inclure les axes de variation en tant que covariables dans les modèles de régression et ainsi

de corriger les p-values du biais de stratification.

IV.1 Analyse d'association

Le type d'analyse centrale à réaliser sur la cohorte GRIV est une analyse **simples marqueurs sur un phénotype binaire** cas versus contrôles (non-progresseurs versus contrôles par exemple). Les tests réalisés sont, au choix, soit un test classique du χ^2 , le test exact de Fisher [40] soit la régression logistique. L'approche de la régression logistique permet d'inclure des covariables comme les axes provenant de l'analyse ACP mais aussi le sexe, d'autres phénotypes externes, ou encore un autre marqueur. Le but étant de tester l'effet spécifique du SNP et de vérifier qu'il est bien indépendant des covariables.

Les **tests multiples** ont été pris en compte en appliquant les corrections de Bonferroni comme principalement demandé par les grands journaux scientifiques. Afin d'identifier des signaux supplémentaires tout en contrôlant le risque de fausses découvertes, nous avons également calculé les q-values par la technique du FDR (False-Discovery Rate). Cette technique étant plus puissante tout en contrôlant le nombre de faux positifs. Cependant l'approche est plus complexe, les méthodes assez diverses et la confiance accordée à ce type de correction est de ce fait moindre. Cependant dans le cadre de maladies multifactorielles -comme le SIDA- dans lesquelles plusieurs gènes sont impliqués, le FDR offre une meilleure perspective sur les résultats « génome entier » que la loi du « tout ou rien » du seuil de Bonferroni. La méthode de FDR d' « estimation locale » (local base estimating) a été appliquée dans nos études avec un seuil de 25%.

Pour chaque SNP passant le seuil statistique de significativité, le contrôle qualité a été intégralement re-vérifié. Puis, pour chaque signal identifié, nous avons vérifié que la fréquence allélique dans une population séropositive était similaire à celle de la population contrôle (*e.g.* vérification dans la population PR pour un signal de LTNP), afin de confirmer que l'association observée reflète bien la progression vers le SIDA et non l'infection VIH-1.

Enfin, nous avons eu l'opportunité de combiner nos données génomiques avec celles d'autres équipes (Euro-CHAVI, ACS, MACS…) et de réaliser des méta-analyses. Pour chaque SNP, les p-values obtenues dans les différentes études ont été combinées en une p-value unique selon la méthode Fisher, la méthode des z-score ou enfin la méthode inverse de la variance suivant le modèle fixe ou aléatoire (fixe s'il n'y avait pas trop d'hétérogénéité entre les cohortes et aléatoire sinon).

Correction de la stratification

Une excellente revue sur les méthodes statistiques pour l'analyse d'association a été faite par Balding [27] en 2006 et reprend la démarche complète.

IV.2 Infrastructures informatiques

IV.2.1 Bases de données

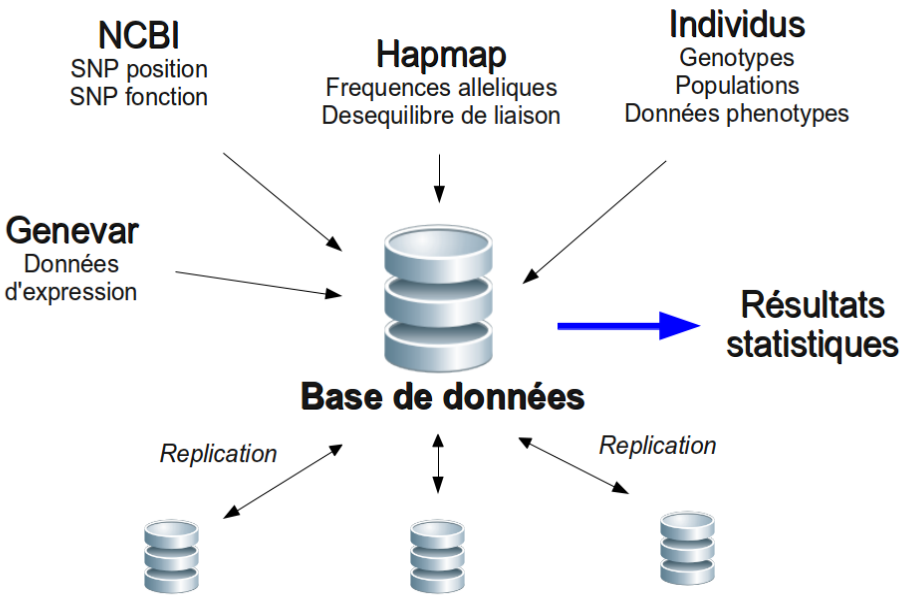


Figure 26 : Représentation schématique des différentes sources d'information nécessaires en même temps pour l'analyse bioinformatique des données génomiques « génome entier »

L'immense masse de données provenant des individus et de leur génotype, des bases de données publiques (Hapmap, dbSNP, Genevar...) ainsi que la masse de résultats provenant des analyses nécessite la mise en place d'une base de données structurées pour permettre à tous les utilisateurs du projet de consulter les informations qui lui sont nécessaires (figure 26). Une base de données a été construite de plus de 20 tables à l'aide du logiciel open-source mysql [135] et l'interrogation de la base fut interfacée via php et des script cgi en perl ou en C.

Ainsi, cela permet de faciliter le traitement de requêtes comme :

Correction de la stratification

« Je veux tous les SNPs :

1. ayant une p-value inférieure à 10^{-4} dans la comparaison des non-progresseurs de GRIV avec les contrôles DESIR
2. ayant une p-value inférieure à 10^{-2} dans la base de données d'expression Genevar
3. qui sont placés dans le promoteur d'un gène donné d'après la base de données dbSNP

IV.2.2 Grappe de calcul

Les nombreux calculs nécessaires (haplotypage et imputation notamment) sont rapidement devenus impossibles sur un seul ordinateur personnel, j'ai donc du mettre en place une grappe de calcul de plusieurs machines en partenariat avec le service informatique du CNAM.

Le but étant de centraliser la gestion des ressources et le traitement des lancements des différents programmes lancés par les différents utilisateurs composant l'équipe. Pour cela la soumission de traitement au «cluster» de calcul est gérée par le gestionnaire Torque (issu du projet original PBS). Torque fournit à la fois un gestionnaire de ressources qui permet de contrôler la charge de chacun des nœuds de la grappe, un gestionnaire de «batch» qui est chargé de transmettre les traitements soumis par les utilisateurs et un ordonnanceur qui permet de programmer l'exécution des traitements sur chacun des nœuds en fonction des ressources disponibles.

Le système NFS constitué de 8To d'espace disque a parallèlement été mise en place.

Le cluster est actuellement constitué de 7 nœuds de calcul représentant l'équivalent de 64 processeurs, toutes ces machines étant basées sur des architectures multiple-processeurs multiple-cœurs INTEL 64 bits. Plus précisément la grappe est constituée de :

- bioinfo10 (8coeurs,32Go de RAM) où se localise le serveur NFS
- bioinfo13 (8coeurs,32Go de RAM)
- bioinfo21 (8coeurs,16Go de RAM)
- bioinfo22 (8coeurs,16Go de RAM)
- bioinfo23 (8coeurs,16Go de RAM)
- bioinfo04 (12coeurs,24Go de RAM)
- bioinfo08 (12coeurs,24Go de RAM)

Correction de la stratification

L'accès aux serveurs se fait via bioinfo10 à l'aide de la commande qsub. De nombreuses autres commandes permettent de suivre l'évolution des jobs, leurs niveaux de priorité, etc.

V. Logiciels utilisés

Tout au long de ma thèse j'ai du utiliser ou programmer de nombreux logiciels pour le traitement des données, pour l'analyse statistique, pour la représentation et l'analyse des résultats et enfin pour la mise en place des infrastructure informatiques.

Contrôle qualité	BEADSTUDIO
	PLINK 1.07[1] [136]
	QCTOOL[7]
	GENABEL [137]
	HAPLOVIEW [138]
Stratification	EIGENSTRAT 4.2 [139]
	STRUCTURE [44]
	PLINK 1.07[1]
	GENABEL [137]
Tagging SNPs	HAPLOVIEW (TAGGER) [19]
	TAGSTER [140]
	FASTTAGGER [141]
Haplotypage	SHAPEIT[2]
	PHASE 2.1 [50]
	FASTPHASE [142]
	PL-EM
	HAPLOVIEW [138]
	PLINK [136]
Voies de signalisation « Pathways »	INTERSNP [143]
	MAGENTA [144]
Imputation	IMPUTE[4] [73]
	MACH [67]
Analyse simples marqueurs	PLINK
	SNPTEST [73]
	PROBABEL [145]
	SCAGEN[5]

Logiciels utilisés

Analyse multi marqueurs	INTERSNP
	PLINK
	WHAP [146]
	SCAGEN[5]
	BOOST [147]
Graphiques	LOCUSVIEW
	HAPLOVIEW

[1]**PLINK** est un logiciel libre et « open-source » permettant principalement de réaliser des tests d'association sur génome entier, développé par Shaun Purcell au Center for Human Genetic Research (CHGR), au Massachusetts General Hospital (MGH), et au Broad Institiute of Harvard and MIT. Le nombre de fonctions présentes dans PLINK est très impressionnant et complet. Ce fut vraiment le logiciel de référence pour le traitement et l'analyse des génomes entiers.

Il possède non exhaustivement des fonctions :

- De gestion des données (retirer/ajouter des SNPs, mise à jour, reformatage...)
- De statistiques sommaires de contrôle qualité (Hardy-Weinberg, données manquantes...)
- D'estimation d'IBD/IBS
- D'analyse d'association simples marqueurs
- De calcul du déséquilibre de liaison
- D'analyse d'association multi-marqueurs (haplotypes, épistasies)
- De méta analyse

[2]**SHAPE-IT** (http://www.griv.org/shapeit/) est un logiciel développé en C++ au CNAM par Olivier Delaneau sous la supervision de Jean-François Zagury (CNAM) permettant de phaser des données issues du génotypage. Il utilise de manière efficace le modèle de Markov caché. Les caractéristiques notables de ce logiciel sont :

- La complexité linéaire avec le nombre d'individus/SNP

Logiciels utilisés

- La possibilité d'haplotyper de très grandes régions en une seule fois sans recourir au découpage arbitraire des chromosomes.
- Le mélange des données de trios (parents-enfants) et sans relation est possible
- L'incertitude est captée dans des graphes d'incertitude réutilisables.
- L'haplotypage est multi-thread pour encore plus de rapidité d'exécution.

R est un tableur open-source permettant de réaliser la plupart des tests statistiques nécessaires à l'analyse des génomes. En plus des outils internes au logiciel très étoffés, de nombreux packages R sont développés spécifiquement pour l'analyse génétique comme Genabel [137] (qui présente des caractéristiques proches de plink) ou la représentation graphique des résultats comme LocusView. J'ai également utilisé des packages spécifiques aux tests multiples comme le FDR (figure 27).

Package	Type de FDR	Données d'entrée	Auteurs
fdrtool	FDR et local FDR	p-values, z-scores, t-scores	K. Strimmer
mixfdr	FDR et local FDR	Z-scores	O. Muralidharan, B. Efron
BUM / SPLOSH	FDR et local FDR	P-values	S. Pounds
SAGx	FDR et local FDR	P-values	P. Broberg
qvalue	FDR	P-values	A. Dabney and J. D. Storey
nFDR	FDR	P-values	M. Guedj and G. Nuel
multtest	FDR	P-values	K. S. Pollard, Y. Ge, S. Taylor, S. Dudoit
LBE	FDR	P-values	C. Dalmasso
FDR-AME	FDR	P-values	Y. Benjamini, E. Kenigsberg, D. Yekutieli
locfdr	Local FDR	Z-scores	B. Efron, B. B. Turnbull and B. Narasimhan
nomi	Local FDR	Z-scores	G. McLachlan, R. W. Bean
LocalFDR	Local FDR	P-values	J. Aubert
kerfdr	Local FDR	P-values	M. Guedj and G. Nuel
twilight	Local FDR	P-values	S. Scheid
localFDR	Local FDR	P-values	J.G. Liao

Figure 27 : liste des packages R permettant de calculer le False Discovery Rate (FDR)

[3]**Impute version 2** est un logiciel d'imputation et d'haplotypage de génotypes conçu par Bryan Howie. Il phase les données de génotypes (si elles ne sont pas déjà préalablement phasées) et impute les génotypes manquants présent dans les haplotypes de référence comme 1000genomes ou Hapmap. En sortie, il donne des probabilités pour chaque genotype (dosage) et un score d'imputation (info).

Logiciels utilisés

On peut ensuite analyser les associations simples marqueurs de ces données de dosage avec le programme [6]**Snptest.** Les tests implémentés permettent de prendre en compte entre autres:

- De multiples données de phénotype (cas-contrôles, simple ou plusieurs données quantitatives).
- Les covariables désirées
- Les modèles génétiques (additive, dominant, récessif, génotypique et hétérozygote)

[4]**SCAGEN** est un logiciel développé en 2005 en interne écrit en C/C++ permettant d'analyser les données issues des plate-formes de génotypage du CNG à Évry. Il permet:

- De convertir et filtrer les données (données manquantes, incohérences...) issues du séquençage en données de génotypes exploitables.
- D'analyser les marqueurs simples en calculant les odds ratio, et en testant l'égalité des distributions alléliques par les tests de χ^2 et de Fisher, de tester le respect de l'équilibre de Hardy-Weinberg.
- De faire des analyses de combinaisons de marqueurs.
- De faire de l'haplotypage via le logiciel PHASE et faire l'analyse sur la distribution des haplotypes.

Resultats SNP simple NP-PR (Seuil HW=0.70 Seuil Effectif=30 Seuil Fisher=0.900)

Populations	Mode	SNPs	OR	Fisher	Hardy-Weinberg	Effectif	Details
NP-PR	Repartition genotypique	CXCR6_-2264 (1|2)	0.770	0.805510	ok	ok	detail
NP-PR	Dominant	CXCR6_-2264-2	0.770	0.805510	ok	ok	detail
NP-PR	Recessif	CXCR6_-2264-1	1.299	0.805510	ok	ok	detail
NP-PR	Repartition genotypique	CXCR6_-2264 (1|1)	1.299	0.805510	ok	ok	detail
NP-PR	Frequence allelique	CXCR6_-2264-2	0.782	0.811529	ok	ok	detail
NP-PR	Frequence allelique	CXCR6_-2264-1	1.279	0.811529	ok	ok	detail

Resultats SNP simple NP-CTR (Seuil HW=0.70 Seuil Effectif=30 Seuil Fisher=0.900)

Populations	Mode	SNPs	OR	Fisher	Hardy-Weinberg	Effectif	Details
NP-CTR	Repartition genotypique	CXCR6_-2264 (1|2)	0.248	0.000148	ok	ok	detail
NP-CTR	Dominant	CXCR6_-2264-2	0.248	0.000148	ok	ok	detail
NP-CTR	Recessif	CXCR6_-2264-1	4.036	0.000148	ok	ok	detail
NP-CTR	Repartition genotypique	CXCR6_-2264 (1|1)	4.036	0.000148	ok	ok	detail
NP-CTR	Frequence allelique	CXCR6_-2264-2	0.260	0.000184	ok	ok	detail
NP-CTR	Frequence allelique	CXCR6_-2264-1	3.840	0.000184	ok	ok	detail

Résultats

Comme cela a été indiqué dans mes objectifs de thèse, j'ai d'abord travaillé sur l'exploitation des haplotypes dérivés des SNPs et pour cela, j'ai démontré qu'il était plus avantageux d'utiliser le maximum d'informations disponibles sur les SNPs pour calculer de larges haplotypes et d'extraire ensuite les sous-haplotypes pertinents. Cela peut sembler évident aujourd'hui, mais cela ne l'était pas en 2006 au moment de ce travail. L'article publié dont je suis premier auteur est donc présenté dans les résultats obtenus.

J'ai ensuite été le bioinformaticien responsable de l'analyse des données de puces de génotypage sur la cohorte GRIV. Dans ce cadre, j'ai notamment réalisé le travail d'analyse des facteurs génétiques impliqués dans le phénotype non-progresseur non « elite » observé dans la cohorte GRIV. L'article dont je suis premier co-auteur est aussi présenté dans les résultats obtenus pendant ma thèse.

Enfin, depuis un an, je travaille sur le projet international IHAC (International Consortium for HIV Acquisition). Ce projet réunit une quinzaine d'équipes internationales travaillant en génomique du SIDA qui ont accepté de regrouper leurs données pour avoir plus de puissance. Dans un premier temps, c'est le phénotype d'acquisition qui est étudié et je présenterai les résultats préliminaires de cette étude qui portent sur plus de 6000 sujets caucasiens infectés et plus de 7000 contrôles non infectés.

Ces trois axes de résultats sont importants pour moi car ils illustrent bien la progression spectaculaire du domaine en matière de traitement des données génomiques, ce qui a été mon activité de recherche principale depuis cinq ans.

I. Évaluations et améliorations des méthodes d'haplotypage (SUBHAP)

Résumé de la publication

Les analyses génétiques d'association ont pour but de trouver des corrélations entre une maladie et les variants génétiques comme les SNPs ou des combinaisons de SNPs appelés haplotypes. Certains haplotypes sont biologiquement très importants car localisés dans le promoteur ou dans l'exon d'un gène. Ils peuvent modifier l'expression ou la structure d'une protéine et donc être reliés directement à un phénotype de maladie.

Jusqu'à présent la reconstruction des haplotypes par les méthodes d'haplotypage classiques à partir de données de génotypes se faisait en prenant en compte l'unique information des SNPs composant l'haplotype. Nous proposons ici une approche dite « globale » permettant de prendre en compte toute l'information disponible dans une région pour reconstruire spécifiquement le « sous-haplotype » voulu.

Nous avons démontré la pertinence de notre approche à l'aide de données d'haplotypes expérimentales issues du promoteur des gènes GH1 et APOE ainsi que de 10 jeux de données simulées. En utilisant le logiciel PHASE, connu comme le plus précis des logiciels d'haplotypage en 2006, nous avons montré par l'introduction aléatoire des données manquantes dans les haplotypes que l'approche « globale » permettait de diminuer de manière substantielle le taux d'erreur dans la reconstruction haplotypique.

En appliquant cette méthode « globale » sur certains signaux précédemment trouvés dans des analyses simples gènes dans la cohortes GRIV, nous avons démontré qu'ils étaient en partie erronés.

En conclusion nous avons démontré que la méthode « globale » pouvait réduire significativement le taux d'erreur d'haplotypage. Cependant dans le cas où le taux de données manquantes est trop important (>10%) parmi les SNPs en dehors de l'haplotype d'intérêt, il peut encore être avantageux d'utiliser la méthode classique dite « locale » de résolution de l'haplotype.

BMC Genetics

BioMed Central

Research article

Open Access

Computation of haplotypes on SNPs subsets: advantage of the "global method"

Cédric Coulonges[†1,2], Olivier Delaneau[†1,2], Manon Girard[1,2], Hervé Do[1,2], Ronald Adkins[3], Jean-Louis Spadoni[2] and Jean-François Zagury*[1,2]

Address: [1]Equipe génomique, bioinformatique et pathologies du système immunitaire, INSERM U736, 15 rue de l'École de Médecine, 75006 Paris, France, [2]Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, 292 rue Saint-Martin, 75003 Paris, France and [3]Children's Foundation Research Center and Center of Genomics and Bioinformatics, University of Tennessee, Memphis, TN, USA

Email: Cédric Coulonges - coulonge@ccr.jussieu.fr; Olivier Delaneau - olivier.delaneau@gmail.com; Manon Girard - girard_manon@hotmail.com; Hervé Do - hervedo@gmail.com; Ronald Adkins - radkins1@utmem.edu; Jean-Louis Spadoni - jean-louis.spadoni@cnam.fr; Jean-François Zagury* - zagury@cnam.fr

* Corresponding author †Equal contributors

Published: 26 October 2006

BMC Genetics 2006, **7**:50 doi:10.1186/1471-2156-7-50

Received: 21 August 2006
Accepted: 26 October 2006

This article is available from: http://www.biomedcentral.com/1471-2156/7/50

Abstract

Background: Genetic association studies aim at finding correlations between a disease state and genetic variations such as SNPs or combinations of SNPs, termed haplotypes. Some haplotypes have a particular biological meaning such as the ones derived from SNPs located in the promoters, or the ones derived from non synonymous SNPs. All these haplotypes are "subhaplotypes" because they refer only to a part of the SNPs found in the gene. Until now, subhaplotypes were directly computed from the very SNPs chosen to constitute them, without taking into account the rest of the information corresponding to the other SNPs located in the gene. In the present work, we describe an alternative approach, called the "global method", which takes into account all the SNPs known in the region and compare the efficacy of the two "direct" and "global" methods.

Results: We used empirical haplotypes data sets from the *GH1* promoter and the *APOE* gene, and 10 simulated datasets, and randomly introduced in them missing information (from 0% up to 20%) to compare the 2 methods. For each method, we used the PHASE haplotyping software since it was described to be the best. We showed that the use of the "global method" for subhaplotyping leads always to a better error rate than the classical direct haplotyping. The advantage provided by this alternative method increases with the percentage of missing genotyping data (diminution of the average error rate from 25% to less than 10%). We applied the global method software on the GRIV cohort for AIDS genetic associations and some associations previously identified through direct subhaplotyping were found to be erroneous.

Conclusion: The global method for subhaplotyping can reduce, sometimes dramatically, the error rate on patient resolutions and haplotypes frequencies. One should thus use this method in order to minimise the risk of a false interpretation in genetic studies involving subhaplotypes. In practice the global method is always more efficient than the direct method, but a combination method taking into account the level of missing information in each subject appears to be even more interesting when the level of missing information becomes larger (>10%).

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Background

Large-scale genomic studies are becoming a standard nowadays. The exploitation of this huge body of data leads to multiple biological applications and in particular, to the unraveling of new molecular mechanisms for diseases through the identification of genetic associations. Genetic association studies are based on the comparison of genetic markers, the most frequent ones being Single Nucleotide Polymorphisms (or SNPs), between a diseased group versus a healthy group (case-control study). If a statistically significant difference is observed in the frequency of a SNP allele between a group of patients and a group of control subjects, it could mean that the gene or its product is involved in disease development. Association studies must also be performed on haplotypes which are the combination of SNPs in a given locus. Indeed, haplotypes and not only SNPs have already been reported to be associated with complex diseases such as AIDS [1-4], cancer [5-7], or Alzheimer's disease [8].

Experimental methods for haplotyping exist such as long-range haplotyping [9], single-copy DNA genotyping in conjunction with the Mass ARRAY system [10], or clone-based systematic haplotyping [11] but they are not applicable at a large scale level because of cost and time consumption. As an alternative, computational approaches have been developed to derive haplotypes from the SNP genotypic information (the couple of alleles found for each SNP) in a whole population. The most widely used algorithms to infer haplotypes from the unphased genotypic data rely today on statistical approaches such as the expectation-maximization (EM) algorithm or Bayesian coalescence-based algorithms [12,13].

Haplotypes have been the subject of an increasing number of studies in the recent years. Haplotypes information makes it possible to highlight the structure of the genome, notably through haploblocks which correspond to segments of chromosomes unlikely to undergo a crossing-over event [14,15]. In order to spare repeated efforts, an international consortium has undertaken the HapMap project with the aim of providing an exhaustive map including the most important SNPs determining the most frequent haplotypes in each haploblock of the human genome. The Hapmap project could accelerate the detection of SNP alleles or haplotypes associated with a disease phenotype [11].

The inference of haplotypes by computational methods can be very difficult and even sometimes incorrect. Indeed, the number of candidate haplotypes increases exponentially with the number of polymorphic sites, this number being 2^n in a subject with n heterozygous SNPs. Thus, it is not generally possible to solve correctly the equations (infer their haplotypes) for all subjects especially when there are missing data (SNPs whose alleles are unknown for some subjects in the population) which happens in most experiments.

Recent studies have compared the various computational methods to derive haplotypes [16-18]. Among them, the PHASE software [19] seemed to yield better results [13,16,20]. However, when haplotypes involving more than 7 SNPs were estimated from unphased genotypes, the reliability was poor even for PHASE, with an error rate jumping as high as 10%. It can thus become very useful to study haplotypes based on smaller set of SNPs in the population, which we will call here "subhaplotypes", because of the higher degree of experimental reliability (less missing data) and the higher degree of accuracy (for the haplotype computation).

It can also be important to investigate subhaplotypes with regard to their putative biological function: for instance subhaplotypes derived from SNPs in the gene promoter region [21,22], derived from SNPs leading to a protein mutation [21], or derived from tagSNPs [23-26]. Up to now, subhaplotypes derived from a set of selected SNPs in a gene have most often been inferred in a population by using only the genotypic information of these very SNPs in this population. However, an alternative approach could be to estimate the haplotypes from all the SNPs found in the gene and then, from these large haplotypes, extract the subhaplotypes corresponding to the set of the selected SNPs. In the case of missing information among the SNPs this approach might be useful because the missing information can be compensated through the linkage disequilibrium existing with other SNPs in the gene [27]. The first method, based on the direct haplotyping of SNPs of interest, will be called the "direct method". The second method, based on the use of larger haplotypes (haplotypes containing a larger number of SNPs) to infer subhaplotypes will be called the "global method".

The purpose of the present study is to evaluate which subhaplotyping procedure was optimal by comparing them on real and artificial genomic datasets. Such a comparative evaluation has not been performed before and it is particularly important for two reasons: 1. up to now most reports on disease genetic association studies use the "direct method" to estimate subhaplotypes [22,28-30]. 2. Many groups focus only on a limited set of representative SNPs such as tagSNPs to compute haplotypes [31,32] when they could use a larger set of SNPs to compute haplotypes more accurately.

Results

The goal of this study is to compare the two subhaplotyping strategies, "direct" and "global". For the comparison of the two strategies, we have first used two real haplotype

datasets previously determined experimentally: haplotypes determined experimentally on 150 Caucasian subjects in the *GH1* gene and corresponding to 14 SNPs with a MAF>1%, and haplotypes data determined experimentally on 80 subjects of various ethnical backgrounds in the *APOE* gene and corresponding to 9 SNPs with a MAF>1%. These experimentally determined haplotypes have been previously used as test samples by other researchers [16,33,34]. We have also used 10 simulated haplotype datasets artificially generated using a coalescent model on 30 SNPs and 100 individuals using the method of Schaffner et al. [35]. All these datasets are described in more details in Material and Methods. In order to look like real genomic data we have also introduced artificially missing information at various rates (2%, 5%, 10%, 15%, and 20%) in these datasets (see Material and Methods).

For the 2 methods, the computation of estimated haplotypes was done with the PHASE software, previously shown to be more reliable than the other haplotyping software [13,16,20,36]. The comparison of the 2 subhaplotyping methods, "direct" and "global", was performed with the following coefficients: the individual error rate for haplotype assignment (the 2 haplotypes assigned to an individual were correct or not), the similarity error rate [13] which measures the number of mutations required to obtain the real haplotypes for an individual, and the I_F coefficient which compares the estimated haplotype frequencies with the real ones [37]. All these coefficients are extensively described in Material and Methods.

