Analyses 'genome entier' de la cohorte griv de patients à profil extrême du sida

Sigrid Le Clerc

le cnam

ÉCOLE DOCTORALE ARTS ET METIERS

CHAIRE de BIOINFORMATIQUE

THÈSE présentée par :

Sigrid LE CLERC

soutenue le : **17 décembre 2010**

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

Discipline/ Spécialité : Génétique et Bioinformatique

Analyses 'génome entier' de la cohorte GRIV de patients à profil extrême du SIDA

THÈSE dirigée par :

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

RAPPORTEURS :

M. ESTAQUIER Jérôme	Docteur, Université Paris Est-Créteil, INSERM, U955
M. SERRE Jean-Louis	Professeur, Université de Versailles Saint-Quentin-en-Yvelines

---

JURY :

M. THERWATH Amu	Professeur, Université Paris Diderot
Mme JULIER Cécile	Docteur, Institut National de la Santé et de la Recherche Médicale, INSERM, U958

Remerciements

J'exprime toute ma reconnaissance au Pr Jean-François Zagury, mon directeur de thèse, de m'avoir accepté dans son équipe, et de m'avoir confié un projet de thèse qui m'a passionné. Merci pour votre soutien tout au long de cette thèse.

Je suis très reconnaissante au Dr Jérôme Estaquier et au Pr Jean-Louis Serre d'avoir accepté de juger mon travail de thèse en tant que rapporteur.

Je remercie également le Pr Amu Therwath et le Dr Cécile Julier de m'avoir fait l'honneur de participer au jury de cette thèse.

Je remercie aussi les personnes qui m'ont accueillie dans leurs laboratoires et bureaux tout au long de ce travail de thèse : le Pr Amu Therwath et le laboratoire d'oncologie moléculaire, le Dr Ioannis Theodorou, le Dr Wassila Carpentier et la plateforme de génotypage de la Pitié-Salpêtrière, le Pr Christiane Rouzioux et son équipe du laboratoire de virologie de l'Hopital Necker.

Je remercie aussi l'Agence Nationale de Recherche sur le SIDA et le Pr Jean-François Delfraissy pour l'allocation de recherche qui m'a été accordée pendant la durée de ma thèse

Je tiens à remercier chaleureusement toute l'équipe GRIV : Sophie toujours prête à 'shaker ses shoulders', Cédric, Olive, Lieng, Taoufik et le petit dernier Vincent. Je remercie également l'équipe Drug Design ou Satan et ses filles (Matthieu, Hélène, Nesrine et Nathalie), ainsi que l'équipe Cytokines ou le poulailler (Rojo, Hadley, Lucille et Gabriel) et enfin Hervé, Christiane et Jean-Louis. Merci à vous tous pour ces bons moments passés ensemble. Merci également aux stagiaires qui sont passés dans notre équipe et qui ont collaboré aux différents projets : le petit bébé Safa, Edwige, Shakti, Marie, Chirine et Pierre.

Enfin un grand merci à ma famille et mes amis qui m'ont supportée (ce qui n'est pas toujours facile) et soutenue tout au long de cette thèse. Je remercie particulièrement Mohand 'ma tête de chips' préférée qui a toujours été là pour moi et qui m'a apportée un grand soutien durant ces 3 années.

Résumé

L'infection par le VIH-1 est un fléau qui touche encore aujourd'hui plus de 33 millions de personnes dans le monde, principalement en Afrique et en Asie. Des traitements ciblant la réplication virale existent, mais ils sont très chers et ne sont pas disponibles pour tous. Les facteurs génétiques de l'hôte jouent un rôle important dans la résistance/susceptibilité face à l'infection et à la progression vers le SIDA. Les avancées technologiques en génétique et en biologie moléculaire ont rendu possible le développement d'études d'association 'génome entier' qui permettent de chercher des associations statistiques entre des variants génétiques recouvrant l'ensemble du génome et des phénotypes d'évolution particuliers. Sans aucun a priori, il est ainsi possible d'identifier des gènes potentiellement impliqués dans le développement de la maladie. Cette meilleure compréhension des mécanismes moléculaires de la maladie doit permettre à terme le développement rationnel de nouvelles stratégies diagnostiques ou thérapeutiques.

Dans le cadre de ma thèse, j'ai ainsi réalisé deux études d'association 'génome entier' dans le SIDA, en comparant dans un cas les 275 non-progresseurs à long terme de la cohorte GRIV avec une cohorte de contrôles séronégatifs et dans l'autre cas les 85 progresseurs rapides de la cohorte GRIV et une cohorte de contrôles séronégatifs. J'ai réalisé une troisième analyse en exploitant les données issues de trois études 'génome entier' internationales dont la nôtre (France, Pays-Bas, USA), ciblant particulièrement les SNPs de fréquence faible (fréquence de l'allèle mineur, MAF<5%). Ces SNPs sont en effet très défavorisés sur le plan statistique et méritent un traitement à part pour proposer de nouvelles pistes.

Ces approches 'génome entier' ont réaffirmé le rôle central du HLA dans l'infection par le VIH, mais aussi dévoilé de nouveaux gènes candidats très pertinents donnant une nouvelle lumière sur les mécanismes moléculaires de progression suite à l'infection par le VIH-1.

Mots clés : étude d'association 'génome entier', VIH-1, SNP, progression, pathogenèse

Résumé en anglais

HIV-1 pandemy is a pleague that still affects more than 33 million people worldwide, mostly in Africa and Asia. There are treatments targeting efficiently the viral replication, but they are very expensive and not available for all. Host genetic factors play a major role in the resistance/susceptibility to infection and to AIDS progression. Technological advances in genetics and molecular biology have made possible the development of genome-wide association studies (GWAS) for the identification of statistical associations between genetic variants over the entire genome and specific disease progression phenotypes. It is thus possible to identify with no a priori genes potentially involved in disease development. The better understanding of the molecular mechanisms of disease pathogenesis should eventually lead to the rational development of new diagnostic or therapeutic strategies.

During my PhD, I have completed two genome-wide association studies on AIDS, comparing in the one hand 275 long-term non-progressors of the GRIV cohort with a group of seronegative controls and on the other hand 85 rapid progressors of the GRIV cohort with a group of seronegative controls. I have completed a third analysis exploiting data from three international GWAS including ours (France, Netherlands, USA), specifically focusing on low frequency SNPs (minor allele frequency, MAF <5%). These SNPs are indeed statistically disadvantaged by usual analyses and deserve a separate treatment to identify new leads.

These large scale studies have reaffirmed the central role of HLA in HIV-1 infection, but they have also revealed new relevant candidate genes shedding a new light on the molecular mechanisms of disease progression following the infection by HIV-1 .

Keywords : genome wide association study, HIV-1, SNP, progression, pathogenesis

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADCC	Antibody-Directed Cellular Cytotoxicity
ADN	Acide Désoxyribo-Nucléique. ADNg, ADN génomique
ANRS	Agence Nationale de Recherche sur le SIDA
ARN	Acide Ribo-Nucléique. ARNg, ARN génomique ; ARNm, ARN messager ; ARNt, ARN de transfert ; siARN, *silencing* ARN ; shARN, *small hairpin* ARN
ASK-1	Apoptosis Signal regulating Kinase-1
CAF	Cell Antiviral Factor
CDC	Center for Diseases Control
CEPH	Centre d'Etude du Polymorphisme Humain
CNAR	Cell Noncytotoxic Antiviral Response
CNV	Copy Number Variation
CTL	Cytotoxic T Lymphocyte
CypA	Cyclophilin A
DC	Dendritic Cell
DL	Déséquilibre de Liaison
env	enveloppe
gag	group-specific antigen
Gb	gigabase
gp	glycoprotéine
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, aussi appelé CMH, Complexe Majeur d'Histocompatibilité
IFN	Interféron
IL	Interleukine
kb	kilobase
LAV	LymphAdenopathy Virus
LFA	Leukocyte Function Antigen

LTNP	Long-Term Non-Progressor
LTR	Long Terminal Repeat
MAF	Minor allele frequency
Mb	mégabase
MBL	Mannose-Binding Lectin
MIP	Macrophage Inflammatory Protein
nef	Negative Regulation factor
NK	cellule Natural Killer
NNRTI	Non Nucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
pb	paire de bases
PBMC	Peripheral Blood Mononononuclear Cells
PBS	Primer Binding Site
pol	polymerase
PP2A	Protéine Phosphatase 2A
PR	Progresseur Rapide
PRR	Pattern Recognition Receptor
rev	régulateur de l'expression des protéines virales
RFLP	Restriction Fragment Length Polymorphism
RRE	Rev Responsive Element
SIDA	Syndrome d'Immuno-Déficience Acquise
SIV	Simian Immunodeficiency Virus
SNP	Single Nucleotide Polymorphism
TAR	Tat-responsive element
tat	trans-activateur de transcription
TCR	T Cell Receptor
TNF	Tumor Necrosis Factor
TRAIL	TNF-Related Apoptosis-Inducing Ligand
TSG	Tumor Susceptibility Gene
vif	viral infectivity factor
VIH	Virus de l'Immuno-déficience Humaine
VNTR	Variable Number of Tadem Repeat
vpr	viral protein r
vpu	viral protein u

WT	Wild Type

Première Partie :

Introduction

1. Le SIDA : Syndrome d'Immuno-Déficience Acquise

1.1. Découverte de la maladie et de son agent étiologique

1.1.1. Description des premiers cas

Le premier cas de Syndrome d'immuno-déficience acquise (SIDA) fut découvert aux Etats-Unis en 1981. Le Docteur Michael Gottlieb décrit la découverte de symptômes rares chez 4 jeunes homosexuels de Los Angeles : pneumocystoses (pneumonies à *Pneumocystis carinii*), candidoses buccales [1]. Ces symptômes sont associés à une perturbation du système immunitaire (amaigrissement, fièvre et une quantité anormalement basse de lymphocytes T $CD4^+$ dans le sang). Entre juin et août 1981, le centre américain du contrôle et de la prévention (CDC : Center for Disease Control) publie trois rapports décrivant ces mêmes symptômes, des cas anormaux d'infections virales (cytomégalovirus, *herpes simplex*...) et des sarcomes de Kaposi chez des individus homosexuels de Californie et de New York [1-3]. La pneumocystose et le sarcome de Kaposi, alors connus comme des affections rarissimes, avaient été observés pour la plupart exclusivement chez des sujets âgés ou en état d'immunodépression profonde or tous les cas recensés par les médecins en 1981 étaient de jeunes hommes jusqu'alors en bonne santé. A l'époque, la découverte de plusieurs cas au sein de la communauté homosexuelle conduit à l'hypothèse d'une transmission par voie sexuelle. Cependant , dès 1982, ces mêmes symptômes décrits chez des héroïnomanes par injection intraveineuse et chez des hémophiles, suggèrent un mode de transmission par voie sanguine et non par voie sexuelle exclusivement ; on découvrira par la suite la transmission mère enfant [4]. Les examens biologiques révélèrent chez tous les malades une immunodépression caractérisée par une chute significative du taux de lymphocytes T $CD4^+$; caractéristique qui donna naissance à l'appellation actuelle 'syndrome d'immuno-déficience acquise' ou SIDA.

1.1.2. Identification de l'agent étiologique : les VIH

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 [5]. 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 [6]. 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 [7]. Ce virus, appelé VIH-2, 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 [8,9].

D'après les données phylogénétiques, ces deux VIH auraient émergé par zoonose à partir de deux hôtes naturels du SIV (Simian Immunodeficiency Virus) : le Chimpanzé (SIVcpz) et le Sooty Mangabey (SIVsmm) auraient généré respectivement le VIH-1 et le VIH-2. Dans la suite de ce mémoire, mon propos se limitera au VIH-1 [10].

1.1.3. Evolution de la pandémie

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. 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, www.unaids.org, 2009, Figure 1). Les données épidémiologiques montrent clairement l'importance des progrès dans la prévention de nouvelles infections et la diminution du nombre annuel de morts (Figure 1). Cependant, l'accès au traitement reste très coûteux, limitant de ce fait leur accès aux personnes les plus pauvres qui sont les personnes les plus touchées par l'infection par le VIH. Il est donc urgent de rechercher de nouvelles pistes thérapeutiques efficaces et accessibles à tous.

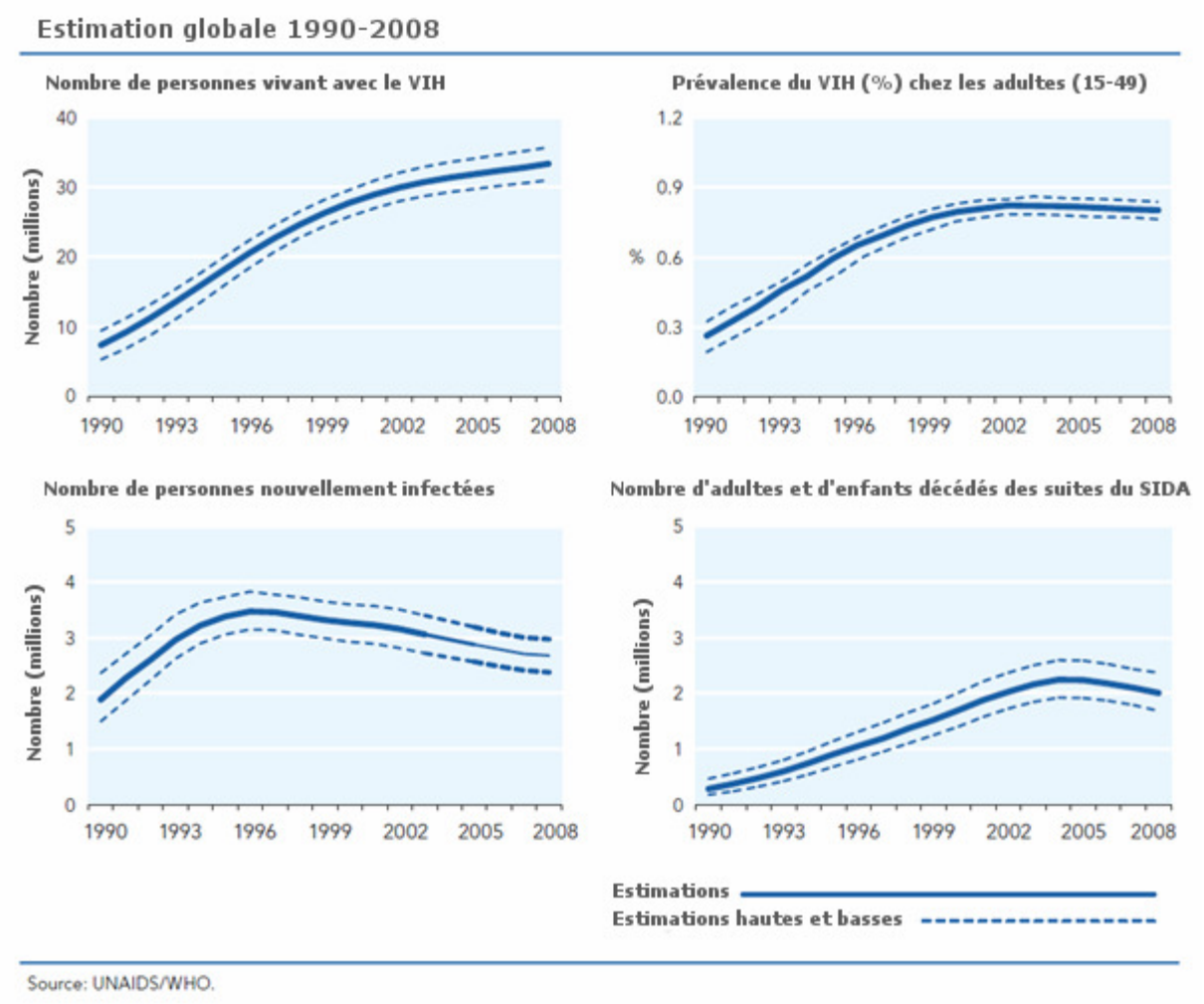


Figure 1 : Estimation globale entre 1990 et 2008 du nombre de personnes vivant avec le VIH, de la prévalence du VIH chez les adultes, du nombre de personnes nouvellement infectées par le VIH et du nombre d'enfants et d'adultes décédés des suites du SIDA.

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 transcritption 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 [11] et de leur intégration dans le génome de l'hôte [12-14]. Cette dernière étape peut conduire (i) à l'incorporation de portions du génome de l'hôte lors de la génération de nouveaux virus [14], (ii) à l'altération de l'activation/inactivation des gènes situés à proximité du site d'intégration (fort pouvoir oncogène des rétrovirus) [12] et (iii) à leur cooptation complète ou partielle dans les cellules germinales hôtes (HERV, Human Endogenous Retrovirus) [13].

1.2.1. Structure du VIH-1

Le virus se présente en microscopie électronique sous forme d'une particule sphérique de 80 à 120 nm, comportant une enveloppe lipidique et une nucléocapside excentrée et cylindrique. La nucléocapside contient : le génome viral, les protéines qui assurent la cohésion de la structure et les protéines dont la fonction enzymatique est nécessaire à la réplication virale (Figure 2).

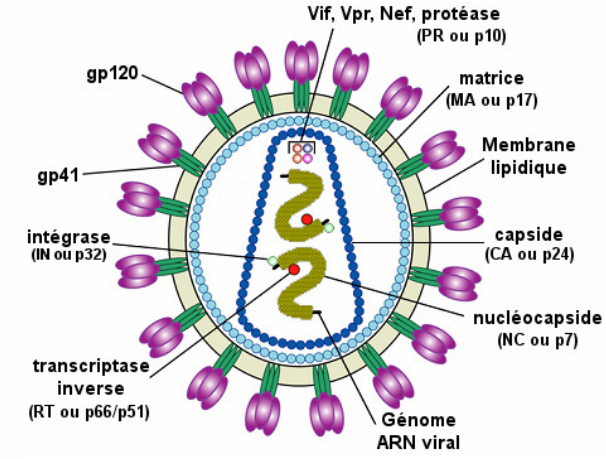


Figure 2 : Structure schématique d'une particule virale du VIH-1.

1.2.2. Structure génétique et protéines constitutives du VIH-1

La taille du génome du VIH-1 est d'environ 10 kb, il est composé de deux molécules identiques d'ARN monocaténaire. Ce génome possède la structure de base des rétrovirus : LTR-gag-pol-env-LTR (figure 3). Les gènes viraux sont au nombre de 9 et codent pour 15 protéines : (i) les gènes *gag*, *pol* et *env* codant pour les protéines de structure, (ii) les gènes *tat*, *rev* et *nef* codant pour les protéines régulatrices, et (iii) les gènes *vpu*, *vpr* et *vif* codant pour les protéines accessoires [15].

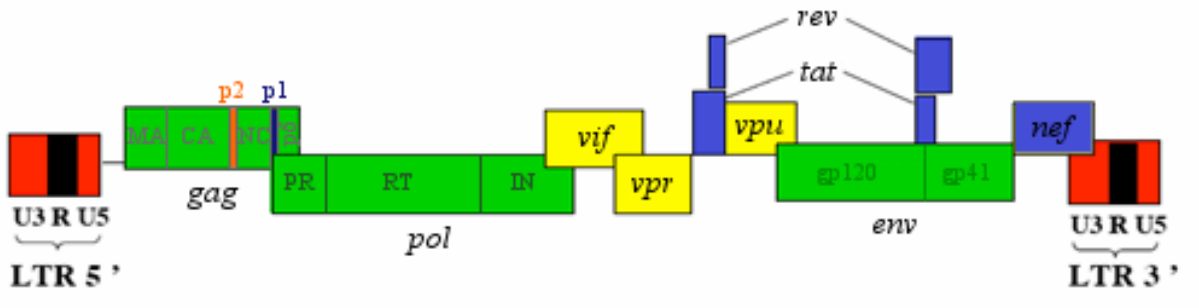


Figure 3 : Structure génomique du VIH-1. MA : matrice ou p17 ; CA : capside ou p24 ; NC : nucléocapside ou p7 ; PR: protéase ou p10 ; RT : transcriptase inverse ou p66/p51 ; IN : intégrase ou p32.

Les **LTR** (*Long Terminal Repeat*) sont des séquences répétées, situées aux extrémités du génome du VIH. Les LTR sont composées de trois régions: U3 (Unique, extrémité 3'), R (Repeated, région centrale) et U5 (Unique, extrémité 5'). Les séquences LTR sont importantes pour la transcription du génome viral. En effet, la transcription débute au niveau de la région R du LTR5', et s'achève au niveau de la région R du LTR3' par la polyadénylation. De plus, le LTR5' possède des sites de fixation pour des facteurs de transcription (NFκB, NFAT, SP1...) [16]. Enfin, à l'extrémité 3' du LTR5', se trouve le site de fixation de l'ARN de transfert ($ARNt^{lys}$) qui permet l'initialisation de la transcription inverse.

Les protéines de structure

Le gène ***gag*** (Group-specific AntiGen) code pour une polyprotéine précurseur de 55 kDa appelée $Pr55^{Gag}$. Cette polyprotéine est clivée par la protéase virale (PR) en plusieurs fragments :

1) protéine de la matrice MA ou p17 : protéine qui s'accumule à la surface interne de la membrane du virus formant ainsi une coque intérieure.

2) protéine de la capside CA ou p24 : protéine constituant le 'core' protéique de forme conique qui contient l'ARN génomique virale.

3) protéine de la nucléocapside NC ou p7 : protéine participant à l'assemblage de l'ARN dans la particule virale. p17 assure la protection de l'ARN vis à vis des nucléases qui entraîneraient sa dégradation.

4) protéine p6 : protéine nécessaire à l'incorporation de la protéine Vpr dans le virus [17], mais aussi à la production de particules virales lors de l'étape de bourgeonnement [18].

5) p1 et p2 : ces deux peptides sont impliqués dans la régulation du taux de clivage de $Pr55^{Gag}$ [19, 20], dans l'assemblage et le bourgeonnement viral [20, 21]

Dans 5% des cas, lors de la traduction de $Pr55^{Gag}$, un décalage dans le cadre de lecture apparaît grâce à la formation d'une structure secondaire impliquant la séquence nucléique qui code pour le peptide p1. Ce décalage entraîne la synthèse d'une large polyprotéine Pr160GagPol contenant les protéines codées par *gag* et *pol* [22]. Le clivage par la protéase virale PR de la partie codée par le gène ***pol*** aboutit à la production de 3 enzymes :

1) protéase PR ou p10 : enzyme responsable de la maturation des protéines virales par clivage des précurseurs Gag et Gag-Pol

2) transcriptase inverse RT ou p66/p51 : enzyme clé responsable de la synthèse d'ADN proviral à partir du génome viral ARN (transcription inverse). Elle possède donc une activité ADN polymérase dont l'initiation requiert la présence du primer $ARNt^{lys}$, ainsi qu'une activité ARNase H.

3) intégrase IN ou p32 : enzyme nécessaire à l'intégration de l'ADN proviral dans le génome cellulaire

Le gène ***env*** (enveloppe) code également pour une polyprotéine précurseur (gp160). Contrairement à Gag et Gag-Pol, cette dernière n'est pas clivée par la protéase virale mais par une protéase cellulaire (la furine). gp160 est une protéine richement glycosylée qui est clivée en deux glycoprotéines d'enveloppe : la glycoprotéine de surface gp120 ou SU et la glycoprotéine transmembranaire gp41 ou TM.

1) gp120 est la protéine de fusion qui participe directement à l'entrée du virus dans la cellule hôte, via son interaction avec la protéine CD4 et les corécepteurs des chimiokines CCR5 et CXCR4 (voir Introduction, chapitre 1.2.3. et la revue [23]). La protéine gp120 est composée de 5 régions conservées (C1-C5) et de 5 régions hypervariables (V1-V5), toutes fortement

glycosylées. La variabilité observée au niveau de la boucle V3 de gp120 conditionne le passage de souches du type monocytotrope au type lymphocytotrope.

2) gp41 est une protéine composée de trois grands domaines : la queue cytoplasmique, la séquence transmembranaire d'ancrage qui fixe le complexe gp120/gp41 dans la membrane et l'ectodomaine contenant les déterminants essentiels à la fusion membranaire entre le virus et la cellule.

Les protéines régulatrices

La protéine **Tat** (Trans-Activateur de Transcription) à localisation nucléaire est un puissant activateur de la transcription des gènes viraux et cellulaires dont *tat* lui-même. En effet, Tat va favoriser l'activité de synthèse de l'ARN polymérase en interagissant avec la séquence TAR (Tat-responsive element) de l'ARN viral. Cette séquence localisée dans la région du R LTR5' [24, 25] est également un site de fixation pour plusieurs facteurs cellulaires capables de se fixer sur TAR-ADN ou TAR-ARN. Outre son rôle principal dans la régulation de la transcription, Tat est sécrétée par des cellules infectées et favorise ainsi l'immunosuppression des cellules infectées ou non infectées, via la sécrétion de cytokines, l'induction de l'apoptose et la diminution de l'expression des molécules de HLA (Human Leukocyte Antigen) de classe I (voir la revue [26]).

La phosphoprotéine **Rev** (Régulateur de l'Expression des protéines Virales) est localisée dans le noyau des cellules infectées. Elle est responsable de l'export rapide des ARNm viraux non épissés ou partiellement épissés en dehors du noyau, favorisant ainsi la synthèse des protéines structurales aux dépens des protéines régulatrices. En effet, un mécanisme naturel empêche la sortie des ARNm non épissés du noyau de la cellule. La fonction de Rev est d'outrepasser ce mécanisme qui empêcherait l'accomplissement du cycle viral. Suite à l'infection, en absence de Rev, la majorité des ARN viraux est épissée, ce qui induit la synthèse précoce des protéines Tat, Rev et Nef qui sont traduites à partir d'ARN multi-épissé. Plus tardivement, la présence de Rev, va permettre la synthèse des protéines structurales Gag, Pol et Env. Le mécanisme d'action de la protéine Rev fait intervenir une séquence de 204 paires de base, RRE (Rev Responsding Element) [27], localisée dans le gène

env. Rev se fixe sur le RRE des ARNm, provoquant ainsi la formation d'un complexe capable d'interagir avec la machinerie d'exportation cellulaire, permettant ainsi l'export des ARNm non épissés ou partiellement épissés [28].

La protéine **Nef** (Negative Regulation Factor) est synthétisée de façon précoce et incorporée dans le virion mature. Du fait de la description d'une meilleure réplication des virions Nef $^{-/-}$ par rapport aux virions sauvages, cette protéine a été initialement baptisée Nef pour facteur de régulation négative de l'expression virale [29, 30]. Cependant, ce rôle a été controversé [31, 32]. Il a été démontré que Nef est impliquée dans l'internalisation et la dégradation des protéines CD4 par endocytose des protéines de surface vers le lysosome [33]. Ce processus limite la fixation de nouvelles particules virales en surface et donc la sur-infection des cellules infectées. Ce phénomène est néfaste pour la réplication virale, car il favorise la reconnaissance et la destruction de la cellule infectée par les cellules cytotoxiques. De façon similaire, Nef régule négativement l'expression en surface des molécules HLA de classe I [34]. Ce mécanisme altère la présentation des antigènes par la cellule infectée et la protège ainsi d'une destruction par le système immunitaire. Enfin, Nef est un facteur important pour l'expansion virale, grâce notamment aux macrophages : Nef active les macrophages et favorise la sécrétion de cytokines et de chimiokines impliquées dans la chimiotaxie, les réponses inflammatoires et l'apoptose. (voir la revue [35])

Les gènes ***vif***, ***vpr*** et ***vpu*** codent pour des protéines accessoires, non indispensables à la réplication virale mais nécessaires à la maturation et au relargage des particules virales infectieuses [36].

La protéine **Vif** (Viral Infectivity Factor) est essentielle pour l'infectivité du virus. Des mutations de ce gène conduisent en effet à la production de particules virales Δ*vif* qui sont jusqu'à 1000 fois moins infectieuses que celles produites par le virus sauvage [37]. La protéine Vif est impliquée dans le contrôle de l'infectivité virale via son implication dans l'inactivation d'un phénomène de résistance naturelle de la cellule hôte face au VIH, faisant intervenir les protéines APOBEC3G et APOBEC3F. Les protéines APOBEC sont des cytidine-déaminases,

enzymes éditrices d'acides nucléiques induisant la déamination des cytosines en uraciles (C en U) au niveau de l'ADN et de l'ARN. L'activité APOBEC entraîne ainsi l'apparition de nombreuses mutations au niveau du matériel génétique viral. L'ADN viral hyper-muté est dégradé par les enzymes de réparation de l'ADN ou aboutit à la synthèse de protéines aberrantes ; Vif va jouer un rôle antagoniste à cette activité antivirale : la présence de Vif entraîne une diminution de la concentration d'APOBEC et de son incorporation dans les virions matures en favorisant son ubiquitinylation et sa dégradation par le protéasome (voir la revue [38]).

Le gène ***vpr*** (Viral Protein R) code pour une protéine conservée chez VIH-1, VIH-2 et SIV. Le rôle de Vpr dans la pathogenèse du SIDA est indéniable, cependant sa fonction exacte durant l'infection reste toujours un sujet de débat. Il a été montré que Vpr est une protéine multifonctionnelle essentielle pour l'infection des macrophages et dans une moindre mesure pour les autres cellules. De plus, après la fusion entre le VIH et la cellule hôte, Vpr est relarguée et sert de protéine cargo pour le transport du complexe de pré-intégration vers le noyau. Vpr favorise également l'activation des promoteurs transcriptionnels des LTR. Enfin Vpr induit aussi un arrêt au stade G2 de la division cellulaire ce qui permet au virus de se répliquer de manière plus efficace.

La protéine **Vpu** est impliquée dans plusieurs phénomènes qui renforcent la pathogénécité virale : la stimulation du relargage des particules virales et la dégradation des protéines CD4. Après l'infection des cellules T, le virus doit faire face à un problème : les protéines CD4 et gp120 sont synthétisées dans le réticulum endoplasmique de la même cellule et peuvent donc interagir et former un complexe. Ce complexe est ciblé par la cellule pour la dégradation et ainsi empêcher la production de virus. Vpu contrecarre ce problème en favorisant la dégradation des protéines CD4. D'autre part, Vpu stimule le relargage des protéines virales, cependant le mécanisme n'est pas encore totalement élucidé. Enfin, Vpu participe à la diminution de l'expression des molécules HLA de classe I en surface de la cellule, limitant ainsi la reconnaissance des cellules infectées par les cellules T cytotoxiques (voir la revue [39]).

1.2.3. Le cycle de réplication virale

La Figure 4 représente les étapes du cycle de réplication virale (voir les revues [40, 41]).

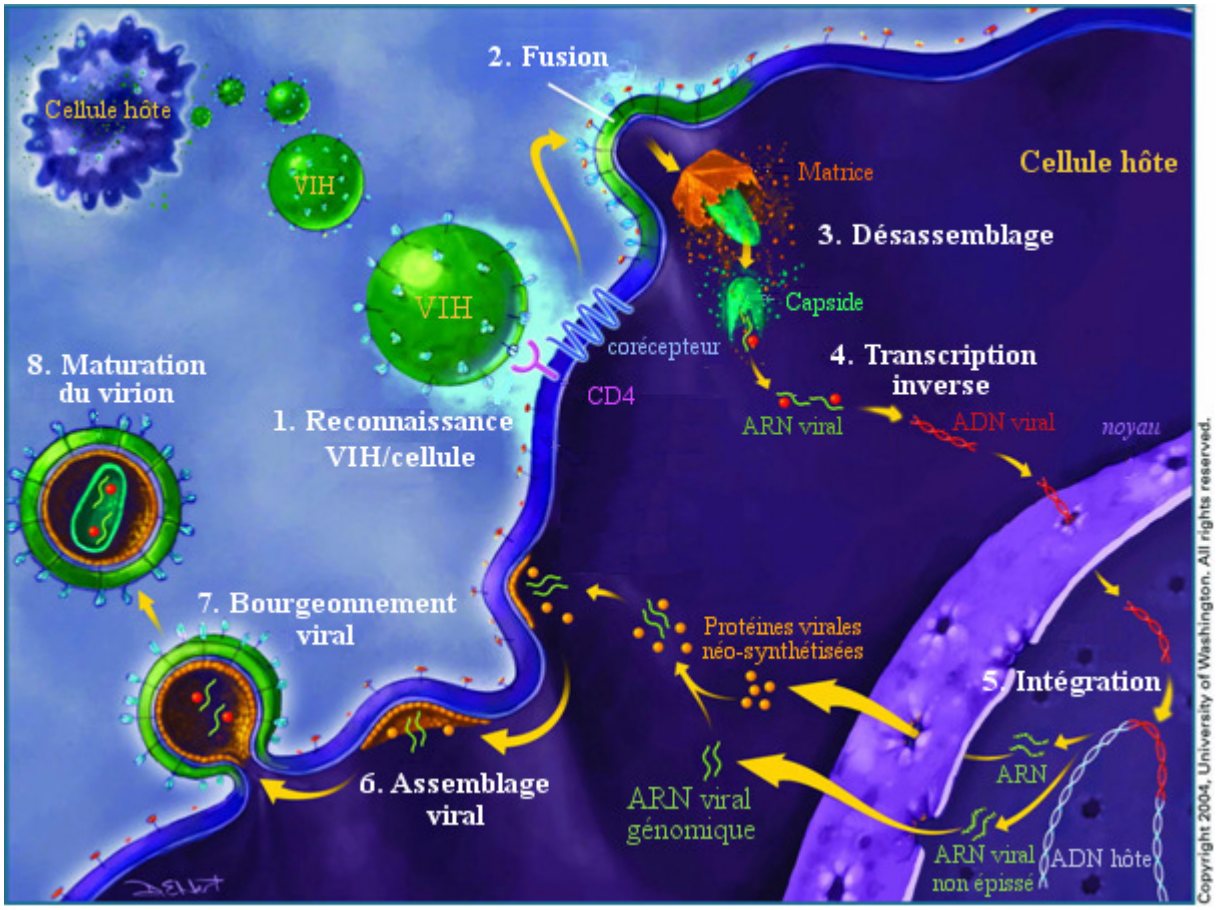


Figure 4 : Cycle de réplication du VIH-1.

Lors de la première étape du cycle de la réplication virale, le virus se lie à la cellule hôte, entraînant la fusion des membranes virales et cellulaires (voir la revue [23]). Cette étape engage la protéine gp120 du VIH-1, la protéine cellulaire CD4 et les récepteurs de chimiokines CCR5 et CXCR4, corécepteurs du VIH-1. En effet, la molécule CD4 a été identifiée comme le récepteur membranaire primaire pour le SIV/VIH, devenant ainsi le premier récepteur viral connu [42, 43]. Un petit segment du domaine extracellulaire N-terminal de CD4 est à l'origine de sa liaison à la queue V1/V2 de gp120 (Figure 5). Cette interaction provoque un changement conformationnel de gp120, exposant sa boucle V3 préalablement enfouie et permettant son association sélective avec un récepteur de chimiokines CCR5 ou CXCR4 [44]. Le complexe ternaire formé entre gp120, CD4 et le corécepteur entraîne un

changement conformationnel de gp41 et la fusion entre les bicouches lipidiques de l'enveloppe virale et de la cellule hôte.

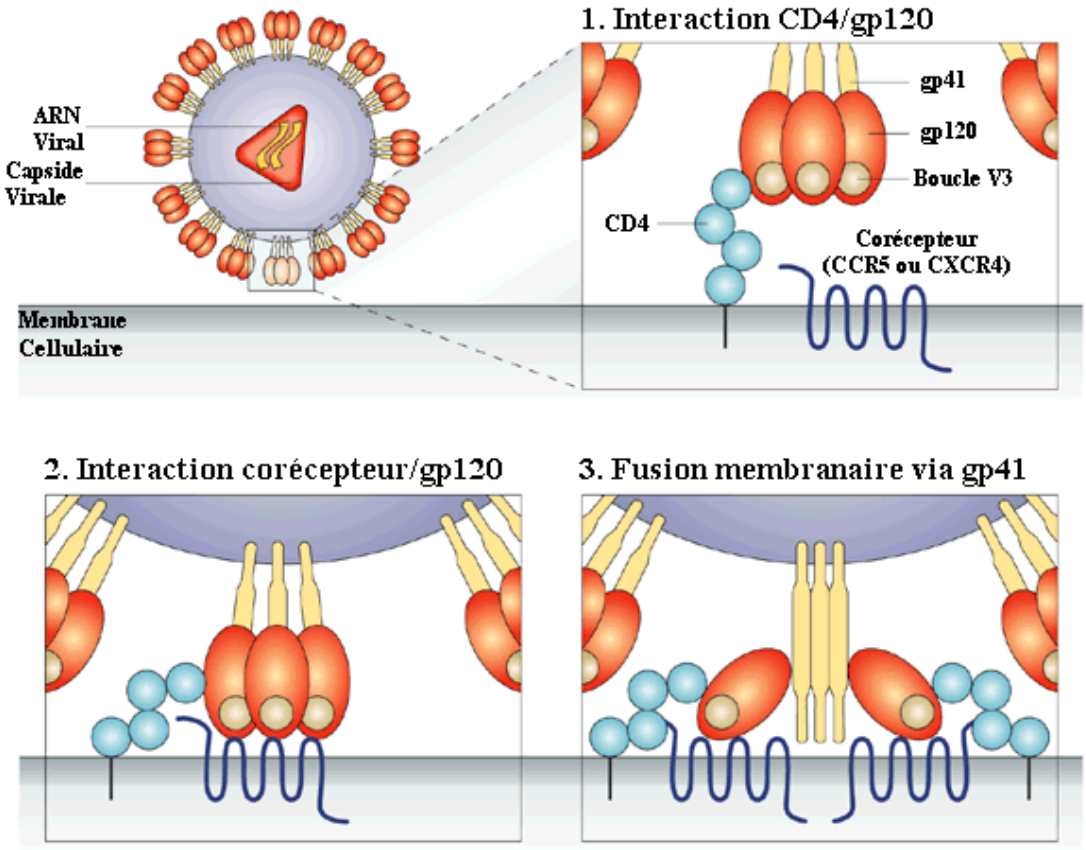


Figure 5 : Représentation schématique des étapes d'entrée du VIH-1 dans la cellule hôte.

La haute variabilité de la boucle V3 de gp120 est responsable du tropisme cellulaire, qui va dépendre de l'affinité différentielle de cette boucle pour les corécepteurs CCR5 et CXCR4 [44]. Le récepteur CCR5 des α-chimiokines RANTES, MIP1α et MIP1β sert de récepteur pour les souches R5 monocyto-tropiques du VIH-1 [45]. Le récepteur CXCR4 de l'α-chimiokine SDF-1 sert de corecepteur pour les souches X4 lympho-tropique [46]. D'autres récepteurs de chimiokines peuvent servir de corécepteurs mineurs à l'entrée du VIH (CCR2, CCR3, CXCR6 GPR15...) [47].

Une fois la fusion opérée, l'ARN génomique viral est libéré hors de la capside dans le cytoplasme cellulaire et la transcription inverse peut avoir lieu. L'ADN proviral peut alors s'intégrer au génome de la cellule hôte ou rester sous forme épisomale : cette étape dépend de

la disponibilité des protéines virales Vpr et intégrases, qui assurent respectivement l'import nucléaire de l'ADN viral [48] et l'intégration dans l'ADN cellulaire [49]. L'intégration serait un processus non aléatoire qui se produirait préférentiellement dans les introns des gènes actifs et plus particulièrement dans les gènes activés après l'infection [50, 51].

Une fois l'intégration réalisée, le virus reste à l'état latent jusqu'à l'activation de la cellule infectée. L'activation induit la synthèse des ARN viraux par la machinerie cellulaire [52] : la fabrication de nouveaux virus est alors initiée. Les ARN viraux sont ensuite exportés vers le cytoplasme, où la synthèse des protéines virales peut avoir lieu. Protéines et ARN viraux néo-synthétisés sont enfin conduits vers la membrane de la cellule pour l'assemblage et le bourgeonnement de nouveaux virus.

A ce stade, le virus est encore immature et non infectieux et nécessite l'intervention de la protéase virale pour achever sa maturation [53]. Le clivage des précurseurs protéiques permet ainsi la libération de milliers de virus matures qui peuvent alors contaminer d'autres cellules.

1.2.4. Tropisme cellulaire

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 : i) la sous population CD4 (dont la déplétion progressive au cours du temps est caractéristique de l'infection VIH); ii) les cellules de la lignée monocytaire-macrophagique (réservoir principal du virus); iii) les cellules dendritiques; iv) 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 [54, 55]. 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* [56, 57].

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 des mécanismes de pénétration du virus dans la cellule, différents de ceux dont nous avons parlé jusqu'ici. Nous pouvons citer en exemple l'entrée dans les cellules (T et macrophages

notamment) du VIH complexé avec des anticorps via les récepteurs du complément et le fragment Fc [58]. Il faut noter que ces processus d'entrée restent limités, comparés au processus médié par CD4.

1.3. Evolution clinique et biologique

La progression de l'infection par le VIH-1 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 6).

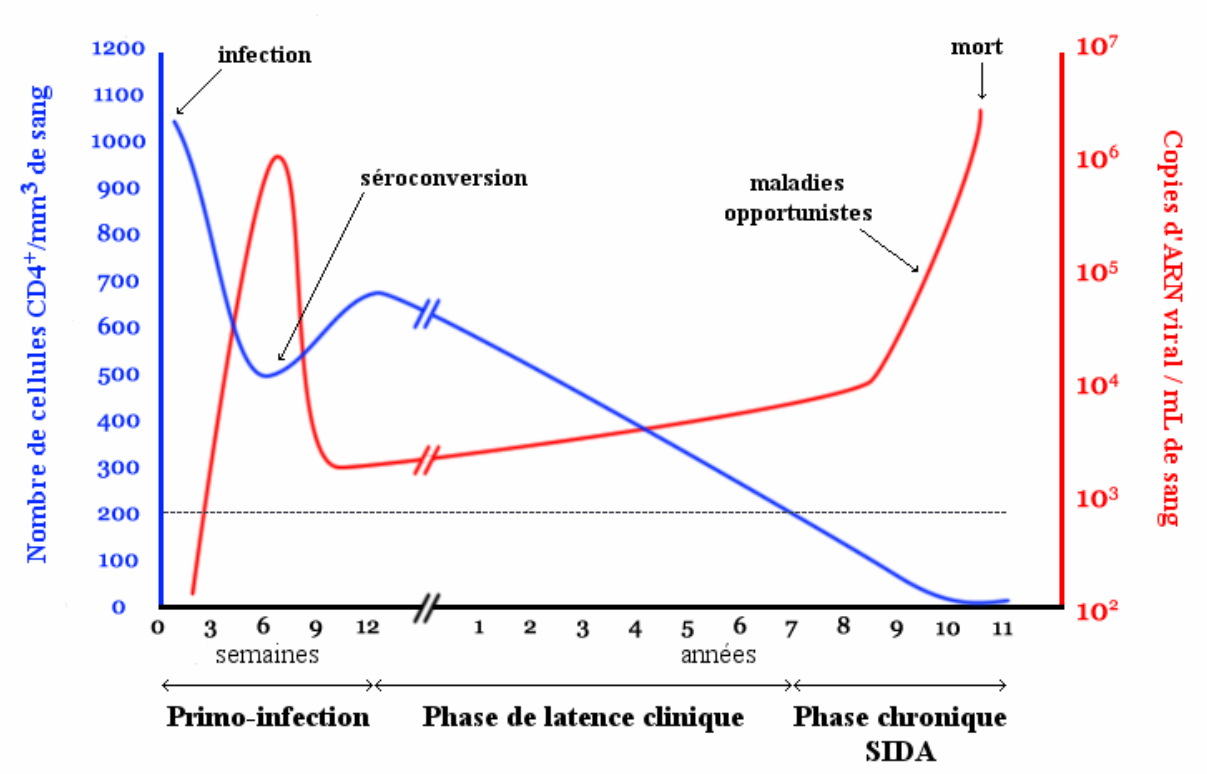


Figure 6 : Profil d'évolution de l'infection par le VIH-1.

1.3.1. 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... Cependant, 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 évènement reflète l'activation du système immunitaire et des réponses humorales et cellulaires, caractérisés 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.

Comme évoqué précédemment (introduction, chapitre 1.2.3), les souches virales isolées à ce stade sont majoritairement R5 monocyto-tropiques, non-cytopathiques et non inductrices de syncytium.

1.3.2. 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 aucun signe clinique lié à l'infection [11]. 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 (Tableau 1), lorsque le seuil de 200 cellules CD4+/mm3 de sang est atteint, le sujet entre dans le stade symptomatique.

1.3.3. Phase symptomatique/SIDA

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.

1.3.4. Réservoirs cellulaires du VIH-1 et persistance virale

Suite à l'infection, la réponse immunitaire initiale a permis la diminution du nombre de particules virales dans le sang et l'entrée des patients dans la phase de latence. Cependant, le virus n'a pas été totalement éliminé et continue à persister dans l'organisme. Les virus responsables d'infections communes (grippe, diarrhée...) sont rapidement éliminés par le système immunitaire. Dans le cas d'une infection par le VIH, le génome viral

rétrotranscrit s'intègre de façon stable dans le génome de cellules à longue durée de vie notamment les cellules T $CD4^+$ dormantes (mémoires ou naïves) et potentiellement dans d'autres types cellulaires (cellules microgliales, dendritiques, progéniteurs hématopoïetiques, NK). La capacité du VIH-1 à persister contribue à l'établissement d'un réservoir à long terme du virus. Du fait de l'absence ou de la faible transcription virale, les cellules réservoirs ne sont pas ciblées par le système immunitaire et survivent pendant de très longues périodes (e.g. 44 mois pour les cellules T $CD4^+$), même lors d'une trithérapie (HAART (Highly Active Anti-Retroviral Therapy), voir introduction, chapitre 1.4.). La réactivation de ces cellules permet la production de virions infectieux et assure la persistance de l'infection virale à long terme. C'est ainsi que l'on peut observer la virémie augmenter chez les patients, moins de deux semaines après l'arrêt du traitement anti-rétroviral [59, 60].

Les macrophages jouent un rôle particulier dans la persistance virale puisqu'ils peuvent passer la barrière hémato-encéphalique et propager ainsi l'infection au système nerveux central, sanctuaire inaccessible au système immunitaire. Les inhibiteurs de transcriptase inverse sont inefficaces sur toutes les cellules réservoirs étant donné que la transcription inverse et l'intégration ont déjà eu lieu, mais les macrophages sont en plus résistants aux inhibiteurs de protéase.

Malgré le développement des traitements qui ont considérablement augmenté la survie et amélioré la qualité de vie des patients, l'éradication du virus chez les patients infectés n'a pas été possible jusqu'à aujourd'hui. Lors de l'émergence des traitements anti-rétroviraux, des estimations optimistes suggéraient une éradication du virus durant les 3 ans de HAART [61]. Actuellement on estime à 60 ans le temps de traitement qui serait nécessaire pour espérer une éradication complète du VIH-1 des patients infectés en assumant le fait que le traitement empêcherait le virus de contaminer de nouvelles cellules [62]. Cette dernière observation souligne l'importance et la nécessité de développer de nouvelles approches pour achever l'éradication du VIH-1 chez les patients infectés.

Tableau 1 : Différents stades de la maladie SIDA selon la classification CDC1993**.**

Catégorie A

Un ou plusieurs des critères suivants chez un adulte ou un adolescent infecté par le VIH et aucun des critères des catégories B et C :

- infection VIH asymptomatique
- lymphadénopathie persistante généralisée
- primo-infection symptomatique

Catégorie B

Une ou plusieurs des manifestations cliniques suivantes chez un adulte ou un adolescent infecté par le VIH et aucune de celles de la catégorie C :

- angiomatose bacillaire
- candidose oropharyngée
- candidose vaginale, persistante, fréquente ou répondant mal au traitement
- dysplasie du col (modérée ou grave), carcinome *in situ*
- syndrome constitutionnel : fièvre (38,5°C) ou diarrhée supérieures à un mois
- leucoplasie chevelue de la langue
- zona récurrent ou envahissant plus d'un dermatome
- purpura thrombocytopénique idiopathique
- listériose
- neuropathie périphérique

Catégorie C

Cette catégorie correspond à la définition du SIDA chez l'adulte. Lorsqu'un sujet a présenté au moins l'une des pathologies suivantes, il est classé définitivement dans la catégorie C :

- candidose bronchique, trachéale ou extrapulmonaire
- candidose de l'œsophage
- cancer invasif du col
- coccidioidomycose disséminée ou extrapulmonaire
- cryptococcose extrapulmonaire
- cryptosporidiose intestinale évoluant depuis plus d'un mois
- infection à CMV (autre que foie, rate, ganglions)
- rétinite à CMV
- encéphalopathie due au VIH
- infection herpétique, ulcères chroniques supérieurs à un mois, ou bronchique, pulmonaire ou oesophagienne
- histoplasmose disséminée ou extrapulmonaire
- isosporidiose intestinale chronique (supérieure à un mois)
- sarcome de Kaposi
- lymphome de Burkitt
- lymphome immunoblastique
- lymphome cérébral primaire
- infection à *Mycobacterium tuberculosis* (intra ou extrapulmonaire)
- infection due à une mycobactérie identifiée ou non, disséminée ou extrapulmonaire
- pneumonie due à *Pneumocystis carinii*
- pneumopathie bactérienne récurrente
- leuco-encéphalite multifocale progressive
- septicémie à salmonelle non typhi récurrente
- syndrome cachectique dû au VIH
- toxoplasmose cérébrale

1.4. Les traitements actuels

Depuis l'introduction de l'AZT en 1987, de nombreux progrès ont été réalisés dans la thérapie anti-rétrovirale notamment avec l'utilisation de la trithérapie ou HAART (Highly Active Antiretroviral Therapy).

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), 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 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 telles 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.

1.5. Profils d'évolution particulière et résistance naturelle au virus

1.5.1. Les LTNP, 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 depuis de nombreuses années ne présentent aucun signe de progression. Ces sujets appelés 'non-progresseurs à long terme' (LTNP) représentent environ 2% des sujets infectés ; ils sont asymptomatiques et présentent un taux de lymphocytes T $CD4^+$ stable sans prise d'aucun traitement anti-rétroviral. Les principaux facteurs influençant la survie à long terme [55] sont :

1) infection par un virus peu réplicatif ou par une souche atténuée avec faible pouvoir réplicatif (mutée pour *nef*),

2) fort taux aux anticorps neutralisants,

3) forte réponse immunitaire cellulaire anti-VIH, notamment via les lymphocytes $CD8^+$ cytotoxiques.

Il existe d'un sous-groupe particulier des LTNP, les '*elite controllers*', qui n'exhibe aucune charge virale ou une charge virale très faible selon les définitions.

1.5.2. Les HEPS, individus hautement exposés et séronégatifs

Il existe un autre profil particulier d'individus face à l'infection par le VIH : en effet, certains sujets après des expositions répétées au VIH ne sont pas infectés : les HEPS (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. Cette résistance pourrait provenir de différents mécanismes :

1) facteurs génétiques entraînant une absence d'expression de CCR5 en surface,

2) anticorps neutralisants,

3) forte réponse antivirale, comme cela été démontré pour les chimiokines [63]

4) réponse immunitaire cellulaire anti-VIH : $CD4^+$ spécifiques, $CD8^+$ CTL, NK et $CD8^+$ non cytotoxiques.

2. Les mécanismes moléculaires de pathogenèse du SIDA

Le système immunitaire de l'hôte réagit face à l'infection par le VIH mais ne parvient pas à éliminer le virus. Le virus a en effet développé de nombreux mécanismes d'échappement au système immunitaire. L'infection par le VIH entraîne ainsi un déficit de l'immunité cellulaire marqué par une baisse de la population lymphocytaire T $CD4^+$, mais également par de nombreuses altérations du système immunitaire, dont les mécanismes ne sont toujours pas complètement élucidés. D'une manière générale, les mécanismes moléculaires de la physiopathogenèse du SIDA restent encore en grande partie à ce jour, un mystère, et les hypothèses sur ces mécanismes sont multiples.

2.1. Effets directs du virus

Dans un premier temps, pour expliquer la baisse de la population T $CD4^+$, les recherches se sont plutôt orientées vers des effets cytopathogènes directs du VIH-1 et des protéines virales.

2.1.1. Effet lytique direct du VIH

En premier lieu, la baisse progressive de la population lymphocytaire T $CD4^+$ pourrait s'expliquer par la lyse directe des cellules infectées [64, 65]. Cependant alors que environ 1% des cellules $CD4^+$ exprime le génome viral chez les sujets encore asymptomatiques, ce taux n'atteint que 10% chez les sujets symptomatiques. Il est donc difficile d'expliquer la disparition des lymphocytes T $CD4^+$ par le seul effet cytopathogène.

2.1.2. Formation de syncytium

En se fixant sur le récepteur CD4 des cellules non infectées, gp120 peut provoquer la fusion entre des lymphocytes $CD4^+$ non infectés et des lymphocytes $CD4^+$ infectés. Il y alors formation de cellules géantes multinucléées ou syncytium. Une observation *in vitro* montre

que les syncytium sont rapidement lysés [66, 67]. Ce phénomène est observé *in vitro*, mais l'existence d'un tel mécanisme *in vivo* reste à démontrer [68].