Finally, we compared the impact of the use of the "global" and the "direct" methods in real genomic data obtained from an AIDS case-control study, the GRIV study, which compares extreme profiles of progression to AIDS with seronegative controls [38].

Comparison of the 2 methods in the GH1 haplotypes dataset

We tested various SNPs subsets of the *GH1* data set to do the comparison of the 2 subhaplotyping methods: we first randomly generated 100 subsets with no missing data for each size of 3, 5, and 7 SNPs. We then created randomly missing data in the genotypic dataset at various rates of 2%, 5%, 10%, 15%, and 20%, 20 times for each rate (a total of 100 genotypic datasets) and then, for each dataset, we generated randomly 20 subsets for each size of 3, 5, and 7 SNPs to compare the 2 methods after introducing missing data (see Material and Methods). Overall, for a given size of SNP subset (3, 5, or 7 out of the 14 SNPs) we tested 100 samples with no missing information, and 2000 samples with missing information.

We compared the global and direct methods on the measure "maximal resolution" (Rmax) corresponding to the haplotypes with the highest probability assigned by PHASE. Interestingly, we noticed in these tests that PHASE always managed to determine at least one possible resolution for each patient. Figure 1 shows graphs giving the mean individual error rate (IER) of both methods according to the rate of missing information. One may observe that the global method appeared to systematically yield a smaller mean error than the direct method (Figure 1). Also, it was not surprising to observe that the level of error of subhaplotype estimates produced by both methods increased with the number of SNPs involved for subhaplotyping and with the level of missing information: a range of 1 to 5% errors with no missing genotypic information to a range of 5 to 25% errors with 20% of missing genotypic information (Figure 1).

Table 1 further analyzes the difference between the 2 methods by presenting the similarity error rate (SimER) and the I_F coefficients (see Material and Methods): the global method clearly yields better results.

Comparison of the 2 methods in other haplotypes datasets

We analyzed in the same way another real haplotype dataset, previously published by Orzack et al. [33]. As shown in Table 2, the global method again yields better results. We also generated a population with artificial haplotypes as described in Schaffner et al. [35], and found similarly that the global method was more accurate (Table 2).

Interestingly, one can see that if the global method is always better, the values of the IER, SER, and I_F coefficients obtained by each method are different between the GH1, ApoE and artificial datasets for each level of missing information (see Table 1 and 2). The genetic structure of the population at stake appears thus to be very important.

Statistical significance

The results shown in Table 1 give the mean values of the error levels, however it does not give the number of times when the global method gets an error level lower than the direct method. We did this computation and found for the GH1 gene that the global method provided a more accurate result in 87% of the tests with no missing information, in 88% of the tests with 2% missing information, in 90% of the tests with 5% missing information, in 92% of the tests with 10% missing information, in 95% of the tests with 15% missing information, in 97% of the tests with 20% missing information. Similar results were obtained for APOE and the simulated SNP data (data not shown).

We also performed ANOVA tests to compare the IER obtained by both methods on each subset of a given size (GH1_3SNPs, GH1_5SNPs, GH1_7SNPs, APOE_4SNPs, SIM1 to SIM10_10 SNPs). The results (data not shown)

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

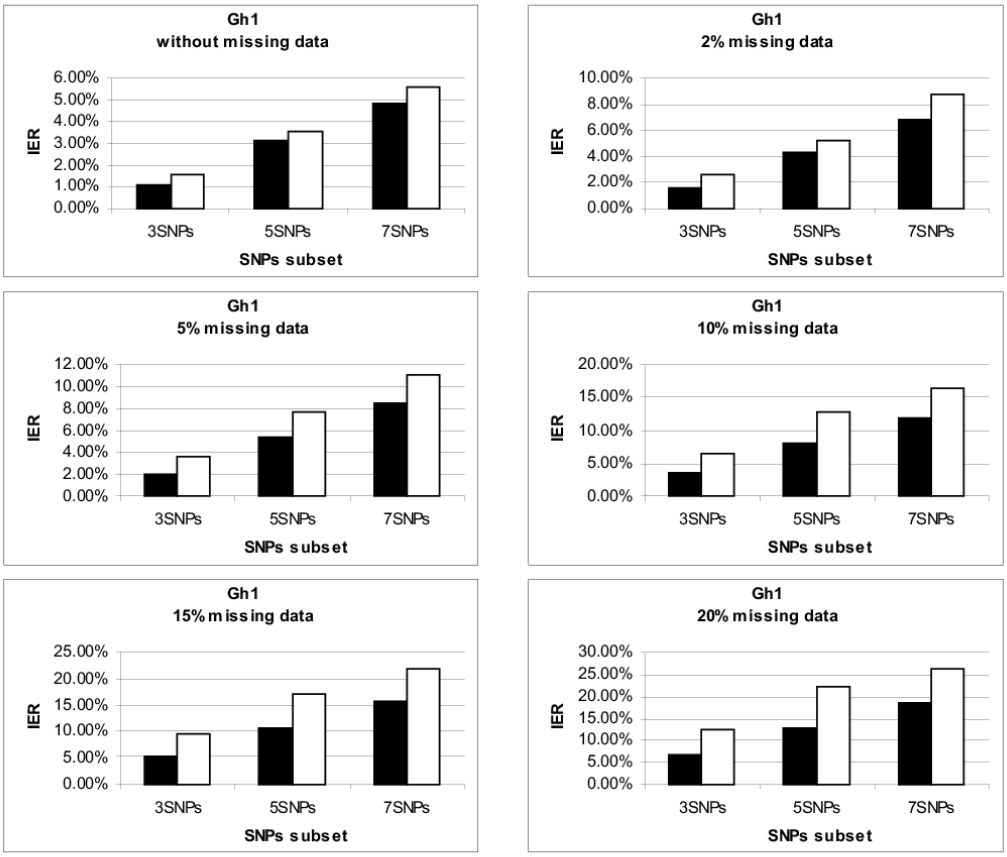


Figure 1
Graphical representation comparing the individual error rates (IER) between the direct and global methods. This figure presents the detailed graphs of the average error rates obtained by the 2 subhaplotyping methods, "direct" (in white) and "global" (in black), when they rely on the resolution with maximum probability (Rmax) produced by PHASE. Each graph corresponds to a different level of missing data introduced in the GH1 genotypic dataset (0%, 2%, 5%, 10%, 15% and 20%) and presents the mean of IER of all the replicates tested. There error rate obtained by the global method is always lower.

show again that the IER obtained by the global method are significantly better ($p < 10^{-4}$) than the IER found by the direct method for all subsets of SNPs.

Use of haplotypes defined through a Rmax cut-off

Since biologists often prefer to work with very clean data, we decided to select the most likely resolutions produced by PHASE. We thus selected those resolutions which exhibited an output probability higher than either 50% or 70% (see Material and Methods). Table 3 shows the results obtained by the direct and the global methods. For both the 50% and the 70% cut-offs, the global method yielded an error rate similar to the local method but it also yielded many more resolutions (Table 3). The global method with a 50% cut-off led to slightly more errors than the global method at 70% cut-off, but it also yielded many more resolutions (Table 3). Finally, when one compares the results obtained with cut-offs (Table 3) with the results obtained by Rmax (Table 1), it seems that the number of resolutions obtained by Rmax (it is always 100%) is higher than the number of resolutions obtained when using a cut-off, however the error rate is not as much

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Table 1: Error rates obtained according to the level of missing information in the GH1 dataset

		GH1			
MD	Method	IER	SimER	I_F	Res rate
0%	**Global_3snp**	**1.12%**	**0.37%**	**0.994**	**100%**
	Local_3snp	1.56%	0.52%	0.9904	100%
	Global_5snp	**3.16%**	**0.65%**	**0.9834**	**100%**
	Local_5snp	3.57%	0.72%	0.9784	100%
	Global_7snp	**4.87%**	**0.74%**	**0.9714**	**100%**
	Local_7snp	5.57%	0.83%	0.9704	100%
2%	**Global_3snp**	**1.63%**	**0.38%**	**0.9937**	**100%**
	Local_3snp	2.57%	0.68%	0.9898	100%
	Global_5snp	**4.34%**	**0.88%**	**0.9826**	**100%**
	Local_5snp	5.26%	1.20%	0.9792	100%
	Global_7snp	**6.83%**	**0.82%**	**0.9693**	**100%**
	Local_7snp	8.70%	1.03%	0.961	100%
5%	**Global_3snp**	**2.03%**	**0.46%**	**0.9934**	**100%**
	Local_3snp	3.65%	0.79%	0.9894	100%
	Global_5snp	**5.40%**	**0.88%**	**0.98**	**100%**
	Local_5snp	7.68%	1.20%	0.972	100%
	Global_7snp	**8.43%**	**1.07%**	**0.967**	**100%**
	Local_7snp	11.00%	1.37%	0.959	100%
10%	**Global_3snp**	**3.57%**	**0.73%**	**0.989**	**100%**
	Local_3snp	6.60%	1.33%	0.983	99.87%
	Global_5snp	**8.06%**	**1.19%**	**0.973**	**100%**
	Local_5snp	12.91%	1.86%	0.962	100%
	Global_7snp	**11.84%**	**1.35%**	**0.959**	**100%**
	Local_7snp	16.41%	1.88%	0.946	100%
15%	**Global_3snp**	**5.21%**	**1.01%**	**0.987**	**100%**
	Local_3snp	9.49%	1.84%	0.977	99.45%
	Global_5snp	**10.67%**	**1.47%**	**0.97**	**100%**
	Local_5snp	17.00%	2.33%	0.953	100%
	Global_7snp	**15.70%**	**1.67%**	**0.953**	**100%**
	Local_7snp	21.87%	2.35%	0.931	100%
20%	**Global_3snp**	**6.65%**	**1.26%**	**0.984**	**98.44%**
	Local_3snp	12.45%	2.41%	0.97	97.02%
	Global_5snp	**12.83%**	**1.72%**	**0.964**	**100%**
	Local_5snp	22.15%	3.02%	0.936	99.60%
	Global_7snp	**18.47%**	**1.92%**	**0.946**	**99.72%**
	Local_7snp	26.44%	2.81%	0.916	100%

Summary of the mean error rates obtained by each subhaplotyping method when they used the Rmax resolution produced by PHASE. The average individual error rate (IER) and similarity error rate (SimER) were computed according to the level of missing data (MD) introduced in the population (0%, 2%, 5%, 10%, 15%, 20%). Tests were performed on randomly selected SNP subsets of size 3, 5, and 7 taken out of the 14 SNPs present in the *GH1* genomic dataset (see text and Material and Methods). This table presents the average of the I_F coefficients which compares the accuracy of the subhaplotypes frequencies found by each subhaplotyping method. The global method fares always better.

different. In other words, the use of cut-offs leads to more accurate resolutions but a smaller percentage of patients gets subhaplotyped.

Combinations of the global method with the direct method according to the relative percentage of missing data

We reasoned that the localization of the missing information in each patient could influence the output on the global method versus that of the direct method. We thus tried a last approach to optimize the quality of the results: combining the global and direct methods when their results for the most probable resolution (Rmax) are different for a given patient (discordant subhaplotypes). For a patient, if the missing information rate was higher among the very SNPs selected for subhaplotyping, the subhaplotype provided by the global method was chosen; otherwise the

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Table 2: Error rates obtained according to the level of missing information in the APOE and simulated datasets

			APOE		
MD	Method	IER	SimER	I_F	Res rate
0%	**Global**	**1.94%**	**0.45%**	**0.986**	**100%**
	Local	4.88%	1.22%	0.97	100%
2%	**Global**	**2.24%**	**0.48%**	**0.986**	**100%**
	Local	5.12%	1.19%	0.972	100%
5%	**Global**	**3.20%**	**0.54%**	**0.987**	**100%**
	Local	5.64%	1.83%	0.978	100%
10%	**Global**	**4.41%**	**0.68%**	**0.979**	**100%**
	Local	6.89%	1.91%	0.972	100%
15%	**Global**	**6.98%**	**1.09%**	**0.974**	**100%**
	Local	10.33%	1.97%	0.964	99.75%
20%	**Global**	**12.35%**	**2.09%**	**0.954**	**100%**
	Local	15.21%	2.54%	0.943	99.21%
			Simulated		
MD	Method	IER	SimER	I_F	Res rate
0%	**Global**	**0.14%**	**0.02%**	**0.996**	**100%**
	Local	0.81%	0.10%	0.989	100%
2%	**Global**	**0.19%**	**0.02%**	**0.989**	**100%**
	Local	1.26%	0.13%	0.982	100%
5%	**Global**	**0.25%**	**0.03%**	**0.982**	**100%**
	Local	1.46%	0.18%	0.975	100%
10%	**Global**	**0.46%**	**0.05%**	**0.968**	**100%**
	Local	2.65%	0.32%	0.961	100%
15%	**Global**	**0.83%**	**0.09%**	**0.954**	**100%**
	Local	4.80%	0.59%	0.947	100%
20%	**Global**	**1.51%**	**0.16%**	**0.941**	**100%**
	Local	8.70%	1.06%	0.934	100%

Summary of the average error rates found when working with subhaplotypes based on 4 SNPs out of 9 in the *APOE* genomic dataset and 10 SNPs out of 30 in the 10 simulated datasets.
IER: individual error rate, Res Rate: resolution rate, SimER: similarity error rate, MD: Missing data, I_F: frequency error rate (see Material and Methods).

subhaplotype provided by the direct method was chosen (see Material and Methods).

As shown in Table 4, the combination method gave a rather good rate of error compared with the global method but there were slightly less patients' haplotypes resolved. Its use appeared most valuable when the number of missing information was higher than 15% (Table 4): the rate of error kept low (less than 7%), while the number of resolved patients remained high (around 90%). The application of this method on the *APOE* gene and on simulated data yielded similar results and conclusions (data not shown).

Application to the analysis of subhaplotypes in an AIDS cohort

GRIV (Genetics of Resistance to immunodeficiency Virus) is a case-control study comparing three groups, HIV-1 seropositive slow progressors (SP), HIV-1 seropositive rapid progressors (RP) and seronegative controls (CTR) [38]. We have previously published the exhaustive genotyping of SNPs from cytokines and cytokine receptors genes in that cohort [2,22]. In these works, we had computed the subhaplotypes derived from promoter SNPs by using the direct method and the comparison of the distribution of these subhaplotypes in the SP, RP, and CTR groups had led to the identification of a few genetic associations with AIDS progression. In the present study, we have recomputed these subhaplotypes with the use of the global method. We found that some positive signals (i.e. associations) found by the direct method have disappeared when using the global method (*IL4 Receptor* and *IL10 Receptor* [22]). On the contrary a test for association that seemed to be negative for the promoter of *IL6* became significant [2]. All these results are summarized in Table 5.

As the global method has very often a lower error rate, we conclude that the positive signals found in these studies

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Table 3: Error rates found by each method when using cut-offs for the probabilities provided by PHASE

		GH1 – cutoff 70% (5 SNPs)		
MD	**Method**	Abs IER	Rel IER	Res rate
0%	**Global**	**2.18%**	**2.24%**	**97.31%**
	Local	2.56%	2.65%	96.54%
2%	**Global**	**2.92%**	**2.99%**	**96.78%**
	Local	2.97%	3.05%	94.78%
5%	**Global**	**3.33%**	**3.49%**	**95.38%**
	Local	3.17%	3.51%	90.42%
10%	**Global**	**4.50%**	**4.92%**	**91.40%**
	Local	4.07%	5.00%	81.50%
15%	Global	5.83%	6.56%	88.88%
	Local	**4.85%**	**6.42%**	**75.54%**
20%	**Global**	**7.06%**	**8.14%**	**86.80%**
	Local	6.00%	8.81%	68.12%
		GH1 – cutoff 50% (5 SNPs)		
MD	**Method**	Abs IER	Rel IER	Res rate
0%	**Global**	**2.99%**	**2.99%**	**99.82%**
	Local	3.50%	3.50%	99.83%
2%	**Global**	**3.82%**	**3.98%**	**99.81%**
	Local	4.18%	4.26%	97.47%
5%	**Global**	**5.03%**	**5.06%**	**99.42%**
	Local	5.08%	5.32%	95.45%
10%	**Global**	**7.17%**	**7.28%**	**98.45%**
	Local	7.24%	8.04%	90.06%
15%	**Global**	**9.41%**	**9.62%**	**97.84%**
	Local	9.50%	10.95%	86.83%
20%	**Global**	**11.25%**	**11.58%**	**97.18%**
	Local	11.96%	14.45%	82.77%

This table presents the summary of average error rates when using a cut-off on resolution probability given by PHASE instead of Rmax when choosing the resolution obtained by each method. The 2 cut-offs chosen were respectively 50% and 70%. The error rate is almost always lower than with Rmax, but the number of assigned haplotypes strongly decreased.
IER: individual error rate, Abs IER: absolute IER, Rel IER: relative IER, Res Rate: resolution rate, MD: Missing Data.

were likely to be artifacts of the direct subhaplotyping while previously negative tests may have missed real associations in AIDS progression.

Discussion and conclusion

In this study, we have confirmed that the error rate found in the resolutions determined by the best haplotyping software known to date, PHASE, could be non negligible even when there were no missing information in the genomic data [13,16,20,36]: it ranged from 1% to 6% according to the selection of SNPs (see Table 1 and 2). Errors were also observed at the level of the haplotypes frequencies (Table 1 and 2). In reality, when dealing with genotypic information obtained experimentally, there is often missing information and our study shows that in that case, the error rate for the estimation of haplotypes can jump even higher, reaching 25% in some instances (Table 1). This has led us to develop an alternative method to estimate haplotypes, the "global method". The rationale of the global method is to use the information contained in other SNPs, which are not used in the direct haplotyping, in order to limit the impact of missing data: for instance, the presence of linkage disequilibrium between SNPs might supplement missing data on certain SNPs.

We performed tests on genomic datasets for which haplotypes had been determined exactly through biological experimentation and also on simulated data. We generated randomly missing genotypic information in these datasets and computed partial haplotypes (subhaplotypes) from subsets of selected SNPs. We found that the global approach, which first computes the haplotypes from all the available SNPs and then extracts the subhap-

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Table 4: Error rates of the combination method

		Combination			
MD	Method	Abs IER	Rel IER	SimER	Res rate
2%	combi_3snp	1.83%	1.84%	0.17%	100.00%
	combi_5snp	2.56%	2.57%	0.36%	99.58%
	combi_7snp	4.89%	4.89%	0.54%	99.23%
5%	combi_3snp	2.72%	2.72%	0.29%	100.00%
	combi_5snp	3.64%	3.69%	0.57%	98.39%
	combi_7snp	6.09%	6.27%	0.60%	97.15%
10%	combi_3snp	3.16%	3.22%	0.45%	98.06%
	combi_5snp	5.60%	5.81%	0.68%	96.40%
	combi_7snp	8.12%	8.54%	0.67%	95.12%
15%	combi_3snp	5.03%	5.09%	0.61%	98.81%
	combi_5snp	5.98%	6.31%	0.82%	94.70%
	combi_7snp	6.62%	7.36%	0.77%	89.95%
20%	combi_3snp	4.09%	4.56%	0.83%	89.78%
	combi_5snp	8.16%	8.70%	0.96%	93.80%
	combi_7snp	6.43%	7.34%	0.90%	87.54%

This table presents the summary of average error rates when using the combination method (see Material and Methods). In that case, we present the example of the results obtained for subsets of 5 SNPs.
IER: individual error rate, Abs IER: absolute IER, Rel IER: relative IER, SimER: similarity error rate, Res Rate: resolution rate, MD: Missing Data.

lotypes corresponding to the selected SNPs, reproducibly led to better estimations with significantly lower error rates (Tables 1 and 2).

Since biologists like to work with exact data, we also tried to work on the resolutions exhibiting a significant reliability as determined by PHASE: resolutions exhibiting a probability higher than 70% or higher than 50%. With this approach, the global method still yielded a lower error rate than the direct method (Table 3). It appears that when one increases the cut-off to assign a resolution the final error rate slightly diminishes while the number of patients being assigned a subhaplotype diminishes rather importantly (Table 3).

We finally tried to combine the global and direct methods for discordant patients (patients for which the haplotypes computed by the direct and global method were different). For that, we used the subhaplotype computed by the direct method if there was less missing information in the SNPs selected for subhaplotyping than in the remaining SNPs, or the global method in the opposite case. We found that this combination method could be a useful compromise when the level of missing information in the population was high: the relative individual error rate was smaller than that of the global method based on Rmax but some patients were not assigned an haplotype (Table 4).

The fact that the global method yields better results than the local method is not a surprise knowing the importance of linkage disequilibrium inside genetic loci. Indeed, Marchini et al. found similarly that for the computation of the r^2 coefficients it was more reliable to use large number of SNPs instead of pairwise comparisons [20].

In practice, if there is not too much missing information (less than 10%), the global method using the PHASE Rmax resolution works well with nearly all subjects being assigned a subhaplotype and with an error rate below 10% (Table 1 and 2). If there is more missing information (more than 10%), it might be interesting to use the combination method knowing that 90% of the subjects are assigned a subhaplotype among which less than 8% have a wrong haplotype (Table 4).

We have demonstrated the practical interest of this new subhaplotyping method in our GRIV genomic dataset: we had previously genotyped the cytokine and cytokine receptors in the GRIV cohort and we had estimated subhaplotypes of the promoter regions by direct subhaplotyping [2,22]. In the present work, we have recomputed the subhaplotypes of the promoter regions using the more precise SUBHAP software: we found that associations previously described for *IL4R*, *IL10R* subhaplotypes did not

BMC Genetics 2006, **7**:50 http://www.biomedcentral.com/1471-2156/7/50

Table 5: Modification of the results obtained in the GRIV case-control study when using the various subhaplotyping methods

Genes	Sub-haplotype	p-value direct Rmax subhap	p-value global Rmax subhap	p-value Combination Rmax subhap
IL10Receptor	Exon	0.026 A.H cases: 100% A.H controls: 100%	*0.103 A.H cases: 100% A.H controls: 100%	0.093 A.H cases: 88% A.H controls: 99%
IL4Receptor	Promoter	0.019 A.H cases: 100% A.H controls: 100%	*0.072 A.H cases: 100% A.H controls: 100%	*0.088 A.H cases: 100% A.H controls: 98%
IL6	Promoter	0.059 A.H cases: 100% A.H controls: 100%	0.012 A.H cases: 100% A.H controls: 100%	*0.009 A.H cases: 82% A.H controls: 90%

*Best method regarding the missing data level.
This table presents the p-values found for the Fisher's exact tests comparing the subhaplotypes distributions between seropositive patients of the GRIV cohort (cases) and seronegative subjects (controls). The subhaplotypes were computed either with the direct Rmax method as previously published [2, 43], or with the global Rmax method, or with the combination Rmax method described in our study. The percentage of missing information was respectively 6.7%, 11.1% and 14.8% for the *IL10R*, *IL4R*, and *IL6* genotypic data. One can see that some signals which were previously published as positive ($p < 0.05$) using the direct method become negative, while some signals which were previously published as negative become much stronger thanks to the novel subhaplotyping methods. For *IL10R* we have a deficit of information for cases and as consequences a lower percentage of assigned haplotypes in the combination method which is more restrictive.
A.H cases: percentage of assigned haplotypes attributed in the tests for cases
A.H control: percentage of assigned haplotypes attributed in the tests for controls.

hold, while signals appeared much stronger for an *IL6* subhaplotype (Table 5).

This work has extensively evaluated the impact of missing data on subhaplotyping and it emphasizes that the level of missing information in the genomic data is a critical issue: the practical impact is not negligible since in our experimental genotyping of the GRIV cohort, the rate of missing data may reach 20% for some SNPs. Such rates have also been widely described in the literature [39-41]. This work also underlines that the genetic structure of the SNPs in the population is an important issue since the error rates may vary from one population to the other (see Table 1 and 2) and it could certainly be interesting to take into account other parameters such as the LD and minor allele frequencies to help optimize the subhaplotyping procedure.

Current genomic studies, such as the Hapmap project, aim at minimizing the number of SNPs necessary to perform genetic associations in complex diseases by using tagSNPs. These studies do not consider the missing information problem inherent to any genotyping experiment which will often prevent the optimal haplotyping of the patients for disease genetic association studies. Our results suggest that if the Hapmap data are evidently very useful in targeting genetic regions of interest, an extensive genotyping with all SNPs in a sensitive region will however likely be needed to infer correct subhaplotypes.

In conclusion, the subhaplotyping method that we described here will allow to improve genetic association studies with complex diseases and take the best advantage of the available genotype data. The global and combination methods are available with Subhap software [42].

Methods

The GH1 haplotypes data set

This haplotypic data set was determined empirically by Haran et al. [34] from 154 patients who were recruited of the British army. The promoter of the growth hormone (*GH1*) gene spans 535 bps, and is very strongly polymorphic with 14 SNPs whose minor allele frequency (MAF) is greater than 1% in the studied population. By cloning and genotyping 154 patients [34], the authors managed to experimentally define 38 different haplotypes based on these 14 SNPs, including 18 haplotypes with a global frequency higher than to 1%. We excluded the only patient implicating a tri-allelic SNP to simplify the calculation: we thus only used 153 patients of this cohort.

The *GH1* gene SNPs presents only one perfect LD and does not include any haploblock (Fig 1) which limits the skewing of the results and makes this genomic dataset more reliable for the comparison of the *direct* and *global* methods.

The APOE haplotypes data set

This haplotypic dataset was determined experimentally by Orzack et al. [33] using a long-range allele-specific PCR on 80 unrelated individuals from 3 ethnic groups: 18 Asian, 19 African and 43 Caucasian individuals. The *APOE* locus is composed of 9 SNPs with MAF>1%. 17 haplotypes were identified experimentally. The level of LD between the *APOE* SNPs was also very low, as for *GH1* polymorphisms. The *GH1* and *APOE* data sets are very useful for

our goal. Indeed, if we show that the global method is more efficient on them, that advantage will be even stronger for common datasets because they generally exhibit more LD.

The simulated haplotypes data set

This haplotypic data was created with COSI package developed by Schaffner et al. [35] based on a coalescent model. We have generated 10 data sets of 30 SNPs on 100 unrelated individuals simulated with constant recombination rate across the region, constant population size, and random mating.

The GRIV data sets

The GRIV cohort is composed of 400 Caucasian HIV-1 positive patients with extreme profiles of progression to AIDS (Slow Progression or Rapid Progression) and has been extensively genotyped by PCR-sequencing on various genes of the immune system [38]. In addition, 400 healthy subjects of similar ethnic origin were also genotyped as controls (CTR). In the present study, we have used the genotypes obtained on genes analyzed in the cohort and previously reported: cytokines and their receptors [2,43]. Unlike the *GH1* and *APOE* data sets previously described, we do not dispose of the real haplotypes for this population.