2.2. Effets indirects viraux-induits

Dans un deuxième temps, les effets cytopathogènes directs du virus n'étant pas suffisants pour expliquer la baisse de la population lymphocytaire T $CD4^+$, les recherches se sont élargies aux effets induits par le virus sur le système immunitaire qui pourraient causer le déficit immunitaire.

2.2.1. Autoimmunité

Les lymphocytes T cytotoxiques (CTL) jouent un rôle prépondérant dans la lutte antivirale de par leur capacité à lyser les cellules cibles. Les cellules $CD4^+$ non infectées peuvent adsorber des molécules gp120 virales solubles et circulantes, et être ainsi reconnues comme étrangères et par conséquent être détruites par les CTL [69 , 70]. Dans cette perspective, les CTL pourraient contribuer directement à l'amplification de l'immunodéficience en détruisant les lymphocytes T $CD4^+$, les lymphocytes B, et les macrophages portant des antigènes du virus à leur surface. Les réponses immunes à médiation cellulaire ont également un effet délétère. Des réponses de type ADCC dues à des anticorps dirigés contre les glycoprotéines gp120 et gp41 du VIH-1 pourraient aussi être responsables de la destruction des lymphocytes $CD4^+$ non infectés qui auraient adsorbé des molécules gp120 [71]. De plus, les lymphocytes non infectés pourraient être détruits par une réponse anti-VIH-1 du fait de l'existence de mimétisme moléculaire des protéines du VIH-1 avec les protéines humaines au niveau de leurs structures primaire et secondaire. En effet, des similitudes de structure secondaire ont été identifiées entre la gp120 et i) les molécules HLA propres à donner lieu à une réaction autoimmune de type allogénique [72 , 73], ii) l'enveloppe du VIH-1 présente des identités peptidiques avec la molécule CD4 (pentapeptide SLWDQ) et la protéine Fas/APO-1 (peptides VEINCTR et FYCNST) [74]. Plus récemment il a été montré que l'infection par le VIH induit l'expression du ligand NKp44L via un motif linéaire NH2-SWSNKS-COOH de la protéine gp41. Les cellules $CD4^+$ exprimant le ligand NKp44L sont très sensibles à l'activité cytotoxique des cellules Natural Killer (NK), et ainsi les NK pourraient intervenir dans le processus d'autoimmunité [75]. Plusieurs hypothèses ont été émises pour expliquer l'induction de l'autoimmunité [76]: i) dérégulation des sous-populations lymphocytaires Th1 et Th2, entraînant une surproduction de cytokines ayant une capacité à induire un processus

autoimmun (par exemple, l'IL10, une cytokine de type Th2) [77]; ii) cytopathogénicité induite par le rétrovirus; iii) hypergammaglobulinémie polyclonale; iv) présence en grande quantité d'anticorps autoréactifs. De plus, il existe un modèle en faveur du développement d'un processus autoimmun du SIDA chez la souris [78].

2.2.2. Les superantigènes

Les superantigènes constituent un groupe de molécules, d'origine bactérienne ou virale, capable d'activer les cellules T de façon non spécifique en se fixant directement sur des sites conservés du CMH (complexe majeur d'histocompatibilité) et du TCR (T-Cell receptor); Les protéines virales du VIH, notamment gp120, pourraient comporter des sites superantigèniques [79, 80]. Cette stimulation induirait la différenciation et la mort des lymphocytes T, de plus elle favoriserait la production de particules virales par des lymphocytes infectés, mécanismes contribuant à la déplétion progressive en lymphocyte T [81, 82]. De façon notable, les superantigènes pourraient mener à un déséquilibre dans le répertoire T, ainsi qu'à une anergie totale (état d'inactivation cellulaire dû à une activation du TCR en l'absence de signal de costimulation), amplifiant l'immunosuppression [54].

2.2.3. Dérèglement de la balance cytokinique Th1/Th2

Les cellules $CD4^+$ effectrices sont séparées en cellule Th1 et Th2 selon le profil de sécrétion cytokinique et les fonctions cellulaires : les cellules $CD4^+$ Th1 sont productrices d'IL2, IFNγ et de TNFβ et associées à une forte inflammation et à l'immunité cellulaire, alors que les cellules $CD4^+$ Th2 sécrètent de l'IL4, IL5, IL6, IL10 et IL13, activatrices de l'immunité humorale. L'IFNγ joue un rôle primordial dans les réponses antivirales et contribue à la mise en place de réponses immunitaires : activation des macrophages, du système du complément, stimulation de l'inflammation en partenariat avec le TNFβ, différenciation des lymphocytes T $CD8^+$ cytotoxiques en coopération avec IL2... Enfin, l'IFNγ s'oppose à la mise en place des réponses Th2, et de la même manière, l'IL4 et l'IL10 secrétées par les cellules Th2 s'opposent à la mise en place des réponses Th1 [54]. Les réponses Th1, importantes pour la réponse antivirale, sont altérées très tôt lors de l'infection VIH au profit des cellules Th2, dans lesquelles le VIH se réplique préférentiellement [83, 84]. Il est important de noter que ce modèle basé sur la dichotomie Th1/Th2 est contesté : (i) une étude a rapporté chez des patients à tous stades de l'infection VIH a) des niveaux d'expression d'IL2 et IL4 quasi-indétectables, et b) des niveaux d'expression d'IFNγ et IL10 stables [85] ; (ii) des cellules T $CD4^+$ n'ayant ni le profil Th1, ni le profil Th2 ont été identifiées (*e.g.* cellules Treg, Th17) [54].

2.2.4. L'apoptose

L'apoptose est un mécanisme physiologique de mort cellulaire. Plusieurs études ont suggéré que la pathogenèse et le stade de progression de l'infection VIH pouvaient être corrélés à une activation de l'apoptose des cellules $CD4^+$ infectées, mais aussi non infectées (effet 'bystander') [86, 87, 88].

Deux voies de signalisation existent, la voie extrinsèque (via les récepteurs de mort) et la voie intrinsèque (via les protéines liées à Bcl-2) :

- Dans le cas de la voie extrinsèque, plusieurs récepteurs de mort sont impliqués comme le récepteur Fas, le récepteur p55 TNFα et les récepteurs TRAIL (TNF-Related Apoptosis-Inducing Ligand). La liaison des ligands (FasL, TNFα, TRAIL) sur ces récepteurs provoque l'activation de la cascade des caspases, protéases contribuant à la mise en œuvre des cytokines, de l'inflammation et de la mort cellulaire.

- Dans le cas de la voie intrinsèque, la mitochondrie joue un rôle clé. En effet, la phase effectrice de l'apoptose comporte l'ouverture des pores de perméabilité des mitochondries. Suite à l'ouverture de ces pores, il y a libération de molécules telles que le cytochrome c et de molécules régulatrices de la famille de bcl-2, soit pro-apoptotiques (Bax et Bid) soit anti-apoptotiques (Bcl-2 et Bcl-x_L).

Lorsque l'on s'intéresse à l'infection par le VIH, la mort accélérée des cellules T peut être liée à différents processus, incluant l'apoptose induite par des protéines virales spécifiques, par l'interaction des protéines d'enveloppe avec CD4 et par l'interaction des cytokines avec des récepteurs spécifiques.

Nous citerons, ci-dessous, un certain nombre de mécanismes impliqués dans l'induction de l'apoptose dans le cas de l'infection par le VIH :

- Les protéines d'enveloppe peuvent causer l'apoptose des cellules infectées et non infectées. La protéine Env exprimée sur la cellule infectée peut interagir avec la molécule CD4 (et éventuellement un corécepteur) disponible sur une cellule voisine et entraîner une fusion totale (formation de syncitia) ou partielle ('baiser de la mort'), menant à la mort des cellules via la surexpression de la protéine pro-apoptotique Bax et la libération consécutive du cytochrome c [54, 68]. La protéine gp120 circulante peut également induire l'apoptose des

cellules non infectées par fixation sur la molécule CD4, causant l'activation de la voie Fas [89, 90] ainsi que l'inhibition de Bcl-2 [91].

- Tat est un trans-activateur de nombreuses voies de signalisation et induit l'apoptose via une surexpression de FasL et de TRAIL et une activation de la voie des caspases [35, 54].

- Nef favorise la mort des cellules infectées et non infectées en activant la voie Fas par production de FasL en surface de la cellule infectée, qui peut interagir avec le récepteur Fas exprimé par les cellules 'bystander' [35, 54, 92].

2.2.5. Activation chronique du sytème immunitaire

Une composante importante de la pathogenèse est l'activation chronique du système immunitaire. Cette activation chronique est accompagnée par la production de cytokines pro-inflammatoires, qui favorisent la production de VIH et l'apoptose des cellules $CD4^+$ et $CD8^+$. L'absence de cette activation chronique chez le non-progresseurs à long terme et dans les infections par SIV non pathogéniques est en accord avec ce concept [54, 56].

Les cellules Th17 sont des cellules $CD4^+$ sécrétrices d'IL17 impliquées dans la médiation de l'inflammation et de l'auto-immunité. Les cellules Th17 sont perdues au niveau de la muqueuse gastro-intestinale lors d'une infection VIH/SIV symptomatique, contrairement à ce qui est observé chez les non-progresseurs à long terme ou chez les singes infectés asymptomatiques. Or, les cellules Th17 préservent l'intégrité de la barrière mucosale. La déplétion en cellules Th17 provoque donc une altération de la muqueuse et favorise la translocation microbienne, à l'origine d'une activation immunitaire chronique et de la progression de l'infection VIH [93].

Plus récemment, la production d' IFNα par les cellules dendritiques plasmacytoides a été impliquée dans l'activation immunitaire [94].

2.2.6. Altération de la présentation antigénique

Comme nous l'avons évoqué précédemment, les protéines virales telles que Nef, Tat et Vpu peuvent induire une diminution de l'expression en surface des molécules HLA de classe I [26, 34, 39] et ainsi protéger les cellules infectées de la destruction par le système immunitaire.

2.2.7. Anticorps anti-VIH

De manière générale, face à une infection virale, les anticorps s'attachent au virus et le neutralisent. Les anticorps neutralisants contre les protéines d'enveloppe gp120 et gp41 du VIH pourraient jouer un rôle majeur dans la réponse immunitaire, cependant le mécanisme de neutralisation dans le cas du VIH reste peu clair.

Les anticorps peuvent aussi être délétères et contribuer à la pathogenèse. En effet, comme nous l'avons vu précédemment, les anti-corps anti-VIH peuvent également participer à la propagation du virus en favorisant son entrée dans les cellules, via les récepteurs du complément et du fragment Fc [54].

3. Etudes génétiques sur cohortes et amélioration de la compréhension des mécanismes de pathogenèse du SIDA

Les hommes se sont intéressés depuis longtemps à leur nature et au fonctionnement de leur organisme. Poussés par la nécessité, nous avons cherché à contrôler les conséquences des maladies sur nos vies. La découverte vers la fin du XX^{ème} siècle, du rôle des microorganismes dans l'étiologie de nombreuses maladies a permis l'émergence des vaccins et des médicaments, permettant ainsi le contrôle d'un grand nombre de maladies. Cependant, dès le XIXe siècle, les médecins avaient constaté, dans les régions d'endémie où les pathogènes persistent très longtemps, qu'il existait toujours une grande variabilité de résistance aux maladies selon les individus, la même pathologie pouvant être mortelle chez les uns, bénigne ou asymptomatique chez les autres. Au cours des années 1920-1950, les données d'épidémiologie génétique ont clairement confirmé que la prédisposition génétique jouait un rôle déterminant dans les maladies infectieuses. La génétique humaine des maladies infectieuses, notamment dans le cas du SIDA, fournit ainsi de nouveaux moyens de diagnostic et de pronostic et ouvre des perspectives préventives et curatives innovantes.

3.1. Introduction et rappels sur les études génétiques

3.1.1. Les polymorphismes de l'ADN

a) Définitions

Le polymorphisme génétique est défini comme l'existence, au sein des individus d'une même espèce, de deux formes (allèles) ou plus d'un locus chromosomique. Généralement, seules les différences observées avec une fréquence supérieure à 1% sont considérées comme des polymorphismes. 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.

b) Les polymorphismes chromosomiques

Les polymorphismes chromosomiques sont des variations structurales liées ou non à des anomalies phénotypiques. Ces variations sont le résultat d'évènements de translocation, inversion, fusion, ou fission de fragments chromosomiques. On peut aussi observer des anomalies portant sur le nombre de chromosomes (*e.g.* trisomie 21) (Figure 7).

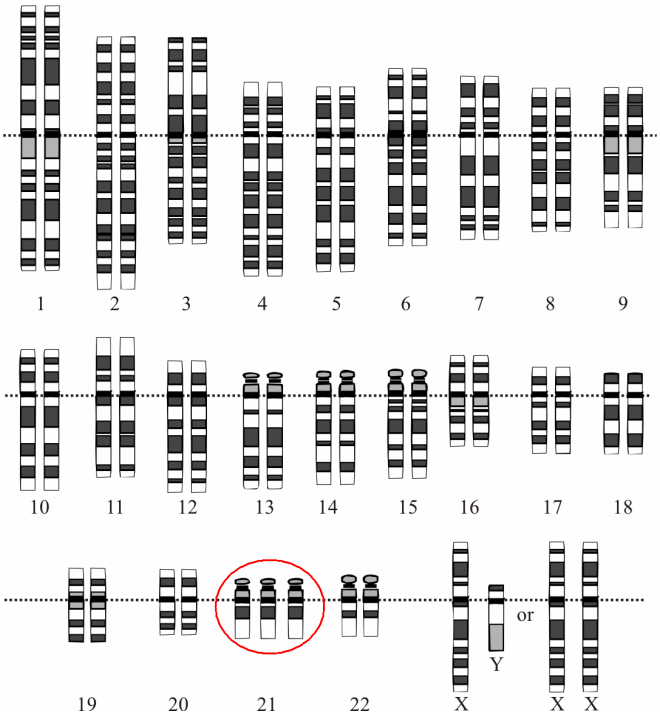


Figure 7 : Représentation schématique des 23 paires de chromosomes chez l'homme, avec une anomalie du nombre d'exemplaires au niveau du chromosome 21.

c) 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 répétition en tandem d'un motif unitaire de taille également variable. Selon la taille du motif et de la répétition on distingue les satellites, les minisatellites et les microsatellites : les microsatellites sont des motifs de 1 à 5 nucléotides répétés 2 à 50 fois (taille totale < 300 pb) (Figure 8) ; les minisatellites 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.

Individu 1 ...ATCCAG**TATA**CAG... 2 répétitions

Individu 2 ...ATCCAG**TATATATA**CAG... 4 répétitions

Figure 8 : Représentation du polymorphisme d'un microsatellite entre 2 individus

d) Les indels

Les indels correspondent à une insertion ou une délétion de nucléotides dans la séquence d'ADN (Figure 9).

Individu 1 ...ATCCAGTCAG...

Individu 2 ...ATCCAGT**CA**CAG...

Figure 9 : représentation d'une insertion de deux nucléotides chez 1 individu

e) Les SNPs

Les SNPs (Single Nucleotide Polymorphism) sont des variations ponctuelles d'un seul nucléotide (Figure 10). Ils sont répartis sur l'ensemble du génome humain et constituent la forme la plus abondante de variations génétiques. Leur nombre est estimé à environ 10 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 l'actuel marqueur génétique de prédilection.

Individu 1 ...ATCCAG**T**CAG...

Individu 2 ...ATCCAG**A**CAG...

Figure 10 : Représentation du polymorphisme d'un nucléotide entre 2 individus

f) 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.

3.1.2. Déséquilibre de liaison et haplotype

a) Notion de déséquilibre de liaison

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 2 loci biallélіques :

- Soit 2 loci :

Au locus A les allèles a_1 et a_2 de fréquence f_{a_1} et f_{a_2}

Au locus B les allèles b_1 et b_2 de fréquence f_{b_1} et f_{b_2}

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

$$\begin{cases} fa_1b_1 = fa_1 \text{ x } fb_1 \\ fa_1b_2 = fa_1 \text{ x } fb_2 \\ fa_2b_1 = fa_2 \text{ x } fb_1 \\ fa_2b_2 = fa_2 \text{ x } fb_2 \end{cases}$$

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 biallélіques :

- Soit 2 loci :

Au locus A les allèles a_1 et a_2 de fréquence f_{a_1} et f_{a_2}

Au locus B les allèles b_1 et b_2 de fréquence f_{b_1} et f_{b_2}

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

$$\begin{cases} f_{a_1b_1} = f_{a_1} \times f_{b_1} - D \\ f_{a_1b_2} = f_{a_1} \times f_{b_2} + D \\ f_{a_2b_1} = f_{a_2} \times f_{b_1} + D \\ f_{a_2b_2} = f_{a_2} \times f_{b_2} - D \end{cases}$$

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 :

$$D' = \begin{cases} \dfrac{D}{\min(f_{a_1b_1}, f_{a_2b_2})} & \text{pour } D \geq 0 \\ \dfrac{D}{\min(f_{a_1b_2}, f_{a_2b_1})} & \text{pour } D < 0 \end{cases} \qquad r^2 = \frac{D^2}{f_{a_1} f_{a_2} f_{b_1} f_{b_2}}$$

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_1 et b_1 sont systématiquement co-transmis -et parallèlement a_2 et b_2-, et seules les combinaisons a_1b_1 et a_2b_2 existent).

b) Notion d'haplotype

Un haplotype correspond à la combinaison d'allèles de 2 ou plusieurs loci polymorphes sur le même chromosome.

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 11).

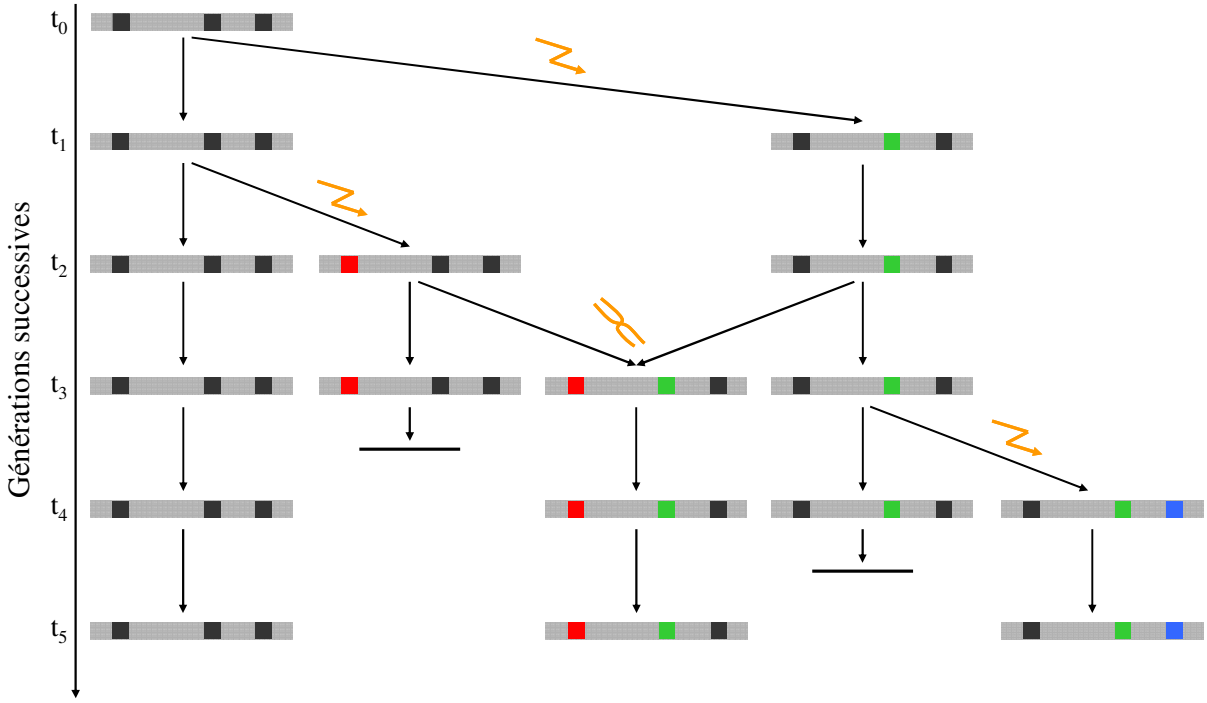


Figure 11 : Représentation schématique de la genèse des haplotypes de 3 SNPs au cours de l'évolution prise à 6 temps précis (t_0-t_5). Les évènements successifs de mutation (⚡) et de recombinaison (⚡) permettent la création de nouveaux SNPs et de nouveaux haplotypes. Les phénomènes de pression de sélection 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 respectivement.

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 de déterminer la phase entre allèles des polymorphismes, en d'autres termes déterminer la combinaison d'allèles au sein d'un chromosome (Figure 12).

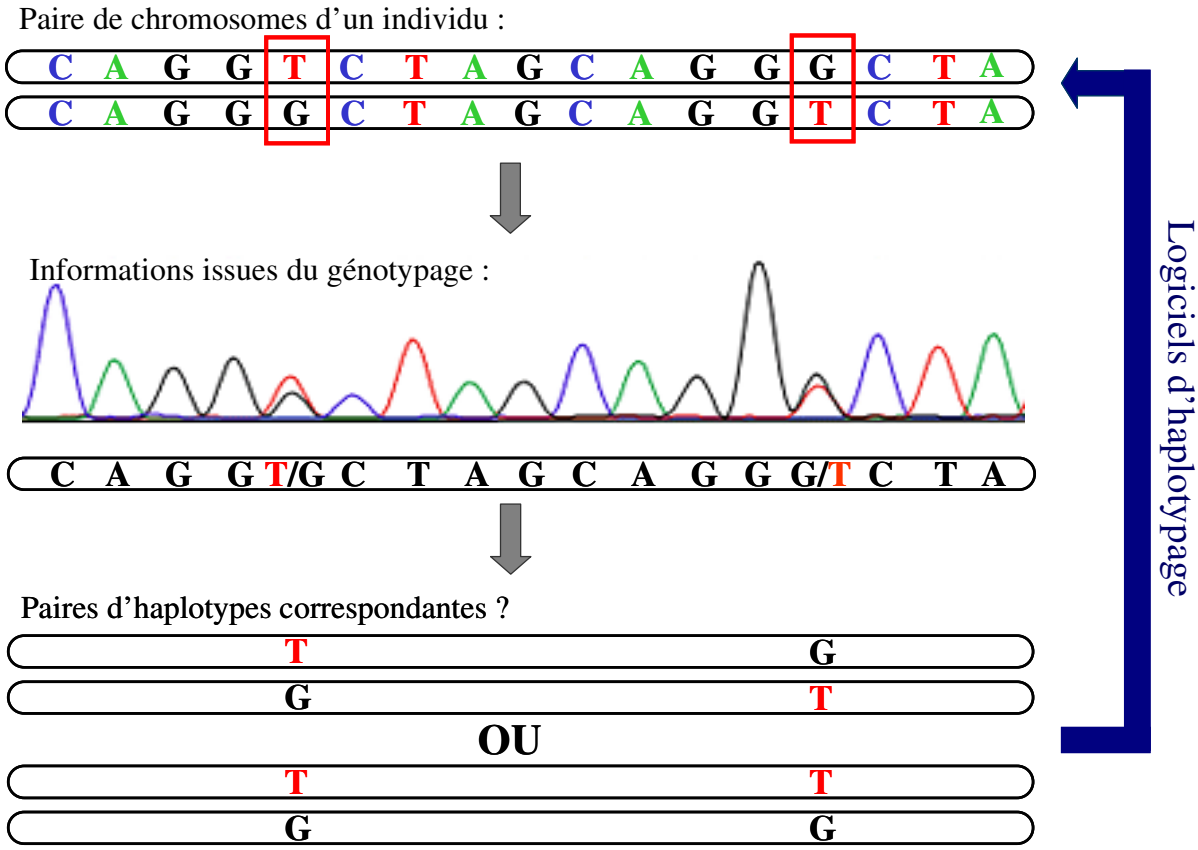


Figure 12 : Problématique de l'haplotypage.

A partir des données issues du génotypage, 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.

3.1.3. Analyse de liaison *vs* Analyse d'association

a) Analyse de liaison

. Les analyses de liaison génétique ont pour but de localiser une région chromosomique présentant une coségrégation 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 [95]. En revanche, ces études sont plus limitées pour la détection des facteurs génétiques impliqués dans les maladies multifactorielles : (i) chaque facteur n'explique qu'une fraction du phénotype, (ii) ces

pathologies ne touchent pas nécessairement plusieurs membres d'une même famille (*e.g.* SIDA).

b) Analyses d'association

Les études d'association cherchent à déceler une association entre un variant génétique et la maladie, à un niveau populationnel et non plus familial uniquement. L'association peut-être directe si le polymorphisme observé est un locus de susceptibilité, ou indirecte si celui-ci se trouve physiquement proche du locus de susceptibilité et que leurs allèles sont statistiquement associés en raison du déséquilibre de liaison. Les études d'association classiques sont des analyses des populations de type cas/contrôle où l'on compare la fréquence d'un allèle particulier entre un échantillon de sujets non apparentés atteints et un échantillon de sujets non apparentés non atteints (Figure 13). Il existe un autre type d'études d'association où l'on s'intéresse à la répartition d'un allèle en fonction d'un phénotype quantitatif au sein d'une population de sujets non apparentés tous atteints.

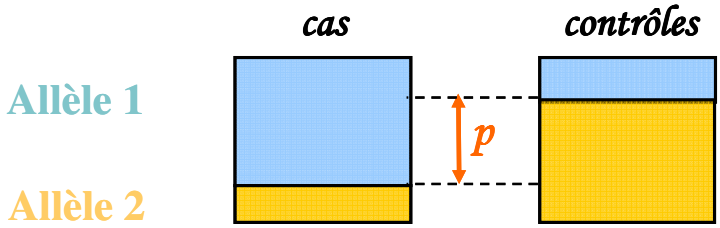


Figure 13 : Représentation schématique du principe d'étude génétique d'association dans le cadre d'une analyse transversale 'cas-contrôle'. *p* = *p-value* = mesure statistique de significativité de l'association.

Dans les deux types d'études, l'évaluation d'une association entre allèle et phénotype se fait par une analyse statistique. La significativité statistique de l'association est évaluée par la mesure '*p*' de *'p-value'*, qui représente la probabilité que le résultat observé soit dû au hasard. Plus concrètement, plus la *p-value* obtenue est faible, moins il y a de risques que l'association observée soit due au hasard et plus il y a de chances qu'elle reflète une réalité biologique. Classiquement, le seuil empirique de significativité de 0,05 (5%) est utilisé lors d'un test statistique.

A première vue, le recrutement de cas et de témoins 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. Il peut en effet, s'avérer difficile d'obtenir les parents de chaque patient, en particulier, lorsque la maladie se développe à un âge avancé. Cependant, le choix

des cas et des témoins peut également soulever certains problèmes. En l'occurrence, il est nécessaire d'assurer l'homogénéité des deux groupes en prenant soin de les assortir sur des covariables telles que l'âge ou le sexe, qui peuvent avoir une influence sur le phénotype observé et ainsi biaiser le facteur génétique que l'on cherche à détecter. Si certains biais comme le sexe ou l'âge sont aisément contrôlables lors du recrutement des individus, d'autres tels que l'origine ethnique le sont plus difficilement.

3.1.4. Approche 'gène candidat' *vs* approche 'génome entier'

a) Approche 'gène candidat'

Les approches 'gène-candidat' consistent à sélectionner un ensemble de gènes dont les fonctions pourraient intervenir dans l'étiologie de la maladie à étudier, et à les tester 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. Même lorsque les connaissances a priori sont larges et que la physiopathologie de la maladie est relativement bien comprise, l'approche 'gène-candidat' n'identifiera qu'une fraction des déterminants génétiques. Ce type d'approche est donc inadapté pour appréhender de manière exhaustive et sans a priori les causes génétiques des maladies. Ce type d'approches a permis de découvrir un grand nombre de gènes impliqués dans des maladies notamment dans le cas du SIDA (voir Introduction, chapitre 3.2.2.).

b) Approche 'génome entier'

Une étude d'association 'génome entier' (ou systématique) investigue une grande partie du génome sans aucun a priori sur l'identité des loci impliqués. Cette approche représente une stratégie impartiale, non dirigée 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 utilisée pour des études de liaison à l'aide des microsatellites 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 multifactorielles, 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 pour Risch et Merikangas [96] l'avenir des études génétiques des maladies complexes.

3.1.5. Emergence des études d'association 'génome entier'

a) HapMap

De part leurs propriétés, les SNPs sont devenus un enjeu majeur dans le développement des études d'associations 'génome entier'. 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, et de plus, ils sont très nombreux et représentent une grande partie de la variabilité du génome humain (voir Introduction, chapitre 3.1.1.). En 2002, le projet HapMap a été mis en place, il vise à référencer les variants génétiques communs que sont les SNPs (http://hapmap.ncbi.nlm.nih.gov/) [97-99]. Ce projet s'est déroulé en trois phases :

Lors de la **phase I** du projet, ils génotypent ~1 million de SNPs chez 270 individus issus de 4 populations d'origine africaine, asiatique et européenne :

- 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 Etats-Unis originaires d'Europe du Nord et de l'Ouest recrutés par le CEPH (Centre d'Etude du Polymorphisme Humain), composés de 30 trios.

Lors de la **phase II**, ils augmentent la densité de la carte génétique en génotypant ~5 800 000 SNPs supplémentaires chez les mêmes individus.

Lors de la **phase III**, ils augmentent la densité de la carte génétique tout en augmentant le nombre de participants :

- 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 Masaï kenyans 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 (fréquence de l'allèle mineur (MAF) > 5%) qui décrit la nature des SNPs, avec leur localisation dans le génome et 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%). 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 12) [100, 101]. 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.

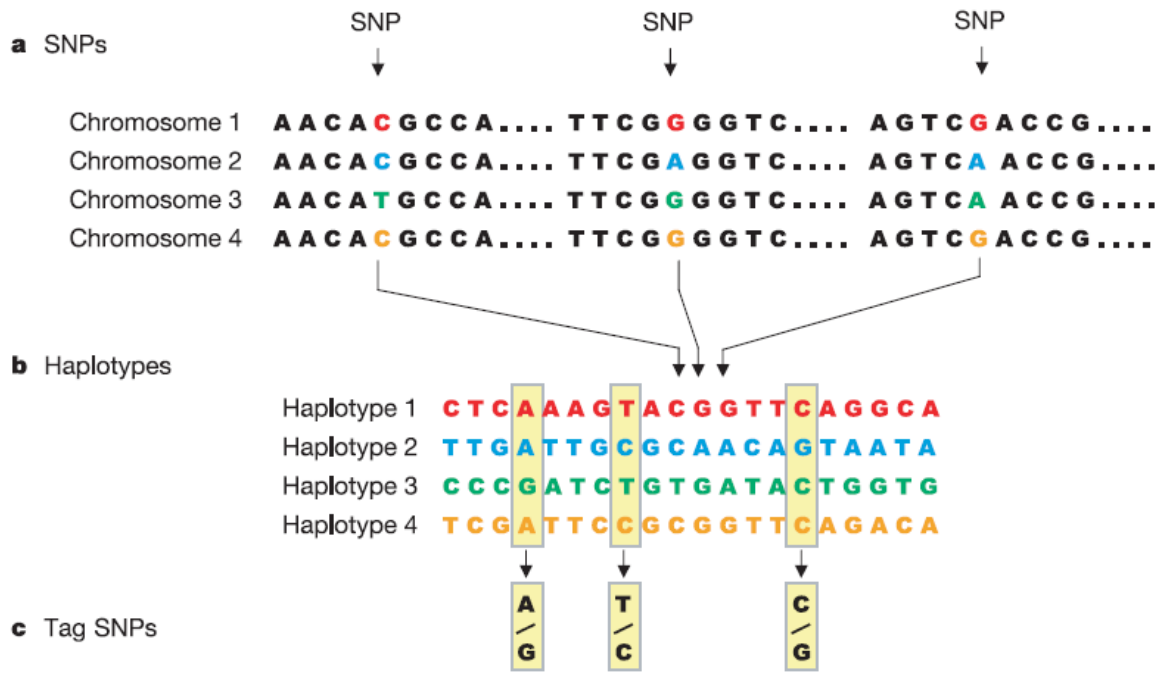


Figure 14 : Illustration de la notion de tagSNP (tirée de [157]). (**a**) Portion de chromosome séquencée dans 4 individus différents. Les SNPs sont représentés en couleur. (**b**) A partir des SNPs, détermination des haplotypes de la région composée ici de 20 SNPs, dont les 3 SNPs du panel a. (**c**) 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.

b) Développement de nouvelles technologies

Le développement de HapMap a permis aux généticiens de tirer profit de l'organisation des SNPs au sein des chromosomes. Les chercheurs peuvent alors examiner l'ensemble du génome à partir d'un nombre restreint de tagSNPs (à peu près 500,000) au lieu d'étudier les 10 millions de SNPs qu'il contient.

Dans cette optique, les entreprises Illumina et Affimetrix 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 moindre coût. La grande différence entre Illumina et Affymetrix repose sur la sélection des tagSNPs, le but étant de minimiser les coûts tout en capturant le maximum d'informations. 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 pas moins de 427 publications (http://www.genome.gov/gwastudies) en seulement quelques années.

c) Et maintenant ?

Lors de la conception des analyses d'association 'génome entier', l'hypothèse qu'une maladie commune est associé à un variant commun [96] était le paradigme prévalent. Le projet HapMap s'était alors axé sur les SNP avec des MAF > 5%, et en conséquence les outils de génotypage mis au point jusqu'ici étaient capables de capturer une fraction du spectre réel des variations génétiques L'implication des SNPs de faible fréquence dans les maladies complexes est aujourd'hui fortement soutenue [102 , 103 , 104]. Dans le but de faciliter l'analyse et la recherche de ces SNPs de faible fréquence, un nouveau projet a vu le jour : le projet 1000 génomes (http://www.1000genomes.org/) dont l'objectif est de séquencer la totalité du génome de 2500 individus originaires de 28 populations différentes, afin d'établir une cartographie fine des variations génétiques (SNP, CNV, indel) du génome humain ayant une fréquence de moins de 1%. Ces nouvelles informations vont permettre de développer des protocoles adaptés et des outils informatiques pour l'analyse des ces variations. Il existe déjà de nouvelles puces de génotypage, plus denses basées sur les données du projet HapMap et sur une partie des données du projet 1000 génomes. Ces puces permettent de cibler jusqu'à 2,5 millions de marqueurs génétiques dont les SNPs de faible fréquence. En conséquence, elles offrent une meilleure couverture de la diversité génétique et permettent aussi l'analyse d'un spectre de fréquences alléliques plus large. Cependant, les limites statistiques constituent une difficulté dans le cas des études 'génome entier', il est donc important de développer des

solutions alternatives pour analyser ce type de données. L'utilisation des nouvelles technologies que constituent les puces de génotypage de forte densité pourrait permettre la découverte de nouveaux variants génétiques impliqués dans l'infection et la progression du VIH-1, et en particulier les polymorphismes de faible fréquence dont l'importance dans les maladies complexes est de plus en plus considérée [102, 104, 105].

3.2. SIDA et génétique

Le SIDA est une maladie infectieuse dont l'évolution est influencée par des facteurs environnementaux (*e.g.* co-infections par des virus hépatite ou herpès) et génétiques viraux (*e.g.* souches X4 plus pathogènes que les souches R5). Cependant, comme nous l'avons évoqué au début de ce chapitre, l'observation de phénotypes hétérogènes en réponse à l'infection par le VIH, et l'existence des populations LTNP et HEPS suggèrent très fortement l'implication des facteurs génétiques de l'hôte dans l'évolution de cette maladie. De ce fait, la recherche en génétique humaine face à l'infection par le VIH s'est développée avec la mise en place de nombreuses cohortes. Les premières analyses génétiques portaient sur des approches du type 'gène candidat' qui se focalisent sur un gène particulier dont on suppose qu'il pourrait jouer un rôle dans la maladie. Nous verrons que les avancées technologiques ont permis le développement des approches 'génome entier' dans le cadre du SIDA.

3.2.1. Les cohortes

A ce jour, la plupart des cohortes existantes sont constituées de patients à tout stade de la maladie non spécifiquement choisis quant à leur type de progression (lente, normale ou rapide). Ces cohortes servent donc de base à des études sur des traits quantitatifs. A titre d'exemple, nous pouvons cité les cohortes ACS (Amsterdam Cohort *Studies*), MACS (Multicenter Aids Cohort Study), SHCS (Swiss HIV Cohort Study)

Après quelques années de recul, le suivi des sujets séropositifs a permis de distinguer trois profils d'évolution face à l'infection par le VIH-1 :

- Les non-progresseurs à long terme (incluant les '*elite controllers*')
- Les progresseurs rapides
- Les sujets non infectés malgré de nombreuses expositions au virus.

La stratégie consistant à trier les patients au départ afin de comparer les facteurs génétiques des populations à phénotypes extrêmes a pu ainsi voir le jour. Cette approche transversale est plus informative qu'une approche classique car si des différences existent,

elles se verront a fortiori plus facilement en comparant les extrêmes. Les cohortes extrêmes, du type de la cohorte GRIV, sont très efficaces pour les approches de génomique de type 'haut-débit' car, sur le plan statistique et économique, elles permettent d'analyser moins de patients sur plus de gènes [106].

Dans le contexte actuel avec le développement des analyse 'génome entier', ces cohortes de patients à profils extrêmes se développent et particulièrement celles composées d'*'elite controllers'*. Ces patients étant capables de contrôler spontanément la réplication du VIH, la découverte des mécanismes sous-jacents permettrait d'envisager le développement de nouveaux vaccins.

3.2.2. Associations génétiques identifiées dans le SIDA par les approches 'gène candidat'

L'approche génétique classique se focalise sur l'étude des polymorphismes d'un gène suspecté d'intervenir dans la pathologie étudiée. Dans le cadre du SIDA, maladie virale caractérisée par un affaiblissement du système immunitaire, les gènes ciblés prioritairement ont été les gènes participant au cycle de réplication virale et les gènes de l'immunité : reconnaissance de l'antigène (*e.g.* système HLA), cytokines et récepteurs... Ce type d'approche a permis de décrire de nombreuses associations génétiques dans l'infection VIH, dont les principales sont résumées dans ce chapitre (voir également les revues [107, 108]).

a) Gènes intervenant dans le cycle de réplication du VIH-1

Entrée du virus dans la cellule

Le récepteur de chimiokines CCR5 est un corécepteur du VIH pour son entrée dans la cellule. Une délétion de 32pb a été identifiée au niveau de la région codante [109, 110] : cette délétion aboutit à la synthèse d'une protéine tronquée, non fonctionnelle et non transportée en surface de la cellule. L'allèle **CCR5-Δ32** n'a été observé que dans les populations d'origine européenne. Les individus homozygotes pour la délétion (**Δ**32/**Δ**32) sont protégés de l'infection par les souches R5 et représentent ~1-2% de la population européenne. Les individus hétérozygotes **Δ**32/WT présentent une expression de CCR5 en surface cellulaire diminuée, une réplication virale et une progression vers le SIDA ralenties.

Un autre polymorphisme de *CCR5* a été découvert au niveau du promoteur : l'allèle **CCR5-P1** est un haplotype composé de 10 SNPs qui augmenterait l'expression de *CCR5* et qui est associé à une progression plus rapide vers le SIDA [111, 112].

Le gène *CCR2* est localisé à proximité du gène *CCR5* (14kb) et code pour un récepteur de chimiokine qui peut servir de corécepteur mineur pour l'entrée des souches R5 du VIH. Le variant exonique **CCR2-64I** a été impliqué dans une progression vers le SIDA retardée [113], mais l'explication fonctionnelle reste peu claire : (i) ce variant protéique pourrait interagir avec la protéine CCR5 et favoriser sa séquestration intracellulaire ; (ii) il pourrait interagir avec CXCR4 et retarder la transition CCR5→CXCR4 dans l'usage du corécepteur, étape clé pour la déplétion des lymphocytes T $CD4^+$.

Les ligands des récepteurs de chimiokines peuvent bloquer l'entrée du VIH par compétition directe avec le virus pour la liaison au corécepteur, et/ou par réduction de l'expression du corécepteur en surface due à son internalisation. Ainsi, le polymorphisme **SDF1-3'A** du gène *SDF-1* (ou *CXCL12*), codant pour le principal ligand de CXCR4, a été associé à une progression vers le SIDA retardée [114]. L'explication fonctionnelle de cette association reste floue : (i) ce variant pourrait accroître la transcription du gène *SDF-1* et retarder ainsi la transition CCR5→CXCR4 dans l'usage du corécepteur ; (ii) ce variant pourrait agir en synergie avec CCR2.

Ces quatre associations ont été largement étudiées et confirmées dans de nombreuses cohortes (voir les revues [115, 116]), dont la cohorte GRIV [114, 117]

Le polymorphisme intronique **In1.1C** du gène *RANTES* (ou *CCL5*), codant pour le principal ligand de CCR5, a été corrélé à une progression accélérée dans les populations d'origine européenne et africaine [118]. Ce polymorphisme, localisé dans un élément régulateur, favorise une diminution de la transcription de *RANTES*, laissant ainsi CCR5 non occupé et disponible pour la liaison avec le VIH.

Une région du chromosome 17, riche en gènes codant pour des ligands de récepteurs de chimiokines, a également été étudiée : CCL2 (ou MCP1), ligand de CCR2 ; CCL7 (ou MCP3) et CCL11 (eotaxin), ligands de CCR3. L'haplotype 7 de cette région ***CCL2-CCL7-CCL11*** a été associé à la susceptibilité à l'infection [119]. Cet haplotype stimule en effet la migration des cellules immunitaires vers les sites d'infection, et contribue ainsi à la propagation du virus.

Désassemblage de la capside virale

TRIM5α est un facteur cellulaire de restriction du VIH qui interagit avec les protéines de la capside virale et favorise ainsi son désassemblage prématuré. Ce processus d'inhibition espèce-spécifique a été initialement identifié chez les primates non humains, et s'est par la

suite avéré d'une efficacité plus limitée chez l'homme. Certains polymorphismes du gène *TRIM5α* humain (SNPs et haplotypes) impliquant les acides aminés R136Q et H43Y ont été associés à une modulation du pouvoir antiviral de TRIM5α *in vitro*, mais les données *in vivo* suggèrent seulement un effet modeste de ces variants sur l'issue de la maladie [120].

La cyclophiline A (CypA) est impliquée dans le repliement des protéines et pourrait également jouer un rôle dans l'immunosuppression médiée par la cyclosporine A. Le SNP **CypA 1650-AA** a été corrélé à une progression plus lente vers le SIDA *in vivo*, mais également à une réplication virale plus faible *ex vivo* [121]. CypA est impliquée dans le désassemblage de la capside et pourrait agir comme un co-facteur de TRIM5α. Cette enzyme participe également à l'assemblage de nouveaux virions via son interaction avec les protéines de la capside.

Transcription inverse

Comme évoqué précédemment (Introduction, chapitre 1.2.2.), la protéine APOBEC3G est une cytidine-déaminase, enzyme éditrice d'acides nucléiques (C en U) responsable de l'apparition de nombreuses mutations au niveau du matériel génétique viral. L'ADN viral hyper-muté devient alors plus sensible à la dégradation, ou donne naissance à des protéines virales aberrantes. Ce processus naturel de résistance au VIH est cependant contrecarré par la protéine Vif qui favorise la dégradation d'APOBEC-3G par le protéasome. Le variant **APOBEC3G-186R** a été associé à une progression accélérée vers le SIDA au sein de la population Africaine [122, 123].

Bourgeonnement viral

Comme décrit ci-dessus, le SNP **CypA 1650-AA**, corrélé à une progression ralentie, est impliqué dans l'assemblage et le désassemblage de la capside, et participe ainsi à la formation de nouveaux virions [121].

Le gène *TSG101* (Tumor Susceptibility Gene) code pour un régulateur négatif du cycle cellulaire. Le polymorphisme **TSG101_-183C** a été corrélé à une progression accélérée, et pourrait également influencer la réplication virale *in vitro* [121]. TSG101 interagit en effet avec la protéine virale p6 et participe ainsi à l'assemblage et au bourgeonnement de nouvelles particules virales.

b) Gènes de l'immunité

HLA

Le HLA_I joue un rôle majeur dans la pathogenèse du VIH-1. En effet, les différents allèles HLA_I lient les antigènes viraux qu'ils présentent aux cellules T $CD8^+$ de façon différente et initient des réponses CTL de force variable. Les molécules HLA_I gouvernent ainsi la réponse immunitaire aux antigènes viraux. La région *HLA* du chromosome 6 a donc été étudiée de façon extensive et de nombreuses associations avec la progression ont été révélées, notamment grâce aux travaux de Kaslow, Carrington, Hendel et Magierowska [124], [125]. Les principaux résultats sont résumés dans la revue [126] et dans le Tableau 2.

Tableau 2 : Résumé des principales associations identifiées dans la région *HLA* avec la progression vers le SIDA.

SUSCEPTIBILITÉ	PROTECTION
Homozygotie HLA-B*35-Px HLA-B*22 Haplotype ancestral B*8.1 + HLA-A*29, HLA-B*53	HLA-B*57 HLA-B*27 HLA-B*14 HLA-A*2/6802

KIR

Outre la présentation des antigènes aux TCR des cellules T $CD8^+$, les molécules HLA de classe I sont également impliquées dans la présentation des antigènes aux récepteurs KIR activateurs et inhibiteurs des cellules NK. Les molécules HLA de classe I influencent donc également l'activation des cellules NK, cellules essentielles de la défense antivirale. Le locus des gènes KIR du chromosome 19 a été investigué et à l'origine de la découverte de nouvelles associations avec la progression (voir la revue [127]).

En présence de son ligand HLA-Bw4, l'allèle **KIR3DS1** (récepteur activateur) facilite la clearance du VIH et est corrélé à une progression vers le SIDA plus lente. A l'inverse, en l'absence de son ligand, KIR3DS1 est corrélé à une progression plus rapide [128].

L'homozygotie **KIR3DL1** et l'hétérozygotie **KIR2DL2/KIR2DL3** ont été associées en l'absence de leur ligand respectif, HLA-Bw4 et HLA-C1, à la résistance à l'infection VIH chez des femmes africaines prostituées HEPS [129].

Cytokines et récepteurs

Le système des cytokines et de leurs récepteurs étant un élément fondamental des réponses immunitaires, leurs gènes ont été largement explorés. Les principales associations identifiées sont résumées dans le Tableau 3.

Tableau 3 : Résumé des principaux gènes de cytokines et de leurs récepteurs pour lesquels des polymorphismes ont été associés avec la progression vers le SIDA.

SUSCEPTIBILITÉ	**PROTECTION**	**Effets de protection ET de susceptibilité**
IL1, IL2, IL6, IL10, IL18, IL4Rα TNFα, TNFβ IFNαR1	FNα, IFNβ IL16, IL1Rα	IL4 IFNγ

Le catalogue précédent des associations génétiques identifiées dans le SIDA par les approches 'gène candidat' n'est pas exhaustif. Notre équipe, grâce à la cohorte GRIV, a participé à la découverte de nombreuses [114, 117, 130-136].

3.2.3. Associations génétiques identifiées dans le SIDA par les approches 'génome entier'

Depuis quelques années, la recherche sur le SIDA a pu bénéficier des nouvelles approches 'génome entier'. Durant ma thèse, pas moins de 8 études de ce type ont été publiées dont deux sont l'objet de cette thèse.

a) La Cohorte Euro-CHAVI

La première parait en 2007, cette étude est basée sur la cohorte Euro-CHAVI [137] qui réunit 486 patients séropositifs d'origine européenne (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 : (i) charge virale plasmatique stable au cours de la phase asymptomatique ; (ii) 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.

L'**étude selon la charge virale** a révélé deux signaux passant le seuil de significativité statistique 'génome entier' de Bonferroni (p=9,3x10^{-8}) : rs2395029 (p=9,36x10^{-12}) et rs9264942 (p=3,77x10^{-9}), tous deux dans la région *HLA* du chromosome 6. Le SNP rs2395029 est situé dans le gène *HCP5* (*HLA Complex P5*) et l'allèle rs2395029-G est associé à une charge virale plus faible. Ce polymorphisme est en fort déséquilibre de liaison avec *HLA-B*57*, allèle protecteur précédemment connu et impliqué dans le contrôle de l'infection virale (voir Introduction, chapitre 3.2.2.) [138]: l'effet bénéfique observé peut ainsi être dû à *HLA-B*57* et/ou au variant *HCP5*. *HCP5* est un rétrovirus endogène humain (HERV) présentant des homologies de séquence avec les gènes *pol* [139] et est exprimé dans les lymphocytes [140]. Ce gène constitue donc un bon candidat, qui pourrait interagir avec le VIH via un mécanisme d'ARN antisens. Le SNP rs9264942 est localisé ~35kb en 5' du gène *HLA-C* et l'allèle rs9264942-C est associé à une charge virale réduite. Cet effet protecteur est indépendant de l'effet *HCP5/HLA-B*57* et est corrélé à une expression accrue du gène *HLA-C*.

L'**étude selon le phénotype de progression** n'a permis de mettre en évidence aucun SNP respectant le seuil de significativité statistique. Cependant, les premiers résultats concernent également des SNPs du chromosome 6 au niveau du locus *ZNRD1/RNF39* (p=3,89x10^{-7}). Ces variants représentent des effets indépendants des deux polymorphismes précédents, situés à plus d'1Mb. Le gène *RNF39* est peu caractérisé, et le gène *ZNRD1* code pour une sous-unité de l'ARN polymérase I. Puisque les SNPs identifiés sont associés à la modulation de l'expression de *ZNRD1*, une hypothèse de mécanisme pourrait être une restriction du VIH par ce gène.

Cette première étude 'génome entier' dans le cadre du SIDA a permis de souligner le rôle central de la région *HLA* dans le contrôle de l'infection.

En 2009, le groupe à l'origine de la cohorte Euro-CHAVI, augmente le nombre de patients, passant de 486 à 2554 et publie une seconde analyse 'génome entier' [141], tout en conservant les mêmes phénotypes. Ils espèrent augmenter la puissance statistique et découvrir de nouvelles associations. Nous évoquerons ici les principaux résultats :

L'**étude selon la charge virale** confirme l'importance des SNPs de *HCP5* (rs2305029, p=4.5x10^{-35}) et du *HLA-C* (rs9264942, 5,9x10^{-32}) dans les différences de niveau de charge virale entre les patients. En addition de ces 2 signaux, 86 SNPs passent le seuil de significativité statistique de Bonferroni (p<5x10^{-8}), tous localisés dans la région du CMH. Par des analyses supplémentaires sur 331 SNPs du CMH avec des *p-values* inférieurs à 1x10^{-4}, ils

démontrent un effet indépendant de 4 SNPs de cette région : rs259919, rs9468692, rs9266409, rs8192591. L'exploration des allèles *HLA* à 4 chiffres, selon un modèle de régression prenant en covariables les 6 signaux précédents, a révélé 4 associations supplémentaires : HLA-A*3201, HLA-B*1302, HLA-B*2705, et HLA-B*3502. Au final, ces 6 SNPs et 4 allèles de la région *HLA* expliquent 12% de la variabilité de la charge virale dans cette cohorte. Ce résultat souligne la présence d'effet distinct au sein de ce locus, sans pour autant affirmer l'implication directe des SNPs mis en évidence.

Dans l'**étude selon le phénotype de progression,** les meilleurs résultats sont encore obtenues pour les SNPs de *HCP5* (rs2305029, p=1,2x10^{-11}) et de *HLA-C* (rs9264942, p=6,4x10^{-12}). Ils démontrent cependant que ces effets sur la progression sont en grande partie dus à l'impact de ces SNPs sur la charge virale. D'autres SNPs passent le seuil de siginificativité de Bonferroni : rs9261174, rs3869068, rs2074480, rs7758512, rs9261129, rs2301753 et rs2074479 (p=1,8x10^{-8}). Ces SNPs sont tous en fort déséquilibre de liaison et localisés autour des gènes *ZNRD1* et *RNF39* dans la zone du CMH. Ces associations sont totalement indépendantes de la charge virale et le variant causal n'est pas déterminé.

Enfin, 27 SNPs, précédemment associés au contrôle du VIH-1 lors d'études gènes candidats, ont été génotypés et analysés. Le variant *CCR5-Δ32* est associé très fortement avec la charge virale (p=1,7x10^{-10}) et avec le phénotype de progression (p=3,5x10^{-7}), mais explique seulement 1,7% de la variabilité.

Cette seconde analyse 'génome entier' a permis de réaffirmer l'importance du CMH, mais a aussi démontré l'existence de plusieurs signaux indépendants au sein de ce locus. De plus cette étude a permis de souligner l'importance de *CCR5* dans l'infection par le VIH

b) La cohorte PRIMO

Trois études ont été entreprises et financées par l'ANRS (Agence Nationale de Recherche sur le SIDA) sur plusieurs cohortes françaises, dont 2 sont le sujet de cette thèse. De ce fait nous n'aborderons que l'analyse réalisée sur la cohorte PRIMO (605 patients génotypés), publiée en 2008 [142]. Cette analyse est basée sur deux phénotypes distincts : le niveau d'ARN viral plasmatique et le niveau d'ADN viral dans les PBMC (peripheral blood mononononuclear 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).