Creation of missing data

The data set from the *GH1* study was a complete data set. In order to study the influence of missing data on the accuracy of the results, missing data were artificially generated inside the *GH1* data set. To be more realistic, missing data was distributed randomly across genomic datasets. We applied similar levels of missing data to the *GH1*, *APOE* and *simulated* datasets: 2%, 5%, 10%, 15% and 20%.

Haplotyping software

We have chosen to use the PHASE software [13,19] to infer haplotypes. Indeed, many studies some of which performed on the *GH1* datasets have compared the different haplotyping algorithms and came to the conclusion that the PHASE algorithm performed better [16, 17, 44, 45] with a lower error rate and a higher number of solved patients. PHASE is based on a Markov chain of Monte Carlo with a recombination model based on the decay of LD with distance. The PHASE parameters were optimized using the empirical haplotypes of the *GH1* promoter: the thinning interval (steps through the Markov chain per iteration) and the number of runs (of the algorithm) didn't seem to alter significantly the results. The number of iterations on these data which apparently yielded the lowest error rate and the best number of inferred haplotypes was 100 iterations and 1000 burn-in (100, 500 and 10000 iterations were tested). The other parameters were set by default.

Subjects with more than 50% missing information were removed in order to avoid estimating haplotypes when there was too much data lacking.

Subhaplotyping methods tested

The direct method

A subset containing only the genotypes of the selected SNPs was extracted from the whole data set for all the individuals. The haplotypes for these SNPs were then inferred by haplotyping this data set with the PHASE software.

The global method

The haplotypes were first inferred with the PHASE algorithm from the whole data set containing all the SNPs genotypes in each gene. This initial haplotyping provides for each patient the diplotype derived from all the SNPs and encompasses automatically the SNPs selected for subhaplotyping. The subhaplotypes corresponding to the selected SNPs could then be extracted directly from this global data set, forming the subhaplotype data set.

The combination method

When the two methods disagree on the resolution for one patient, a resolution was chosen after assessing which method was the most reliable.

In the case of the combination method based on the Rmax resolution, the choice of the resolution depended on the rate of missing information in the SNPs used to estimate the subhaplotype. If the missing information was higher than 30% both in SNPs composing the subhaplotype and in the remaining ones used for the global method, we considered that the patient's haplotypes could not be solved.

Comparison of the resolutions found by each method with the real subhaplotypes

The results obtained when using each subhaplotyping method were compared to the real subhaplotypes as determined experimentally. This comparison was done by using various coefficients measuring the error rate that are described in the paragraphs hereafter. We have used the ANOVA model to test if there was any statistical difference between the error rates obtained from the two methods.

IER and SimER: error rates for haplotype assignments

The resolution rate (Res Rate) is the proportion of individuals for which a diplotype was found by the subhaplotyping method. Res Rate thus ranges from 0 to a maximum of 1 when all patients are assigned an haplotype.

The individual error rate (IER) is the proportion of subjects whose inferred diplotype is not correct. In case the Res Rate was <1, we called relative IER (Rel IER) the proportion of subjects whose inferred diplotype is not correct among all the subjects who were assigned a diplotype. In case the Res Rate was <1, we called absolute IER (Abs IER) the proportion of subjects whose inferred diplotype was not correct among the whole population.

The similarity error rate (SimER) is another measure of similarity between the estimated haplotypes and real haplotypes, which was developed by Stephens et al. [13]: it is based on the percentage of errors found at the level of SNPs for each haplotype.

I_F: error rate for the frequencies of the attributed haplotypes

I_F [37] measure how closely the inferred and empirical haplotype frequencies correspond and is given by:

$$I_F = 1 - \frac{1}{2}\sum_{k=1}^{h} \left| p_{ek} - p_{tk} \right|$$

where p_{ek} and p_{tk} are the inferred and empirically determined frequencies for the k*th* haplotype, and h is the number of haplotypes. I_F range from 0 to a maximum value of 1 when the frequencies match perfectly

Authors' contributions

CC and OD have worked on developing the methods and programs used in this study under the direct supervision of JFZ who conceived the study. RA, MG and JLS worked on evaluating the accuracy of PHASE in the tested datasets. HD provided the real GRIV genotypic data. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr Horan for giving us access to empirical haplotypes data from the *GH1* promoter and Dr Orzack for giving us access to the empirical haplotypes data from the *APOE* gene and. We thank all the patients and medical staff who kindly collaborated with the GRIV study. This project was funded in part by Sidaction and by ACV development foundation. O. Delaneau and Herve Do have a fellowship from the French Ministry of Research and Education (MNERT).

References

1. Hendel H, Caillat-Zucman S, Lebuanec H, Carrington M, O'Brien S, Andrieu JM, Schachter F, Zagury D, Rappaport J, Winkler C, Nelson GW, Zagury JF: **New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS.** *J Immunol* 1999, **162(11):**6942-6946.
2. Vasilescu A, Heath SC, Ivanova R, Hendel H, Do H, Mazoyer A, Khadivpour E, Goutalier FX, Khalili K, Rappaport J, Lathrop GM, Matsuda F, Zagury JF: **Genomic analysis of Th1-Th2 cytokine genes in an AIDS cohort: identification of IL4 and IL10 haplotypes associated with the disease progression.** *Genes Immun* 2003, **4(6):**441-449.
3. Winkler CA, Hendel H, Carrington M, Smith MW, Nelson GW, O'Brien S J, Phair J, Vlahov D, Jacobson LP, Rappaport J, Vasilescu A, Bertin-Maghit S, An P, Lu W, Andrieu JM, Schachter F, Therwath A, Zagury JF: **Dominant effects of CCR2-CCR5 haplotypes in HIV-1 disease progression.** *J Acquir Immune Defic Syndr* 2004, **37(4):**1534-1538.
4. Flores-Villanueva PO, Hendel H, Caillat-Zucman S, Rappaport J, Burgos-Tiburcio A, Bertin-Maghit S, Ruiz-Morales JA, Teran ME, Rodriguez-Tafur J, Zagury JF: **Associations of MHC ancestral haplotypes with resistance/susceptibility to AIDS disease development.** *J Immunol* 2003, **170(4):**1925-1929.
5. Wagner K, Hemminki K, Israelsson E, Grzybowska E, Klaes R, Chen B, Butkiewicz D, Pamula J, Pekala W, Forsti A: **Association of polymorphisms and haplotypes in the human growth hormone 1 (GH1) gene with breast cancer.** *Endocr Relat Cancer* 2005, **12(4):**917-928.
6. Bonilla C, Mason T, Long L, Ahaghotu C, Chen W, Zhao A, Coulibaly A, Bennett F, Aiken W, Tullock T, Coard K, Freeman V, Kittles RA: **E-cadherin polymorphisms and haplotypes influence risk for prostate cancer.** *Prostate* 2006, **66(5):**546-556.
7. Sweeney C, Curtin K, Murtaugh MA, Caan BJ, Potter JD, Slattery ML: **Haplotype analysis of common vitamin d receptor variants and colon and rectal cancers.** *Cancer Epidemiol Biomarkers Prev* 2006, **15(4):**744-749.
8. Saleheen D: **Haplotype analysis in VEGF gene and increased risk of Alzheimer's disease.** *Ann Neurol* 2005, **58(3):**488; author reply 488-9.
9. Zhang K, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM: **Long-range polony haplotyping of individual human chromosome molecules.** *Nat Genet* 2006, **38(3):**382-387.
10. Ding C, Cantor CR: **Direct molecular haplotyping of long-range genomic DNA with M1-PCR.** *Proc Natl Acad Sci U S A* 2003, **100(13):**7449-7453.
11. Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S: **Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes.** *Genome Res* 2003, **13(12):**2717-2724.
12. Niu T, Qin ZS, Xu X, Liu JS: **Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **70(1):**157-169.
13. Stephens M, Donnelly P: **A comparison of bayesian methods for haplotype reconstruction from population genotype data.** *Am J Hum Genet* 2003, **73(5):**1162-1169.
14. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296(5576):**2225-2229.
15. Stumpf MP: **Haplotype diversity and the block structure of linkage disequilibrium.** *Trends Genet* 2002, **18(5):**226-228.
16. Adkins RM: **Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset.** *BMC Genet* 2004, **5:**22.
17. Niu T: **Algorithms for inferring haplotypes.** *Genet Epidemiol* 2004, **27(4):**334-347.
18. Sabbagh A, Darlu P: **Inferring haplotypes at the NAT2 locus: the computational approach.** *BMC Genet* 2005, **6(1):**30.
19. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68(4):**978-989.
20. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P: **A comparison of phasing algorithms for trios and unrelated individuals.** *Am J Hum Genet* 2006, **78(3):**437-450.
21. Diop G, Spadoni JL, Do H, Hirtzig T, Coulonges C, Labib T, Issing W, Rappaport J, Therwath A, Lathrop M, Matsuda F, Zagury JF: **Genomic approach of AIDS pathogenesis: exhaustive genotyping of the TNFR1 gene in a French AIDS cohort.** *Biomed Pharmacother* 2005, **59(8):**474-480.
22. Do H, Vasilescu A, Diop G, Hirtzig T, Coulonges C, Labib T, Heath SC, Spadoni JL, Therwath A, Lathrop M, Matsuda F, Zagury JF: **Associations of the IL2Ralpha, IL4Ralpha, IL10Ralpha, and IFN (gamma) R1 cytokine receptor genes with AIDS progression in a French AIDS cohort.** *Immunogenetics* 2006, **58(2-3):**89-98.
23. Burkett KM, Ghadessi M, McNeney B, Graham J, Daley D: **A comparison of five methods for selecting tagging single-nucleotide polymorphisms.** *BMC Genet* 2005, **6 Suppl 1:**S71.
24. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide**

polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004, **74(1):**106-120.
25. Carlson CS, Heagerty PJ, Hatsukami TS, Richter RJ, Ranchalis J, Lewis J, Bacus TJ, McKinstry LA, Schellenberg GD, Rieder M, Nickerson D, Furlong CE, Chait A, Jarvik GP: **TagSNP analyses of the PON gene cluster: effects on PON1 activity, LDL oxidative susceptibility, and vascular disease.** *J Lipid Res* 2006, **47(5):**1014-1024.
26. Howie BN, Carlson CS, Rieder MJ, Nickerson DA: **Efficient selection of tagging single-nucleotide polymorphisms in multiple populations.** *Hum Genet* 2006, **120(1):**58-68.
27. Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N, Crocker N, Boyd PR, Zaykin DV, Purvis IJ: **Effectiveness of computational methods in haplotype prediction.** *Hum Genet* 2002, **110(2):**148-156.
28. Cheong HS, Shin HD, Lee SO, Park BL, Choi YH, Lim GI, Uh ST, Kim YH, Lee JY, Lee JK, Kim HT, Ryu HJ, Kim KK, Han BG, Kim JW, Kimm K, Oh B, Park CS: **Polymorphisms in interleukin 8 and its receptors (IL8, IL8RA and IL8RB) and association of common IL8 receptor variants with peripheral blood eosinophil counts.** *J Hum Genet* 2006.
29. Nunez C, Alecsandru D, Varade J, Polanco I, Maluenda C, Fernandez-Arquero M, de la Concha EG, Urcelay E, Martinez A: **Interleukin-10 haplotypes in Celiac Disease in the Spanish population.** *BMC Med Genet* 2006, **7:**32.
30. Tregouet DA, Barbaux S, Escolano S, Tahri N, Golmard JL, Tiret L, Cambien F: **Specific haplotypes of the P-selectin gene are associated with myocardial infarction.** *Hum Mol Genet* 2002, **11(17):**2015-2023.
31. Cousin E, Deleuze JF, Genin E: **Selection of SNP subsets for association studies in candidate genes: comparison of the power of different strategies to detect single disease susceptibility locus effects.** *BMC Genet* 2006, **7:**20.
32. Stram DO: **Tag SNP selection for association studies.** *Genet Epidemiol* 2004, **27(4):**365-374.
33. Orzack SH, Gusfield D, Olson J, Nesbitt S, Subrahmanyan L, Stanton VPJ: **Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes.** *Genetics* 2003, **165(2):**915-928.
34. Horan M, Millar DS, Hedderich J, Lewis G, Newsway V, Mo N, Fryklund L, Procter AM, Krawczak M, Cooper DN: **Human growth hormone 1 (GH1) gene expression: complex haplotype-dependent influence of polymorphic variation in the proximal promoter and locus control region.** *Hum Mutat* 2003, **21(4):**408-423.
35. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: **Calibrating a coalescent simulation of human genome sequence variation.** *Genome Res* 2005, **15(11):**1576-1583.
36. Xu H, Wu X, Spitz MR, Shete S: **Comparison of haplotype inference methods using genotypic data from unrelated individuals.** *Hum Hered* 2004, **58(2):**63-68.
37. Excoffier L, Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12(5):**921-927.
38. Hendel H, Cho YY, Gauthier N, Rappaport J, Schachter F, Zagury JF: **Contribution of cohort studies in understanding HIV pathogenesis: introduction of the GRIV cohort and preliminary results.** *Biomed Pharmacother* 1996, **50(10):**480-487.
39. Li J, Jiang T: **Computing the minimum recombinant haplotype configuration from incomplete genotype data on a pedigree by integer linear programming.** *J Comput Biol* 2005, **12(6):**719-739.
40. Su SC, Kuo CC, Chen T: **Inference of missing SNPs and information quantity measurements for haplotype blocks.** *Bioinformatics* 2005, **21(9):**2001-2007.
41. Padilla MA, Algina J: **Type I Error Rates For A One Factor Within-Subjects Design With Missing Values.** *J Mod Appl Stat Methods* 2004, **3(2):**406-416.
42. Coulonges C, Delaneau O: **Subhap [http://www.griv.org/software/].** 2006.
43. Inbar E, Yakir B, Darvasi A: **An efficient haplotyping method with DNA pools.** *Nucleic Acids Res* 2002, **30(15):**e76.
44. Salem RM, Wessel J, Schork NJ: **A comprehensive literature review of haplotyping software and methods for use with unrelated individuals.** *Hum Genomics* 2005, **2(1):**39-66.

II. Analyse de la cohorte GRIV

A partir de 2007, le génotypage des patients de la cohorte GRIV par les puces de génotypage Illumina 317K a permis de réaliser plusieurs études « génome entier » pour lesquelles j'ai réalisé l'analyse bioinformatique des données génétiques et qui ont abouti à plusieurs publications de l'équipe [148-151]. Ces publications, ainsi que celles produites par les groupes internationaux compétiteurs [121, 152], ont toutes pointées vers une seule région statistiquement significative du génome : la région HLA du chromosome 6.

Dans le cadre de la présentation des résultats de ma thèse, j'ai choisi de présenter la publication sur le phénotype non-progresseur non « elite », non pas parce que j'en suis premier co-auteur, mais surtout parce que c'est le premier signal hors région HLA qui a été identifié par analyse « genome entier ». Ce signal ne tombe pas dans une région totalement inconnue des experts du VIH-1 (région des co-récepteurs du VIH-1) mais il n'en reste pas moins intéressant.

Résumé de la publication

Découverte d'un récepteur de chimiokine *CXCR6* impliqué dans la non-progression à long terme

Le résultat majeur issu des précédentes études d'association « génome entier » (GWAS) est le polymorphisme rs2395029 *HCP5/HLA-B*57*, impliqué dans le contrôle de la charge virale et dans la non progression à long terme (LTNP). Or, seule une minorité des LTNP de la cohorte GRIV porte l'allèle protecteur rs2395029-G et contrôle la charge virale à des niveaux très faibles. Afin d'identifier des facteurs génétiques ayant un impact sur le phénotype LTNP sans nécessairement contrôler la charge virale, nous avons ré-analysé les données génomiques de LTNP de la cohorte GRIV en excluant au préalable les individus « elites controllers » (contrôlant leur charge virale plasmatique à un niveau <100 copies/mL).

Ainsi, la comparaison des 186 LTNP non « elites » avec 697 individus contrôles séronégatifs a mis en avant le SNP rs2234358 du gène *CXCR6* ($p=2,5x10^{-7}$, OR=0,85). Nous avons démontré l'indépendance de ce signal du chromosome 3 vis-à-vis des polymorphismes du gène voisin CCR5, bien connus pour leur rôle dans l'infection VIH-1. Ce résultat a pu être répliqué dans 3 cohortes indépendantes, et la p-value combinée entre ces cohortes atteint le seuil de significativité statistique « génome entier » ($p_{combinée}=9,7x10^{-10}$). Enfin, l'extraction et la compilation de l'ensemble

des LTNP non « elites » des quatre cohortes, et leur comparaison avec les contrôles séronégatifs a permis de souligner la spécificité de ce signal pour la LTNP ($p=2,5x10^{-8}$).

A ce jour, le SNP rs2234358 est le seul signal identifié et confirmé par étude « génome entier » en dehors de la région HLA et passant le seuil de significativité statistique « génome entier ». *CXCR6* étant un co-récepteur mineur dans le cadre de l'infection VIH-1 -contrairement à l'infection SIV-, son rôle de médiateur dans l'inflammation pourrait être impliqué dans la pathogenèse du SIDA.

MAJOR ARTICLE

Multiple-Cohort Genetic Association Study Reveals CXCR6 as a New Chemokine Receptor Involved in Long-Term Nonprogression to AIDS

Sophie Limou[a], Cédric Coulonges[a], Joshua T. Herbeck, Daniëlle van Manen, Ping An, Sigrid Le Clerc, Olivier Delaneau, Gora Diop, Lieng Taing, Matthieu Montes, Angélique B. van't Wout, Geoffrey S. Gottlieb, Amu Therwath, Christine Rouzioux, Jean-François Delfraissy, Jean-Daniel Lelièvre, Yves Lévy, Serge Hercberg, Christian Dina, John Phair, Sharyne Donfield, James J. Goedert, Susan Buchbinder, Jérôme Estaquier, François Schächter, Ivo Gut, Philippe Froguel, James I. Mullins,[a] Hanneke Schuitemaker,[a] Cheryl Winkler,[a] and Jean-François Zagury[a]

Background. The compilation of previous genomewide association studies of AIDS shows a major polymorphism in the *HCP5* gene associated with both control of the viral load and long-term nonprogression (LTNP) to AIDS.

Methods. To look for genetic variants that affect LTNP without necessary control of the viral load, we reanalyzed the genomewide data of the unique LTNP Genomics of Resistance to Immunodeficiency Virus (GRIV) cohort by excluding "elite controller" patients, who were controlling the viral load at very low levels (<100 copies/mL).

Results. The rs2234358 polymorphism in the *CXCR6* gene was the strongest signal ($P = 2.5 \times 10^{-7}$; odds ratio, 1.85) obtained for the genomewide association study comparing the 186 GRIV LTNPs who were not elite controllers with 697 uninfected control subjects. This association was replicated in 3 additional independent European studies, reaching genomewide significance of $P_{\text{combined}} = 9.7 \times 10^{-10}$. This association with LTNP is independent of the *CCR2–CCR5* locus and the *HCP5* polymorphisms.

Conclusions. The statistical significance, the replication, and the magnitude of the association demonstrate that CXCR6 is likely involved in the molecular etiology of AIDS and, in particular, in LTNP, emphasizing the power of extreme-phenotype cohorts. CXCR6 is a chemokine receptor that is known as a minor coreceptor in human immunodeficiency virus type 1 infection but could participate in disease progression through its role as a mediator of inflammation.

Previous genomewide association studies (GWASs) of AIDS have revealed a major association involving a genetic polymorphism within the human leukocyte antigen region, the rs2395029 *HCP5* single-nucleotide polymorphism (SNP), which tracks *HLA-B*5701*. This SNP was associated with viral load control through analysis of human immunodeficiency virus type 1 (HIV-1) seroconverters [1, 2] and by the comparison of patients with long-term nonprogression (LTNPs) with uninfected control subjects as well [3]. LTNPs are a small percentage (1%–5%) of HIV-1 seroconverters [4–6] and thus constitute a powerful contrasting tool to unravel new genetic factors associated with AIDS progression. Of the LTNPs in the Genomics of Resistance to Immunodeficiency Virus (GRIV) cohort, patients carrying the *HCP5* rs2395029-G allele exhibited a significantly lower viral load than the rest of the cohort [3]. Only a minority (ie, 25%) of the GRIV LTNPs exhibited effective viral load control (ie, a very low viral load of <100 copies/mL). Because viral load is known to account for only 34% of the variability in the time to a CD4 T cell decrease of <200 cells/μL [7], we decided to perform a new analysis of the genomewide data for

Received 19 February 2010; accepted 9 April 2010; electronically published 12 August 2010.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Potential conflicts of interest: none reported.

Financial support: see the Acknowledgments section.

[a] These authors have equally contributed to the work.

Author affiliations are listed at the end of the text.

Reprints or correspondence: Jean-François Zagury, 292 rue Saint Martin, 75003 Paris, France (zagury@cnam.fr).

The Journal of Infectious Diseases 2010;202(6):908–915

0022-1899/2010/20206-0014$15.00
DOI: 10.1086/655782

GRIV LTNPs by excluding these "elite controller" subjects (subjects who had a viral load of <100 copies/mL). The aim of the current study was thus to focus on genetic variations affecting LTNP without necessarily controlling the viral load at a very low level. The result is that we have indeed identified a new specific signal in the *CXCR6* gene and have replicated this finding in 3 additional independent cohorts of European descent.

METHODS

The GRIV Study: Participants, Genotyping, and Analysis

The GRIV cohort. The GRIV study cohort and methods were described in detail in previously published work on the genomewide association study of LTNPs [3]. The GRIV cohort was established in France in 1995 to generate a large collection of DNAs for genetic studies seeking to identify host genes associated with rapid and LTNP to AIDS [8–11]. Only white people who were of European descent and were living in France were eligible for enrollment, to reduce confounding by population substructure. The LTNPs were seroprevalent subjects who were included on the basis of their main clinical outcomes, CD4 T cell count, and time to disease progression: asymptomatic HIV-1 infection for >8 years, no receipt of antiretroviral treatment, and a CD4 T cell count consistently >500 cells/mm^3.

Among those in the LTNP group ($n = 275$), viral load (ie, the plasma HIV-1 RNA concentration) at the time of inclusion could be assessed for 248 individuals. Of these 248 individuals, 186 had a viral load >100 copies/mL. All subjects provided written, informed consent before their enrollment in the GRIV genetic association study.

The control group. The Data from an Epidemiological Study on Insulin Resistance Syndrome (DESIR) program was a 9-year follow-up study designed to clarify the development of the insulin resistance syndrome. During 1994–1996, subjects were recruited from volunteers insured by the French social security system, which offers periodic health examinations free of charge [12]. This control group was comprised of 697 nonobese and normoglycemic individuals, and all were French, of European descent, and HIV-1 seronegative.

Genotyping method and quality control. The GRIV cohort and the control group were genotyped using the Illumina Infinium II HumanHap300 BeadChips (Illumina). Genotyping quality was assessed using BeadStudio software (version 3.1; Illumina). Missing data (>2%), low minor allele frequency (<1%), and deviations from Hardy-Weinberg equilibrium in the control group ($P < 1.0 \times 10^{-3}$) were excluded from analysis during these quality control steps. Moreover, identification of potential population stratification was identified using Structure software (version 2.2) [13], by producing a quantile-quantile plot (see Figure A1*A* in the Appendix, which appears only in the electronic version of the *Journal*) and by computing the genomic inflation factor λ. Overall, little effect of stratification was observed, and 283,637 SNPs could be tested statistically for association with LTNP.

Statistical analysis. For each SNP, we performed a standard case-control analysis, using Fisher's exact tests (with Plink software [14]) to compare allelic distributions between LTNPs and the control subjects. Bonferroni corrections were made to account for multiple comparisons.

SNP imputation. Untyped SNPs present in the HapMap database of chromosome 3 were imputed for all GRIV patients and control subjects, by use of Impute software (version 2.1) [15] and the HapMap release 21 phased data for the white population (CEU) as the reference panel [16].

CXCR6 ***genotyping by PCR sequencing.*** Primers and conditions used for polymerase chain reaction (PCR) amplifications were standard. Sequencing reactions were performed according to the Dye Terminator method by use of an ABI Prism 3730XL DNA Analyzer (Applied Biosystems). Alignment, SNP discovery, and genotyping were performed using the software Genalys, which was developed by the Commissariat à l'Énergie Atomique/Centre National de Génotypage [17].

Haplotype inference. Haplotype inference was obtained using the rapid and accurate Shape-IT algorithm [18].

Bioinformatics exploration. To further explore the associations observed, we tried to identify modifications in messenger RNA (mRNA) expression (*Genevar* [19] and Dixon [20] databases), splicing (NetGene2 [21]), polyadenylation (polyAH [22] and polyApred [23]), and transcription factor–binding sites (SignalScan [24], TESS [25], and TFSearch [26], derived from the TRANSFAC database).

Replication in the Amsterdam Cohort Study: Participants, Genotyping, and Analysis

The Amsterdam Cohort Study (ACS) participants and methods were described in detail elsewhere [27]. In the present study, 335 HIV-1–infected homosexual men from ACS were analyzed for the course of HIV-1 infection using AIDS-related death as an end point. AIDS-related death is defined as death with AIDS-related malignancy, death with AIDS opportunistic infections, or death with an AIDS-related cause not specified by the treating physician.

The ACS rs2234358 genotyping data were obtained using Illumina Infinium II HumanHap300 BeadChips (Illumina). Quality control filters were applied to ensure reliable genotyping data. Potential population stratification was also analyzed using Structure software (version 2.2) [13], and 19 participants were thus excluded from further analyses ($n = 316$) because they differed significantly from the HapMap white population.

Statistical analysis was performed by Kaplan-Meier survival analysis and determination of the log rank *P* value under the

genotypic model, by use of SPSS software (version 16.0; SPSS) and the R package [28].

Because the viral load (ie, the plasma HIV-1 RNA concentration) and the CD4 T cell count were assessed during routine clinical follow-up, we could identify the ACS LTNPs who matched the GRIV definition and exhibited a viral load of >100 copies/mL ($n = 31$). LTNP status was easily determined for seroconverters because the date of seroconversion was known, and this was also the case for seroprevalent subjects, because the time of seropositivity was imputed (on average, at 18 months before enrollment).

Replication in the Multicenter AIDS Cohort Study: Participants, Genotyping, and Analysis

Multicenter AIDS Cohort Study (MACS) participants and methods previously have been described in detail elsewhere [29]. GWAS data were collected from 156 HIV-1–infected white homosexual men, with time to clinical AIDS used as an end point. This panel was chosen to be enriched with extreme AIDS progression phenotypes.

The MACS rs2234358 genotyping data were obtained using the Affymetrix GeneChip Human Mapping 500K Array (Affymetrix), in which the rs2234358 SNP is tagged by rs4682799 ($r^2 = 1$). Quality control filters were applied to ensure reliable genotyping data, and population stratification was also checked.

Statistical analysis was performed by Kaplan-Meier survival analysis and Cox proportional regression determination of the *P* value under the genotypic model using the R package.

As with the ACS, viral load (the plasma HIV-1 RNA concentration) and CD4 T cell count were assessed during routine clinical follow-up. We could extract 59 MACS LTNPs, selected from among seroconverters and seroprevalent subjects, who matched the GRIV definition and exhibited a viral load of >100 copies/mL.