Analyse du niveau de l'ARN viral : le seul SNP passant le seuil de significativité de Bonferroni est le rs10484554 ($3,58x10^{-9}$) 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 impliquée dans la régulation de la prolifération et la différenciation cellulaire [143], alors que *KLF3* coderait pour un facteur de transcription [144].

Analyse du 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 (False Discovery rate) avec un taux d'erreur de seulement 1,4%.

Cette analyse a permis de confirmer l'importance du CMH dans le niveau d'ARN viral, mais la nouveauté réside dans la mise en évidence de l'importance de cette région dans le niveau d'ADN viral.

c) La cohorte MACS156

En 2009, une analyse sur la cohorte MACS (Multicenter AIDS Cohort Study) voit le jour [145]. Dans un premier temps, ils ont sélectionné au sein de cette cohorte, un groupe de 156 patients enrichis en individus présentant des profils de progression extrême (progresseurs rapides, non-progresseurs à long terme, progresseurs modérés). Ils ont analysé leurs données de génotypage selon 2 stratégies : (i) Analyse par catégorie en comparant la répartition des génotypes dans les progresseurs rapides, les non-progresseurs à long terme et les progresseurs modérés, (ii) Analyse de la répartition des génotypes en fonction du temps pour atteindre la phase SIDA (AIDS87, la phase SIDA selon la classification du Center for Disease Control datant de 1987). Dans un second temps, ils ont sélectionné un groupe de SNPs pour effectuer une réplication dans une cohorte indépendante de 590 séroconvertis VIH-1

La première étape de cette analyse a permis de confirmer le signal sur le gène *HCP5*. **La réplication dans une cohorte indépendante** a mis en évidence 3 SNPs proches du gène *PROX1* (Tableau 4). *PROX1* code pour une protéine dont la fonction biologique peut facilement être reliée à l'infection par le VIH-1, notamment dans son rôle de régulateur négatif de l'expression de l'IFNγ dans les cellules T [146].

Cette étude 'génome entier' a ainsi dévoilé un nouveau candidat potentiel pour les mécanismes moléculaires de pathogenèse de l'infection VIH.

	SNP		
	rs17762192	rs1367951	rs17762150
Analyse MACS156	$7.13x10^{-5}$	$6.69x10^{-5}$	$3.64x10^{-5}$
réplication	$4.80x10^{-4}$	$5.90x10^{-4}$	$8.30x10^{-4}$

Tableau 4 : *p-values* obtenues pour les 3 SNPs à proximité de *PROX1* lors de l'analyse sur la cohorte MACS 156, puis lors de la réplication sur une cohorte indépendante.

d) La cohorte Afro-Americaine

Toujours en 2009, une étude d'association 'génome entier' a été réalisée sur des patients afro-américain (N=515), provenant des cohortes DoD HIV NHS (United States military Department of Defense Human Immunodeficiency Virus Natural History Study) et MACS (Multicenter AIDS Cohort Study) [147]. Les génotypes obtenus sur ces patients ont été étudiés selon un phénotype : charge virale plasmatique stable au cours de la phase asymptomatique.

Aucun signal ne passe le seuil de significativité. Néanmoins, par une analyse plus fine sur un sous groupe de patients, ils ont montré que le SNP du CMH avec la meilleur *p-value* (rs2523608, p=2,3x10^{-6}) était très fortement lié à l'allotype *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 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 [148], mais cette GWAS a mis en avant cet allèle comme l'association la plus significative affectant le contrôle viral dans cette population.

e) Les non-progresseurs à long terme non 'élites' de la cohorte GRIV

En parallèle de mon travail au sein du CNAM, une autre étudiante de notre laboratoire a réalisé des travaux sur les facteurs génétiques influençant la non-progression à long terme sans nécessairement contrôler la charge virale. Pour cela une étude d'association 'génome entier' a été menée sur la comparaison de 186 non-progresseurs à long terme non 'élites' issus de la cohorte GRIV (*i.e.* avec une charge virale plasmatique >100 copies/mL) avec 697 contrôles séronégatifs [149].

La plus forte association a été obtenue pour le SNP rs2234358 avec une *p-value* proche du seuil de significativité 'génome entier' : p=2,5x10^{-7}, OR=1,85. Ce polymorphisme du gène ***CXCR6*** est localisé dans une région du chromosome 3 riche en gènes codant pour des récepteurs de chimiokines, et est notamment distant de 422kb par rapport au gène *CCR5* [109-113]. Ce signal représente donc une nouvelle association avec la LTNP, indépendante des résultats précédemment connus des loci *CCR2-CCR5* et *HCP5/HLA-B*57*. Le signal rs2234358 a été répliqué dans 3 cohortes indépendantes, également d'origine européenne et évaluant un phénotype de progression vers le SIDA : (i) la cohorte hollandaise ACS (n=316) [150]; (ii) un sous-groupe de la cohorte américaine MACS enrichi en phénotypes extrêmes (n=156) [145]; (iii) une compilation de plusieurs cohortes américaines USA HIV-1 (n=556) [151]. La combinaison des *p-values* obtenues dans les 4 études atteint le seuil de significativité statistique 'génome entier' : $p_{combinée}$=9,7x10^{-10}.

Les LTNP respectant la définition GRIV présentant une charge virale >100 copies/mL ont été extrait des cohortes ACS et MACS. Dans ces sous-groupes, la fréquence de l'allèle rs2234358-T est autour de 40%, ce qui est similaire à ce qui est observé dans les LTNP non 'élites' de GRIV (36,83%), alors que dans les sous-groupes non LTNP et dans les groupes contrôles, la fréquence est d'environ 50%. Après vérification de la stratification éventuelle de l'ensemble des groupes, nous avons ajouté les LTNP non 'élites' de ACS et MACS à ceux de la cohorte GRIV. La comparaison statistique de ce groupe élargi (n=276) au groupe de contrôle permet d'atteindre le seuil de significativité statistique (p=2,1x10^{-8}).

L'étude du profil haplotypique du gène a révélé plusieurs haplotypes du promoteur en fort déséquilibre de liaison (r^2=0,97) avec rs2234358, polymorphisme de la région 3'UTR situé à 42pb du codon STOP. 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.

Le travail 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 prévention chez les LTNP. CXCR6 est un récepteur de chimiokines connu comme étant un corécepteur majeur du SIV, et seulement mineur dans le cadre du VIH-1 [152]. CXCR6 est également un médiateur de l'inflammation impliqué dans la migration des cellules Th1 [153, 154], et dans l'activation des cellules NKT [155]. Des explorations fonctionnelles devraient permettre d'améliorer la compréhension des mécanismes moléculaires impliqués dans la progression de l'infection VIH-1.

4. Objectifs de ma thèse

J'ai rejoint l'équipe du Pr Zagury au cours de l'année 2007 pour réaliser ma thèse. Durant mon stage de Master 2 au Centre National de Génotypage, je me suis familiarisée avec les techniques de génotypage haut débit. Cette formation correspondait aux besoins du Pr Zagury pour engager une étude 'génome entier' sur la cohorte GRIV.

Lorsque j'ai débuté ma thèse, les avancées technologiques en matière de génotypage haut débit des SNPs avait déclenché un engouement autour des études 'génome entier'. En effet, cette approche jusqu'ici inaccessible était porteuse de nombreux espoirs pour la découverte de facteurs génétiques associés à des pathologies. En revanche, aucune étude 'génome entier' n'avait encore été publiée dans le cadre du SIDA. Notre équipe disposant de la plus grande cohorte au monde de patients à profils extrêmes (non-progresseurs à long terme et progresseurs rapides) de progression vers le SIDA -la cohorte GRIV- nous avons entrepris de réaliser l'étude 'génome entier' des sujets de la cohorte GRIV. L'objectif principal de ce travail était l'identification de nouvelles associations génétiques avec la progression de l'infection VIH-1, dans l'espoir d'améliorer la compréhension des mécanismes moléculaires de pathogenèse et de pouvoir proposer de nouvelles cibles diagnostiques et thérapeutiques.

Il était prévu que mon projet de thèse se déroule selon les étapes suivantes :

1. Génotypage par puces de l'ensemble de la cohorte GRIV,
2. Exploitation des données génomiques issues des puces de génotypage,
3. Identification et interprétation biologique d'associations génétiques avec la non-progression à long terme,
4. Identification et interprétation biologique d'associations génétiques avec la progression rapide,
5. Développement d'approches alternatives pour l'exploitation des données génomiques de la cohorte GRIV.

Seconde partie :

Matériel et Méthodes

1. La cohorte GRIV et les populations contrôles

La **cohorte GRIV** (Génomique de la Résistance face à l'Infection par le VIH-1) a été établie en France depuis 1995, afin de mener des études génétiques pour identifier des facteurs génétiques de l'hôte influençant la progression vers le SIDA. Afin de réduire les biais de stratification de population, et l'influence des facteurs environnementaux et génétiques viraux, seuls les individus d'origine européenne vivant en France sont éligibles pour le recrutement : tous les sujets vivent dans un environnement similaire et sont infectés par des souches B du VIH-1.

La cohorte est constituée de deux populations à profil extrême de progression : 100 **progresseurs rapides** (PR) et 300 **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 présentent un taux de cellules T $CD4^+$ supérieur à 500/mm^3 en absence de tout traitement antirétroviral (Figure 15).

Les groupes PR et LTNP génotypés 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).

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, et représente un outil de travail exceptionnel pour l'identification de facteurs génétiques associés à la progression de l'infection VIH-1 [106, 135, 136].

Le projet GRIV se place dans un contexte d'étude génétique « cas-contrôle » : les individus extrêmes PR et LTNP sont comparés à des sujets séronégatifs de même origine ethnique (**cohortes contrôles SU.VI.MAX** [156] et **DESIR** [157].

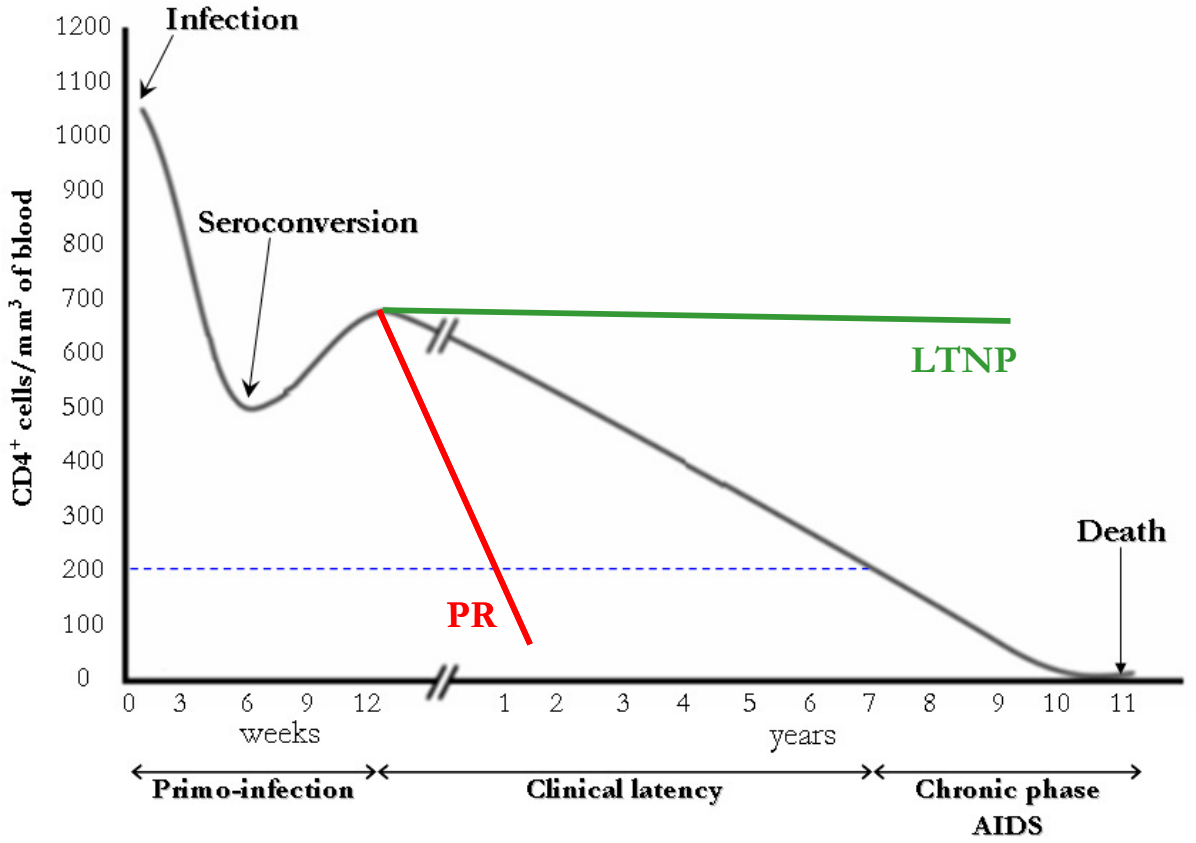


Figure 15 : Description des profils extrêmes de progression de la cohorte GRIV.

L'étude SU.VI.MAX (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 représentatifs de cette étude, 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.

L'étude DESIR (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.

2. Génotypage

Les puces Illumina HumanHap300 permettent le génotypage de deux 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 $r^2 \geq 0,8$ [144]. Dans cette optique, les puces Illumina HumanHap300 ciblent des tagSNPs fréquents (>5%) de la Phase I de HapMap avec un seuil de $r^2 \geq 0,8$ 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 $r^2 \geq 0,7$ 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 : (i) l'ensemble du protocole est standardisé et automatisé (utilisation de kits et de robots), ce qui offre une grande robustesse et reproductibilité des résultats ; (ii) la consommation d'ADN est faible (750ng) pour la quantité de SNPs génotypés ; (iii) les SNPs ciblés sont répartis sur l'ensemble du génome de façon gène-centré, 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 microbilles de verre de 3µm de diamètre, et permettent le génotypage simultané de deux individus grâce à une séparation centrale hermétique. Chaque microbille 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.

La procédure expérimentale s'étale sur trois jours et requiert initialement seulement 750ng d'ADN génomique (ADNg) par échantillon soit 15 µl concentré à 50ng/µL (Figure 16 à 19).

- Le premier jour d'expérimentation est consacré à l'amplification de l'ADNg (Figure 16) : tout d'abord, l'ADN est dénaturé par l'ajout de NaOH à 0,1N (10min à 22°C), et neutralisé à l'aide d'une solution de neutralisation Illumina®. Un Master Mix d'amplification Illumina® est ajouté aux échantillons. L'ADN est alors incubé à 37°C pendant 20-24h. Cette étape permet d'augmenter la quantité d'ADN génomique (ADNg) (Figure 16).

Figure 16 : Premier jour d'expérimentation (procédé Illumina Infinium II).

- Le second jour (Figure 17) débute par la fragmentation de l'ADNg selon un processus enzymatique contrôlé Illumina® à 37°C pendant 1h. L'ADN fragmenté est ensuite précipité grâce à du 2-propanol et une solution de précipitation Illumina® contenant des sels (1h à 4°C). La plaque est alors centrifugée et le surnageant (contenant dNTP, enzyme...) est éliminé. La plaque contenant les culots est laissée à température ambiante pendant 1h. L'ADN purifié est ensuite remis en suspension dans un tampon d'hybridation (1h à 48°C), et dénaturé par la chaleur (20min à 95°C). Deux échantillons sont alors déposés sur chaque puce et les puces sont incubées à 48°C pendant 16-24h : au cours de cette l'étape d'hybridation, les fragments d'ADN se fixent de façon spécifique au niveau des oligonulcéotides fixés sur les billes des puces (Figure 17).

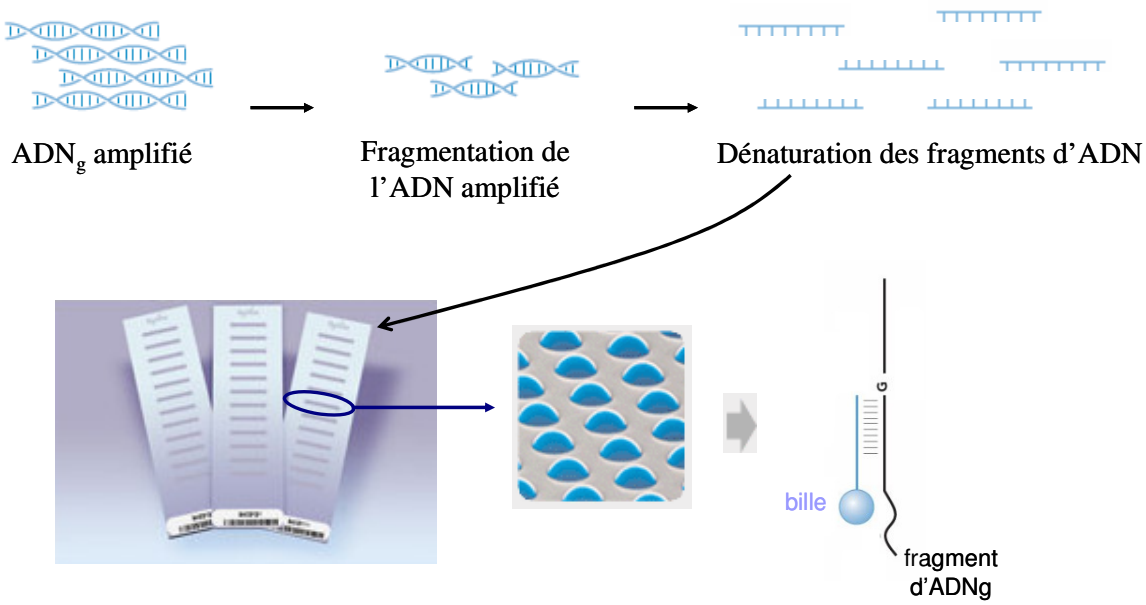


Figure 17 : Second jour d'expérimentation (procédé Illumina Infinium II).

- Lors du troisième jour (Figure 18), toutes les étapes sont réalisées dans un bain-marie à 44°C. Les fragments non hybridés ou fixés de façon aspécifique sont éliminés par lavages avec un de tampon Illumina®. L'étape d'extension

simple base est réalisée à l'aide de 3 solutions : 2 solutions tampons Illumina®, le Master Mix contenant les ddNTP couplés à des fluorochromes et la polymérase. Les oligonuclétides fixés sur les billes sont spécifiques de la région strictement en amont du SNP, et ainsi, les ddNTP marqués se fixent en fonction de l'allèle au niveau de la position du SNP. L'incubation avec le formamide 95%/EDTA 1mM pendant 7min permet d'éliminer l'ADN hybridé. Le marquage est ensuite amplifié en ajoutant des anticorps eux aussi couplés à des fluorochromes et spécifiques de chacun des ddNTP. Enfin, la puce est lavée et séchée avant d'être placée dans le scanner pour révéler la fluorescence, et déterminer les génotypes de l'ensemble des SNPs (Figures 18 et 19).

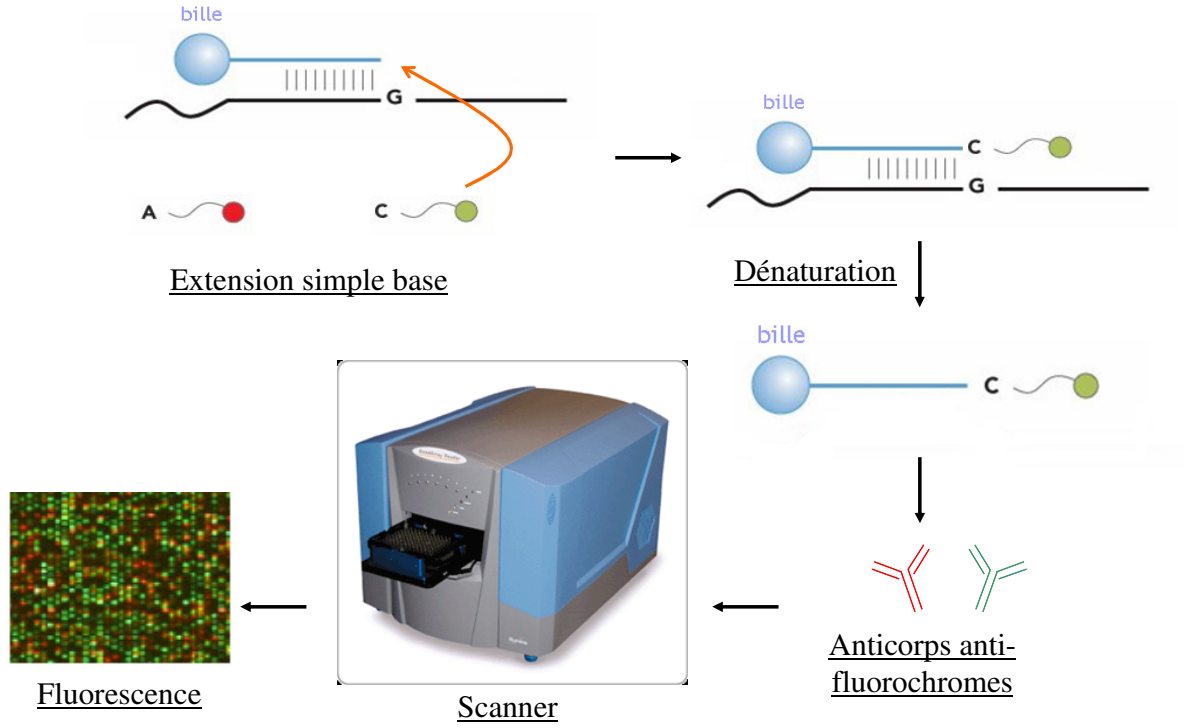


Figure 18 : Troisième jour d'expérimentation (procédé Illumina Infinium II).

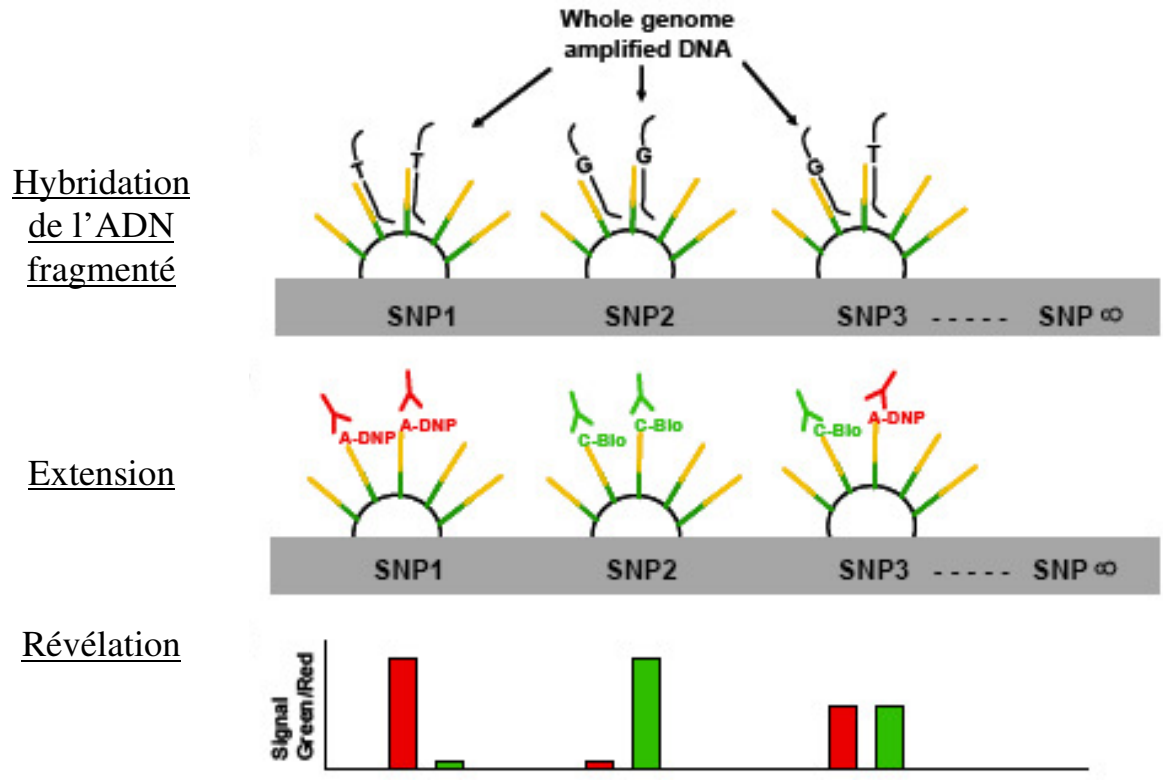


Figure 19 : Obtention des génotypes à partir de la fluorescence : dans le cas d'un génotype homozygote (SNP1 et SNP2) : les fragments d'ADN hybridés au niveau de l'amorce ne porteront qu'un seul allèle ; les bases couplées à un fluorochrome, ajoutées lors de l'extension, seront d'un seul type ; un seul type de fluorescence (rouge ou verte) sera émise lors de la révélation, traduisant l'homozygotie du génotype. Dans le cas d'un génotype hétérozygote G/T (SNP3) : les fragments d'ADN hybridés porteront soit l'allèle G, soit l'allèle T ; les bases ajoutées à l'amorce lors de l'extension seront donc A couplée à un fluorochrome rouge ou C couplée à un fluorochrome vert ; la fluorescence rouge et verte seront émises lors de la révélation donnant une couleur orangée et traduisant un génotype hétérozygote G/T.

N.B. : Dans le procédé Illumina Infinium II, 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.

3. Contrôle qualité

3.1. Analyse BeadStudio

Les données brutes issues du génotypage ont été analysées à l'aide du logiciel Illumina BeadStudio v3.1 et filtrées selon plusieurs paramètres. Dans un premier temps les génotypes sont attribués d'après une classification fournit par Illumina, générée sur une population caucasienne. Cette étape assure la robustesse des génotypes attribués. Ensuite, les individus avec un « call rate » (pourcentage de SNPs génotypés par individu) inférieur à 95% sont éliminés. D'autre part, les SNPs avec un « call frequency » (pourcentage d'individus génotypés par SNP) inférieur à 99% sont re-classifiés. En d'autre terme, un algorithme du logiciel va attribué les génotypes en fonction des intensités obtenues pour chacun des individus sur un SNP donné. Après la re-classification, les individus avec un 'call rate' inférieur à 97% sont supprimés. Les étapes de classification peuvent induire des erreurs dans l'attribution des génotypes, qui peuvent être évitées en suivant la procédure mise en place par Illumina (http://www.illumina.com/downloads/GTDataAnalysis_TechNote.pdf). Cette procédure permet d'évaluer la qualité de la re-classification selon différents critères, qui peuvent être corrigés manuellement si nécessaire. Enfin, les SNPs avec un 'call frequency' inférieur à 98% (*i.e.* un taux de données manquantes >2%) sont exclus. L'ensemble de ces étapes assure des données de génotypage fiable avec peu de données manquantes.