Replication in the USA HIV-1 Cohort: Participants, Genotyping, and Analysis

USA HIV-1 cohort patients and methods previously were described in detail elsewhere [30]. For this study, 556 HIV-1 seroconverters of European ancestry were collected from 4 USA-based natural history HIV/AIDS longitudinal cohorts (MACS, San Francisco City Clinic Cohort, Multicenter Hemophilia Cohort Study, and Hemophilia Growth and Development Study), with AIDS-related death used as an end point. Of importance, the 556 USA HIV-1–infected participants did not include subjects overlapping with the 156 MACS participants.

The USA HIV-1 cohort rs2234358 genotyping data were obtained using commercial TaqMan genotype assays (with assay ID C_1929536_1; Applied Biosystems). Conformity to the genotype frequencies expected under Hardy-Weinberg equilibrium was checked.

Statistical analysis was performed by Kaplan-Meier survival analysis and Cox proportional regression for determination of the *P* value under the genotypic model using the statistical SAS package (version 9.13; SAS Institute).

Independence from the *CCR2–CCR5* Locus and *HCP5* Polymorphisms

The genotypic data were available for the known *CCR2–CCR5* locus and *HCP5* polymorphisms in the GRIV, MACS, and USA HIV-1 cohorts, and it was thus possible to assess the independence of the rs2234358 SNP from these polymorphisms.

For the GRIV cohort, multivariate logistic regression analysis was used to adjust effects of covariates CCR2–64I, CCR5-Δ32, CCR5-P1, and *HCP5* rs2395029. The same approach was done for the MACS and USA HIV-1 cohorts but by fitting to the data a linear model instead of a logistic model. The independent effect of the rs2234358 SNP on disease phenotype was confirmed by adjusting the model with these covariates: the *P* values that were obtained were similar with and without the covariate analysis.

RESULTS

After quality control tests, a case-control analysis using Fisher's exact tests was performed to compare allelic distributions of the 283,637 SNPs between the GRIV LTNPs exhibiting a detectable viral load (>100 copies/mL) ($n = 186$) and uninfected controls ($n = 697$) (see Methods). The strongest association was found for rs2234358, with a *P* value close to the 1.7×10^{-7} Bonferroni threshold for genomewide significance (see Figure A1*B* in the Appendix, which appears only in the electronic version of the *Journal*): $P = 2.5 \times 10^{-7}$ (odds ratio [OR], 1.85 [95% confidence interval {CI}, 1.46–2.36]). The rs2234358-T allele was associated with not being an LTNP (36.83% in LTNPs vs. 51.94% in controls) (Figure 1*A*). This allele was not associated with acquisition of HIV-1 infection, because its frequency was similar in seropositive and control groups: 51.16% among GRIV rapid progressors [31], 48.59% in the ACS, 51.92% in the MACS, 48% in the USA HIV-1 cohort, 54.9% in the Dutch control population (CTR_{ACS}), 48.3% in the Illumina controls ($CTR_{Illumina}$), and 50.9% in HapMap CEU (see Figure A2 in the Appendix, which appears only in the electronic version of the *Journal*).

The rs2234358 SNP lies within the *CXCR6* gene in a region of chromosome 3 that is rich in genes encoding chemokine receptors, and it is notably positioned 422 kb from the well-known *CCR5* gene [32] (Figure 2*A*). To eliminate possible tracking effects, we evaluated a potential association between the *CXCR6* signal and the *CCR2–CCR5* haplotypes (Δ32, P1, and 64I) previously associated with the course of HIV-1 disease [9, 32–33]. First, the rs2234358 SNP had no linkage disequilibrium (LD, $r^2 < 0.1$) with any of the CCR5-Δ32, CCR5-P1,

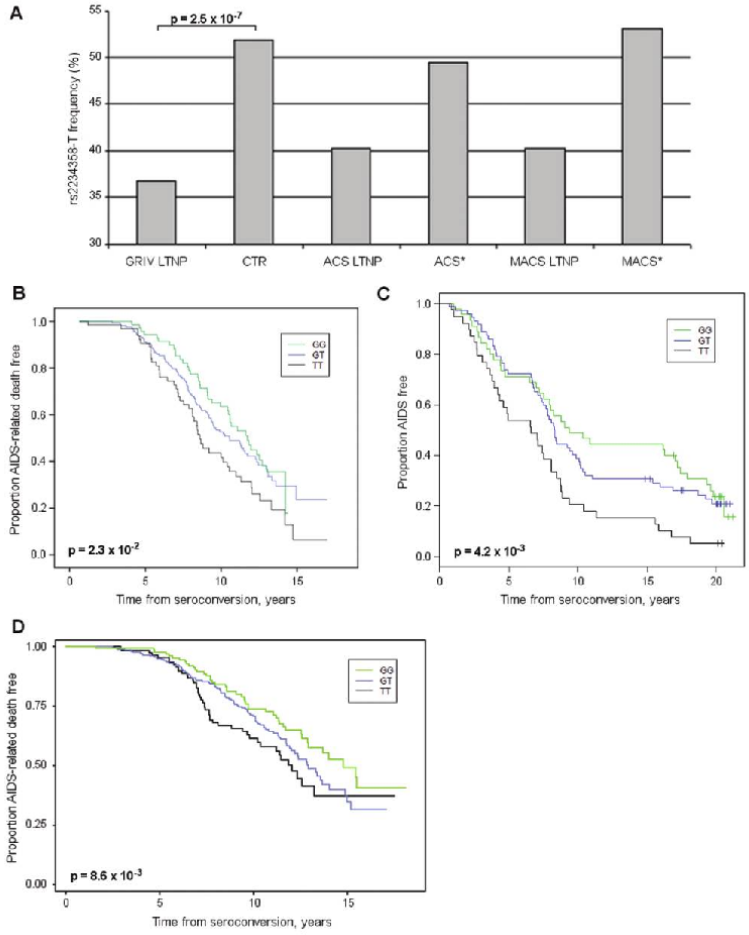


Figure 1. Effect of rs2234358 in the Genomics of Resistance to Immunodeficiency Virus (GRIV), Amsterdam Cohort Study (ACS), Multicenter AIDS Cohort Study (MACS), and USA HIV-1 study groups. *A,* Allelic frequency of rs2234358-T in the GRIV long-term nonprogressor (LTNP) population ($n = 186$) and the control group (CTR) ($n = 697$). Frequencies are also given for the 31 ACS subjects with LTNP (ACS LTNPs), for the remaining 285 ACS participants (ACS*), for the 59 MACS subjects with LTNP (MACS LTNPs), and for the remaining 97 MACS participants (MACS*). *B,* Kaplan-Meier survival curve derived from the ACS cohort for the time to AIDS-related death. Genotypes GG (*green*) ($n = 76$), GT (*blue*) ($n = 171$), and TT (*black*) ($n = 69$). *C,* Kaplan-Meier survival curve derived from the MACS cohort for time to clinical AIDS. Genotypes GG (*green*) ($n = 45$), GT (*blue*) ($n = 72$), and TT (*black*) ($n = 39$). *D,* Kaplan-Meier survival curve derived from the USA HIV-1 cohort for time to AIDS-related death. Genotypes GG (*green*) ($n = 140$), GT (*blue*) ($n = 297$), and TT (*black*) ($n = 119$).

or CCR2–64I haplotypes. Second, we could not find any epistatic effects between rs2234358 and any of these haplotypes, by use of either Plink software [14] or logistic regression using *CCR2-CCR5* haplotypes as covariates (version 2.1) (see Methods). Third, the HapMap LD for whites did not reveal any SNP with a high LD ($r^2 > 0.9$) beyond the *CXCR6* locus. Of note, we also did not observe an epistatic effect between rs2234358 and the chromosome 6 rs2395029 *HCP5/HLA-B*5701* signal. This *CXCR6* signal thus represents a new association with LTNP, independent from the well-known *CCR2–CCR5* and *HCP5/HLA-B*5701* associations. Of interest, we inferred the SNP distribution over the entire chromosome 3, using Impute software (see Methods). Instead of the 20,000 genotyped SNPs present in the Illumina HumanHap300 BeadChip in chromosome 3, a

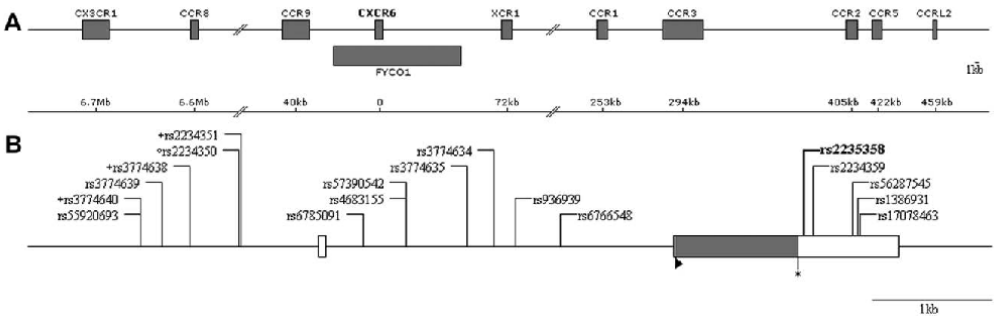


Figure 2. *A,* Genetic map of the *CXCR6* gene region. *CXCR6* is localized within the 14th intron of a predicted gene, *FYCO1*. *B,* Genetic map of the *CXCR6* gene. Exons and untranslated regions are symbolized by full and empty rectangles, respectively. The positions of the ATG and STOP codons are indicated by a triangle (▶) and an asterisk (*), respectively. All single-nucleotide polymorphisms (SNPs) covered by the polymerase chain reaction sequencing study are represented, and the rs2234358 SNP of interest is shown in boldface type. The 3 promoter haplotypes in high linkage disequilibrium (LD) with rs2234358 ($r^2 = 0.97$) correspond to 2-SNP haplotypes composed of the rs2234350 SNP (°), with either one of the SNPs denoted by the + symbol (these 3 latter SNPs exhibit $r^2 = 0.99$).

total of 176,000 SNPs could be imputed for which we could not identify a *P* value better than the one exhibited for rs2234358. The attributable risk for rs2234358-T variant is very strong, because it explains 12% of the prevention of LTNP. For comparison, the attributable risk for CCR5-Δ32 is 5.1% in the GRIV LTNP cohort.

The rs2234358 signal was replicated in 3 independent additional cohort studies of white people of European descent that also evaluated for AIDS progression phenotype (after removal of stratification outliers; see Methods): (1) the European ACS cohort ($n = 316$) ($P = 2.3 \times 10^{-2}$) (Figure 1*B*), (2) a European descent subgroup of the American MACS cohort enriched in extreme phenotypes ($n = 156$) ($P = 4.2 \times 10^{-3}$) (Figure 1*C*), and (3) a pool of European American HIV-1 cohorts ($n = 556$) ($P = 8.6 \times 10^{-3}$) (Figure 1*D* and Table A1, the latter of which may be found in the Appendix, which appears only in the electronic version of the *Journal*). As shown in Figure 1, the rs2234358-T allele favored progression in all of these cohorts, which is in agreement with a prevention of LTNP. Overall, the combined *P* value computed by the Fisher method between the 4 cohorts (GRIV, ACS, MACS, and USA HIV-1) passed the Bonferroni genomewide significance threshold: $P_{\text{combined}} = 9.7 \times 10^{-10}$.

It was surprising to observe significant but rather weak *P* values in all cohorts except the GRIV cohort, so we assessed whether the effect was specifically amplified in the LTNP subpopulation. We identified 31 and 59 LTNPs fulfilling the GRIV definition and with a detectable viral load (>100 copies/mL) in the ACS and MACS cohorts, respectively. In these groups, the rs2234358-T allele frequency was ~40%, which is similar to that found among the GRIV LTNPs (Figure 1*A*). Because no LTNP from these 3 cohorts differed significantly from the HapMap white population, according to the Structure analysis [13] (see Figure A3*B* in the Appendix, which appears only in the electronic version of the *Journal*), the ACS and MACS LTNPs were added to those in the GRIV cohort, and we computed a *P* value comparing this extended LTNP case group ($n = 276$) with the control group ($n = 697$). The *P* value again reached genomewide significance: $P = 2.1 \times 10^{-8}$ (OR, 1.77 [95% CI, 1.44–2.18]), confirming the association of rs2234358-T with prevention of LTNP (Table A1, which may be found in the Appendix, which appears only in the electronic version of the *Journal*). Of importance, several additional control groups were tested and exhibited a similar allele frequency for rs2234358-T (see Figure A2 in the Appendix, which appears only in the electronic version of the *Journal*).

To further explore this association, we resequenced the entire *CXCR6* gene to detect additional variants (Figure 2*B* and Table A2, the latter of which may be found in the Appendix, which appears only in the electronic version of the *Journal*): rs2234358 remained the SNP exhibiting the strongest association. Interestingly, using Shape-IT software (version 2.0) to compute haplotypes [18], we found several haplotypes comprising *CXCR6* promoter SNPs in high LD ($r^2 = 0.97$) with rs2234358 (Figure 2*B*).

The rs2234358 SNP is located in the 3′ untranslated region of *CXCR6,* located 42 bp downstream from the termination codon (Figure 2*B*), and could thus influence gene expression, mRNA stability, mRNA regulation, or mRNA splicing. According to the Dixon or *Genevar* mRNA expression databases, none of the genotyped SNPs are predicted to affect *CXCR6* or any other chromosome 3 gene expression, and bioinformatics methods failed to predict a modification of splicing or polyadenylation sites (see Methods). Nevertheless, we identified sev-

eral putative transcription factor–binding sites containing the SNPs included in promoter haplotypes in high LD with rs2234358 (see Methods). Further experiments are required to determine the causative genetic variants and the biological mechanisms at stake.

DISCUSSION

Because the major signal identified in previous AIDS GWASs was associated with control of viral replication, we reanalyzed the genomewide data of the French GRIV LTNP cohort by excluding elite controller patients (ie, patients with a viral load of <100 copies/mL). The comparison of 186 LTNPs exhibiting a viral load of >100 copies/mL with 697 uninfected controls highlighted a strong association for the *CXCR6* rs2234358 ($P = 2.5 \times 10^{-7}$). This new signal was replicated by a candidate SNP approach in 3 additional independent European descent cohorts (including 316, 156, and 556 subjects), and the combined *P* value of the 4 cohorts reached the genomewide significance threshold: $P_{combined} = 9.7 \times 10^{-10}$. This chromosome 3 association is independent from the well-known neighboring *CCR2–CCR5* locus, is not linked with the control of viral load (the GRIV LTNP groups carrying the various rs2234358 genotypes exhibit a similar mean viral load) ($P = .72$, data not shown), and exhibits a high attributable risk of LTNP of 12%.

This study presents the first non-HLA–replicated association obtained through a GWAS approach. The *P* value for rs2234358 is very strong in the GRIV cohort, and this signal was confirmed in 3 independent cohorts but with weaker *P* values. The specific design of the LTNP phenotype can explain this discrepancy. Indeed, the extraction of LTNPs with a viral load >100 copies/mL from the ACS and MACS cohorts confirmed the strength of this common SNP association with LTNP: ~40% versus ~50% in several uninfected control groups (see Figure 1*A* and Figure A2 in the Appendix, the latter of which appears only in the electronic version of the *Journal*). It emphasizes the critical importance of cohort design and the particular power of extreme phenotypes [5, 34–35], particularly in light of a recent powerful GWAS involving >2500 patients, which solely identified chromosome 6–related signals [36].

The finding of a new chemokine receptor genetic variant contributing to a differential progression to AIDS is not so much of a surprise, because the chemokine system is a major weapon of the early host defense system against infectious diseases and comprises >100 members. An exonic *CXCR6* variant present in African Americans (but absent in Europeans) was previously associated with *Pneumocystis carinii* pneumonia–mediated progression to AIDS [37]. Our *CXCR6* genetic variant is not exonic, and its biological effect should rather be a modulation of *CXCR6* expression. CXCR6, known as a minor HIV-1 coreceptor [38] and mediator of inflammation [39, 40], is notably expressed in organs (thymus, gut, and bone marrow) and in immune cells [41], which are important for HIV-1 infection. It is involved in the trafficking of effector T cells mediating type 1 inflammation [39] and in the activation and homeostasis of natural killer T cells [42], known to be an important bridge between innate and adaptive immune responses. Interestingly, in simian immunodeficiency virus (SIV) infection, it has been proposed that interleukin-17–secreting natural killer T cells could compensate for the Th17 defect in the gut, because they are essential for controlling mucosal barrier integrity and microbial translocation [43, 44]. These hypotheses are compatible with a major role of CXCR6 as an inflammation mediator in AIDS [39, 40], but they deserve further functional/biological research to enhance our understanding of the molecular pathways to HIV-1 LTNP.

At a time when HIV-1 entry inhibitors such as CCR5 and CXCR4 antagonists are being developed, the identification of a molecular mechanism of AIDS pathogenesis involving a new chemokine receptor is of particular interest and opens new insights for therapeutic drug targets and prediction of HIV-1 progression.

AUTHOR AFFILIATIONS

Chaire de Bioinformatique, Conservatoire National des Arts et Métiers (S.L., C.C., S.L.C., O.D., G.D., L.T., M.M., F.S., and J.-F.Z.), [2]Agence Nationale de Recherches sur le SIDA et les Hépatites Virales Genomic Group (French Agency for Research on AIDS and Hepatitis) (C.C., S.L.C., C.R., J.-F.D., and J.-F.Z.), and [3]Laboratoire d'Oncologie Moléculaire, Université Paris 7, Paris (A.T.), [4]Université Paris 12, Institut National de la Santé et de la Recherche Médicale (lNSERM) U955 (S.L., S.L.C., J.-D.L., Y.L., J.E., and J.-F.Z.), and [5]Assistance Publique Hôpitaux de Paris, Hôpital Henri Mondor, Créteil (J.E.), [6]Commissariat à l'Énergie Atomique/Institut de Génomique, Centre National de Génotypage, Evry (S.L. and I.G.), [7]Unite Mixte de Recherche (UMR) U557 INSERM/U1125 Inra/Conservatoire National des Arts et Métiers/UP13, Centre de Recherche en Nutrition Humaine Ile-de-France, Santé-Médecine-Biologie Humaine Paris 13, Bobigny (S.H.), and [8]UMR Centre National de la Recherche Scientifique 8090, Institut Pasteur de Lille, Lille, France (C.D. and P.F.); [9]Department of Experimental Immunology, Sanquin Research, Landsteiner Laboratory, Center for Infectious Diseases and Immunity Amsterdam Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (D.v.M. and A.B.v.W.); [10]Genomic Medicine, Hammersmith Hospital, Imperial College London, London, United Kingdom (P.F.); [11]Department of Microbiology, University of Washington School of Medicine, Seattle (J.T.H., G.S.G., and J.I.M.); [12]Laboratory of Genomic Diversity, Science Applications International Corporation–Frederick, National Cancer Institute–Frederick, Frederick (P.A. and C.W.), and [13]Infections and Immunoepidemiology Branch, National Cancer Institute–Bethesda, Di-

vision of Cancer Epidemiology and Genetics, Rockville, Maryland (J.J.G.); [14]Feinberg School of Medicine, Division of Infectious Diseases, Northwestern University, Chicago, Illinois (J.P.); [15]Department of Biostatistics, Rho, Chapel Hill, North Carolina (S.D.); [16]San Francisco Department of Public Health, HIV Research Section, San Francisco, California (S.B.).

Acknowledgments

We thank all the patients and medical staff who have kindly collaborated with the Genomics of Resistance to Immunodeficiency Virus project. S.L. benefits from a fellowship from the French Ministry of Education, Technology and Research, and S.L.C. benefits from a fellowship of Agence Nationale de Recherches sur le SIDA et les Hépatites Virales Genomic Group.

Financial support. Agence Nationale de Recherche sur le SIDA, Sidaction, Innovation 2007 program of Conservatoire National des Arts et Métiers, AIDS Cancer Vaccine Development Foundation, Neovacs SA, Vaxconsulting, and R37 (grant AI047734) from the US National Institutes of Health. This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health (NIH; contract HHSN261200800001E) and in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The authors acknowledge funding from the Netherlands Organization for Scientific Research (TOP; registration number 9120.6046). The Hemophilia Growth and Development Study is funded by the National Institutes of Health, National Institute of Child Health and Human Development, 1 R01 HD41224. The Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration between the Amsterdam Health Service, the Academic Medical Center of the University of Amsterdam, Sanquin Research, and the University Medical Center Utrecht, are part of the Netherlands HIV Monitoring Foundation and are financially supported by the Netherlands National Institute for Public Health and the Environment.

References

1. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. Science **2007**; 317: 944–7.
2. Dalmasso C, Carpentier W, Meyer L, et al. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. PLoS ONE **2008**; 3:e3907.
3. Limou S, Le Clerc S, Coulonges C, et al. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). J Infect Dis **2009**; 199:419–26.
4. Grabar S, Selinger-Leneman H, Abgrall S, Pialoux G, Weiss L, Costagliola D. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. AIDS **2009**; 23:1163–9.
5. Huber C, Pons O, Hendel H, et al. Genomic studies in AIDS: problems and answers. Development of a statistical model integrating both longitudinal cohort studies and transversal observations of extreme cases. Biomed Pharmacother **2003**; 57:25–33.
6. Petrucci A, Dorrucci M, Alliegro MB, et al. How many HIV-infected individuals may be defined as long-term nonprogressors? A report from the Italian Seroconversion Study. Italian Seroconversion Study Group (ISS). J Acquir Immune Defic Syndr Hum Retrovirol **1997**; 14:243–8.
7. Mellors JW, Margolick JB, Phair JP, et al. Prognostic value of HIV-1 RNA, CD4 cell count, and CD4 cell count slope for progression to AIDS and death in untreated HIV-1 infection. JAMA **2007**; 297:2349–50.
8. Rappaport J, Cho YY, Hendel H, Schwartz EJ, Schachter F, Zagury JF. 32 bp CCR-5 gene deletion and resistance to fast progression in HIV-1 infected heterozygotes. Lancet **1997**; 349:922–3.
9. Winkler CA, Hendel H, Carrington M, et al. Dominant effects of CCR2-CCR5 haplotypes in HIV-1 disease progression. J Acquir Immune Defic Syndr **2004**; 37:1534–8.
10. Hendel H, Caillat-Zucman S, Lebuanec H, et al. New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. J Immunol **1999**; 162:6942–6.
11. Flores-Villanueva PO, Hendel H, Caillat-Zucman S, et al. Associations of MHC ancestral haplotypes with resistance/susceptibility to AIDS disease development. J Immunol **2003**; 170:1925–9.
12. Balkau B. An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome [in French]. Rev Epidemiol Sante Publique **1996**; 44:373–5.
13. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics **2000**; 155:945–59.
14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **2007**; 81:559–75.
15. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet **2007**; 39:906–13.
16. International HapMap Project. http://www.hapmap.org. Accessed 22 July 2010.
17. Takahashi M, Matsuda F, Margetic N, Lathrop M. Automated identification of single nucleotide polymorphisms from sequencing data. J Bioinform Comput Biol **2003**; 1:253–65.
18. Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics **2008**; 9:540.
19. Ge D, Zhang K, Need AC, et al. WGAViewer: software for genomic annotation of whole genome association studies. Genome Res **2008**; 18:640–3.
20. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. Nat Genet **2007**; 39:1202–7.
21. NetGene2 Server. http://www.cbs.dtu.dk/services/NetGene2/. Accessed 22 July 2010.
22. SoftBerry.com. http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter. Accessed 30 July 2010.
23. polyApred. http://www.imtech.res.in/raghava/polyapred/submission.html. Accessed 22 July 2010.
24. WWW Signal Scan. http://www-bimas.cit.nih.gov/molbio/signal/. Accessed 22 July 2010.
25. Transcription Element Search System. http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME. Accessed 22 July 2010.
26. TFSearch: Searching transcription factor binding sites (ver 1.3). http://www.cbrc.jp/research/db/TFSEARCH.html. Accessed 22 July 2010.
27. van Manen D, Kootstra NA, Boeser-Nunnink B, Handulle MA, van't Wout AB, Schuitemaker H. Association of HLA-C and HCP5 gene regions with the clinical course of HIV-1 infection. AIDS **2009**; 23: 19–28.
28. R Project. http://www.r-project.org. Accessed 22 July 2010.
29. Herbeck JT, Gottlieb GS, Winkler CA, et al. Multi-stage genome-wide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. J Infect Dis **2010**; 201:618–26.
30. An P, Duggal P, Wang LH, et al. Polymorphisms of CUL5 are associated with $CD4^+$ T cell loss in HIV-1 infected individuals. PLoS Genet **2007**; 3: e19.
31. Le Clerc S, Limou S, Coulonges C, et al. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). J Infect Dis **2009**; 200:1194–201.
32. Dean M, Carrington M, Winkler C, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science **1996**; 273:1856–62.

33. Martin MP, Dean M, Smith MW, et al. Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science **1998**; 282:1907–11.
34. Froguel P, Blakemore AI. The power of the extreme in elucidating obesity. N Engl J Med **2008**; 359:891–3.
35. Zhang G, Nebert DW, Chakraborty R, Jin L. Statistical power of association using the extreme discordant phenotype design. Pharmacogenet Genomics **2006**; 16:401–13.
36. Fellay J, Ge D, Shianna KV, et al. Common genetic variation and the control of HIV-1 in humans. PLoS Genet **2009**; 5:e1000791.
37. Duggal P, An P, Beaty TH, et al. Genetic influence of CXCR6 chemokine receptor alleles on PCP-mediated AIDS progression among African Americans. Genes Immun **2003**; 4:245–50.
38. Deng HK, Unutmaz D, KewalRamani VN, Littman DR. Expression cloning of new receptors used by simian and human immunodeficiency viruses. Nature **1997**; 388:296–300.
39. Kim CH, Kunkel EJ, Boisvert J, et al. Bonzo/CXCR6 expression defines type 1-polarized T-cell subsets with extralymphoid tissue homing potential. J Clin Investig **2001**; 107:595–601.
40. Landrø L, Damås JK, Halvorsen B, et al. CXCL16 in HIV infection—a link between inflammation and viral replication. Eur J Clin Invest **2009**; 39:1017–24.
41. Koprak S, Matheravidathu S, Springer M, Gould S, Dumont FJ. Down-regulation of cell surface CXCR6 expression during T cell activation is predominantly mediated by calcineurin. Cell Immunol **2003**; 223: 1–12.
42. Germanov E, Veinotte L, Cullen R, Chamberlain E, Butcher EC, Johnston B. Critical role for the chemokine receptor CXCR6 in homeostasis and activation of CD1d-restricted NKT cells. *J Immunol* **2008**; 181: 81–91.
43. Campillo-Gimenez L, Cumont MC, Fay M, et al. AIDS progression is associated with the emergence of IL-17-producing cells early after simian immunodeficiency virus infection. J Immunol **2010**; 184:984–92.
44. Raffatellu M, Santos RL, Verhoeven DE, et al. Simian immunodeficiency virus-induced mucosal interleukin-17 deficiency promotes *Salmonella* dissemination from the gut. Nat Med **2008**; 14:421–8.