3.2. Déviation de l'équilibre d'Hardy-Weinberg

La loi de Hardy-Weinberg postule que la diversité génétique de la population se maintient et tend vers un équilibre stable des fréquences des allèles et des génotypes au cours des générations [158-160]. Cet équilibre est observé si les hypothèses suivantes sont respectées : la population étudiée est de taille infinie; absence de migration, mutation et sélection ; la panmixie (rencontre aléatoire des individus) et la pangamie (rencontre aléatoire des gamètes).

Dans ce cas d'équilibre, pour un SNP bi-allélique a_1/a_2 où fa_1 est la fréquence de l'allèle a_1 et $fa_2 = (1- fa_1)$ est la fréquence de l'allèle a_2, alors :

- la fréquence du génotype homozygote a_1/a_1 est de fa_1^2,
- la fréquence du génotype hétérozygote a_1/a_2 est de $2fa_1\, fa_2$,

- la fréquence du génotype homozygote a_2/a_2 est de fa_2^2.

En pratique, cette loi est bien respectée, et un écart à l'équilibre d'Hardy-Weinberg dans un groupe de patients pour un SNP donné suggère un effet biologique, tandis qu'une déviation dans un groupe contrôle suggère généralement une erreur de génotypage. La déviation de l'équilibre de Hardy-Weinberg a été évaluée pour chaque SNP dans chaque groupe en comparant la répartition des fréquences génotypiques observées et des fréquences génotypiques théoriques en utilisant un test statistique exact [161]. Les SNPs déviant de cet équilibre dans la population contrôle ($p<10^{-3}$) ont été exclus.

3.3. SNPs de fréquence faible

L'élimination des SNPs de faible fréquence est une étape classique du contrôle 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 ont donc été éliminés.

4. Analyse statistique

Pour chaque SNP, une analyse classique « cas-contrôle » a été réalisée en utilisant le **test exact de Fisher**, implémenté dans le logiciel PLINK (http://pngu.mgh.harvard.edu/~purcell /plink/) [162], afin de comparer la distribution allélique entre un groupe cas (LTNP ou PR) et le groupe 'contrôle'.

La réalisation de **tests multiples** a été prise en compte en appliquant les corrections de Bonferroni. Afin d'identifier des signaux supplémentaires tout en contrôlant le risque de fausses découvertes, nous avons également calculé la *q-value* de FDR (False-Discovery Rate) pour chaque *p-value* : la *q-value* estime la proportion de faux-positifs en-dessous d'un seuil de *p-value* [163]. Le FDR est plus puissant que la méthode classique de Bonferroni [164-166], puisqu'il permet l'identification de plus de signaux vrai-positifs. Dans le cadre de maladies polyfactorielles - 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 individuellement 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 [167].

5. Etude de stratification des populations

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** v2.2 [168]. 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.

Toujours dans l'optique d'éviter l'identification de fausses associations reflétant une stratification de nos populations, nous avons utilisé la technique des Genomic Control. Un graphe **quantile-quantile** a été tracé en reportant les *p-values* obtenues lors des tests statistiques en fonction des *p-values* attendues afin de visualiser si la distribution observée dévie de la distribution théorique. Enfin, le **facteur d'inflation génomique,** évaluant l'écart à la distribution théorique, a été calculé [169].

Plus récemment, nous avons également utilisé la méthode Eigenstrat, permettant de détecter et de corriger la stratification d'une population [170]. Cette méthode, basée sur une analyse en composantes principales, permettant de modéliser les différences ancestrales selon des axes continus de variation. Une analyse statistique selon un modèle de régression permet d'inclure les principaux axes en tant que covariables, et ainsi de corriger les *p-values*. La mise en œuvre de cette méthodologie a confirmé les résultats obtenus dans nos trois GWAS.

6. Approfondissement des associations identifiées

6.1. Déséquilibre de liaison (DL)

Pour chaque SNP associé significativement à la progression de l'infection VIH-1, nous avons recherché l'ensemble des SNPs en DL ($r^2 \geq 0,8$) d'après les données du projet HapMap pour la population d'origine européenne afin d'identifier les gènes possiblement liés à l'association. Un SNP est affecté à un gène s'il se situe dans le gène à ±2kb ; sinon, il est considéré comme intergénique.

6.2. Inférence des haplotypes

Lorsqu'un signal est identifié au sein d'un gène, nous avons exploré l'association éventuelle des haplotypes de ce gène avec les phénotypes de progression. Les haplotypes ont été inférés à l'aide du logiciel Shape-IT, algorithme le plus rapide et le plus efficace actuellement [171].

6.3. Exploration bioinformatique

Pour chaque SNP d'intérêt, nous avons exploré différentes bases de données afin d'établir d'éventuelles associations avec un mécanisme biologique.

Deux **bases de données d'expression**, disponibles publiquement, permettent d'établir des corrélations entre polymorphisme et variation du niveau d'expression des ARNm : *Genevar* [172 , 173] et Dixon [174].

Des informations sont également disponibles sur des données d'**épissage** (*NetGene2*, http://www.cbs.dtu.dk/services/NetGene2/), de **polyadénylation** (*polyAH*, http://linux1.soft-berry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter et *polyApred*, http://www.imtech.res.in/raghava/polyapred/submission.html), et de **sites de fixation de facteurs de transcription** (*SignalScan*, http://www-bimas.cit.nih.gov/molbio/signal/, *TESS*, http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME, et *TFSearch*, http://www.cbrc.jp/research/db/TFSEARCH.html, dérivé de la base de données TRANSFAC).

Troisième partie :

Résultats

1. Etude 'génome entier' sur les non-progresseurs à long terme de la cohorte GRIV

ARTICLE 1

Genomewide Association Study of an AIDS Nonprogression Cohort Emphasizes the Role Played by HLA Genes (ANRS Genomewide Association Study 02)

Sigrid Le Clerc*, Sophie Limou*, Cédric Coulonges, Wassila Carpentier, Christian Dina, Olivier Delaneau, Taoufik Labib, Lieng Taing, Rob Sladek, ANRS Genomic Group, Christiane Deveau, Rojo Ratsimandresy, Matthieu Montes, Jean-Louis Spadoni, Jean-Daniel Lelièvre, Yves Lévy, Amu Therwath, François Schächter, Fumihiko Matsuda, Ivo Gut, Philippe Froguel, Jean-François Delfraissy, Serge Hercberg, and Jean-François Zagury

J Infect Dis. 2009 ; 199(3):419-26.

RESUME

Afin d'élucider les facteurs de prédisposition de progression vers le SIDA, nous avons entrepris une analyse 'génome entier' en comparant 275 non-progresseurs à long terme séropositifs pour le VIH de la cohorte GRIV avec une cohorte contrôle composé de 1352 individus séronégatifs.

La plus forte association est obtenue pour le SNP rs2395029 localisé dans le gène *HCP5* (p=6,79x10^{-10}; odd ratio: 3,47). Cette association avec le SNP rs2395029 avait déjà été identifiée lors d'une précédente analyse 'génome entier' [137]. Ce SNP est en fort déséquilibre de liaison avec des SNPs localisés dans les gènes *HLA-B, MICB, TNF,* et avec plusieurs autres SNPs et haplotypes du locus *HLA*. L'étude des covariables a montré l'absence d'association avec l'allèle rs2395029-G chez les femmes de la cohorte GRIV suggérant un effet uniquement chez les hommes. Cependant, ce résultat nécessite une confirmation dans une cohorte indépendante.

Dans un deuxième temps, nous avons réalisé une méta-analyse de nos données avec celles générées par le groupe Euro-CHAVI [137]. Cette démarche a permis de confirmer

l'implication du SNP localisé dans le géne HCP5 (p=3,9x10^{-19}) et d'identifier plusieurs nouvelles associations, toutes impliquant des gènes de la région *HLA* : *TNF, RDBP, BAT1–5, PSORS1C1,* and *HLA-C.*

Enfin, une étude stratifiée par les génotypes de *HCP5* rs2395029 a mis en évidence un effet indépendant du gène *ZNRD1*, également localisé dans le locus *HLA*. Nous avons montré que les SNP identifiés ne corrèlent pas avec la charge virale, ce qui suggère un rôle dans la progression. De plus, dans une récente étude 'génome entier' d'ARN interférent, la protéine ZNRD1 a été identifiée parmi les 273 protéines nécessaires à la réplication et à l'infection VIH-1 [175].

Cette étude constitue la première étude génome entier réalisée sur une cohorte de non-progresseurs VIH^+. Cette analyse souligne le potentiel de certains gènes du *HLA* dans le contrôle de la progression de la maladie rapidement après l'infection.

MAJOR ARTICLE

Genomewide Association Study of an AIDS-Nonprogression Cohort Emphasizes the Role Played by *HLA* Genes (ANRS Genomewide Association Study 02)

Sophie Limou,[1,2,4,5,a] Sigrid Le Clerc,[1,2,4,a] Cédric Coulonges,[1,2] Wassila Carpentier,[2] Christian Dina,[6] Olivier Delaneau,[1] Taoufik Labib,[1,4] Lieng Taing,[1] Rob Sladek,[8] ANRS Genomic Group,[2,b] Christiane Deveau,[2] Rojo Ratsimandresey,[1] Matthieu Montes,[1] Jean-Louis Spadoni,[1] Jean-Daniel Lelièvre,[4] Yves Lévy,[4] Amu Therwath,[3] François Schächter,[1] Fumihiko Matsuda,[9] Ivo Gut,[5] Philippe Froguel,[6,10] Jean-François Delfraissy,[2] Serge Hercberg,[7] and Jean-François Zagury[1,2,4]

[1]Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, [2]Agence Nationale de Recherche sur le SIDA et les Hépatites Virales (ANRS) Genomic Group, and [3]Laboratoire d'Oncologie Moléculaire, Université Paris 7, Paris, and [4]Henri Mondor Hospital, Institut National de la Santé et de la Recherche Médicale (INSERM) U841, Créteil, [5]Commissariat à L'Énergie Atomique/Institut de Génomique, Centre National de Génotypage, Evry, [6]Unité Mixte de Recherche (UMR) Centre National de la Recherche Scientifique 8090, Institut Pasteur de Lille, Lille, and [7]UMR U557 INSERM/U1125 L'Institut National de la Recherche Agronomique/Conservatoire National des Arts et Métiers/Université Paris 13, Centre de Recherche en Nutrition Humaine Ile-de-France, Santé-Médecine-Biologie Humaine Paris 13, Bobigny, France; [8]Department of Human Genetics, Faculty of Medicine, McGill University, and Génome Québec Innovation Centre, Montreal, Canada; [9]INSERM U852, Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [10]Genomic Medicine, Hammersmith Hospital, Imperial College London, London, United Kingdom

To elucidate the genetic factors predisposing to AIDS progression, we analyzed a unique cohort of 275 human immunodeficiency virus (HIV) type 1–seropositive nonprogressor patients in relation to a control group of 1352 seronegative individuals in a genomewide association study (GWAS). The strongest association was obtained for *HCP5* rs2395029 ($P = 6.79 \times 10^{-10}$; odds ratio, 3.47) and was possibly linked to an effect of sex. Interestingly, this single-nucleotide polymorphism (SNP) was in high linkage disequilibrium with *HLA-B, MICB, TNF,* and several other HLA locus SNPs and haplotypes. A meta-analysis of our genomic data combined with data from the previously conducted Euro-CHAVI (Center for HIV/AIDS Vaccine Immunology) GWAS confirmed the *HCP5* signal ($P = 3.02 \times 10^{-19}$) and identified several new associations, all of them involving *HLA* genes: *MICB, TNF, RDBP, BAT1–5, PSORS1C1,* and *HLA-C*. Finally, stratification by *HCP5* rs2395029 genotypes emphasized an independent role for *ZNRD1,* also in the HLA locus, and this finding was confirmed by experimental data. The present study, the first GWAS of HIV-1 nonprogressors, underscores the potential for some *HLA* genes to control disease progression soon after infection.

After 25 years of intensive research, there is still no definitive cure or vaccine for AIDS, and innovative strategies to fight HIV-1 infection are needed. Nowadays, genotyping by high-density arrays scanning the whole genome allows discovery of unsuspected genetic risk factors that influence the pathogenesis of disease [1]. This systematic genetic approach should reveal new leads for strategies targeting AIDS, given that associations based on a candidate gene approach accounted for no more than 10% of the genetic risk factors influencing disease progression [2]. Recently, a genomewide association study (GWAS) based on a European multicenter seroconverter HIV-1 cohort, the Euro-CHAVI (Center for HIV/AIDS Vaccine Immunology) cohort, identified 2 alleles in *HCP5* and *HLA-C* that explained nearly 15% of the variation in the viral load set point [3]. Although genomic studies of AIDS usually rely on seroconverter patients who display all stages of disease, our rationale was that the extreme nonprogression phenotype of the

Received 9 October 2008; accepted 3 November 2008; electronically published 10 January 2009.

Potential conflicts of interest: none reported.

Financial support: Agence Nationale de Recherche sur le SIDA et les Hépatites Virales (ANRS); Innovation 2007 program of the Conservatoire National des Arts et Métiers; AIDS Cancer Vaccine Development Foundation; Neovacs SA; Vaxconsulting. S.L. benefits from a fellowship from the French Ministry of Education, Technology, and Research, and S.L.C. benefits from a fellowship from ANRS.

[a] S.L. and S.L.C. contributed equally to this work.

[b] The ANRS Genomic Group oversees the AIDS genomic projects of the ANRS; study group members are listed at the end of the text.

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

The Journal of Infectious Diseases 2009; 199:419–26

0022-1899/2009/19903-0018$15.00
DOI: 10.1086/596067

Table 1. Fifty best results obtained for the comparison between nonprogressors and control subjects.

The table is available in its entirety in the online edition of the *Journal of Infectious Diseases.*

GRIV (Genomics of Resistance to Immunodeficiency Virus) cohort could bring even more contrast to the attempt to identify genetic effects. Previous gene-candidate analyses have shown the power of this unique design, notably for the *HLA* and *CCR5* genes [4, 5].

METHODS

The GRIV cohort. The GRIV cohort was established in France in 1995 to generate a large collection of DNA for genetic studies to identify host genes associated with nonprogression to AIDS [4, 6]. Only white people of European descent living in France were eligible for enrollment to reduce confounding by population substructure. These criteria limit the influence of ethnic and environmental factors (all subjects live in a similar environment and are infected by B strains) and emphasize the genetic makeup of each individual in determining the various patterns of progression. Nonprogressors were included on the basis of the main clinical outcomes, CD4 T cell count and time to disease progression; inclusion criteria were asymptomatic HIV-1 infection for >8 years, no receipt of treatment, and a CD4 T cell count consistently remaining >500 cells/mm^3. Viral load was not part of the GRIV inclusion criteria; however, the values at inclusion were obtained and used to assess potential correlations with genotypes. DNA was obtained from fresh peripheral blood mononuclear cells or from Epstein-Barr virus–transformed cell lines. The nonprogressors group ($n = 275$) was composed of 201 men and 74 women whose ages at inclusion ranged from 19 to 62 years (median, 35 years). At inclusion, the median CD4 T cell count was 706 cells/mm^3 among the nonprogressors (minimum and maximum values, 501 and 2298 cells/mm^3). All patients provided written informed consent before enrollment in the GRIV genetic association study.

The seropositive control population. To determine whether positive signals corresponded either to an association with nonprogression or to an association with HIV-1 infection, we needed a group of seropositive control subjects who were not nonprogressors. For that, we used 86 white French subjects who qualified as rapid progressors to AIDS (i.e., a CD4 T cell count decreasing to <300 cells/mm^3 within 3 years of seroconversion). This control group was composed of 74 men and 12 women aged from 21 to 55 years (median, 32 years). The median CD4 T cell count of this seropositive control population was 230 cells/mm^3 (minimum and maximum values, 20 and 297 cells/mm^3). Viral loads were not available.

The SU.VI.MAX control group. The SU.VI.MAX (Supplémentation en Vitamines et Minéraux Antioxydants) study was a randomized, double-blind, placebo-controlled and primary-prevention trial designed to test the efficacy of daily supplements of antioxidant vitamins and minerals at nutrition-level doses in reducing the frequency of several major health problems in industrialized countries, especially the main causes of premature death, cancers and cardiovascular diseases. This cohort study was started in 1994 in France and was composed of 12,735 subjects [7]. The control group genotyped in the present study comprised 1352 representative SU.VI.MAX participants, all white persons living in France who were HIV-1 seronegative. This control cohort was composed of 525 men and 827 women, with a mean age of 53.1 and 48.5 years, respectively.

Genotyping method. Genotyping was performed for the GRIV cohort and the control groups by means of Infinium II HumanHap300 BeadChips (Illumina). The genomic DNA (750 ng) was whole-genome amplified, fragmented, denatured, and hybridized on prepared HumanHap300 BeadChips for a minimum of 16 h at 48°C. Nonspecifically hybridized fragments were removed by washing, and the remaining specifically hybridized DNA was fluorescently labeled by a single base-extension reaction and was detected using a BeadArray scanner (Illumina). Normalized bead-intensity data obtained for each sample were loaded into BeadStudio software (version 3.1; Illumina), which converted the fluorescence intensities into single-nucleotide polymorphism (SNP) genotypes.

Quality control. Using the BeadStudio software, we analyzed the crude genotyping data, and SNPs were filtered according to the following parameters. First, samples with a call rate (percentage of SNPs genotyped by sample) <95% in the Illumina clusters were deleted. Second, the SNPs having a call frequency (percentage of samples genotyped by SNP) <99% were reclustered. Third, after reclustering, samples with a call rate <97% were deleted. The clustering step can create SNP genotyping errors, which can be prevented by following the Illumina quality-control procedure (see http://www.illumina.com/downloads/GTDataAnalysis_TechNote.pdf). This method evaluates the quality of the newly created clusters according to several criteria, which can be manually checked and corrected as necessary. By this Illumina procedure, 1300 SNPs were excluded. Finally, after all the quality-control steps, the 15,731 SNPs with a call frequency <98% (>2% of missing data) were excluded. This quality-control procedure ensures reliable genotyping data with few missing data.

Hardy-Weinberg equilibrium analysis was performed for each SNP in each group using an exact statistical test [8] implemented in PLINK software (available at: http://pngu.mgh.harvard.edu/~purcell/plink/) [9]. Deviation from Hardy-Weinberg equilibrium in a group of patients suggests that the

The figure is available in its entirety in the online edition of the *Journal of Infectious Diseases.*

Figure 1. Quantile-quantile plot for expected *(red)* vs. observed *(black) P* values from the comparison of nonprogressors with control subjects.

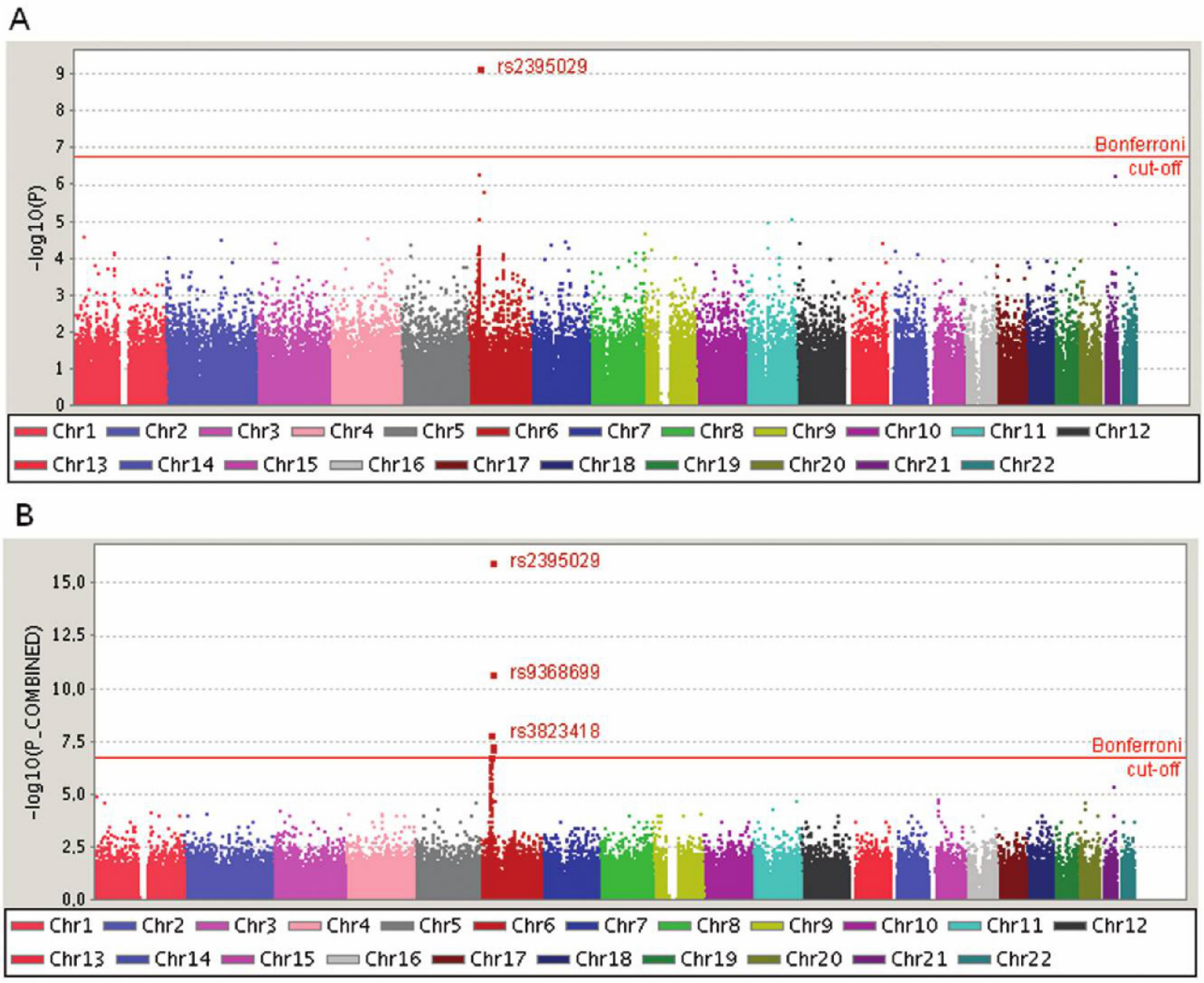


Figure 2. Distribution along the human autosomes of $-\log_{10}(P$ values) obtained for the comparison of nonprogressors with control subjects *(A)* and for the meta-analysis of the GRIV and Euro-CHAVI studies *(B)*. For the latter plot, we used the classical Fisher method, which allows combining *P* values obtained in 2 independent studies. The red line marks the Bonferroni threshold. Chr, chromosome.

SNP has a biological effect, while deviation in the control group or in all groups suggests a systematic error in genotyping. The 1475 SNPs that were not in Hardy-Weinberg equilibrium in the control group ($P < 1.0 \times 10^{-3}$) were rejected in this way.

A total of 235 SNPs with a low minor allelic frequency (<1%) in the global population were also filtered.

Linkage disequilibrium. For each SNP exhibiting a significant association, we looked for the other SNPs in linkage disequilibrium ($r^2 \geq 0.8$) in the HapMap population of Western European ancestry (CEU, HapMap data Release 21a/phase II January 2007, on NCBI B35 assembly, dbSNP125; available at: http://www.hapmap.org) to identify the genes possibly involved with the associations. A SNP was assigned to a gene if it was located in the gene or in the 2-kb flanking regions (potential regulatory sequence); otherwise, it was considered intergenic.

Statistical analysis. For each SNP, we performed a standard case-control analysis using Fisher's exact test (with PLINK software) to compare allelic distributions between the nonprogression group and the control group. To take into account the multiple comparisons, we computed the Bonferroni corrections. For all the SNPs meeting the statistical threshold (table 1), the quality of genotyping was individually rechecked with the BeadStudio software. We also checked that the allelic frequencies in the seropositive control population were similar to those in the seronegative SU.VI.MAX control population for those

Table 2. Comparative analysis of the single-nucleotide polymorphisms (SNPs) found to be highly significant by the GRIV genomewide association study (GWAS) and at least 1 other GWAS.

The table is available in its entirety in the online edition of the *Journal of Infectious Diseases*.

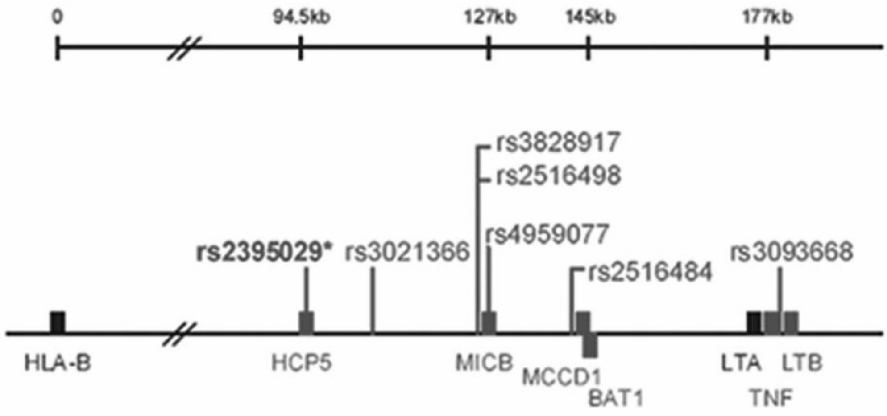


Figure 3. Complexity of the *HCP5* association. Shown is a genetic map of the HLA region. Table 3 lists single-nucleotide polymorphisms (SNPs) and haplotypes of several genes in strong linkage disequilibrium with the *HCP5* rs2395029 SNP, which is marked by an asterisk in this figure.

SNPs of interest, confirming that the observed associations were indeed linked to nonprogression.

Meta-analysis of the GRIV and Euro-CHAVI studies. A total of 286,529 SNPs were found to be in common between the GRIV GWAS and the previous GWAS of AIDS, the Euro-CHAVI study [3]. The genotypes obtained by the Euro-CHAVI study were not directly available, but the *P* values obtained for each SNP for the viral load end point could be obtained from the supporting online material for Fellay et al. [3] (available at: http://www.sciencemag.org/cgi/content/full/1143767/DC1). The *P* values obtained for each SNP in our study and in the Euro-CHAVI study were combined to provide a single probability value using the classical Fisher method [10]. We could not adjust the combined *P* values for opposite allelic effects (i.e., assign $P = 1$ if the odds ratios went in opposite directions), because the detailed allelic information for the Euro-CHAVI study was not available. We could compute this meta-analysis only for the Euro-CHAVI viral load end point data, because the *P* values for the progression to AIDS end point were not available.