III. Méta-analyse des données IHAC

Le but du projet IHAC (International HIV Acquisition Consortium) est de trouver des facteurs génétiques associés à la susceptibilité de la contamination par le virus VIH. Ce projet a débuté en 2009 et les analyses sont menées de manière conjointe par deux centres d'analyse américain et européen, l'un au Broad Institute (Boston, USA), l'autre au Conservatoire National des Arts et Métiers (Paris) et j'en suis le responsable opérationnel sous la direction du Pr. Zagury. Le projet n'a pas encore abouti à une publication mais des résultats préliminaires ont déjà été présentés aux membres du consortium.

Ce projet fait intervenir une vingtaine de cohortes génotypées sur des plates-formes différentes et est, à ce titre, impressionnant.

Voici un bilan des étapes réalisées et des premiers résultats.

III.1 Collecte des données et contrôle qualité (2009-2011)

Après l'envoi par chaque intervenant de ses données, j'ai du réaliser un contrôle qualité:

- Éliminer les **SNPs** avec des données manquantes > 5%, des fréquences alléliques < 1% et dont l'équilibre de Hardy-Weinberg n'est pas respecté avec une p-value < 10^{-5}
- Éliminer les **individus** qui sont trop éloignés des autres par stratification, ceux qui ont des données manquantes > 5%, une hétérozygosité avec un coefficient de consanguinité > 0,1.

Une fois les données réunies il fut également nécessaire de vérifier (par le calcul de l'IBS) qu'il n'y avait pas de doublons entre les cohortes.

III.2 Constitution des groupes cas-contrôles

Enfin une dernière étape constituait à un « appariement » le plus homogène possible des cas et des contrôles stratifiés par origines ethniques et par puce de génotypage (pour ne pas perdre trop de SNP dans le regroupement).

Grâce à des analyses en composantes principales, il fut possible de former 6 groupes cas-contrôles assez homogènes comme en atteste les Q-Q plots (figure 28).

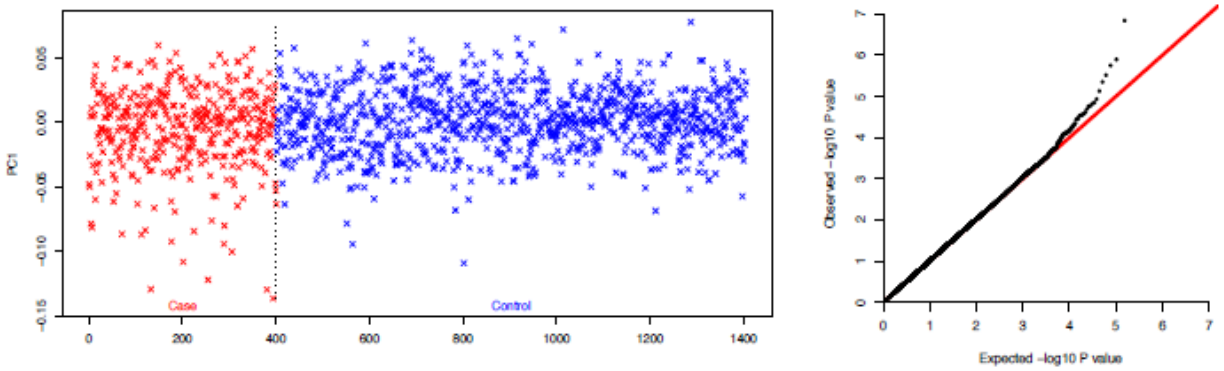


Figure 28 : Analyse en composante principale et représentation des patients sur le 1er axe, on observe une certaine homogénéité des patients. Le Q-Q plot d'association (sur les SNPs non imputés) ne présente pas de déviation conséquente

III.3 Imputation et tests d'association

III.3.1 Approche générale « classique »

Les données de chaque groupe cas-contrôles ont ensuite suivi un « pipeline » d'analyse comme suit :

1. Détermination des covariables pour prendre en compte la stratification résiduelle [153] entre les populations cas et contrôles avec le logiciel EIGENSTRAT [154]
2. Pré-haplotypage des données de cas et de contrôles ensemble pour éviter un biais dans l'algorithme qui pourrait converger vers des résolutions d'haplotypes différentes si les cas étaient haplotypés séparément des contrôles avec le logiciel SHAPEIT [51, 52]
3. Dernière phase d'imputation à l'aide des dernières données de référence 1000genomes (phase 1 interim) disponibles regroupant 1094 individus et environ 37 millions de SNPs à l'aide du logiciel IMPUTE2 [155].

Ce traitement en pipeline sur chaque groupe cas-contrôle a permis d'obtenir en moyenne 8,5 millions de SNPs bien imputés (info>0.8) par groupe cas-contrôles, dont 6,5 millions de SNPs bien imputés (info>0.8) communs à tous les groupes (voir figure 29 pour la qualités d'imputation).

4. L'analyse des associations s'est faite d'abord groupe par groupe en utilisant le logiciel SNPTEST dans les modes allélique, additif, dominant, récessif et en utilisant une régression logistique incluant les composantes principales calculées en 1) ci-dessus.

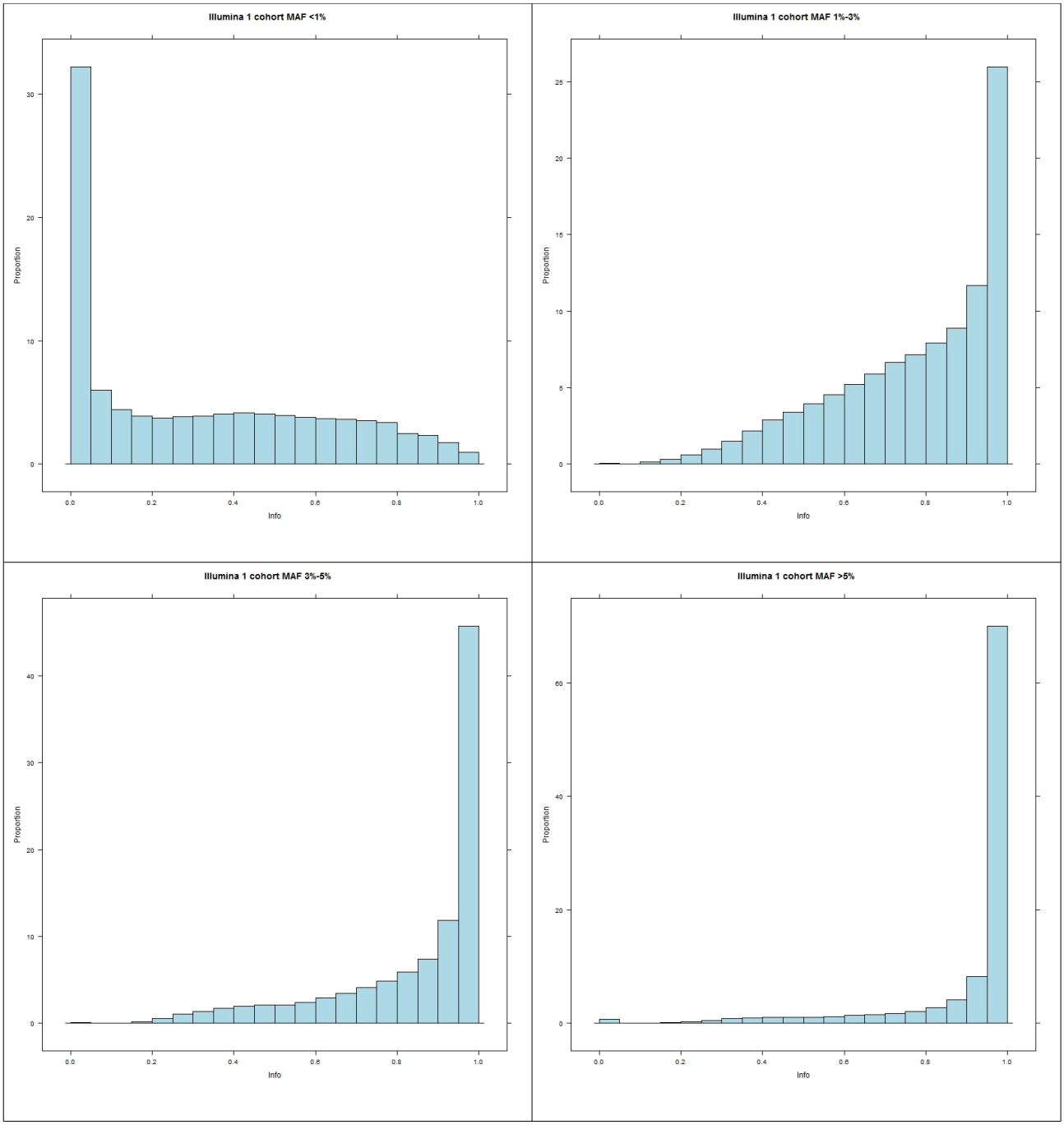


Figure 29 : Histogramme sur la qualité d'imputation (info score) suivant les groupes de fréquence alléliques mineures (MAF) respectivement <1%, entre 1 et 3%, entre 3 et 5% et supérieur à 5%. On observe que l'imputation est très mauvaise pour les SNPs rares inférieurs à 1% mais s'améliore rapidement pour devenir excellente ensuite

Nous avons ensuite procédé à deux approches complémentaires :

1. Un calcul direct de méta-analyse des 6 cohortes en utilisant le logiciel META [17, 156] et en filtrant par la qualité d'imputation représentée par l'info score (info>0,8) par la méthode du z-score et de l'inverse de la variance (modèle « fixe » si homogène et modèle « aléatoire » sinon). Ceci aboutit à comparer environ 6,5 millions de marqueurs (figure 30).

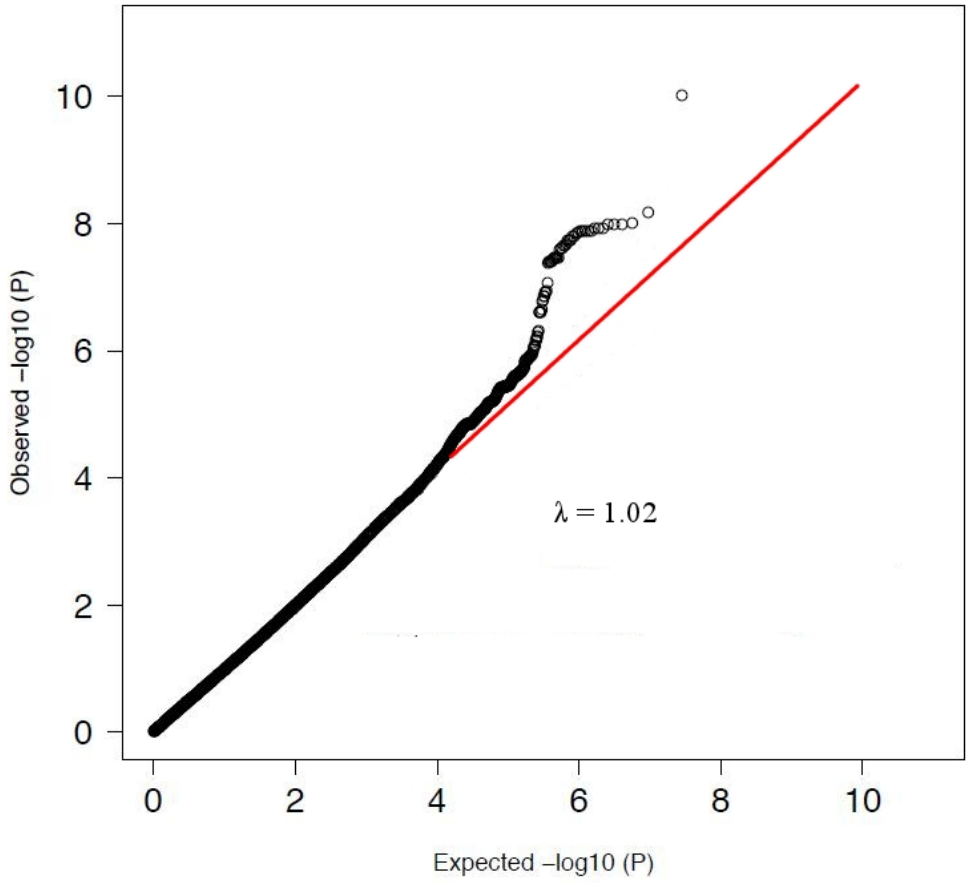


Figure 30 : Q-Q plot des 6,5 millions de SNPs dans la méta-analyse finale en filtrant les SNPs avec un seuil d'info score d'imputation à 0,8 dans chaque cohorte. On voit que le facteur d'inflation λ après correction par l'ajout des composantes principales en covariables ne dénote pas une stratification notable

2. Un calcul lié au nombre de fois où le signal est répliqué (p<0,05) dans les 6 groupes cas-contrôles avec un effet allant dans le même sens (avec le même filtre de qualité d'imputation à info>0,8). Il est possible de calculer le nombre de signaux prévisionnels satisfaisant de tels critères et de voir si le nombre de signaux observés est supérieur.

Cette deuxième approche a l'avantage de prendre en compte le fait qu'un signal biologique

significatif risque de ne pas être fort et qu'il est donc intéressant de rechercher plutôt sa réplication éventuelle [157].

III.4 Résultats obtenus

Après méta-analyse, un seul SNP passe le seuil de significativité fixé à 5.10^{-8} et il est dans la région HLA dans le gène MICA (figure 31). Cette région génétique a été plusieurs fois décrite comme étant associée au contrôle de la charge virale et à la non progression [61, 118, 121, 148, 158]. Cela laisse à penser qu'on pourrait observer un biais dû à une trop forte proportion de patients contrôleurs du virus dans les cohortes composant IHAC et cela est suggéré par le fait qu'on retrouve le signal très fortement dans les deux cohortes américaines composées de sujets contrôleurs du virus.

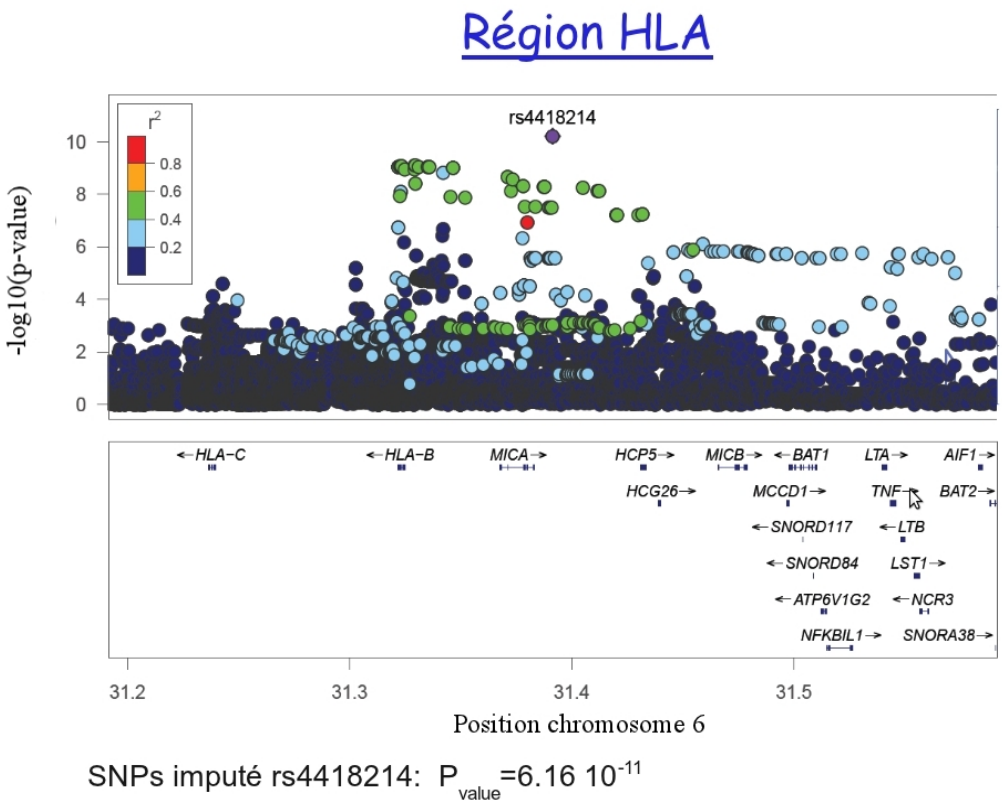


Figure 31 : Manhattan-plot de la région HLA dans la méta-analyse du projet IHAC et mise en avant des blocs de déséquilibre de liaison dans cette région significativement associée avec l'acquisition du VIH-1. En abscisse la position sur le chromosome 6 des SNPs et en ordonnée $-\log10(p_{combinée})$

Méta-analyse des données IHAC

Pour la deuxième approche, nous avons recherché pour chaque SNP combien de fois on le retrouvait avec une p-value inférieure à 0,05 dans chacune des 6 cohortes (ou à défaut dans au moins 5), ceci avec le même sens de l'effet. Cette approche n'est pas conventionnelle mais il s'agit plutôt d'une méthode exploratoire. Un renfort statistique pourrait cependant arriver dans un futur proche:

• Grâce à l'analyse des autres cohortes présentes dans IHAC (Afro Américaines et Africaines) dans le cas où les régions seraient retrouvées significatives. L'espoir est cependant faible tant la structure du génome chez les africains est différente (beaucoup plus de diversité et moins de déséquilibre de liaison) rendant les analyses « génome entier » difficiles [159].

• Grâce à l'évaluation par permutation du nombre de fois où l'on retrouve au moins 5 fois une association dans les 6 cohortes suivant l'hypothèse nulle. En théorie:

$$P(X \geq k) = C_n^k p_1 p_2^{k-1} \simeq 2{,}9.10^{-7} \; avec \; p_1 = 0{,}05 \;\; p_2 = 0.025 \; n = 6 \; et \; k = 5$$

En ayant 6,5 millions de SNPs (en réalité plutôt 8 à 10 fois moins testés réellement du fait des forts blocs de déséquilibre de liaison), nous obtenons 3 signaux indépendants passant ce seuil. A ce stade, nous avons déjà cherché à identifier une pertinence biologique de ces signaux.

rsid	chromosome	position	P_value	MAF	
rs76514342	1	245727311	3.71E-06	0.04503	**KIF26B**
rs55892969	11	245729902	1.29E-05	0.0440356	
rs4823317	22	46043466	2.21E-07	0.460209	

rsid	dutch	french	ill1	ill2	usA1	usA2
rs76514342	0.015945	0.038628	0.0055815	0.62224	0.019611	0.046509
rs55892969	0.022441	0.049808	0.015106	0.7384	0.03079	0.027622
rs4823317	0.020707	0.029809	0.045634	0.0086757	0.001227	0.44453

Tableau : Nous avons ici la liste des 3 SNPs répondant aux critères de réplication dans la deuxième approche avec la $p_{combinée}$ et la p-value individuelle dans chacune des 6 cohortes (Dutch, French, Illumina 1, Illumina 2, USA 1 et USA 2). En rouge le gène « biologiquement » intéressant KIF26B.

Nous avons particulièrement remarqué une région sur les 3, située sur le chromosome 1, dans un gène KIF26B, les deux autres signaux étant situés dans des régions intergéniques. Ce gène est essentiel au développement du rein car la protéine produite régule l'adhésion des cellules mésenchymateuses aux cellules adjacentes des bourgeons urétéraux, possiblement via une interaction avec MYH10 [160]. Il est intéressant de noter que MYH10 code pour une chaîne lourde de la myosine et il a été démontré que les chaînes lourdes de la myosine pouvaient être clivées par

une protéase du VIH-1 [161]. La protéine KIF26B est exprimée dans les tissus embryonnaires, mais aussi dans les muqueuses, ce qui pourrait suggérer un rôle face à la pénétration du virus dans l'organisme.

Discussion

Les trois chapitres de mes résultats reflètent bien l'évolution très rapide de l'analyse génétique des cohortes. Il y a cinq ans, j'étudiais encore quelques dizaines de marqueurs sur quelques gènes candidats avec des outils rudimentaires de bioinformatique et avec à disposition de faibles bases de données. Petit à petit, avec l'arrivée des puces de génotypage, il fut possible d'obtenir des centaines de milliers de marqueurs sur des centaines d'individus. Un progrès immense qui arrivait avec son lot de défis à relever.

Le premier défi est statistique avec la prise en main de notions diverses notamment de comparaisons de distribution. Par ailleurs, d'un point de vue bioinformatique, il a fallu développer des outils nécessaires aux analyses d'association, comprendre des concepts importants comme l'haplotypage ou explorer la structure du génome humain. L'essor de la technologie a permis l'émergence de bases de données publiques regroupant des dizaines puis des centaines d'individus à travers le projet Hapmap [14, 15] puis le projet 1000 génomes [1] permettant de comparer encore plus de marqueurs grâce à l'imputation et permettant les méta-analyses [162, 163], augmentant encore la puissance de découverte de nouveaux signaux associés à des maladies [66].

En seulement quelques années des centaines d'études et des milliers de SNPs se sont retrouvés associés aux maladies grâce à ces nouvelles techniques de biologie moléculaire à haut-débit représentées par les puces de génotypage. Nous sommes au cœur d'une révolution sans précédent dans la compréhension des facteurs génétiques pouvant influencer les maladies. Rares sont les maladies où les facteurs de risque n'ont pas été recherchés, du diabète au SIDA en passant par la schizophrénie, les cancers ou les maladies cardiovasculaires... Il suffit de feuilleter le catalogue des GWAS pour se perdre dans les résultats (http://www.genome.gov/gwastudies/).

Malgré cela, la stratégie « génome entier » a atteint ses limites. En effet, celle-ci est basée sur le dogme « variants communs pour maladies communes » qui n'a permis d'expliquer qu'une fraction de la variabilité génétique des patients face aux pathologies [164]. Il reste encore de nombreuses pistes à explorer comme l'influence des variants rares, des marqueurs autres que les SNPs ou l'épigénétique. C'est encore un domaine qui évolue très rapidement au gré des technologies toujours plus performantes.

I. Subhap: un précurseur

Pour se replacer dans le contexte de l'époque, nous avions affaire à des données génomiques sur des jeux de données restreints à quelques dizaines de SNPs centrés sur une région ou un gène. Les logiciels d'haplotypage étaient déjà relativement fiables, même si depuis, des efforts sensibles d'efficacité ont été entrepris notamment au travers du logiciel SHAPE-IT développé au CNAM.

Cependant, il manquait de la pratique sur des données existantes pour évaluer l'impact des différentes stratégies sur la qualité de l'inférence des haplotypes. SUBHAP partait de l'idée que plus le jeu de données était grand, plus on avait d'informations (notamment de déséquilibre de liaison) et plus l'inférence serait exacte. Pour démontrer ceci, nous avions artificiellement inséré des données manquantes à divers taux et observé que l'impact d'une approche « globale » englobant toute l'information disponible sur la région apportait un bénéfice substantiel dans la qualité de l'haplotypage.

En effet, le génome est structuré en blocs de déséquilibre de liaison et les SNPs sont pour la plupart corrélés les uns avec les autres. La totalité de l'information nécessaire à reconstruire un haplotype ne se limite donc pas aux seuls SNPs qui le composent. Cela peut paraître évident aujourd'hui mais cela ne l'était pas encore réellement à l'époque.

La première application de cette idée se retrouve dans le logiciel SHAPE-IT capable d'inférer l'haplotype sur un chromosome entier sans procéder à un découpage arbitraire préalable, améliorant sensiblement la technique.

Aujourd'hui, la méthode SUBHAP apparaît comme un précurseur de l'imputation à grande échelle du génome. En effet, la disponibilité des haplotypes de Hapmap puis de 1000genomes sert de base à l'inférence des données manquantes, y compris les SNPs non présents dans le jeu de données initial. Plusieurs logiciels dérivés des logiciels d'haplotypage [54] sont spécifiquement dédiés à cette tâche d'imputation (MACH, BEAGLE, IMPUTE). SUBHAP apparaît donc comme un précurseur de l'idée suivante, à savoir que toute l'information disponible doit être utilisée pour compléter les données manquantes.

En ajoutant le gain de puissance sur une cohorte unique à la multiplication des méta-analyses, l'imputation est devenue une étape très utilisée en analyse d'association confirmant la pertinence de l'approche que nous avons proposée.

II. Bilan des études d'association «génome entier» sur la cohorte GRIV

L'étude génome entier que j'ai réalisé sur le phénotype non-progresseur non « elite » fait suite à deux autres études réalisées sur les phénotypes non-progresseur et progresseur rapide dans cette même cohorte. Pour mémoire, ces deux études sont résumées dans les deux paragraphes suivants:

II.1 Étude de la non progression à long terme

Nous avons réalisé une étude d'association « génome entier » à l'aide de puces Illumina HumanHap300 basée sur la comparaison de 275 non-progresseurs à long terme avec 1352 contrôles séronégatifs [148]. La seule association passant le seuil de significativité statistique après correction de Bonferroni concerne le polymorphisme **rs2395029** du gène *HCP5*, localisé dans la région HLA du chromosome 6 ($p=6,79x10^{-10}$).

Le gène *HCP5* codant pour un rétrovirus endogène humain (HERV) présentant des homologies de séquence avec les gènes pol, il pourrait agir en tant qu'ARN antisens interférant avec la réplication virale. Cependant, rs2395029 se situe dans une région génétique complexe caractérisée par de très nombreux déséquilibres de liaison et il existe de nombreux allèles causaux candidats (SNPs et haplotypes), impliquant des gènes majeurs de l'immunité (*HLA-B*57, MICB, BAT1, LTB, TNF*), pour expliquer cette association. L'allèle *HLA-B*57* est associé au contrôle de la réplication du VIH-1 et de la progression vers le SIDA. *MICB* est un ligand pour les cellules T $CD8^+$ et NK, cellules clés pour la réponse anti-VIH. *BAT1* est un composant essentiel de la machinerie d'épissage et d'export de l'ARN et est également un régulateur négatif de cytokines pro-inflammatoires. *LTB* est un modulateur essentiel de l'inflammation. Enfin, le *TNF* est une cytokine pro-inflammatoire, ayant été largement explorée dans l'étude de l'infection VIH. D'un point de vue biologique, ces gènes sont donc tous de bons candidats pour la pathogenèse du VIH, et d'autres études génétiques de la région (séquençage, haplotypage...) et des explorations fonctionnelles seront certainement nécessaires pour permettre d'améliorer la compréhension de cette association.

Cette étude constitue ainsi la première étude « génome entier » sur des patients VIH-1 non-progresseurs à long terme et souligne le rôle majeur des gènes de la région HLA dans ce phénotype extrême.

II.2 Étude de la progression rapide

A l'aide de puces Illumina HumanHap300 nous avons comparé 85 progresseurs rapides avec 1352 contrôles séronégatifs [149].