Identification of population stratification. To correct for possible population stratification at the intercontinental level,

Table 3. Single-nucleotide polymorphisms (SNPs) and haplotypes of several genes in strong linkage disequilibrium ($r^2 \geq 0.8$ for HapMap data on white individuals) with the *HCP5* rs2395029 SNP (see figure 3).

Gene	Allele	Location	r^2
HCP5	rs2395029	Exon (Val112Gly)	. . .
HLA-B	HLA-B*5701	. . .	1.00
MICB	rs2516498	5'LR	0.83
	rs3828917	5'LR	1.00
	rs4959077	Intron	1.00
	rs2534654, rs2246626 (G-C)	Intron/3'LR	1.00
	rs3828917, rs1051788 (T-G)	5'LR/exon (Asp136Asn)	1.00
TNF	rs3093668	3'LR	0.83
	rs3093661	Intron	0.83
	rs3093661, rs4645843 (A-C)	Intron/exon (Pro84Leu)	0.83
	rs1799964, rs1800630, rs1800750 (C-C-G)	5'LR/5'LR/5'LR	0.83
MCCD1	rs2516484	5'LR	0.83
BAT1	rs2516484	3'LR	0.83
LTB	rs3093668	3'LR	0.83
	rs3093559, rs3093553 (C-G)	3'LR/intron	0.83
	rs3093726, rs3093559 (C-G)	3'LR/3'LR	0.83

NOTE. The *HCP5* rs2395029 SNP is marked by an asterisk in figure 3. To compute the linkage disequilibrium between the haplotypes and the *HCP5* SNP, we limited ourselves to haplotypes composed of 2 or 3 SNPs derived from the known HapMap SNPs in this HLA region. The list is only a small sample of the numerous haplotypes with $r^2 \geq 0.8$ identified. 5'LR corresponds to the 5' part within 2 kb of the gene; 3'LR corresponds to the 3' part within 0.5 kb of the gene.

Table 4. Influence of sex on the *HCP5* and *C6orf48* associations with nonprogression.

The table is available in its entirety in the online edition of the *Journal of Infectious Diseases*.

Table 6. Fifty best combined *P* values obtained by the meta-analysis of the GRIV and Euro-CHAVI studies.

The table is available in its entirety in the online edition of the *Journal of Infectious Diseases*.

case and control genotypes were analyzed using STRUCTURE software (version 2.2) [11]. We selected a set of 328 SNPs that were informative for ancestral origin (*F* statistics fixation index > 0.2) on the basis of the Perlegen data set and that were separated by 5 Mb to avoid linkage disequilibrium. We also included genotypes obtained from unrelated individuals representing the 3 populations studied by the HapMap project, to better separate the nonprogressor and control individuals according to their continental origin. All case and control subjects fell within the range of the white individuals from HapMap.

To avoid spurious associations resulting from possible population stratification or genotyping errors, a quantile-quantile plot was produced by plotting the ranked values of the test statistics against the approximated expected order statistic (figure 1). We also computed the genomic inflation factor λ [12]. The result ($\lambda = 1.064$), along with the quantile-quantile plot, suggested little overall effect of stratification.

RESULTS AND DISCUSSION

Using the Illumina HumanHap300 BeadChips, we performed a GWAS by comparing our nonprogression group ($n = 275$) with a control group ($n = 1352$) from the SU.VI.MAX cohort. After the different quality-control tests (see Methods), a total of 291,119 autosomal SNPs were tested for association with nonprogression. For each SNP, Fisher's exact test was performed comparing the allelic frequencies in the case group versus those in the control group, and the resulting *P* values were adjusted by the Bonferroni correction. Figure 2*A* depicts the distribution of the *P* values along the chromosomes, and table 1 presents the most significant signals.

The sole association remaining after the Bonferroni adjustments was the *HCP5* rs2395029-G allele ($P = 6.79 \times 10^{-10}$; odds ratio, 3.47 [95% confidence interval, 2.39–5.04]) (table 1) located on chromosome 6 (figure 2*A*). This *HCP5* SNP was previously identified by Fellay et al. [3], who hypothesized that, because *HCP5* encodes a human endogenous retrovirus with sequence homology to HIV-1 *pol* [13], it may act as antisense RNA interfering with HIV-1 replication. This SNP is also in absolute linkage disequilibrium with the HLA-B*5701 allele [14], which has been associated with the control of HIV-1 replication and disease progression [15]. Interestingly, this SNP was also shown to be a major signal in a GWAS of psoriasis and psoriatic arthritis [16] (table 2). Figure 3 and table 3 show that this SNP could be tracking through linkage disequilibrium causal alleles in other major genes of the HLA locus, including *MICB, BAT1, LTB,* and *TNF*. MICB is a ligand for CD8 T cells and natural killer cells, which are key players in the anti–HIV-1 immune response. BAT1 is an essential component for splicing and RNA export [17] but is also known as a negative regulator of the inflammatory cytokines tumor necrosis factor (TNF), interleukin (IL)–1, and IL-6 [18]. Lymphotoxin β (LTB) is an inflammatory

Table 5. Best *P* values obtained by the meta-analysis of the GRIV and Euro-CHAVI studies.

SNP	Chr	Chr position	A1	A2	Allelic frequency (A1), % NP	CTR	SCP	Fisher $P_{NP\text{-}CTR}$	Fisher $P_{Euro\text{-}CHAVI}$	$P_{combined}$	Gene(s)/LD
rs2395029	6	31539759	G	T	8.9	2.7	2.3	6.79×10^{-10}	9.36×10^{-12}	3.02×10^{-19}	*HCP5*, intergenic, *MICB, MCCD1, BAT1, LTB,TNF*
rs9368699	6	31910520	C	T	8.2	3.2	3.5	5.27×10^{-07}	1.20×10^{-06}	1.84×10^{-11}	*C6orf48, RDBP, TNXB, BAT2, BAT3, LY6G5C, BAT5*
rs3823418	6	31208921	A	G	24.2	16.8	16.9	5.72×10^{-05}	1.11×10^{-05}	1.40×10^{-08}	*PSORS1C1*
rs2248462	6	31554775	A	G	31.4	24.2	25.8	5.61×10^{-04}	3.61×10^{-06}	4.26×10^{-08}	Intergenic, *MICB*
rs2516509	6	31557973	G	A	31.1	24.1	25.4	6.56×10^{-04}	3.61×10^{-06}	4.95×10^{-08}	Intergenic
rs10484554	6	31382534	T	C	18.3	13.3	11.1	3.78×10^{-03}	8.06×10^{-07}	6.27×10^{-08}	Intergenic, *HLA-C*
rs3815087	6	31201566	T	C	27.5	20.9	19.8	1.04×10^{-03}	7.09×10^{-06}	1.46×10^{-07}	*PSORS1C1*, intergenic

NOTE. This table presents the *P* values obtained by the classical Fisher method that met the Bonferroni threshold. For each SNP, the chromosome (Chr), the chromosome position, the allelic frequencies in the various populations (nonprogressors [NP], control subjects [CTR], and seropositive control population [SCP]), the *P* values obtained in each study, and the combined *P* value are shown. The seropositive control population is a group of 86 HIV-1–seropositive patients who are not nonprogressors and allows genetic associations with nonprogression to be distinguished from associations with HIV-1 infection. The gene or genes corresponding to the SNP or SNPs in linkage disequilibrium (LD; $r^2 \geq 0.8$ for HapMap data on white individuals) are also informed. A SNP was assigned to a gene if it was located in the gene or in the 2-kb flanking regions (potential regulatory sequence).

modulator essential for the development of lymphoid, dendritic, and natural killer cells [19]. TNF is a key proinflammatory cytokine that has been widely investigated in HIV-1 infection [20]. From a biological standpoint, all these genes are critical for immunity and, as such, are good candidates to intervene in the pathogenesis of HIV-1 infection. Indeed, they have all been associated with various immune-related diseases [21–24]. Overall, the complex genetic pattern of this region makes it difficult to discriminate between a specific signal alone or one in combination (i.e., haplotypes): as shown in figure 3 and table 3, several SNPs or haplotypes in the genes *HLA-B, MICB, BAT1, LTB, TNF,* and *MCCD1* are in high or full linkage disequilibrium with *HCP5* rs2395029.

To complete our analysis of the *HCP5* SNP association, we explored the influence of covariables, such as *CCR5-Δ32* and *CCR5-P1* haplotypes [5], the HIV-1 infection mode (mucosal or parenteral), and sex. No effect was observed except for a sex influence: the rs2395029-G frequency was 4.05% in nonprogressor women versus 10.70% in nonprogressor men ($P = 1.71 \times 10^{-2}$), whereas, in control subjects, the frequency was close to 3% in men and in women (table 4). Such interaction between genetic factors and sex have been previously described for both HLA and non-HLA genetic associations with other pathologies [25]. The lack of association for *HCP5* in women requires confirmation in a cohort containing a sufficiently large number of women, because the alleles of interest have a frequency of only 3%.

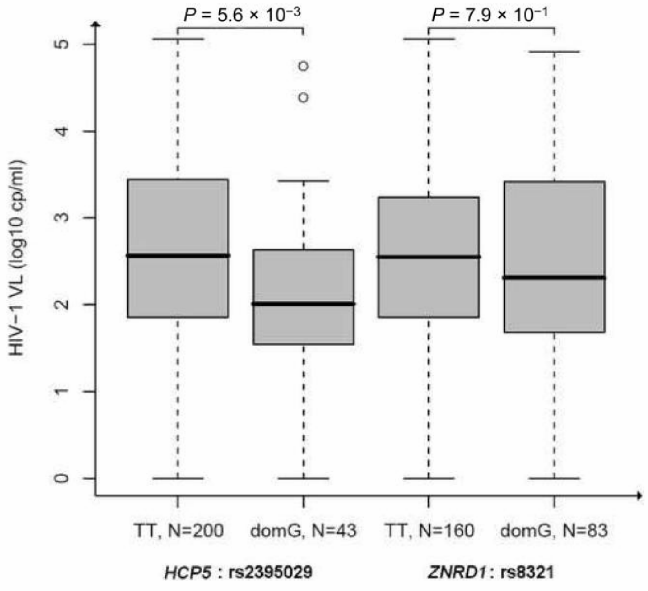


Figure 4. Correlation between genotypes and viral load. The box plots represent the viral load at inclusion of nonprogressor subjects carrying various *HCP5* and *ZNRD1* genotypes and were compared using Student's *t* test. A significantly lower viral load was found in the nonprogressor subjects carrying the *HCP5* rs2395029-GT genotype, compared with that in the ones with *HCP5* rs2395029-TT. No significant difference was observed between *ZNRD1* rs8321 genotypes (TT vs. dominant G [domG]) or for genotypes of *ZNRD1* (rs9261290, rs1245371, and rs259940; data not shown). rs8321 and rs9261290 were identified by our combined analysis with the Euro-CHAVI study, and rs1245371 and rs259940 were identified by our analysis of *HCP5*-independent signals. VL, viral load.

The figure is available in its entirety in the online edition of the *Journal of Infectious Diseases.*

Figure 5. Linkage disequilibrium map presenting single-nucleotide polymorphisms (SNPs) of the chromosome 6 HLA locus exhibiting strong *P* values in the meta-analysis of the GRIV and Euro-CHAVI studies.

The data from the sole AIDS GWAS of the Euro-CHAVI cohort published to date were available [3], and thus we performed a meta-analysis by combining their *P* values with ours using the classical Fisher method (see Methods). Figure 2*B* presents the distribution of the combined *P* values along the autosomes and several associations surpassing the Bonferroni threshold were found, all located on chromosome 6 (table 5) (see table 6 for an extended list of *P* values).

As expected, the strongest signal was obtained for the *HCP5* rs2395029 SNP ($P_{combined} = 3.02 \times 10^{-19}$). The second strongest association was obtained for the *C6orf48* rs9368699 SNP ($P_{combined} = 1.84 \times 10^{-11}$), which was in linkage disequilibrium with *HCP5* rs2395029 ($r^2 = 0.68$) and with several SNPs located in *HLA* genes, such as *TNXB, BAT2, BAT3,* and *RDBP,* suggesting a wide range of possible biological effects that could explain this association. For instance, tenascin XB (TNXB) is an extracellular matrix protein previously associated with systemic lupus erythematosus [26], HLA-B–associated transcript (BAT) 2 is a potential splicing factor previously associated with rheumatoid arthritis [27], BAT3 is an essential regulator of apoptosis and p53-mediated responses to genotoxic stress [28], and *RDBP* encodes a subunit of the negative elongation factor complex known to repress HIV-1 transcription elongation driven by Tat [29]. Then, the *PSORS1C1* gene exhibited 2 significant SNPs, rs3823418 and rs3815087 ($P_{combined} = 1.4 \times 10^{-8}$ and $P_{combined} = 1.46 \times 10^{-7}$, respectively), in partial linkage disequilibrium with each other ($r^2 = 0.66$ for HapMap data on white individuals). *PSORS1C1* is a psoriasis-susceptibility candidate gene [30]. The intergenic SNP rs2248462, which could not be assigned to a specific gene, exhibited a strong association ($P_{combined} = 4.26 \times 10^{-8}$) that could be explained by the linkage disequilibrium with the *MICB* gene discussed above. Finally, the combined analysis underlined the *HLA-C*–related SNP rs10484554 ($P_{combined} = 6.27 \times 10^{-8}$), which was also identified in the GWAS of psoriasis and psoriatic arthritis [16] (see table 2). This gene was widely discussed in the Euro-CHAVI GWAS for the association found with the rs9264942 SNP, which is unfortunately not present in the Illumina HumanHap300 BeadChip [3]. The sex dependence was still observed for the *C6orf48* rs9368699 SNP but not for the following SNPs (table 4).

The inclusion criteria for the Euro-CHAVI study were based on viral load during the asymptomatic set point period of infection, and for the GRIV study they were based on maintenance of CD4 T cell counts over time. The individuals with a low viral

load after infection are likely to be the ones with a stable high CD4 T cell count. Indeed, we found that, in the nonprogressors carrying the *HCP5* rs2395029-G allele, viral load was significantly lower ($P = 5.6 \times 10^{-3}$) than in the other nonprogressors (figure 4). It is thus not surprising to identify common genetic signals between these 2 studies, even though these cohorts were assembled independently. Notably, of the 50 best signals found in this meta-analysis, 46 originated from the HLA locus, emphasizing the massive role played by HLA in the nonprogression phenotype (table 6). The presence of these strong associations is a cross-validation of both cohorts and also emphasizes that the HLA locus is critical for the early control of HIV-1 replication and disease nonprogression. Reciprocally, in our GWAS alone, 31 of the 50 best signals were not from chromosome 6 (table 1) and were not found in the meta-analysis (table 6), suggesting that positive signals outside the HLA locus may be associated with the nonprogression phenotype without influencing viral load. This observation is in line with findings from a recent study by Mellors et al. [31], which showed that the viral load was predictive at a 34% level for the time to reach a CD4 T cell count <200 cells/mm^3.

Because the *HCP5* rs2395029-G allele was present in only 17.8% of the nonprogressor subjects, we reanalyzed the data in the nonprogressor and control individuals not carrying that allele in order to identify *HCP5*-independent signals. Dramatically, most of the signals from chromosome 6 disappeared because of the genetic linkage with the *HCP5* rs2395029-G. However, the strongest signals were still found in the HLA region, with 2 SNPs of the *ZNRD1/RNF39* region, rs1245371 and rs259940 ($P = 9.21 \times 10^{-7}$ and $P = 2.04 \times 10^{-6}$), in linkage disequilibrium. These 2 SNPs are genetically independent from the *HCP5* SNP ($r^2 = 0.06$) (figure 5). Interestingly, the *ZNRD1/RNF39* locus was also identified by our meta-analysis (for rs8321, $P = 4.66 \times 10^{-7}$; for rs9261290, $P = 5.11 \times 10^{-7}$) (table 6) and by the Euro-CHAVI study (for rs3869068, $P = 3.89 \times 10^{-7}$) with the progression-to-AIDS end point (defined as the time elapsed until treatment initiation or until reaching a CD4 T cell count <350 cells/mm^3). Unlike the *HCP5* rs2395029 SNP, none of the *ZNRD1/RNF39* SNPs alleles seemed to correlate with viral load (figure 4), suggesting that this locus influences disease progression. Functionally, the Genevar expression database identified an association between several *ZNRD1/RNF39* alleles and the differential expression of *ZNRD1* (table 2). Finally, we found that a recent genomewide RNA interference study identified zinc ribbon domain containing (ZNRD) 1 among the 273 proteins required for HIV-1 infection and replication [32], suggesting that this RNA polymerase I subunit is an active component in the *ZNRD1/RNF39* region.

In conclusion, the major novelty of the present GWAS of AIDS was the investigation of a cohort with the extreme HIV-1 nonprogression phenotype, in contrast to the usual seroconverter cohorts. We replicated the major role played by the *HCP5* gene of the HLA region in chromosome 6 previously reported in the Euro-CHAVI GWAS [3], a role that could be explained by linkage disequilibrium with other major genes of the HLA locus, such as *HLA-B, MICB, TNF, LTB,* and *BAT1*. The sex dependence of the *HCP5* SNP in our work is striking because it is in high linkage disequilibrium with *HLA-B*57,* which has been investigated for years. It was likely not observed before because most AIDS cohorts are deficient in women; however, this important observation needs confirmation. We then computed a meta-analysis with the previous GWAS and put forward new associations in the same locus: *C6orf48* (in linkage disequilibrium with *RDBP, TNXB,* and *BAT*), *PSORS1C1, MICB,* and *HLA-C*. The HLA region comes first in our AIDS-nonprogression genetic study; however, given that this region presents a complex pattern of high linkage disequilibrium and that all the genes identified display a strong relevance to immunology and AIDS, it is difficult to discriminate which one(s) is(are) the causal variant(s). More refined studies will be needed to discriminate which mechanisms and which HLA locus genes are at stake. Our study, however, has suggested an independent role for the *ZNRD1* gene in disease progression. Overall, our results underline the potential for controlling disease progression and/or viral replication by some *HLA* gene variants soon after infection. Notably, 2 major SNPs identified in our meta-analysis—*HCP5* and *HLA-C*—also had the strongest signals observed in the GWAS of psoriasis and psoriatic arthritis [16].

Because of the large amount of data generated in the present GWAS, statistical cutoffs were required to minimize false discovery, and many true positives with lower *P* values were likely missed but remain candidates of interest. As a reminder, allow us to state that the published *P* values from various cohorts for the widely recognized association between *CCR5*-Δ*32* and AIDS progression have all been in the range of 1×10^{-2} to 1×10^{-4} and would not be seen by the current genomewide studies. This latter observation emphasizes the need to analyze more patients and perform more meta-analyses to extract additional signals from the large pool of genes screened.

ANRS GENOMIC GROUP

The ANRS Genomic Group is composed of Prof. Jean-François Delfraissy (Agence Nationale de Recherche sur le SIDA et les Hépatites Virales, Paris), Dr. Laurence Meyer (Hôpital Kremlin-Bicêtre, France), Prof. Philippe Broët (Hôpital Kremlin-Bicêtre, France), Dr. Cyril Dalmasso (Hôpital Kremlin-Bicêtre, France), Dr. Wassila Carpentier (Hôpital La Salpêtrière, Paris), Prof. Patrice Debré (Hôpital La Salpêtrière, Paris), Dr. Ioannis Théodorou (Hôpital La Salpêtrière, Paris), Prof. Christine Rouzioux (Hôpital Necker, Paris), Cédric Coulonges (Conservatoire National des Arts et Métiers, Paris), Sigrid Le Clerc (Conservatoire National des Arts et Métiers, Paris), Sophie Limou (Conservatoire National des Arts et Métiers, Paris), and Prof. Jean-

François Zagury (Conservatoire National des Arts et Métiers, Paris).

Acknowledgments

We are grateful to all the patients and medical staff who have kindly collaborated with the GRIV project.

References

1. Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD. Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov **2008**; 7:221–30.
2. O'Brien SJ, Nelson GW. Human genes that limit AIDS. Nat Genet **2004**; 36:565–74.
3. 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.
4. 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.
5. 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.
6. 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.
7. Hercberg S, Galan P, Preziosi P, 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 Mineraux AntioXydants Study. Int J Vitam Nutr Res **1998**; 68:3–20.
8. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet **2005**; 76:887–93.
9. 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.
10. Fisher R. Statistical methods for research workers. Edinburgh: Oliver & Boyd, **1932**.
11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics **2000**; 155:945–59.
12. Devlin B, Roeder K. Genomic control for association studies. Biometrics **1999**; 55:997–1004.
13. Kulski JK, Dawkins RL. The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16. Immunogenetics **1999**; 49:404–12.
14. de Bakker PI, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet **2006**; 38:1166–72.
15. Stephens HA. HIV-1 diversity versus HLA class I polymorphism. Trends Immunol **2005**; 26:41–7.
16. Liu Y, Helms C, Liao W, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet **2008**; 4:e1000041.
17. Reed R, Hurt E. A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell **2002**; 108:523–31.
18. Allcock RJ, Williams JH, Price P. The central MHC gene, BAT1, may encode a protein that down-regulates cytokine production. Genes Cells **2001**; 6:487–94.
19. Rennert PD, Browning JL, Mebius R, Mackay F, Hochman PS. Surface lymphotoxin alpha/beta complex is required for the development of peripheral lymphoid organs. J Exp Med **1996**; 184:1999–2006.
20. Kedzierska K, Crowe SM. Cytokines and HIV-1: interactions and clinical implications. Antivir Chem Chemother **2001**; 12:133–50.
21. Kilding R, Iles MM, Timms JM, Worthington J, Wilson AG. Additional genetic susceptibility for rheumatoid arthritis telomeric of the DRB1 locus. Arthritis Rheum **2004**; 50:763–9.
22. Kula D, Jurecka-Tuleja B, Gubala E, Krawczyk A, Szpak S, Jarzab M. Association of polymorphism of LTalpha and TNF genes with Graves' disease. Folia Histochem Cytobiol **2001**; 39(Suppl 2):77–8.
23. Rodriguez-Rodero S, Rodrigo L, Fdez-Morera JL, et al. MHC class I chain-related gene B promoter polymorphisms and celiac disease. Hum Immunol **2006**; 67:208–14.
24. Vandenbroeck K. Cytokine gene polymorphisms in multifactorial conditions. Taylor & Francis Group, ed. Boca Raton: Taylor & Francis Group, **2006**.
25. Ordovas JM. Gender, a significant factor in the cross talk between genes, environment, and health. Gend Med **2007**; 4(Suppl B):S111–22.
26. Kamatani Y, Matsuda K, Ohishi T, et al. Identification of a significant association of a single nucleotide polymorphism in TNXB with systemic lupus erythematosus in a Japanese population. J Hum Genet **2008**; 53: 64–73.
27. Singal DP, Li J, Lei K. Genetics of rheumatoid arthritis (RA): two separate regions in the major histocompatibility complex contribute to susceptibility to RA. Immunol Lett **1999**; 69:301–6.
28. Sasaki T, Gan EC, Wakeham A, Kornbluth S, Mak TW, Okada H. HLA-B-associated transcript 3 (Bat3)/Scythe is essential for p300-mediated acetylation of p53. Genes Dev **2007**; 21:848–61.
29. Fujinaga K, Irwin D, Huang Y, Taube R, Kurosu T, Peterlin BM. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. Mol Cell Biol **2004**; 24:787–95.
30. Holm SJ, Carlen LM, Mallbris L, Stahle-Backdahl M, O'Brien KP. Polymorphisms in the SEEK1 and SPR1 genes on 6p21.3 associate with psoriasis in the Swedish population. Exp Dermatol **2003**; 12:435–44.
31. 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.
32. Brass AL, Dykxhoorn DM, Benita Y, et al. Identification of host proteins required for HIV infection through a functional genomic screen. Science **2008**; 319:921–6.

Genomewide Association Study of an AIDS Non-Progression Cohort Emphasizes the Role Played by *HLA* Genes

Sigrid Le Clerc[1,2,3,4*], Sophie Limou[2,3*], Cédric Coulonges[1,3], Wassila Carpentier[3], Christian Dina[5], Olivier Delaneau[1], Taoufik Labib[1,2], Lieng Taing[1], Rob Sladek[6], ANRS Genomic Group[3], Christiane Deveau[3], Rojo Ratsimandresey[1], Matthieu Montes[1], Jean-Louis Spadoni[1], Jean-Daniel Lelièvre[2], Yves Lévy[2], Amu Therwath[7], François Schächter[1], Fumihiko Matsuda[8], Ivo Gut[4], Philippe Froguel[5,9], Jean-François Delfraissy[3], Serge Hercberg[10], Jean-François Zagury[1,2,3#]

SUPPLEMENTARY ONLINE CONTENT

Figure 1: Quantile-quantile plot for expected (red) *vs.* observed (black) *p-values* from the non-progressors-controls comparison. X axis: -$\log_{10}$(observed *p-values*); Y axis: -$\log_{10}$(expected *p-values* under the null hypothesis). This quantile-quantile plot shows several *p-values* exceeding the expectations and we assessed their statistical significance with the Bonferroni threshold (see *Main text*).

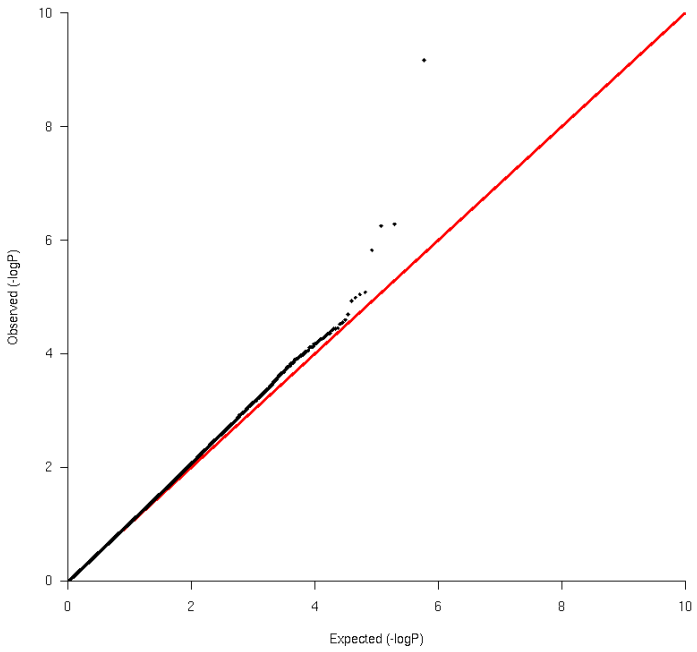


Figure 5: Linkage disequilibrium map presenting SNPs of the chromosome 6 HLA locus exhibiting strong *p-values* in the meta-analysis of GRIV and Euro-CHAVI. The linkage disequilibrium plots (r^2 and D') were drawn using the WGA Viewer software. (http://www.genome.duke.edu/centers/pg2/index_html/downloads/AnnotationSoftware).