Du fait de l'effectif réduit de notre population de progresseurs rapides, aucun signal n'a atteint le seuil de significativité statistique après correction de Bonferroni. Nous avons cependant réalisé une approche basée sur la méthode du False Discovery Rate (FDR) plus puissante. En prenant un seuil acceptable de 25% (un quart des signaux déclarés associés étant des faux positifs), nous trouvons quatre gènes associés.

L'association la plus forte est liée au gène *PRMT6* (**rs4118325**, $p=6,09x10^{-7}$, OR=0,24) mais aussi le gène *SOX5* (**rs1522232**, $p=4,29x10^{-6}$, OR=0,45), *RXRG* (**rs10800098**, $p=4,29x10^{-6}$, OR=3,29) et *TGFBRAP1* (**rs1020064**, $p=4,29x10^{-6}$, OR=0,34). Enfin, deux autres signaux non situés à proximité d'un gène (±100kb) ont été significativement associés à la progression rapide.

SOX5 est un facteur de régulation impliqué dans la voie de signalisation du *TGFβ* lors de la chondrogenèse, et aucune donnée ne permet de relier aisément cette protéine à la pathogenèse du VIH-1. *RXRG,* codant pour un récepteur nucléaire de l'acide rétinoïque, est impliqué dans la répression de la transcription du VIH-1. *PRMT6* est une arginine N-méthyltransférase pouvant méthyler les protéines Tat et Rev du VIH-1, et la molécule HMGA1, protéine non-histone notamment impliquée dans la régulation de la transcription et dans l'intégration des rétrovirus dans le génome hôte. Ces modifications altèrent les fonctions des protéines virales et pourraient altérer l'intégration du provirus dans le génome. *TGFBRAP1* est impliquée dans la voie de signalisation du *TGFβ*, cytokine immunosuppressive pléiotropique.

Ces signaux de progression rapide mettent l'accent sur l'importance du contrôle de la réplication virale et sur la voie de signalisation du *TGFβ* et offrent de nouvelles perspectives pour la compréhension des mécanismes de pathogenèse du VIH-1. La différence complète de résultats avec l'étude sur la non progression s'explique par une différence importante dans le phénotype observé. De plus, il tend à nous révéler qu'un signal favorisant la non progression n'aurait pas d'effet sur la progression rapide (et inversement).

II.3 Travaux sur les non-progresseurs à long terme non « elite »

Lors de l'étude sur la non progression à long terme de Limou et al. [148], nous avions montré que les sujets porteurs du variant rs2395029-C du gène *HCP5* avait une charge virale significativement plus faible que les autres. De plus, Fellay et al. [121] avaient aussi identifié ce variant comme associé au contrôle de la charge virale. sachant que la plupart des non-progresseurs avaient une charge virale non négligeable, nous nous sommes donc intéressés aux facteurs génétiques influençant la non progression à long terme sans nécessairement contrôler la charge virale.

Pour cela nous avons repris nos données de puces Illumina HumanHap300 et basé notre étude « génome entier » sur la comparaison de 186 non-progresseurs à long terme non « elite » (*i.e.* avec une charge virale plasmatique >100 copies/mL) avec 697 contrôles séronégatifs. La plus forte association a été obtenue pour le SNP **rs2234358** ($p=2,5x10^{-7}$, OR=1,85). Après avoir contrôlé que ce résultat était indépendant à la fois des haplotypes *CCR2-CCR5* (Δ32, P1, et 64I) et à la fois du locus ***HCP5/HLA-B*57***, nous avons pu le répliquer dans plusieurs autres cohortes également d'origine européenne et évaluant un phénotype de progression vers le SIDA ($p_{combinée}=9,7x10^{-10}$).

Ce signal représente donc une nouvelle association avec la progression à long terme, indépendante des résultats précédemment connus des loci *CCR2-CCR5* et *HCP5/HLA-B*57* et a montré l'importance critique de la composition des cohortes et la puissance des phénotypes extrêmes pour la découverte de nouveaux signaux. Ce SNP est localisé sur deux gènes: *CXCR6* et *FYCO1*. Aucun SNP n'étant en fort déséquilibre de liaison au-delà du locus *CXCR6/FYCO1*, nous pouvons émettre l'hypothèse que ce locus serait impliqué directement dans la pathogenèse de l'infection VIH-1. L'étude du profil haplotypique du gène a révélé plusieurs haplotypes du promoteur en fort déséquilibre de liaison avec rs2234358. L'exploration des bases de données bioinformatiques (expression des ARNm, sites d'épissage, de polyadénylation, de liaison de facteurs de transcription) a suggéré d'éventuels sites de fixation de facteurs de transcription au niveau de certains SNPs composant les haplotypes du promoteur en déséquilibre de liaison avec rs2234358. Des expérimentations approfondies sont nécessaires pour déterminer précisément le variant causal [165] et le mécanisme biologique en jeu.

Le travail que nous avons réalisé présente la première association répliquée en dehors de la région HLA obtenue par une approche « génome entier ». Le risque attribuable de cette association rs2234358 est très élevé puisqu'elle explique 12% de la LTNP. Par comparaison, le risque

attribuable de CCR5-Δ32 est de 5,1% et celui de *HCP5* est de 15% dans la cohorte GRIV de LTNP. Un autre variant génétique de *CXCR6*, présent dans la population afro-américaine et absent de la population européenne, a été précédemment associé à la progression de la pneumonie à *Pneumocystis carinii* intervenant dans le SIDA [166] suggérant le rôle potentiel de *CXCR6* dans la pathogenèse du VIH-1. Compte-tenu du rôle documenté de *CXCR6* dans le SIDA, nous avons considéré que le signal avait plus de chances d'être lié à *CXCR6* qu'à *FYCO1*.

L'implication de *CXCR6* dans la progression à long terme n'est pas vraiment une surprise. En effet ce gène est un récepteur de chimiokines connu comme étant un co-récepteur majeur du SIV, et mineur dans le cadre du VIH-1 [167]. Il est également connu pour être un médiateur de l'inflammation, pour être impliqué dans la régulation des cellules T dans l'inflammation [168] et dans l'activation de l'homéostasie des lymphocytes T « Natural Killer » [169].

Cette étude montre aussi qu'un effet faible ou modéré, lorsqu'il est répliqué dans plusieurs cohortes, peut aboutir à l'identification d'une association significative avec la maladie. La puissance des méta-analyses ainsi révélée permet d'insuffler un grand espoir d'identifier de nouveaux signaux associés au SIDA en combinant de nombreuses cohortes, ce qui a été entrepris dans le projet IHAC dont je rediscute un peu plus loin.

III. Comparaison des associations obtenues avec les autres études génétiques sur le SIDA

III.1 Comparaison avec les approches « gènes candidats »

Les approches « gènes candidats » ont apporté une grande quantité d'informations sur les gènes de l'hôte pouvant jouer un rôle dans l'infection par le VIH-1 (voir Introduction). Ces méthodes ont permis de confirmer ou d'infirmer par réplication sur différentes cohortes l'implication des gènes étudiés, permettant ainsi d'éclaircir certains des mécanismes de pathogenèse. Lorsque l'on compare les résultats obtenus par l'approche « gène candidat » à ceux des études d'association « génome entier », on retrouve essentiellement la région HLA en commun et particulièrement l'allèle HLA-B*57. Pour ces deux types d'approches, la complexité de cette région due au déséquilibre de liaison, rend encore difficile, à ce jour, la discrimination du ou des variants causaux.

Comparaison des associations obtenues avec les autres études génétiques sur le SIDA

L'approche « génome entier » sur les progresseurs rapides de la cohorte GRIV a mis en lumière des signaux dans des loci peu soupçonnés jusqu'à présent dans la progression rapide avec les gènes PRMT6, SOX5, RXRG et TGFBRAP1 [149] ou dans la non-progression avec le gène RICH2 [150]. Ces résultats soulignent l'intérêt des études « génome entier » dans la découverte de nouveaux signaux associés à la progression vers le SIDA en s'affranchissant des a priori biologiques.

Le cas de *CXCR6* est un cas particulier car ce gène avait été séquencé en « gène candidat » par notre équipe en 2004. Nous nous étions focalisés, à l'époque, sur les régions exoniques, et nous n'avions pas trouvé d'association aussi forte que celle trouvée dans l'étude génome entier. Sans l'étude « génome entier », cela aurait pu conduire à une conclusion erronée sur l'absence de signal associé à la maladie dans GRIV.

Il faut également souligner que des signaux comme le CCR5-Δ32, retrouvés par la plupart des équipes, n'ont pas été vus par l'approche « génome entier ». Cela s'explique simplement par le fait que le variant CCR5-Δ32 n'est pas représenté sur les puces. Ceci présente un point faible des puces car les insertions/délétions génotypées sur les gènes candidats sont souvent responsables de maladies ou à défaut des facteurs de risque [105, 106, 115, 170, 171].

III.2 Comparaison avec les autres études « génome entier » publiées

Cette première vague d'études « génome entier » - menées dans le cadre du SIDA - a mis en lumière l'importance des régions du HLA du chromosome 6, et des récepteurs de chimiokines du chromosome 3 (locus CCR2-CCR5 et gène *CXCR6*) dans l'évolution différentielle de l'infection par le VIH-1. Ces deux régions génétiques sont les seules, à ce jour, présentant des associations répliquées ayant atteint le seuil de significativité « génome entier ». A noter que les signaux rs2395029 et rs9264942 ont également été confirmés dans des études SNP candidats sur des cohortes indépendantes [133, 172, 173]. Il est important de souligner que plusieurs signaux ne passant pas les seuils statistiques classiques, représentent malgré tout de bons candidats pouvant intervenir dans la pathogenèse du SIDA.

Toutes les cohortes étudiées dans ces études « génome entier » ont été collectées selon des critères différents (charge virale, cellules T $CD4^+$, trithérapie, séroconvertis, séroprévalents, profils

extrêmes, hommes/femmes, mode d'infection homosexuel, consommateur de drogues en intraveineuse...), ciblent des phénotypes différents (qualitatifs avec les phénotypes extrêmes : LTNP, PR, « elite controllers » et quantitatifs : charge virale, cellules T $CD4^+$, temps jusqu'au développement du SIDA...), sont de tailles diverses (45 à 2554 individus), concernent des populations d'origines variées (européenne : Angleterre, Australie, France, USA... ; afro-américaine), et ont été génotypées sur différentes plate-formes (Illumina, Affymetrix, à l'aide de puces ciblant de 300K à 1M de SNPs).

Toutes ces différences ne représentent pas uniquement une faiblesse pour l'étude de l'infection par le VIH-1. En effet, toutes ces cohortes sont complémentaires. Elles permettent de cibler un spectre plus large de facteurs génétiques impliqués à différents moments, dans différents compartiments ou contextes de l'infection par le VIH-1. Il est donc important de continuer à étudier et comparer l'ensemble de ces cohortes, et de développer de nouvelles cohortes bien définies, ciblant de nouvelles populations (africaines, asiatiques...), de nouveaux phénotypes...

Malgré ces disparités, le signal *HCP5/HLA-B*57* a été identifié dans toutes les études pour lesquelles le SNP était génotypé (*i.e.* plates-formes Illumina), accentuant l'importance capitale de ce locus pour le contrôle de l'infection par le VIH-1.

De nombreux gènes découverts grâce au études « génome entier » se révèlent souvent être de potentiels gènes candidats, même si cela est parfois fait a posteriori, d'où l'idée de réduire le nombre de marqueurs à étudier (et diminuer la baisse de puissance qu'impliquent les tests multiples) ou l'enrichissement en p-values de certaines voies de signalisation (« pathways ») spécifiques [38, 144, 174, 175].

Enfin, l'analyse du signal de non-progression *CXCR6* présent dans la cohorte GRIV et répliqué avec de faibles p-values dans les cohortes ACS et MACS156 démontre la pertinence et la puissance de l'utilisation des phénotypes extrêmes. Cette même étude révèle l'importance des méta-analyses pour trouver de nouveaux signaux expliquant la variabilité phénotypique dans le SIDA. Ainsi, dans un but de gagner en puissance statistique, l'augmentation du nombre de patients apparaît comme étant l'étape naturelle.

IV. Le projet IHAC

Comme nous l'avons vu précédemment, les méta-analyses peuvent se révéler très utiles pour trouver de nouveaux signaux expliquant la variabilité phénotypique dans le SIDA. Cette approche a par ailleurs été validée dans de nombreuses autres pathologies [22, 65, 163, 176, 177]. Dans ce contexte, le projet international IHAC (International HIV-1 Acquisition Consortium) suscite un espoir important dans la communauté génomique du SIDA. Ce projet est réalisable grâce à l'imputation des marqueurs qui permet la comparaison de plusieurs études sur différentes plates-formes. Il accroît dans certains cas la force d'un signal observé sur un SNP génotypé.

Pour pouvoir mener à bien ce projet, il a fallu mettre en place une infrastructure informatique sans précédent capable d'imputer cette masse d'information correspondant à plus de 13000 patients avec un panel de référence de 1000 individus sur 37 millions de SNPs disponibles issu de la base de données 1000genomes. Le premier problème à surmonter fut le temps de calcul, par la méthode classique et via les logiciels d'imputation existants, il aurait fallu plusieurs mois voire plusieurs années pour arriver à nos fins. Il a donc fallu utiliser les dernières recherches dans le domaine qui suggéraient de passer plutôt par une étape de pré-haplotypage des données permettant en quelques semaines d'obtenir les résultats que nous voulions.

La première déception fut de n'obtenir qu'un seul signal passant les seuils de significativité (de Bonferroni), représenté par un SNP situé dans la région HLA non loin du rs2395029 retrouvé dans presque toutes les études « génome entier » sur des cohortes européennes. Bien que le phénotype étudié soit l'acquisition du virus VIH-1, on peut observer que la sur-représentation de patients non-progresseurs dans plusieurs cohortes est la cause de cette association apparente, et donc que ce signal ne n'est a priori qu'un faux positif dû à ce biais.

Il y a donc lieu de chercher une autre heuristique et nous avons pensé à une deuxième approche complémentaire consistant à chercher les signaux répliqués, même faiblement, dans au moins cinq cohortes. Sur les trois découverts, deux sont intergéniques mais l'un d'eux est dans un gêne KIF26B qui peut être un candidat intéressant puisqu'il pourrait intervenir dans la pénétration du virus dans les muqueuses.

Néanmoins au bilan global, les résultats sont assez faibles. L'espoir repose désormais sur l'arrivée de nouveaux panels de référence regroupant à la fois des SNPs et des insertions/délétions (indels) au sein des patients de 1000genomes. De nombreux indels étant associés à des maladies,

comme celui déjà bien connu de CCR5-Δ32 dans le SIDA, on peut penser qu'ils nous aideront a déterminer des variants réellement causaux et au mieux nous révélerons un nouveau signal. Une autre stratégie envisageable est de tester l'effet de plusieurs marqueurs à travers l'épistasie, les voies de signalisation ou les effets haplotypiques.

Une autre perspective que constitue la mise en commun prochaine des données cliniques pourrait être plus prometteuse. L'étude d'association sur la progression clinique et sur la charge virale pourrait en effet être plus informative car elle caractérise mieux la relation hôte/virus. Le choix des cohortes et des phénotypes étudiés sera critique et plusieurs possibilités devront être envisagées :

- profil clinique extrême comme les LTNP, les progresseurs rapides, les non-progresseurs non « elite », les non-progresseurs « elite »
- l'analyse de données de survie sur le temps avant SIDA selon CDC1993 ou chute de CD4<400, charge virale, etc.

V. Critiques des analyses « génome entier »

Le succès des analyses d'association « génome entier » dans l'identification des facteurs de risques associés aux maladies a été massif : depuis cinq ans, il ne se passe pratiquement pas une semaine sans qu'une nouvelle publication vienne s'ajouter à la liste des marqueurs liés à la susceptibilité face aux maladies. Pour la première fois, dans l'histoire, nous avons le pouvoir d'identifier précisément les variations génétiques entre les individus contribuant aux variations phénotypiques liées aux maladies ouvrant l'ère de la médecine personnalisée.

Cependant, en dépit des succès indéniables qui ont coûté des centaines de millions de dollars, certaines failles apparaissent aussi [178]. En effet, beaucoup de variants génétiques associés aux maladies et notamment au SIDA restent inconnus. Au delà de l'approche « classique » d'analyse il convient d'aller plus loin [179] et de trouver de nouvelles technologies.

V.1 Indels

Ce type de variants a été peu exploré à grande échelle alors qu'ils ont le potentiel biologique de porter une part non négligeable de la variabilité génétique. Ils sont certainement fondamentaux dans la recherche d'association avec les maladies [180]. Les puces de génotypage ne permettent pas

de caractériser directement les indels même s'ils ont de bonnes chances d'être causaux dans de très nombreux cas (près de la moitié des SNPs du GWAS catalogue sont en déséquilibre de liaison total avec un indel). Cependant, les indels ont été séquencés sur les patients du projet 1000genomes et devraient pouvoir être partiellement imputés à partir des données de puces dans un avenir proche.

Le séquençage des indels constitue donc une approche complémentaire aux analyses de SNPs.

V.2 Allèles avec des effets faibles

L'analyse de centaines de milliers de variants est à la fois la force et la faiblesse des études «génome entier». La masse de résultats impose des corrections statistiques pour les tests multiples qui noient les vrais positifs dans la masse des exclus.

La solution repose sur l'augmentation du nombre d'individus dans les cohortes. Dans le SIDA la relative faible taille des cohortes empêche d'observer des effets très faibles (OR<1,5) alors que dans d'autres maladies, comme le diabète, les cliniciens ont pu constituer des cohortes de plusieurs dizaines de milliers d'individus et observer des effets impliquant des OR aussi petits que 1,2.

Multiplier autant que possible les méta-analyses et la mise en commun des patients devient plus que jamais indispensable pour aller plus loin.

V.3 Variants rares

Le dogme « **maladie commune, variant commun** » qui est la base des analyses « génome entier » par puces de génotypage part du principe que la majorité des facteurs de risques associés aux maladies sont attribuables à un nombre relativement faible de variants communs.

Au départ les informations disponibles par exemple dans Hapmap, reposaient sur les marqueurs communs plus faciles à identifier que les plus rares. Aussi, dans le but de capter la plus grande proportion de la variation génétique, les concepteurs de puces se sont basés sur ces marqueurs. Malheureusement on s'est aperçu qu'une part importante de la variabilité génétique ne pouvait être expliquée par les études « génome entier » dans le SIDA. Néanmoins Le Clerc et al. [150], avec une approche pertinente se focalisant sur les SNPs de faibles fréquences [181] (MAF <5%), a mis en avant un nouveau gène potentiellement associé à la maladie.

Critiques des analyses « génome entier »

On peut penser qu'augmenter la taille des cohortes pourrait être une bonne solution, d'autant qu'avec le projet 1000genomes, l'information des variants plus rares apparaît. Mais le problème est que les SNPs communs des puces de génotypage ne représentent pas toujours bien les variants rares et l'imputation peut être difficile. Il conviendrait donc de faire de nouvelles puces qui se baseraient non plus sur les SNPs commun de Hapmap mais aussi sur les SNPs plus rares de 1000genomes ou plus radicalement de faire du séquençage du génome entier pour obtenir un catalogue complet de toutes les variations existantes dans les cas et les contrôles [182]. La vitesse importante de baisse des coûts de ces technologies rendent ce scénario de plus en plus faisable. Néanmoins, l'interprétation s'avère compliquée et oblige le statisticien à prendre comme hypothèse que les **variants rares sont fortement pénétrants voire mendéliens** (si on possède l'allèle on présente forcement le phénotype).

Une partie de l'explication peut résider dans ces variants rares incluant des CNVs, des indels et des SNPs rares (par convention on appelle rare un variant dont la MAF dans la population générale est inférieure à 1%), ils sont aujourd'hui fortement suspectés de jouer un rôle important dans les maladies multi-factorielles. Les méthodes d'analyse actuelles se focalisent sur l'influence collective de plusieurs marqueurs rares dans une même région [183]. En effet il est très difficile de lier la maladie à un variant très rare seul du fait de la faible puissance des tests statistiques. On se focalise également sur des variants susceptibles d'avoir un rôle biologique important en séquençant plus particulièrement l'exome des patients. En effet, une mutation dans une protéine aura vraisemblablement un effet fort et un impact important sur la maladie. Les premières études basées sur le séquençage de l'exome commencent à paraître, preuve de l'importance de cette approche. On peut citer la schizophrénie avec la découverte d'une mutation *de novo* rare présente chez les malades [170], une mutation du gène ZNF644 chez les grands myopes [184], une mutation du gène BCOR dans la leucémie myéloïde [185] ou encore une mutation du gène MYH6 dans les maladies des sinus [186].

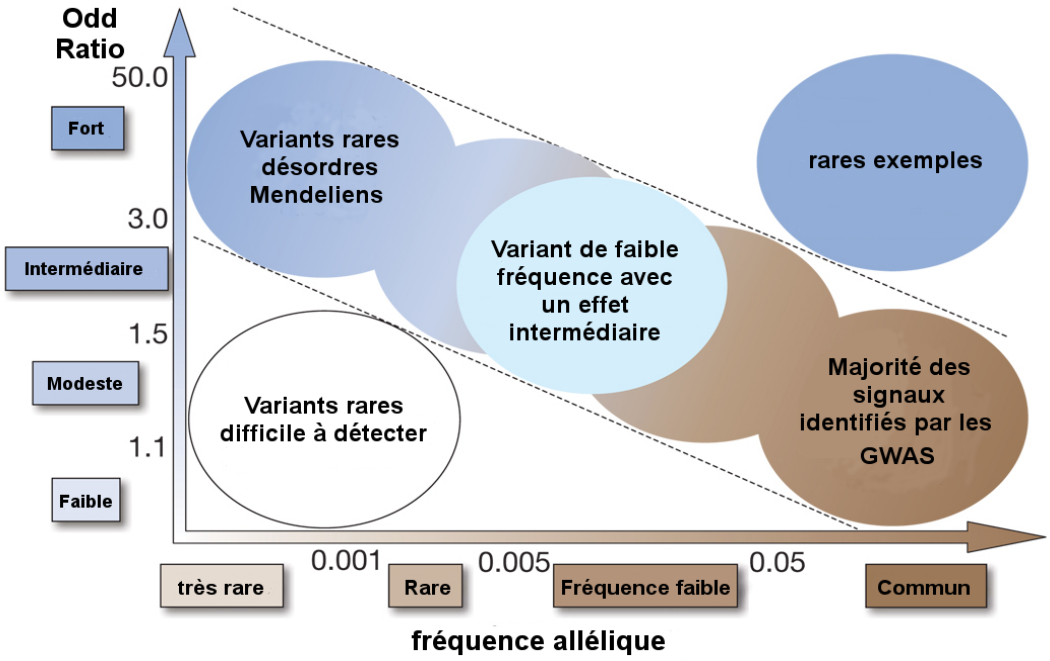


Figure 32 : Classification des marqueurs génétiques associés aux maladies en fonction de leur fréquence et de leur pénétrance dans la maladie

V.4 Épistasies et approche multi-marqueurs

La plupart du temps, on approche les maladies par une analyse simples marqueurs, chaque facteur de risque agissant indépendamment des autres. Ainsi la présence de deux facteurs de risque chez un individu additionnerait leurs effets [187, 188]. Or on sait bien, qu'en réalité, l'interaction des gènes joue un rôle essentielle en génétique humaine [189], dans ce cas la présence de deux allèles combinés aurait un effet supérieur. En allant plus loin, la combinaison de plusieurs marqueurs avec un faible effet peut aboutir à un grand effet combiné. Cette approche est souvent ignorée dans les analyses d'association à cause des restrictions statistiques des tests multiples [190].

Le développement de méthodes statistiques/bioinformatiques intégrant plusieurs marqueurs à la fois constitue sans aucun doute une voie de recherche prioritaire, mais à ce jour aucune méthode de ce type ne fait encore consensus.

Il en va de même avec l'analyse des haplotypes qui malgré leur importance ne sont que rarement étudiés dans les études « génome entier », il ne faut pas oublier que nous sommes des organismes diploïdes [56].

V.5 CNVs

Les avancées dans la connaissance du génome ont mis en lumière l'importance des autres marqueurs de variabilité comme les CNVs (Copy Number Variation) qui sont des grandes insertions ou délétions. Ces marqueurs se retrouvent dans toutes les populations sans affecter nécessairement la santé des patients. De par les publications sorties sur le sujet, il semble probable qu'une part non négligeable de la variabilité génétique face aux maladies soit contenue dans ces CNVs .

Théoriquement il est possible de retrouver ces CNVs à l'aide des puces de génotypage en observant l'intensité de fluorescence des marqueurs [191]. La condition est que la densité de SNPs soit suffisamment importante, ce qui en pratique ne s'est pas révélé être le cas avec les puces Illumina 300K utilisées dans la cohorte GRIV. En effet, il est documenté [192] que cette couverture est largement insuffisante pour détecter avec précision ces marqueurs.

Le développement des puces de génotypage et le séquençage capable de repérer spécifiquement ces CNVs constituent l'approche d'avenir dans ce domaine.

V.6 Épigénétique

L'information contenue dans l'ADN ne se limite pas à sa séquence. En effet, des modifications de méthylation des cytosines ou des histones contenues dans la molécule d'ADN modulent l'expression des gènes. Ces caractéristiques épigénétiques sont héritables de ses parents et peuvent apparaître spontanément en réponse à l'environnement. On ne sait pas encore formellement comment ces variations « épigénétiques » influencent la susceptibilité aux maladies.

Toutes les technologies utilisées en génome entier (y compris le séquençage) ne peuvent détecter ce type de variations. Des techniques sont développées pour détecter à grande échelle ces variants épigénétiques grâce à des technologies dédiées, il existe même des **analyses « épigénome entier »** [193] (EWAS).

V.7 Transcriptome, protéomes, métabolome

Plus généralement, il existe plusieurs domaines périphériques et descendant à l'analyse du génome qui peuvent constituer des pistes complémentaires non négligeables dans la compréhension des pathologies. On peut citer :

Critiques des analyses « génome entier »

- Le transcriptome représentant l'ensemble des ARNs messagers issus de l'expression d'une partie du génome dans une cellule ou un tissu donné. Cela est possible grâce à des puces à ADN.
- Son analogue le protéome qui représente l'ensemble des protéines présentes dans la cellule.
- Plus récemment le métabolome qui étudie la présence des petites molécules dite métabolites telles que les hormones qui peuvent être trouvées dans un échantillon biologique.

Ces systèmes contrairement au génome ont la particularité d'être dynamiques et de changer au cours du temps.

V.8 Hétérogénéité des maladies

Les maladies sont définies principalement par des catalogues de symptômes qui peuvent provenir de multiples causes génétiques distinctes. Si, à l'extrême, chaque patient présente des symptômes pour une unique cause génétique distincte des autres, la stratégie d'analyse d'association sur cohorte s'effondre. Le statisticien et le généticien ne peuvent pallier à cette problématique. C'est aux **cliniciens** et aux **médecins** d'arriver à constituer des cohortes au phénotype homogène selon des critères bien définis. Dans le projet IHAC nous avons déjà discuté du choix fondamental des phénotypes pour l'analyse de la progression vers le SIDA.

VI. Evolution de la recherche en génétique

VI.1 Séquençage « nouvelle génération » (NGS)

Les avancées technologiques récentes ont permis le développement du séquençage haut débit, rendant ainsi possible le séquençage entier du génome humain ou de l'ensemble des exons (exome). Ces méthodes permettent d'obtenir les informations génotypiques d'un grand nombre de marqueurs, notamment les marqueurs rares avec des **MAF inférieures à 1%**. Aujourd'hui, il est possible de réaliser une analyse fine sur l'ensemble du génome ou de l'exome et sur un nombre assez élevé d'individus. De plus, ce type d'approche offre la possibilité de détecter, en plus des SNPs, d'autres polymorphismes comme les **CNVs** ou les **indels**, offrant ainsi une couverture plus complète de la diversité entre les individus et de nouvelles informations sur la structure du génome [194].

Evolution de la recherche en génétique

On considère qu'à partir de 20X, on a une bonne couverture du génome d'un individu et qu'à 50-100X on atteint une couverture optimale. Dans tous les cas, plus la couverture sera importante et plus on aura de chances de couvrir au moins une fois chaque région du génome. Ainsi on pourra corriger les erreurs des plates-formes et enfin il sera possible de découvrir les variants structuraux comme les insertions/délétions.

Il existe 4 aspects qui représentent un enjeu pour le développement et la réussite du séquençage haut-débit: bioinformatique, informatique, statistique et financier:

- La **bioinformatique** doit être capable d'identifier à partir des données de séquençage, les « reads » de taille variable (quelques dizaines à quelques centaines de paires de bases) et les variations entre les patients.
- Les ressources **informatiques** afin de gérer les données de séquençage « génome entier » sont énormes. Des améliorations informatiques en terme de gestion et de traitement des données (transfert, stockage, capacité de calcul pour l'alignement des séquences…) sont nécessaires pour améliorer la qualité des résultats.
- D'un point de vue **statistique** être capable d'identifier les marqueurs associés aux maladies à l'aide de nouvelles approches (tests multiples, variants rares...).
- Enfin le **coût** des technologies de séquençage reste très élevé même si cela change très rapidement.

Il existe aujourd'hui 3 technologies issues de 3 compagnies capables de séquencer la totalité du génome ainsi que l'exome humain : Illumina Solexa, Biosystems Solid et Roche 454.

Les techniques de séquençage sont différentes entre ces 3 méthodes. En terme de coût par mégabase, les plates-formes Illumina et Solid sont nettement supérieures. Roche 454 offre des « reads » beaucoup plus longs ce qui est un avantage ensuite pour l'alignement et la reconstruction du génome mais en contrepartie son coût est plus élevé [195].

Actuellement, la machine Illumina Hi Seq peut produire 20 millions de paires de bases pour 1500$, ce qui signifie que pour une couverture optimale, le séquençage de l'exome coûte environ 500$ tandis que le génome entier environ 1000-2000$. Il y a encore un an cela coûtait dix fois plus cher. On peut imaginer que bientôt cela reviendra au prix des puces de génotypage d'il y a seulement quelques années.

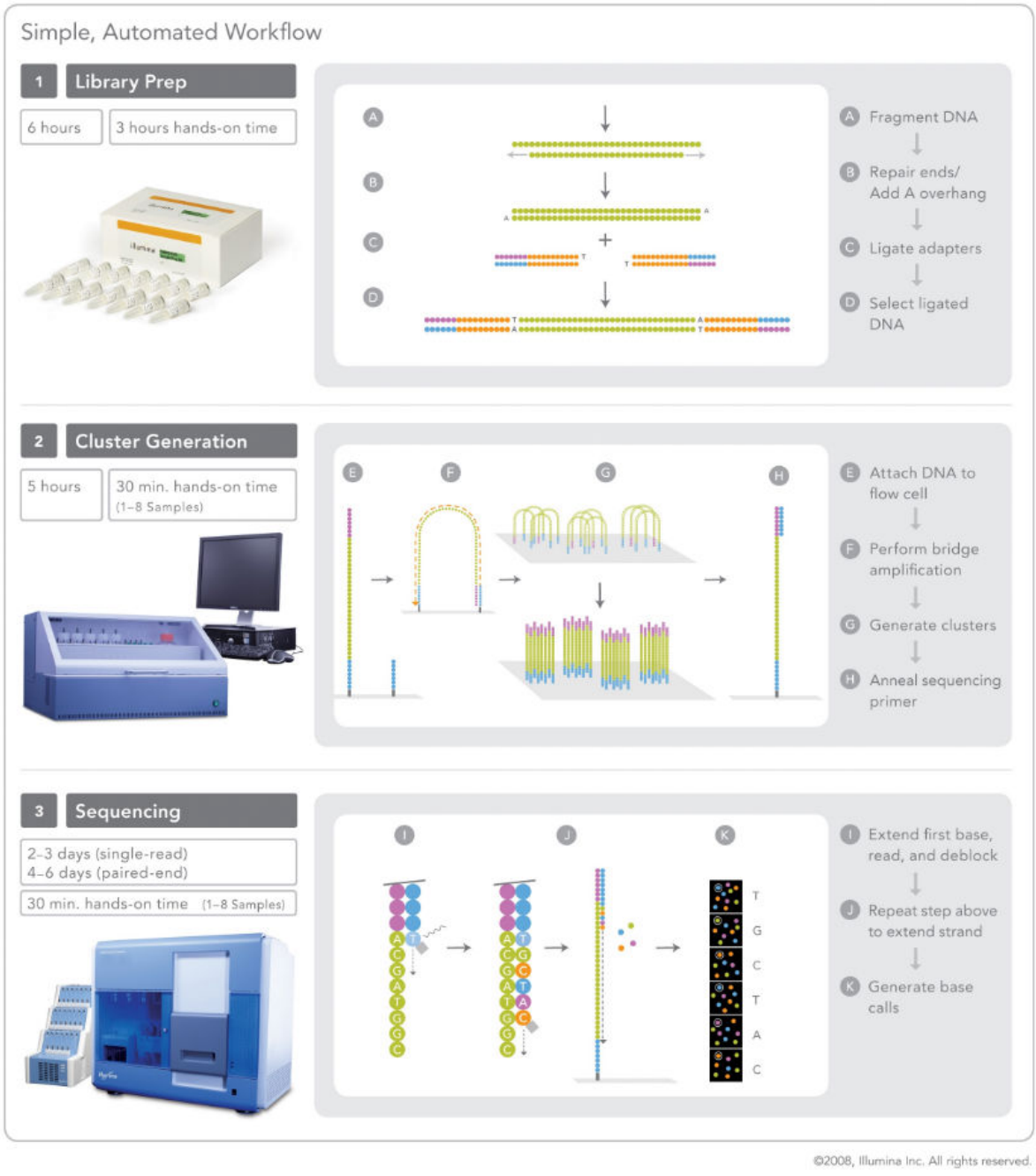


Figure 33 : Représentation schématique des différentes étapes de séquençage sur Illumina Solexo (extrait de Illumina's Genome Analyser Iix)

Le choix du séquençage de l'exome peut apparaître judicieux eu égard au moindre coût même si le prix du séquençage du génome entier diminue fortement. Cette approche présente également l'avantage de permettre un meilleur séquençage de chaque région ciblée avec une meilleure couverture [196, 197]. L'exome ne représente qu'une fraction faible du génome, se

focalise sur des régions transcrites dont les variations sont directement interprétables [198, 199]. Il permet d'étudier tous les variants fonctionnels, de plus en plus répertoriés dans les bases de données telles que dbNSFP [200] et a déjà fait ses preuves notamment dans les maladies mendéliennes [201-205]. Si l'hypothèse que des variants rares peuvent provoquer des désordres mendéliens dans toutes les maladies communes, cette stratégie peut s'avérer payante [206]. Néanmoins, il serait intéressant de développer une approche incluant d'autres zones fonctionnelles telles que les régions promotrices des gènes.

Les puces de génotypage 2,5M Illumina actuelles, même si elles ne tiennent pas encore compte de toutes les données du projet 1000genomes, peuvent apporter une solution moins coûteuse et moins fastidieuse que le séquençage, ainsi qu'une fiabilité des résultats éprouvée. Lorsque l'on regarde les SNPs présents dans les exons sur la puce 2,5M Illumina, ils ne sont qu'au nombre de 90 000. En fait l'approche par puce de génotypage cible les variants avec un effet relativement faible tandis que la stratégie de séquençage cible les mutations rares avec un effet fort, elles sont donc en réalité complémentaires. Preuve de cette complémentarité, le projet 1000genomes a utilisé cette double approche (séquençage et génotypage sur puces Illumina).

Enfin, une dernière approche plus récente, réduisant fortement les coûts, consiste à séquencer le génome entier avec une faible couverture (0,5X à 4X) tout en utilisant l'information de 1000genomes pour imputer les données manquantes [207]. Cette technique présente l'avantage de la rapidité et du coût sans mettre en péril la fiabilité des résultats. C'est d'ailleurs la stratégie adoptée par le consortium 1000genomes [1].

VI.2 Perspectives pour la génomique du SIDA

Dans la discussion des analyses « génome entier », nous avons évoqué plusieurs pistes possibles. Parmi celles-ci, le séquençage apparaît prioritaire permettant la caractérisation des variants rares à effet fort. En effet, outre le recrutement de nouveaux patients, la perspective du séquençage pourrait permettre d'aller plus loin dans la découverte de variants associés à la maladie.

L'étude de la cohorte GRIV a déjà confirmé l'intérêt des phénotypes extrêmes [119, 120, 208], y compris avec des populations de taille réduite, et la cohorte GRIV s'avère donc prometteuse dans ce type d'approche.

Déjà, le groupe Euro-CHAVI a entrepris le séquençage de patients à profils de progression extrême (non-progresseurs à long terme et progresseurs rapides) ce qui pourrait permettre la

détection de variants rares importants dans la compréhension des mécanismes de pathogenèse [209-211].

Les « elite controllers » du VIH présentent un phénotype très rare dans la population (moins de 1% des personnes infectées) avec des effets très forts dans la région HLA notamment sur le HLA-B*57 [148]. Une piste possible serait d'explorer par séquençage les « elite controllers » non HLA-B*57 qui pourraient constituer une population de choix pour ce type d'approche dans la même démarche que celle ayant conduit à la découverte du gène *CXCR6* [151].

Les progresseurs rapides de GRIV, ayant montré de forts odds-ratio en analyse « génome entier » pourraient eux aussi constituer une cohorte intéressante dans la mesure où ce phénotype, qui semble très informatif et homogène, n'est que très peu répliqué dans le monde.

Aujourd'hui, si l'on souhaite procéder à du séquençage pour identifier l'effet de variants très rares au niveau d'un gène, il semble que l'**analyse de l'exome** et de populations extrêmes soit le bon compromis sur le plan information/coût.

VI.3 Perspective de ces technologies : vers une carte d'identité génétique ?

A chaque étape de la découverte des variants associés aux maladies, on se rapproche de la compréhension des mécanismes, et ainsi une analyse individuelle des facteurs génétiques associés aux maladies devient possible. Des sociétés privées proposent déjà ces analyses moyennant quelques centaines de dollars comme 23andMe, deCODEme and Navigenics. Cependant la pertinence de laisser des sociétés privées dévoiler au grand public des particularités connues de son génome suscite de vifs débats [212]. En effet, les modifications qu'un individu apportera à son mode de vie en fonction de l'analyse de ses données génomiques risquent d'être contre-productives.

Excepté pour les maladies monogéniques, la faiblesse de la majorité des effets associés jusqu'ici envoie un mauvais message au patient. Quel est l'impact de savoir qu'on a 1,5 fois plus de chance que la population globale d'avoir telle ou telle pathologie ? Ou à l'inverse 2 fois moins de chance ? D'autant que l'impact environnemental sur les risques associés aux maladies est souvent bien supérieur. Si un malade sait qu'il a deux fois moins de chance d'avoir un cancer du poumon ne risque-t-il pas de continuer à fumer par exemple ?

Cependant, avec l'avancée des technologies notamment de séquençage et leur exploitation

bioinformatique, des marqueurs ayant un effet plus fort sur la susceptibilité sont attendus et une compréhension plus complète de la composante génétique des maladies devrait devenir réalité. Cela ouvre la voie à une carte génétique individuelle, **clé de traitements mieux ciblés** [213]. La diminution croissante du coût de l'analyse du génome individuel devrait permettre à terme de personnaliser les traitements [214] en tenant compte des réactions aux médicaments, au profil d'évolution supposé dans la maladie qui diffère selon les individus. Il est donc probable qu'à terme, avec le faible coût des analyses génétiques et l'amélioration des outils prédictifs, une carte d'identité génétique puisse être proposée systématiquement. Le législateur devra certainement mettre au point un cadre légal et pratique optimal pour le développement de tels outils.

Conclusion

Avec l'arrivée de plus en plus massive des données génomiques de la cohorte GRIV, d'abord des données SNPs sur gènes candidats puis sur puces de génotypage Illumina, il nous est devenu indispensable d'utiliser la bioinformatique. Comprendre les problématiques biologiques, analyser les besoins pour ensuite y répondre est devenu un enjeu fondamental dans la réussite de notre projet d'analyse génétique.

Les travaux exposés dans cette thèse correspondent à une approche moderne de la bioinformatique, les progrès des technologies m'obligeant à évoluer tout au long de ces cinq dernières années. Au cours de cette thèse, j'ai participé aux études d'association « gène candidat » puis à l'analyse du « génome entier » réalisées sur les patients de la cohorte GRIV et enfin à une méta-analyse internationale nécessitant une grande adaptation.

Pour cela, j'ai été amené tout au long de ma thèse à élaborer de nouveaux outils (notamment à travers le logiciel d'haplotypage SUBHAP), à utiliser les logiciels existants en bioinformatique et statistique et à mettre en place des infrastructures utiles à tous les acteurs du projet.

Les outils que j'ai mis en place ont ainsi permis de mettre en avant l'importance de la région HLA, riche en gènes de l'immunité, dans le contrôle de la charge virale et dans la non progression à long terme vers le SIDA. Par une approche originale sur les phénotypes, j'ai pu trouver le seul signal en dehors du locus HLA, identifié par étude « génome entier » et répliqué dans plusieurs cohortes, au niveau du gène *CXCR6*.

Enfin, dans le cadre du projet IHAC, la mise en place d'un « pipeline » complexe d'imputation et de méta-analyse des données génomiques issues de nombreuses cohortes m'a permis, en plus d'avoir réussi à appréhender les techniques, de mettre en avant un nouveau gène candidat, KIF26B.

Ce travail s'inscrit dans la perspective d'une meilleure compréhension des maillons moléculaires impliqués dans la pathogenèse du VIH-1 qui sont encore mal élucidés à ce jour. L'identification de nouveaux mécanismes pourrait permettre de définir rationnellement de nouvelles cibles thérapeutiques et ainsi aider au développement d'un

traitement efficace du SIDA. Elle permettrait de développer des outils diagnostiques qui aideront les médecins à mieux prédire les évolutions de leurs patients.

Pour conclure, cette thèse aura constitué pour moi un apprentissage particulièrement enrichissant en abordant un sujet multi-disciplinaire à l'interface de la biologie, de la génétique, des statistiques et de l'informatique. J'ai eu la joie de participer aux recherches dans un domaine en pleine effervescence qui n'en est encore qu'à ses balbutiements avec encore de nombreuses mutations technologiques et scientifiques à venir.

Références bibliographiques

1. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010, 467, 1061-73.
2. Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C.L., Furtado, M.R., Kidd, J.R. et al.. SNPs for a universal individual identification panel. *Hum. Genet.* 2010, 127, 315-24.
3. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B. et al.. The structure of haplotype blocks in the human genome. *Science* 2002, 296, 2225-9.
4. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S.. High-resolution haplotype structure in the human genome. *Nat. Genet.* 2001, 29, 229-32.
5. Smith, M.W., Dean, M., Carrington, M., Winkler, C., Huttley, G.A., Lomb, D.A. et al.. Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort. *Science* 1997, 277, 959-65.
6. Bellanné-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S. et al.. Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992, 70, 1059-68.
7. Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G. et al.. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992, 359, 380-7.
8. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P. et al.. A second-generation linkage map of the human genome. *Nature* 1992, 359, 794-801.
9. Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P. et al.. The 1993-94 Généthon human genetic linkage map. *Nat. Genet.* 1994, 7, 246-339.
10. Cohen, D., Chumakov, I. & Weissenbach, J.. [First generation of the physical map of the human genome]. *C. R. Acad. Sci. III, Sci. Vie* 1993, 316, 1484-8.
11. Venter, J.C.. The sequence of the human genome. *Science* 2001, 291, 1304-51.
12. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004, 431, 931-45.
13. The International HapMap Project. . *Nature* 2003, 426, 789-96.
14. The International Hapmap Project. A haplotype map of the human genome. *Nature* 2005, 437, 1299-320.
15. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L. et al.. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010, 467, 52-8.
16. Liu, Z. & Lin, S.. Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet. Epidemiol.* 2005, 29, 353-64.
17. Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L. et al.. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* 2010, 42, 436-40.
18. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. et al.. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 2008, 9, 356-69.

19. de Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. & Altshuler, D.. Efficiency and power in genetic association studies. *Nat. Genet.* 2005, 37, 1217-23.
20. Stram, D.O.. Tag SNP selection for association studies. *Genet. Epidemiol.* 2004, 27, 365-74.
21. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J., Sackler, R.S., Haynes, C. et al.. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005, 308, 385-9.
22. International Multiple Sclerosis Genetics Consortium. Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* 2007, 357, 851-62.
23. Meyre, D., Delplanque, J., Chèvre, J., Lecoeur, C., Lobbens, S., Gallina, S. et al.. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* 2009, 41, 157-9.
24. Mira, M.T., Alcais, A., di Pietrantonio, T., Thuc, N.V., Phuong, M.C., Abel, L. et al.. Segregation of HLA/TNF region is linked to leprosy clinical spectrum in families displaying mixed leprosy subtypes. *Genes Immun.* 2003, 4, 67-73.
25. Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K.. A comprehensive review of genetic association studies. *Genet. Med.* 2002, 4, 45-61.
26. Hirschhorn, J.N. & Daly, M.J.. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 2005, 6, 95-108.
27. Balding, D.J.. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 2006, 7, 781-91.
28. Risch, N. & Merikangas, K.. The future of genetic studies of complex human diseases. *Science* 1996, 273, 1516-7.
29. Fridley, B.L. & Biernacka, J.M.. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur. J. Hum. Genet.* 2011, 19, 837-43.
30. Jorgenson, E. & Witte, J.S.. A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.* 2006, 7, 885-91.
31. Richardson, K., Lai, C., Parnell, L.D., Lee, Y. & Ordovas, J.M.. A Genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. *BMC Genomics* 2011, 12, 504.
32. Xu, Z. & Taylor, J.A.. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* 2009, 37, W600-5.
33. Yang, T., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E. et al.. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* 2010, 26, 2474-6.
34. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C.C. et al.. A genome-wide association study of global gene expression. *Nat. Genet.* 2007, 39, 1202-7.
35. Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J. et al.. Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 2008, 319, 921-6.
36. König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M.C., Irelan, J.T. et al.. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 2008, 135, 49-60.
37. Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C. et al.. Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe* 2008, 4, 495-504.
38. Wang, K., Li, M. & Hakonarson, H.. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 2010, 11, 843-54.
39. Ziegler, A.. Genome-wide association studies: quality control and population-based measures. *Genet. Epidemiol.* 2009, 33 Suppl 1, S45-50.
40. Fisher, R.. Edinburgh. *Statistical Methods for Research Workers* 1932.
41. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R.. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* 2005, 76, 887-93.
42. Lander, E.S. & Schork, N.J.. Genetic dissection of complex traits. *Science* 1994, 265, 2037-48
43. Devlin, B. & Roeder, K.. Genomic control for association studies. *Biometrics* 1999, 55, 997-

1004.
44. Falush, D., Stephens, M. & Pritchard, J.K.. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003, 164, 1567-87.
45. Wang, L. & Xu, Y.. Haplotype inference by maximum parsimony. *Bioinformatics* 2003, 19, 1773-80.
46. Kimmel, G. & Shamir, R.. The incomplete perfect phylogeny haplotype problem. *J Bioinform Comput Biol* 2005, 3, 359-84.
47. Excoffier, L. & Slatkin, M.. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 1995, 12, 921-7.
48. Fallin, D. & Schork, N.J.. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.* 2000, 67, 947-59.
49. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E. et al.. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 2006, 78, 437-50.
50. Stephens, M., Smith, N.J. & Donnelly, P.. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 2001, 68, 978-89.
51. Delaneau, O., Coulonges, C., Boelle, P., Nelson, G., Spadoni, J. & Zagury, J.. ISHAPE: new rapid and accurate software for haplotyping. *BMC Bioinformatics* 2007, 8, 205.
52. Delaneau, O., Coulonges, C. & Zagury, J.. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 2008, 9, 540.
53. Adkins, R.M.. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genet.* 2004, 5, 22.
54. Browning, S.R. & Browning, B.L.. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 2011, 12, 703-14.
55. Clark, A.G.. The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* 2004, 27, 321-33.
56. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J.. The importance of phase information for human genomics. *Nat. Rev. Genet.* 2011, 12, 215-23.
57. Grisoni, M., Proust, C., Alanne, M., DeSuremain, M., Salomaa, V., Kuulasmaa, K. et al.. Haplotypic analysis of tag SNPs of the interleukin-18 gene in relation to cardiovascular disease events: the MORGAM Project. *Eur. J. Hum. Genet.* 2008, 16, 1512-20.
58. Pickard, B.S., Christoforou, A., Thomson, P.A., Fawkes, A., Evans, K.L., Morris, S.W. et al.. Interacting haplotypes at the NPAS3 locus alter risk of schizophrenia and bipolar disorder. *Mol. Psychiatry* 2009, 14, 874-84.
59. Trégouët, D., König, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S. et al.. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* 2009, 41, 283-5.
60. Winkler, C.A., Hendel, H., Carrington, M., Smith, M.W., Nelson, G.W., O'brien, S.J. et al.. Dominant effects of CCR2-CCR5 haplotypes in HIV-1 disease progression. *J. Acquir. Immune Defic. Syndr.* 2004, 37, 1534-8.
61. Flores-Villanueva, P.O., Hendel, H., Caillat-Zucman, S., Rappaport, J., Burgos-Tiburcio, A., Bertin-Maghit, S. et al.. Associations of MHC ancestral haplotypes with resistance/susceptibility to AIDS disease development. *J. Immunol.* 2003, 170, 1925-9.
62. Saleheen, D.. Haplotype analysis in VEGF gene and increased risk of Alzheimer's disease. *Ann. Neurol.* 2005, 58, 488; author reply 488-9.
63. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. et al.. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 2001, 69, 138-47.
64. Evangelou, E., Maraganore, D.M. & Ioannidis, J.P.A.. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE* 2007, 2, e196.
65. de Bakker, P.I.W., Neale, B.M. & Daly, M.J.. Meta-analysis of genome-wide association

studies. *Cold Spring Harb Protoc* 2010, 2010, pdb.top81.
66. Guan, Y. & Stephens, M.. Practical issues in imputation-based association mapping. *PLoS Genet.* 2008, 4, e1000279.
67. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R.. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 2010, 34, 816-34.
68. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P.. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 2007, 39, 906-13.
69. Servin, B. & Stephens, M.. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 2007, 3, e114.
70. Browning, S.R.. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* 2008, 124, 439-50.
71. Li, Y., Willer, C., Sanna, S. & Abecasis, G.. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009, 10, 387-406.
72. Pei, Y., Zhang, L., Li, J. & Deng, H.. Analyses and comparison of imputation-based association methods. *PLoS ONE* 2010, 5, e10827.
73. Marchini, J. & Howie, B.. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 2010, 11, 499-511.
74. Howie, B., Marchini, J. & Stephens Matthew. Genotype Imputation with Thousands of Genomes. *G3 journal* 2011, 1, 457-470.
75. Gottlieb, M.S.. Pneumocystis pneumonia--Los Angeles. 1981. *Am J Public Health* 1981, 96, 980-1; discussion 982-3.
76. Barré-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J. et al.. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 1983, 220, 868-71.
77. Gallo, R.C., Salahuddin, S.Z., Popovic, M., Shearer, G.M., Kaplan, M., Haynes, B.F. et al.. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* 1984, 224, 500-3.
78. Clavel, F., Guétard, D., Brun-Vézinet, F., Chamaret, S., Rey, M.A., Santos-Ferreira, M.O. et al.. Isolation of a new human retrovirus from West African patients with AIDS. *Science* 1986, 233, 343-6.
79. Marlink, R., Kanki, P., Thior, I., Travers, K., Eisen, G., Siby, T. et al.. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science* 1994, 265, 1587-90.
80. Adjorlolo-Johnson, G., De Cock, K.M., Ekpini, E., Vetter, K.M., Sibailly, T., Brattegaard, K. et al.. Prospective comparison of mother-to-child transmission of HIV-1 and HIV-2 in Abidjan, Ivory Coast. *JAMA* 1994, 272, 462-6.
81. Coffin, J.M.. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 1995, 267, 483-9.
82. Bister, K. & Jansen, H.W.. Oncogenes in retroviruses and cells: biochemistry and molecular genetics. *Adv. Cancer Res.* 1986, 47, 99-188.
83. Jern, P. & Coffin, J.M.. Effects of retroviruses on host genome function. *Annu. Rev. Genet.* 2008, 42, 709-32.
84. Swain, A. & Coffin, J.M.. Mechanism of transduction by retroviruses. *Science* 1992, 255, 841-5.
85. Freed, E.O.. HIV-1 and the host cell: an intimate association. *Trends Microbiol.* 2004, 12, 170-7.
86. Cullen, B.R.. Regulation of human immunodeficiency virus replication. *Annu. Rev. Microbiol.* 1991, 45, 219-50.
87. Levy, J.A.. HIV pathogenesis: 25 years of progress and persistent challenges. *AIDS* 2009, 23, 147-60.
88. Munier, S., Borjabad, A., Lemaire, M., Mariot, V. & Hazan, U.. In vitro infection of human primary adipose cells with HIV-1: a reassessment. *AIDS* 2003, 17, 2537-9.

89. Hazan, U., Romero, I.A., Cancello, R., Valente, S., Perrin, V., Mariot, V. et al.. Human adipose cells express CD4, CXCR4, and CCR5 [corrected] receptors: a new target cell type for the immunodeficiency virus-1?. *FASEB J.* 2002, 16, 1254-6.
90. Hahn, B.H., Shaw, G.M., De Cock, K.M. & Sharp, P.M.. AIDS as a zoonosis: scientific and public health implications. *Science* 2000, 287, 607-14.
91. Weissman, J.B.. Availability of AZT for treatment of AIDS. *N. Engl. J. Med.* 1987, 316, 1158.
92. Peytavin, G., Calvez, V. & Katlama, C.. [CCR5 antagonists: a new class of antiretrovirals]. *Therapie* 2009, 64, 9-16.
93. Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L. & Costagliola, D.. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. *AIDS* 2009, 23, 1163-9.
94. Huff, B.. Who are the elite controllers?. *GMHC Treat Issues* 2005, 19, 12.
95. Saksena, N.K., Rodes, B., Wang, B. & Soriano, V.. Elite HIV controllers: myth or reality?. *AIDS Rev* 2007, 9, 195-207.
96. Pereyra, F., Addo, M.M., Kaufmann, D.E., Liu, Y., Miura, T., Rathod, A. et al.. Genetic and immunologic heterogeneity among persons who control HIV infection in the absence of therapy. *J. Infect. Dis.* 2008, 197, 563-71.
97. Kulkarni, P.S., Butera, S.T. & Duerr, A.C.. Resistance to HIV-1 infection: lessons learned from studies of highly exposed persistently seronegative (HEPS) individuals. *AIDS Rev* 2003, 5, 87-103.
98. Yang, C., Li, M., Limpakarnjanarat, K., Young, N.L., Hodge, T., Butera, S.T. et al.. Polymorphisms in the CCR5 coding and noncoding regions among HIV type 1-exposed, persistently seronegative female sex-workers from Thailand. *AIDS Res. Hum. Retroviruses* 2003, 19, 661-5.
99. Hendel, H., Hénon, N., Lebuanec, H., Lachgar, A., Poncelet, H., Caillat-Zucman, S. et al.. Distinctive effects of CCR5, CCR2, and SDF1 genetic polymorphisms in AIDS progression. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 1998, 19, 381-6.
100. Ioannidis, J.P., Rosenberg, P.S., Goedert, J.J., Ashton, L.J., Benfield, T.L., Buchbinder, S.P. et al.. Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: An international meta-analysis of individual-patient data. *Ann. Intern. Med.* 2001, 135, 782-95.
101. Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E. et al.. Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation. *Blood* 2011, 117, 2791-9.
102. Labrecque, J., Metz, M., Lau, G., Darkes, M.C., Wong, R.S.Y., Bogucki, D. et al.. HIV-1 entry inhibition by small-molecule CCR5 antagonists: a combined molecular modeling and mutant study using a high-throughput assay. *Virology* 2011, 413, 231-43.
103. Scherer, L.J. & Rossi, J.J.. Ex vivo gene therapy for HIV-1 treatment. *Hum. Mol. Genet.* 2011, 20, R100-7.
104. Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C.M. et al.. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996, 382, 722-5.
105. Dean, M., Carrington, M., Winkler, C., Huttley, G.A., Smith, M.W., Allikmets, R. et al.. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE . *Science* 1996, 273, 1856-62.
106. Rappaport, J., Cho, Y.Y., Hendel, H., Schwartz, E.J., Schachter, F. & Zagury, J.F.. 32 bp CCR-5 gene deletion and resistance to fast progression in HIV-1 infected heterozygotes. *Lancet* 1997, 349, 922-3.
107. McDermott, D.H., Zimmerman, P.A., Guignard, F., Kleeberger, C.A., Leitman, S.F. & Murphy, P.M.. CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). *Lancet* 1998, 352, 866-70.

108. Martin, M.P., Dean, M., Smith, M.W., Winkler, C., Gerrard, B., Michael, N.L. et al.. Genetic acceleration of AIDS progression by a promoter variant of CCR5. *Science* 1998, 282, 1907-11.
109. Carrington, M., Nelson, G.W., Martin, M.P., Kissner, T., Vlahov, D., Goedert, J.J. et al.. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 1999, 283, 1748-52.
110. Gao, X., Nelson, G.W., Karacki, P., Martin, M.P., Phair, J., Kaslow, R. et al.. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* 2001, 344, 1668-75.
111. Hendel, H., Caillat-Zucman, S., Lebuanec, H., Carrington, M., O'Brien, S., Andrieu, J.M. et al.. New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS. *J. Immunol.* 1999, 162, 6942-6.
112. Diop, G., Spadoni, J., Do, H., Hirtzig, T., Coulonges, C., Labib, T. et al.. Genomic approach of AIDS pathogenesis: exhaustive genotyping of the TNFR1 gene in a French AIDS cohort. *Biomed. Pharmacother.* 2005, 59, 474-80.
113. Diop, G., Hirtzig, T., Do, H., Coulonges, C., Vasilescu, A., Labib, T. et al.. Exhaustive genotyping of the interferon alpha receptor 1 (IFNAR1) gene and association of an IFNAR1 protein variant with AIDS progression or susceptibility to HIV-1 infection in a French AIDS cohort. *Biomed. Pharmacother.* 2006, 60, 569-77.
114. Do, H., Vasilescu, A., Diop, G., Hirtzig, T., Heath, S.C., Coulonges, C. et al.. Exhaustive genotyping of the CEM15 (APOBEC3G) gene and absence of association with AIDS progression in a French cohort. *J. Infect. Dis.* 2005, 191, 159-63.
115. Do, H., Vasilescu, A., Diop, G., Hirtzig, T., Coulonges, C., Labib, T. et al.. Associations of the IL2Ralpha, IL4Ralpha, IL10Ralpha, and IFN (gamma) R1 cytokine receptor genes with AIDS progression in a French AIDS cohort. *Immunogenetics* 2006, 58, 89-98.
116. Do, H., Vasilescu, A., Carpentier, W., Meyer, L., Diop, G., Hirtzig, T. et al.. Exhaustive genotyping of the interleukin-1 family genes and associations with AIDS progression in a French cohort. *J. Infect. Dis.* 2006, 194, 1492-504.
117. Limou, S., Coulonges, C., Foglio, M., Heath, S., Diop, G., Leclerc, S. et al.. Exploration of associations between phospholipase A2 gene family polymorphisms and AIDS progression using the SNPlex method. *Biomed. Pharmacother.* 2008, 62, 31-40.
118. An, P. & Winkler, C.A.. Host genes associated with HIV/AIDS: advances in gene discovery. *Trends Genet.* 2010, 26, 119-31.
119. Hendel, H., Cho, Y.Y., Gauthier, N., Rappaport, J., Schächter, F. & Zagury, J.F.. Contribution of cohort studies in understanding HIV pathogenesis: introduction of the GRIV cohort and preliminary results. *Biomed. Pharmacother.* 1996, 50, 480-7.
120. Zhang, G., Nebert, D.W., Chakraborty, R. & Jin, L.. Statistical power of association using the extreme discordant phenotype design. *Pharmacogenet. Genomics* 2006, 16, 401-13.
121. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M. et al.. A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007, 317, 944-7.
122. Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M. et al.. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. *PLoS ONE* 2008, 3, e3907.
123. White, R.A., Pasztor, L.M., Richardson, P.M. & Zon, L.I.. The gene encoding TBC1D1 with homology to the tre-2/USP6 oncogene, BUB2, and cdc16 maps to mouse chromosome 5 and human chromosome 4. *Cytogenet. Cell Genet.* 2000, 89, 272-5.
124. Suske, G., Bruford, E. & Philipsen, S.. Mammalian SP/KLF transcription factors: bring in the family. *Genomics* 2005, 85, 551-6.
125. Herbeck, J.T., Gottlieb, G.S., Winkler, C.A., Nelson, G.W., An, P., Maust, B.S. et al.. Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. *J. Infect. Dis.* 2010, 201, 618-26.
126. Wang, L., Zhu, J., Shan, S., Qin, Y., Kong, Y., Liu, J. et al.. Repression of interferon-gamma expression in T cells by Prospero-related homeobox protein. *Cell Res.* 2008, 18, 911-20.

127. Pelak, K., Goldstein, D.B., Walley, N.M., Fellay, J., Ge, D., Shianna, K.V. et al.. Host determinants of HIV-1 control in African Americans. *J. Infect. Dis.* 2010, 201, 1141-9.
128. Shrestha, S., Aissani, B., Song, W., Wilson, C.M., Kaslow, R.A. & Tang, J.. Host genetics and HIV-1 viral load set-point in African-Americans. *AIDS* 2009, 23, 673-7.
129. Bol, S.M., Moerland, P.D., Limou, S., van Remmerden, Y., Coulonges, C., van Manen, D. et al.. Genome-wide association study identifies single nucleotide polymorphism in DYRK1A associated with replication of HIV-1 in monocyte-derived macrophages. *PLoS ONE* 2011, 6, e17190.
130. Hercberg, S., Galan, P., Preziosi, P., Roussel, A.M., Arnaud, J., Richard, M.J. et al.. Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Minéraux AntioXydants Stud. *Int J Vitam Nutr Res* 1998, 68, 3-20.
131. Balkau, B.. [An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome]. *Rev Epidemiol Sante Publique* 1996, 44, 373-5.
132. van Manen, D., Delaneau, O., Kootstra, N.A., Boeser-Nunnink, B.D., Limou, S., Bol, S.M. et al.. Genome-wide association scan in HIV-1-infected individuals identifying variants influencing disease course. *PLoS ONE* 2011, 6, e22208.
133. van Manen, D., Kootstra, N.A., Boeser-Nunnink, B., Handulle, M.A., van't Wout, A.B. & Schuitemaker, H.. Association of HLA-C and HCP5 gene regions with the clinical course of HIV-1 infection. *AIDS* 2009, 23, 19-28.
134. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P.. The effects of human population structure on large genetic association studies. *Nat. Genet.* 2004, 36, 512-7.
135. Dubois, P., Hinz, S. & Carsten, P.. MySQL - Guide officiel. *(ISBN 978-2-7440-1782-7)* 2004, , .
136. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D. et al.. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007, 81, 559-75.
137. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M.. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007, 23, 1294-6.
138. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J.. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, 21, 263-5.
139. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D.. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 2006, 38, 904-9.
140. Xu, Z., Kaplan, N.L. & Taylor, J.A.. TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics* 2007, 23, 3254-5.
141. Liu, G., Wang, Y. & Wong, L.. FastTagger: an efficient algorithm for genome-wide tag SNP selection using multi-marker linkage disequilibrium. *BMC Bioinformatics* 2010, 11, 66.
142. Scheet, P. & Stephens, M.. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 2006, 78, 629-44.
143. Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P. & Becker, T.. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics* 2009, 25, 3275-81.
144. Segrè, A.V., DIAGRAM Consortium, Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D.. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* 2010, 6, .
145. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M.. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 2010, 11, 134.
146. Purcell, S., Daly, M.J. & Sham, P.C.. WHAP: haplotype-based association analysis. *Bioinformatics* 2007, 23, 255-6.

147. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L.S. et al.. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* 2010, 87, 325-40.
148. Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O. et al.. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* 2009, 199, 419-26.
149. Le Clerc, S., Limou, S., Coulonges, C., Carpentier, W., Dina, C., Taing, L. et al.. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.* 2009, 200, 1194-201.
150. Le Clerc, S., Coulonges, C., Delaneau, O., Van Manen, D., Herbeck, J.T., Limou, S. et al.. Screening Low Frequency SNPS From Genome Wide Association Study Reveals A New Risk Allele for Progression to Aids. *J. Acquir. Immune Defic. Syndr.* 2010, , .
151. Limou, S., Coulonges, C., Herbeck, J.T., van Manen, D., An, P., Le Clerc, S. et al.. Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J. Infect. Dis.* 2010, 202, 908-15.
152. Fellay, J., Ge, D., Shianna, K.V., Colombo, S., Ledergerber, B., Cirulli, E.T. et al.. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* 2009, 5, e1000791.
153. Javed, A., Drineas, P., Mahoney, M.W. & Paschou, P.. Efficient Genomewide Selection of PCA-Correlated tSNPs for Genotype Imputation. *Ann. Hum. Genet.* 2011, 75, 707-22.
154. Patterson, N., Price, A.L. & Reich, D.. Population structure and eigenanalysis. *PLoS Genet.* 2006, 2, e190.
155. Howie, B.N., Donnelly, P. & Marchini, J.. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009, 5, e1000529.
156. Jiao, S., Hsu, L., Hutter, C.M. & Peters, U.. The use of imputed values in the meta-analysis of genome-wide association studies. *Genet. Epidemiol.* 2011, 35, 597-605.
157. Goldstein, D.B.. Genomics and biology come together to fight HIV. *PLoS Biol.* 2008, 6, e76.
158. Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I.W., Walker, B.D. et al.. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 2010, 330, 1551-7.
159. Teo, Y., Small, K.S. & Kwiatkowski, D.P.. Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.* 2010, 11, 149-60.
160. Uchiyama, Y., Sakaguchi, M., Terabayashi, T., Inenaga, T., Inoue, S., Kobayashi, C. et al.. Kif26b, a kinesin family gene, regulates adhesion of the embryonic kidney mesenchyme. *Proc. Natl. Acad. Sci. U.S.A.* 2010, 107, 9240-5.
161. Shoeman, R.L., Sachse, C., Höner, B., Mothes, E., Kaufmann, M. & Traub, P.. Cleavage of human and mouse cytoskeletal and sarcomeric proteins by human immunodeficiency virus type 1 protease. Actin, desmin, myosin, and tropomyosin. *Am. J. Pathol.* 1993, 142, 221-30.
162. de Bakker, P.I.W., Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S. & Voight, B.F.. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 2008, 17, R122-8.
163. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T. et al.. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* 2008, 40, 638-45.
164. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J. et al.. Finding the missing heritability of complex diseases. *Nature* 2009, 461, 747-53.
165. Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B. & Hakonarson, H.. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* 2010, 86, 730-42.
166. Duggal, P., An, P., Beaty, T.H., Strathdee, S.A., Farzadegan, H., Markham, R.B. et al.. Genetic influence of CXCR6 chemokine receptor alleles on PCP-mediated AIDS progression among African Americans. *Genes Immun.* 2003, 4, 245-50.

167. Deng, H.K., Unutmaz, D., KewalRamani, V.N. & Littman, D.R.. Expression cloning of new receptors used by simian and human immunodeficiency viruses. *Nature* 1997, 388, 296-300.
168. Kim, C.H., Kunkel, E.J., Boisvert, J., Johnston, B., Campbell, J.J., Genovese, M.C. et al.. Bonzo/CXCR6 expression defines type 1-polarized T-cell subsets with extralymphoid tissue homing potential. *J. Clin. Invest.* 2001, 107, 595-601.
169. Germanov, E., Veinotte, L., Cullen, R., Chamberlain, E., Butcher, E.C. & Johnston, B.. Critical role for the chemokine receptor CXCR6 in homeostasis and activation of CD1d-restricted NKT cells. *J. Immunol.* 2008, 181, 81-91.
170. Xu, Y., Peng, B., Fu, Y. & Amos, C.I.. Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics* 2011, 12, 331.
171. Gao, M., Wang, C., Sima, X. & Han, X.. NFKB1 -94 Insertion/Deletion ATTG Polymorphism Contributes to Risk of Systemic Lupus Erythematosus. *DNA Cell Biol.* 2011, , .
172. Catano, G., Kulkarni, H., He, W., Marconi, V.C., Agan, B.K., Landrum, M. et al.. HIV-1 disease-influencing effects associated with ZNRD1, HCP5 and HLA-C alleles are attributable mainly to either HLA-A10 or HLA-B*57 alleles. *PLoS ONE* 2008, 3, e3636.
173. Trachtenberg, E., Bhattacharya, T., Ladner, M., Phair, J., Erlich, H. & Wolinsky, S.. The HLA-B/-C haplotype block contains major determinants for host control of HIV. *Genes Immun.* 2009, 10, 673-7.
174. Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A., Jarvelin, M., Balding, D. et al.. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* 2009, 4, e8068.
175. O'Dushlaine, C., Kenny, E., Heron, E.A., Segurado, R., Gill, M., Morris, D.W. et al.. The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* 2009, 25, 2762-3.
176. Bates, J.S., Lessard, C.J., Leon, J.M., Nguyen, T., Battiest, L.J., Rodgers, J. et al.. Meta-analysis and imputation identifies a 109 kb risk haplotype spanning TNFAIP3 associated with lupus nephritis and hematologic manifestations. *Genes Immun.* 2009, 10, 470-7.
177. International Parkinson Disease Genomics Consortium. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 2011, 377, 641-9.
178. Ku, C.S., Loy, E.Y., Pawitan, Y. & Chia, K.S.. The pursuit of genome-wide association studies: where are we now?. *J. Hum. Genet.* 2010, 55, 195-206.
179. Cantor, R.M., Lange, K. & Sinsheimer, J.S.. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 2010, 86, 6-22.
180. Zia, A. & Moses, A.M.. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics* 2011, 12, 299.
181. Panagiotou, O.A., Evangelou, E. & Ioannidis, J.P.A.. Genome-wide significant associations for variants with minor allele frequency of 5% or less--an overview: A HuGE review. *Am. J. Epidemiol.* 2010, 172, 869-89.
182. Lango, H. & Weedon, M.N.. What will whole genome searches for susceptibility genes for common complex disease offer to clinical practice?. *J. Intern. Med.* 2008, 263, 16-27.
183. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J.. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 2010, 11, 773-85.
184. Shi, Y., Li, Y., Zhang, D., Zhang, H., Li, Y., Lu, F. et al.. Exome sequencing identifies ZNF644 mutations in high myopia. *PLoS Genet.* 2011, 7, e1002084.
185. Grossmann, V., Tiacci, E., Holmes, A.B., Kohlmann, A., Martelli, M.P., Kern, W. et al.. Whole-exome sequencing identifies mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* 2011, , .
186. Holm, H., Gudbjartsson, D.F., Sulem, P., Masson, G., Helgadottir, H.T., Zanon, C. et al.. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* 2011, 43, 316-20.
187. Frankel, W.N. & Schork, N.J.. Who's afraid of epistasis?. *Nat. Genet.* 1996, 14, 371-3.

188. Cordell, H.J.. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* 2009, 10, 392-404.
189. Phillips, P.C.. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* 2008, 9, 855-67.
190. Becker, T., Herold, C., Meesters, C., Mattheisen, M. & Baur, M.P.. Significance levels in genome-wide interaction analysis (GWIA). *Ann. Hum. Genet.* 2011, 75, 29-35.
191. Yau, C. & Holmes, C.C.. CNV discovery using SNP genotyping arrays. *Cytogenet. Genome Res.* 2008, 123, 307-12.
192. Carter, N.P.. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 2007, 39, S16-21.
193. Rakyan, V.K., Down, T.A., Balding, D.J. & Beck, S.. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 2011, 12, 529-41.
194. Siu, H., Zhu, Y., Jin, L. & Xiong, M.. Implication of next-generation sequencing on association studies. *BMC Genomics* 2011, 12, 322.
195. Glenn, T.C.. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011, 11, 759-69.
196. Clark, M.J., Chen, R., Lam, H.Y.K., Karczewski, K.J., Chen, R., Euskirchen, G. et al.. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* 2011, 29, 908-14.
197. Singleton, A.B.. Exome sequencing: a transformative technology. *Lancet Neurol* 2011, 10, 942-6.
198. Sanders, S.S.. Whole-exome sequencing: a powerful technique for identifying novel genes of complex disorders. *Clin. Genet.* 2011, 79, 132-3.
199. Zhang, X., Li, M. & Zhang, X.. [Exome sequencing and its application]. *Yi Chuan* 2011, 33, 847-56.
200. Liu, X., Jian, X. & Boerwinkle, E.. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 2011, 32, 894-9.
201. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. et al.. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 2011, 12, 745-55.
202. Bonnefond, A., Durand, E., Sand, O., De Graeve, F., Gallina, S., Busiah, K. et al.. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS ONE* 2010, 5, e13630.
203. Ku, C., Naidoo, N. & Pawitan, Y.. Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* 2011, 129, 351-70.
204. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M. et al.. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 2010, 42, 30-5.
205. Biesecker, L.G.. Exome sequencing makes medical genomics a reality. *Nat. Genet.* 2010, 42, 13-4.
206. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J.. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* 2009, 10, 241-51.
207. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R.. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 2011, 21, 940-51.
208. Froguel, P. & Blakemore, A.I.F.. The power of the extreme in elucidating obesity. *N. Engl. J. Med.* 2008, 359, 891-3.
209. Bodmer, W. & Bonilla, C.. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 2008, 40, 695-701.
210. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R.. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 2007, 80, 727-39.
211. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R. & Amos, C.I.. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 2008, 82,

100-12.
212. Kaye, J.. The regulation of direct-to-consumer genetic tests. *Hum. Mol. Genet.* 2008, 17, R180-3.
213. Vijverberg, S.J.H., Pieters, T. & Cornel, M.C.. Ethical and social issues in pharmacogenomics testing. *Curr. Pharm. Des.* 2010, 16, 245-52.
214. Daly, A.K.. Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.* 2010, 11, 241-6.

Liste des publications

1. Diop, G., Spadoni, J., Do, H., Hirtzig, T., **Coulonges, C**., Labib, T. et al.. Genomic approach of AIDS pathogenesis: exhaustive genotyping of the TNFR1 gene in a French AIDS cohort. *Biomed. Pharmacother.* 2005, 59, 474-80.
2. Do, H., Vasilescu, A., Diop, G., Hirtzig, T., Heath, S.C., **Coulonges, C**. et al.. Exhaustive genotyping of the CEM15 gene and absence of association with AIDS progression in a French cohort. *J. Infect. Dis.* 2005, 191, 159-63.
3. **Coulonges, C.**, Delaneau, O., Girard, M., Do, H., Adkins, R., Spadoni, J. et al.. Computation of haplotypes on SNPs subsets: advantage of the "global method". *BMC Genet.* 2006, 7, 50.
4. Diop, G., Hirtzig, T., Do, H., **Coulonges, C**., Vasilescu, A., Labib, T. et al.. Exhaustive genotyping of the interferon alpha receptor 1 (IFNAR1) gene and association of an IFNAR1 protein variant with AIDS progression or susceptibility to HIV-1 infection in a French AIDS cohort. *Biomed. Pharmacother.* 2006, 60, 569-77.
5. Do, H., Vasilescu, A., Carpentier, W., Meyer, L., Diop, G., Hirtzig, T., **Coulonges, C**. et al.. Exhaustive genotyping of the interleukin-1 family genes and associations with AIDS progression in a French cohort. *J. Infect. Dis.* 2006, 194, 1492-504.
6. Do, H., Vasilescu, A., Diop, G., Hirtzig, T., **Coulonges, C**., Labib, T. et al.. Associations of the IL2Ralpha, IL4Ralpha, IL10Ralpha, and IFN (gamma) R1 cytokine receptor genes with AIDS progression in a French AIDS cohort. *Immunogenetics* 2006, 58, 89-98.
7. Delaneau, O., **Coulonges, C**., Boelle, P., Nelson, G., Spadoni, J. & Zagury, J.. ISHAPE: new rapid and accurate software for haplotyping. *BMC Bioinformatics* 2007, 8, 205.

8. Delaneau, O., **Coulonges, C**. & Zagury, J.. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 2008, 9, 540.

9. Limou, S., **Coulonges, C**., Foglio, M., Heath, S., Diop, G., Leclerc, S. et al.. Exploration of associations between phospholipase A2 gene family polymorphisms and AIDS progression using the SNPlex method. *Biomed. Pharmacother.* 2008, 62, 31-40.

10. Le Clerc, S., Limou, S., **Coulonges, C**., Carpentier, W., Dina, C., Taing, L. et al.. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.* 2009, 200, 1194-201.

11. Limou, S., Le Clerc, S., **Coulonges, C**., Carpentier, W., Dina, C., Delaneau, O. et al.. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* 2009, 199, 419-26.

12. Le Clerc, S., **Coulonges, C**., Delaneau, O., Van Manen, D., Herbeck, J.T., Limou, S. et al.. Screening Low Frequency SNPS From Genome Wide Association Study Reveals A New Risk Allele for Progression to Aids. *J. Acquir. Immune Defic. Syndr.* 2010, , .

13. **Coulonges, C.***, Limou, S.*, Herbeck, J.T., van Manen, D., An, P., Le Clerc, S. et al.. Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J. Infect. Dis.* 2010, 202, 908-15. *1er co-auteurs avec une contribution identique.

14. Bol, S.M., Moerland, P.D., Limou, S., van Remmerden, Y., **Coulonges, C**., van Manen, D. et al.. Genome-wide association study identifies single nucleotide polymorphism in DYRK1A associated with replication of HIV-1 in monocyte-derived macrophages. *PLoS ONE* 2011, 6, e17190.

Liste des communications orales

1. Cedric Coulonges, Olivier Delaneau, Jean-Francois Zagury. Computation of haplotypes on SNPs subsets: advantage of the "global method". *JOBIM*. Bordeaux. 2006.
2. Cedric Coulonges. A new genomic database applied on GRIV cohort. 3*th Ermenonville International Workshop.* 2008
3. Cedric Coulonges. Haplotypes approaches for the GRIV cohort analyses. 4*th Ermenonville International Workshop.* 2009
4. Cedric Coulonges. CNV discovery using Illumina 314K SNP array. 5*th Ermenonville International Workshop.* 2010
5. Cedric Coulonges. Practical aspects in Genome Wide Association Studies (GWAS). *Journée thématique BILGWAS Nantes.* 2010
6. Cedric Coulonges. Preliminary results from the IHAC project. 6*th Ermenonville International Workshop* : « Genomics and disease pathogenesis ». 2011

le cnam

Cédric Coulonges

Analyses bioinformatiques dans le cadre de la génomique du SIDA

Résumé

Les technologies actuelles permettent d'explorer le génome entier pour y découvrir des variants génétiques associés aux maladies. Cela implique des outils bioinformatiques adaptés à l'interface de l'informatique, des statistiques et de la biologie. Ma thèse a porté sur l'exploitation bioinformatique des données génomiques issues de la cohorte GRIV du SIDA et du projet international IHAC (International HIV Acquisition Consortium).

Posant les prémices de l'imputation, j'ai d'abord développé le logiciel SUBHAP. Notre équipe a montré que la région HLA était essentielle dans la non progression et le contrôle de la charge virale et cela m'a conduit à étudier le phénotype non-progresseur non « elite ». J'ai ainsi révélé un variant du gène CXCR6 qui, en dehors du HLA, est le seul résultat identifié par approche génome-entier et répliqué.

L'imputation des données du projet IHAC (10000 patients infectés et 15000 contrôles) a été réalisée et des premières associations sont en cours d'exploration.

Mots clés : étude d'association, VIH-1, SNP, haplotypes, imputation, méta-analyse

Résumé en anglais

Nowadays with the newest technologies, the entire genome can be explored to uncover genetic variants which may be linked to diseases. This requires bioinformatics tools which are adequate for studies which are at the border between computing, statistics and biology. My thesis work focused on the bioinformatical analysis of genomic data from the GRIV AIDS cohort and from the IHAC (International HIV Acquisition Consortium) project.

I first laid the foundation for imputation work by developing the SUBHAP software. Our team showed that the HLA region was essential in non-progression and viral charge control. This led me to study the non progressor non elite phenotype. Thus, I uncovered a variant of the CXCR6 gene which is, apart from HLA, the only result identified with a GWAS approach so far and which has been reproduced.

The imputation of data from the IHAC project (10000 infected patients and 15000 control subjects) was also performed and the first associations are now being studied.

Keywords : GWAS, HIV-1, SNP, haplotypes, imputation, meta-analysis