(see next page)

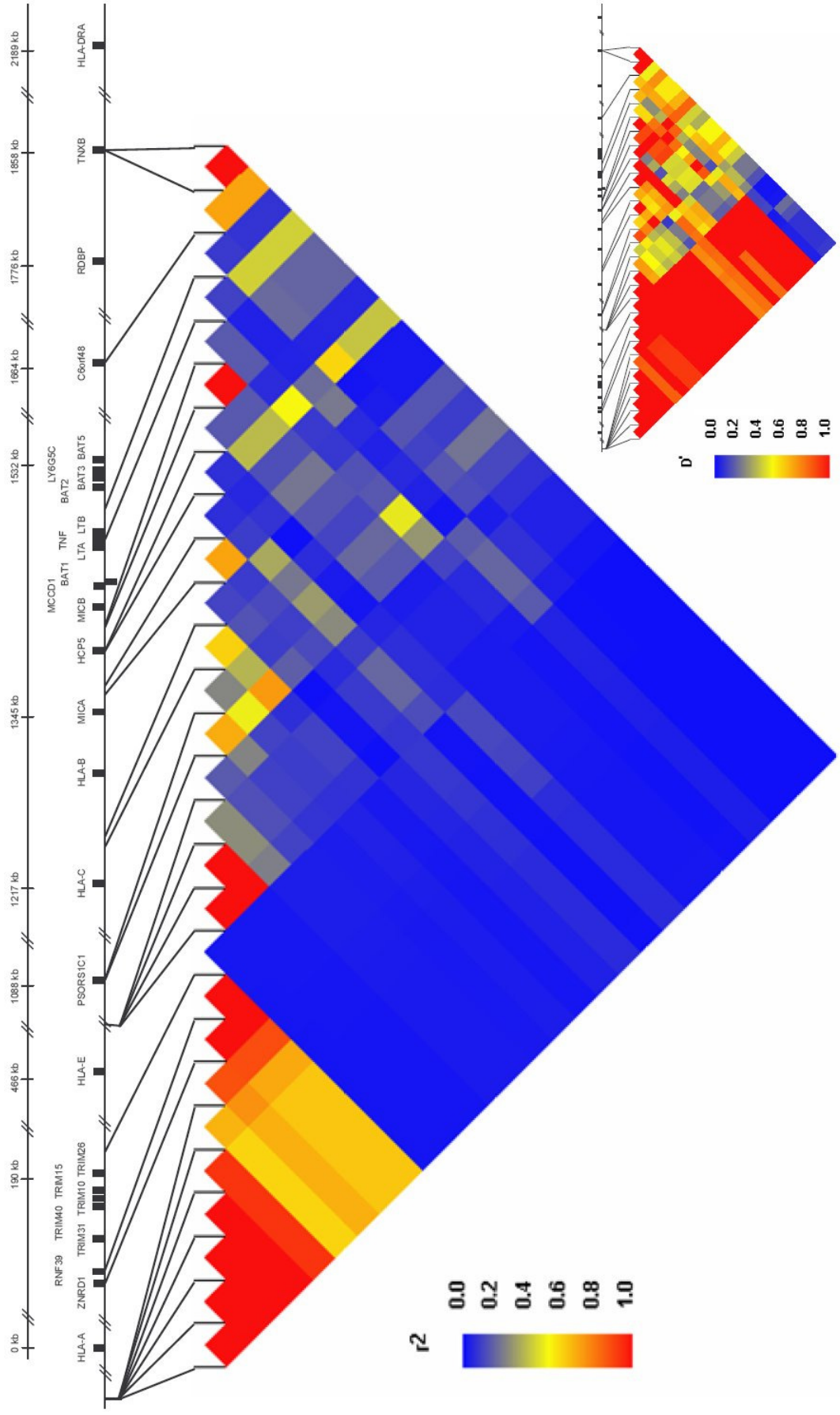

0 kb
190 kb
466 kb
1088 kb
1217 kb
1345 kb
1532 kb
1664 kb
1776 kb
1858 kb
2189 kb
HLA-A
ZNRD1
RNF39
TRIM31
TRIM40
TRIM10
TRIM15
TRIM26
HLA-E
PSORS1C1
HLA-C
HLA-B
MICA
HCP5
MICB
MCCD1
BAT1
LTA
TNF
LTB
LY6G5C
BAT2
BAT3
BAT5
C6orf48
RDBP
TNXB
HLA-DRA
r2
0.0
0.2
0.4
0.6
0.8
1.0
D'
0.0
0.2
0.4
0.6
0.8
1.0

Table 1: **List of the 50 best results obtained for the non-progressors-controls comparison.**

SNP	Chs	A1	A2	Allelic frequency (A1) %			OR (95% CI)	$Pfisher_{NP-CTR}$
				NP	CTR	SCP		
rs2395029	6	G	T	8.9	2.7	2.3	3.47 (2.39-5.04)	6.79E-10
rs9368699	6	C	T	8.2	3.2	3.5	2.74 (1.89-3.99)	5.27E-07
rs2835811	21	A	G	21.3	12.7	12.8	1.86 (1.47-2.35)	5.65E-07
rs3807024	6	G	T	18.7	11.0	9.9	1.87 (1.46-2.39)	1.51E-06
rs2844511	6	T	C	27.3	37.2	32.0	0.63 (0.52-0.77)	8.30E-06
rs2298809	11	C	T	52.4	42.0	40.7	1.52 (1.26-1.83)	8.99E-06
rs1947298	11	A	G	40.7	30.8	37.8	1.54 (1.27-1.86)	1.03E-05
rs2835762	21	T	C	21.6	13.9	9.3	1.70 (1.35-2.15)	1.18E-05
rs7045455	9	T	C	12.4	6.7	5.3	1.96 (1.46-2.64)	2.03E-05
rs12062136	1	A	C	10.0	5.1	7.7	2.08 (1.50-2.89)	2.55E-05
rs7695470	4	T	C	37.8	28.6	32.0	1.52 (1.25-1.84)	2.82E-05
rs10171238	2	C	T	16.4	24.4	24.4	0.60 (0.47-0.77)	2.99E-05
rs3731343	7	T	G	38.0	47.7	45.9	0.67 (0.56-0.81)	3.53E-05
rs7637998	3	C	T	27.3	36.5	37.2	0.65 (0.53-0.80)	3.63E-05
rs2575147	13	G	A	46.7	37.2	36.1	1.48 (1.23-1.78)	3.63E-05
rs2239177	12	C	T	35.3	44.9	51.2	0.67 (0.55-0.81)	3.87E-05
rs1032428	7	T	C	54.0	44.4	40.1	1.47 (1.22-1.77)	4.32E-05
rs12189045	5	C	T	19.6	28.0	24.1	0.63 (0.50-0.79)	4.38E-05
rs2523535	6	C	T	27.6	36.6	38.4	0.66 (0.54-0.81)	4.54E-05
rs532385	6	T	C	23.8	16.4	16.3	1.60 (1.28-1.99)	4.90E-05
rs1376483	11	C	A	52.4	42.9	50.0	1.46 (1.22-1.76)	5.05E-05
rs803105	7	G	A	57.1	47.6	54.7	1.47 (1.22-1.77)	5.35E-05
rs10964665	9	A	G	55.1	45.6	43.5	1.46 (1.22-1.76)	5.46E-05
rs2844513	6	C	T	39.6	49.1	39.5	0.68 (0.56-0.82)	5.50E-05
rs3823418	6	A	G	24.2	16.8	16.9	1.58 (1.27-1.97)	5.72E-05
rs2293732	14	C	A	48.4	39.0	41.3	1.47 (1.22-1.76)	5.96E-05
rs12198173	6	A	G	13.8	8.2	8.1	1.80 (1.36-2.38)	6.34E-05
rs4915129	1	T	C	22.9	31.4	26.7	0.65 (0.52-0.80)	6.60E-05
rs10494045	1	A	C	22.9	31.4	26.7	0.65 (0.52-0.81)	6.61E-05
rs7159	8	A	G	25.8	34.6	35.5	0.66 (0.54-0.81)	6.75E-05
rs7386386	8	T	C	55.8	46.5	51.2	1.45 (1.21-1.75)	6.80E-05
rs4452646	6	T	C	19.3	27.3	22.7	0.64 (0.51-0.80)	7.57E-05
rs13199524	6	T	C	13.6	8.0	7.6	1.81 (1.37-2.39)	7.61E-05
rs9267404	6	G	T	43.3	34.3	34.7	1.46 (1.21-1.76)	7.65E-05
rs2336645	1	A	C	21.5	29.8	26.2	0.64 (0.52-0.80)	7.67E-05
rs1110774	14	T	G	25.3	33.9	26.7	0.66 (0.54-0.81)	7.68E-05
rs13194504	6	A	G	2.0	5.8	5.8	0.33 (0.18-0.62)	7.78E-05
rs4543230	5	C	T	16.4	24.0	20.9	0.62 (0.49-0.79)	8.66E-05
rs10495545	2	T	C	6.7	12.4	8.1	0.51 (0.36-0.73)	8.88E-05
rs10123143	9	T	C	28.4	37.1	31.4	0.67 (0.55-0.82)	9.09E-05
rs10117983	9	C	T	17.1	10.9	9.3	1.68 (1.31-2.17)	9.25E-05
rs7772067	6	G	A	19.3	27.3	22.7	0.64 (0.51-0.80)	9.26E-05

rs1393350	11	A	G	17.5	25.2	21.5	0.63 (0.50-0.80)	9.38E-05
rs6832964	4	C	T	54.9	45.8	43.6	1.44 (1.20-1.74)	9.89E-05
rs11167148	8	G	A	55.8	46.7	51.2	1.44 (1.20-1.74)	1.00E-04
rs13225949	7	A	G	13.8	20.9	14.0	0.60 (0.467-0.78)	1.00E-04
rs12200614	6	C	A	15.5	22.7	17.4	0.62 (0.48-0.79)	1.06E-04
rs4947324	6	T	C	14.7	9.0	11.2	1.75 (1.33-2.29)	1.07E-04
rs870034	12	T	C	51.1	42.0	48.3	1.44 (1.20-1.734)	1.07E-04
rs2844480	6	A	G	28.0	20.3	16.3	1.52 (1.24-1.88)	1.08E-04

Note: The *p-values* were computed with Fisher's exact tests in the allelic frequency mode and presented with their corresponding allelic frequency in the different populations (NP: Non-progressors, CTR: Controls and SCP: Seropositive control population), chromosome location, Odds-Ratios (OR) and 95% confidence interval (95% CI).

Table 2: Comparative analysis of the SNPs found highly significant by the GRIV genome-wide association study and at least one other genome-wide association study.

Allele	Gene	Reference(s)	Association
rs2395029-G	*HCP5*	GRIV	promotes HIV-1 non-progression
		GRIV, Euro-CHAVI	low HIV-1 viral load
		PS/PSA	promotes Psoriasis and Psoriatic Arthritis
HLA-B*5701	*HLA-B*	GRIV, and many studies	promotes HIV-1 non-progression
		GRIV, and many studies	low HIV-1 viral load
rs9368699-C	*C6orf48*	GRIV, Euro-CHAVI[#]	promotes HIV-1 non-progression
		GRIV	low HIV-1 viral load
rs3823418-A	*PSORS1C1*	GRIV, Euro-CHAVI[#]	promotes HIV-1 non-progression
rs2248462-A	intergenic, *MICB*	GRIV, Euro-CHAVI[#]	promotes HIV-1 non-progression
rs2516509-G	intergenic	GRIV, Euro-CHAVI[#]	promotes HIV-1 non-progression
rs10484554-T	intergenic, *HLA-C*	GRIV	promotes HIV-1 non-progression
		GRIV, Euro-CHAVI	low HIV-1 viral load
		PS/PSA	promotes Psoriasis and Psoriatic Arthritis
rs3815087-T	*PSORS1C1*	GRIV, Euro-CHAVI[#]	promotes HIV-1 non-progression
		GRIV	low HIV-1 viral load
rs8321-G	*ZNRD1*	GRIV, Euro-CHAVI*	prevents HIV-1 non-progression
		Genevar database	differential expression of *ZNRD1*
rs9261290-C	*ZNRD1*	GRIV, Euro-CHAVI*	prevents HIV-1 non-progression
		Genevar database	differential expression of *ZNRD1*

Note: Our bibliographic analyses of the best SNPs found in the GRIV genome-wide association study (GWAS) shows that some of them were also found in two additional GWAS: the Euro-CHAVI study (Fellay J et al. A whole-genome association study of major determinants for host control of HIV-1. Science, 2007) and the Psoriasis/Psoriatic Arthritis (PS/PSA) study (Liu Y et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease Loci. PLoS Genet, 2008). In order to have a synthetic view of the knowledge, this table gathers all the known information regarding these SNPs from these GWAS and from other functional studies such as the *Genevar* database. The Sanger Institute *Genevar* database is a genome-wide gene expression project quantifying mRNA expression in immortalized B cell lines from the HapMap samples (http://www.sanger.ac.uk/humgen/genevar). This database allows testing the association of genotypes with gene expression.

[#] For these SNPs also found by the Euro-CHAVI study, the information regarding the polarity of the viral load according to their alleles was not available.

* For these SNPs also found by the Euro-CHAVI study, the information regarding their individual allele effects on disease progression was not available.

Table 4: The *HCP5* and *C6orf48* associations are influenced by gender.

SNP	Chs	A1	A2	Allelic frequency (A1) %						$P_{NP♂-NP♀}$	$OR_{NP♂-NP♀}$ (95% CI)	P_{NP-CTR}	OR_{NP-CTR} (95% CI)	$P_{NP♂-CTR♂}$	$OR_{NP♂-CTR♂}$ (95% CI)	$P_{NP♀-CTR♀}$	$OR_{NP♀-CTR♀}$ (95% CI)
				NP	NP♂	NP♀	CTR	CTR♂	CTR♀								
rs2395029	6	G	T	8.9	10.7	4.1	2.7	2.5	2.9	1.71E-02	0.35 (0.15-0.85)	6.79E-10	3.47 (2.39-5.04)	8.72E-10	4.71 (2.85-7.77)	4.45E-01	1.41 (0.59-3.36)
rs9368699	6	C	T	8.18	10.45	2.03	3.15	3.16	3.14	6.92E-04	0.18 (0.05-0.58)	5.27E-07	2.74 (1.89-3.99)	5.60E-08	3.68 (2.30-5.89)	4.71E-01	1.30 (0.55-3.08)

Note: The *HCP5* (rs2395029) and *C6orf48* (rs9368699) SNP associations with non-progression were only observed in men (♂) and no effect was observed in women (♀). For both SNPs, the allelic frequency in the whole non-progressors (NP) and controls (CTR), and in the male and female subgroups are presented. The *p-values*, Odds-Ratios (OR) and 95% confidence interval (95% CI) are indicated for the NP♂-NP♀, NP-CTR, NP♂-CTR♂ and NP♀-CTR♀ comparisons. The NP group comprised 201 males and 74 females.

Table 6: List of the 50 best combined *p-values* obtained by the meta-analysis of the GRIV and Euro-CHAVI data.

SNP	Chs	Alleles	$Pfisher_{NP-CTR}$	$Pfisher_{E-CHAVI}$	Pcombined
rs2395029	6	G/T	6.79E-10	9.36E-12	3.02E-19
rs9368699	6	C/T	5.27E-07	1.20E-06	1.84E-11
rs3823418	6	A/G	5.72E-05	1.11E-05	1.40E-08
rs2248462	6	A/G	5.61E-04	3.61E-06	4.26E-08
rs2516509	6	A/G	6.56E-04	3.61E-06	4.95E-08
rs10484554	6	C/T	3.78E-03	8.06E-07	6.27E-08
rs3815087	6	C/T	1.04E-03	7.09E-06	1.46E-07
rs2844513	6	C/T	5.50E-05	2.40E-04	2.53E-07
rs3093662	6	A/G	1.26E-04	1.33E-04	3.18E-07
rs13194504	6	A/G	7.78E-05	2.60E-04	3.79E-07
rs12198173	6	A/G	6.34E-05	3.34E-04	3.95E-07
rs2284178	6	C/T	4.27E-04	5.30E-05	4.21E-07
rs3749971	6	C/T	3.03E-04	8.31E-05	4.66E-07
rs8321	6	G/T	3.77E-04	6.68E-05	4.66E-07
rs2844511	6	C/T	8.30E-06	3.08E-03	4.72E-07
rs9261290	6	A/G	2.77E-04	1.00E-04	5.11E-07
rs2894207	6	C/T	1.56E-02	1.82E-06	5.23E-07
rs3130380	6	A/G	3.73E-04	8.27E-05	5.64E-07
rs1235162	6	C/T	1.20E-03	3.07E-05	6.65E-07
rs4324798	6	A/G	1.44E-04	2.60E-04	6.77E-07
rs3131093	6	C/T	1.44E-04	2.60E-04	6.78E-07
rs7756521	6	C/T	2.25E-04	1.99E-04	8.03E-07
rs3117143	6	A/C	1.97E-04	2.60E-04	9.11E-07
rs1233579	6	A/G	2.01E-04	2.60E-04	9.28E-07
rs13199524	6	C/T	7.61E-05	8.22E-04	1.10E-06
rs4713385	6	A/G	2.02E-04	4.48E-04	1.55E-06
rs4713380	6	C/T	2.09E-04	4.48E-04	1.61E-06
rs2844480	6	A/G	1.08E-04	9.64E-04	1.78E-06
rs9295924	6	A/G	2.12E-04	4.92E-04	1.78E-06
rs2516400	6	C/T	1.68E-03	9.74E-05	2.72E-06
rs2746150	6	C/T	7.51E-04	2.25E-04	2.80E-06
rs6457374	6	C/T	6.40E-04	2.69E-04	2.85E-06
rs1051794	6	A/G	1.12E-04	1.76E-03	3.25E-06
rs13437082	6	C/T	1.61E-04	1.38E-03	3.63E-06
rs1265099	6	C/T	1.47E-04	1.65E-03	3.93E-06
rs2835811	21	A/G	5.65E-07	4.55E-01	4.15E-06
rs4711269	6	C/T	1.94E-04	1.38E-03	4.31E-06
rs9378109	6	A/C	6.20E-04	4.76E-04	4.73E-06
rs8192591	6	A/G	1.77E-04	1.88E-03	5.29E-06
rs2523535	6	C/T	4.54E-05	1.02E-02	7.20E-06

rs2523619	6	A/G	7.36E-02	6.32E-06	7.25E-06
rs3132685	6	C/T	8.69E-03	6.73E-05	8.98E-06
rs10484399	6	A/G	3.33E-04	1.85E-03	9.42E-06
rs4131373	1	C/T	9.44E-02	9.14E-06	1.29E-05
rs13194781	6	A/G	5.52E-04	1.85E-03	1.51E-05
rs2844498	6	A/G	1.24E-03	9.35E-04	1.70E-05
rs4778192	15	C/T	3.12E-03	3.79E-04	1.73E-05
rs2523608	6	C/T	9.49E-04	1.31E-03	1.82E-05
rs4947324	6	C/T	1.07E-04	1.23E-02	1.92E-05
rs2298809	11	C/T	8.99E-06	1.52E-01	1.98E-05

Note: For each SNP, the chromosome, the alleles, the *p-values* obtained in each study and the combined *p-value* are presented.

2. Etude 'génome entier' sur les progresseurs rapides de la cohorte GRIV

ARTICLE 2

Genomewide Association Study of a Rapid Progression Cohort Identifies New Susceptibility Alleles for AIDS (ANRS Genomewide Association Study 03)

Sigrid Le Clerc *, Sophie Limou *, Cédric Coulonges, Wassila Carpentier, Christian Dina, Lieng Taing, Olivier Delaneau, Taoufik Labib, Rob Sladek, ANRS Genomic Group, Christiane Deveau, Hélène Guillemain, Rojo Ratsimandresy, Matthieu Montes, Jean-Louis Spadoni, Amu Therwath, François Schächter, Fumihiko Matsuda, Ivo Gut, Jean-Daniel Lelièvre, Yves Lévy, Philippe Froguel, Jean-François Delfraissy, Serge Hercberg, and Jean-François Zagury

J Infect Dis. 2009 ; 200(8):1194-201.

RESUME

Lorsque nous avons entrepris l'étude d'association 'génome entier' sur les progresseurs rapides de la cohorte GRIV, toutes les analyses jusque là n'avaient porté que sur des phénotypes de contrôle de la charge virale [137] ou de non-progression [176].

Nous avons réalisé une étude de type cas-contrôle sur 85 progresseurs rapides de la cohorte GRIV comparé à 1352 individus contrôles sur 291,119 SNPs autosomaux, respectant les critères de contrôle qualité permettant ainsi la découverte de facteurs de prédisposition à progresser rapidement vers la phase SIDA. Nous avons utilisé le False Dicovery Rate (FDR) pour définir notre seuil statistique de significativité

Plusieurs nouvelles associations avec le phénotype de progression rapide ont été identifiées (FDR<25%) : *PRMT6* (p=6,1x10^{-7}; odds ratio : 0,24), *SOX5* (p=1,8x10^{-6}; odds ratio, 0,45), *RXRG* (p=3,9x10^{-6}; odds ratio, 3,29), et *TGFBRAP1* (p=7x10^{-6}; odds ratio: 0,34) ; L'analyse des haplotypes a identifié des SNPs dans les régions promotrices et exoniques potentiellement importantes dans la fonction des gènes *PRMT6* et *TGFBRAP1.*

La pertinence biologique et statistique de ces associations souligne la puissance de l'utilisation des phénotypes extrêmes dans les études d'associations 'génome entier', y compris avec des échantillons de taille limitée. Ces résultats mettent en avant le rôle de la voie de signalisation du TGFβ dans la pathogenèse liée au VIH.

MAJOR ARTICLE

Genomewide Association Study of a Rapid Progression Cohort Identifies New Susceptibility Alleles for AIDS (ANRS Genomewide Association Study 03)

Sigrid Le Clerc,[1,2,4,a] Sophie Limou,[1,2,4,5,a] Cédric Coulonges,[1,2] Wassila Carpentier,[2] Christian Dina,[6] Lieng Taing,[1] Olivier Delaneau,[1] Taoufik Labib,[1,4] Rob Sladek,[8] ANRS Genomic Group,[2,b] Christiane Deveau,[2] Hélène Guillemain,[1] Rojo Ratsimandresey,[1] Matthieu Montes,[1] Jean-Louis Spadoni,[1] Amu Therwath,[3] François Schächter,[1] Fumihiko Matsuda,[9] Ivo Gut,[5] Jean-Daniel Lelièvre,[4] Yves Lévy,[4] Philippe Froguel,[6,10] Jean-François Delfraissy,[2] Serge Hercberg,[7] and Jean-François Zagury[1,2,4]

[1]Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, [2]Agence Nationale de Recherche sur le SIDA (ANRS) Genomic Group, and [3]Laboratoire d'Oncologie Moléculaire, Paris, and [4]Université Paris 12, Institut National de la Santé et de la Recherche Médicale (INSERM) U955, Créteil, [5]Commissariat à L'Energie Atomique/Institut de Génomique, Centre National de Génotypage, Evry, [6]Unité Mixte de Recherche (UMR) Centre National de la Recherche Scientifique 8090, Institut Pasteur de Lille, Lille Cedex, and [7]UMR U557 INSERM/U1125 Institut National de la Recherche Agronomique/Conservatoire National des Arts et Métiers/Université Paris 13, and Centre de Recherche en Nutrition Humaine Ile-de-France, Santé-Médecine-Biologie Humaine (SMBH) Paris 13, Bobigny, France; [8]Department of Human Genetics, Faculty of Medicine, McGill University, and Génome Québec Innovation Centre, Montreal, Canada; [9]INSERM U852, Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [10]Genomic Medicine, Hammersmith Hospital, Imperial College London, London, United Kingdom

Background. Previous genomewide association studies (GWASs) of AIDS have targeted end points based on the control of viral load and disease nonprogression. The discovery of genetic factors that predispose individuals to rapid progression to AIDS should also reveal new insights into the molecular etiology of the pathology.

Methods. We undertook a case-control GWAS of a unique cohort of 85 human immunodeficiency virus type 1 (HIV-1)–infected patients who experienced rapid disease progression, using Illumina HumanHap300 BeadChips. The case group was compared with a control group of 1352 individuals for the 291,119 autosomal single-nucleotide polymorphisms (SNPs) passing the quality control tests, using the false-discovery rate (FDR) statistical method for multitest correction.

Results. Novel associations with rapid progression (FDR, ≤25%) were identified for *PRMT6* ($P = 6.1 \times 10^{-7}$; odds ratio [OR], 0.24), *SOX5* ($P = 1.8 \times 10^{-6}$; OR, 0.45), *RXRG* ($P = 3.9 \times 10^{-6}$; OR, 3.29), and *TGFBRAP1* ($P = 7 \times 10^{-6}$; OR, 0.34). The haplotype analysis identified exonic and promoter SNPs potentially important for *PRMT6* and *TGFBRAP1* function.

Conclusions. The statistical and biological relevance of these associations and their high ORs underscore the power of extreme phenotypes for GWASs, even with a modest sample size. These genetic results emphasize the role of the transforming growth factor β pathway in the pathogenesis of HIV-1 disease. Finally, the wealth of information provided by this study should help unravel new diagnostic and therapeutic targets.

Genomewide association studies (GWASs) may provide new insights into the molecular etiology of complex diseases by discovering unsuspected genetic risk factors and, as a consequence, identify new diagnostic or therapeutic targets [1]. Reports of 3 GWASs of AIDS have already been published [2–4] and described mainly

Received 8 April 2009; accepted 28 May 2009; electronically published 15 September 2009.

Reprints or correspondence: Dr Jean-François Zagury, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France (zagury@cnam.fr).

The Journal of Infectious Diseases 2009;200:1194–201

0022-1899/2009/20008-0003$15.00
DOI: 10.1086/605892

Potential conflicts of interest: none reported.

Financial support: Agence Nationale de Recherche sur le SIDA (ANRS); Innovation 2007 program of Conservatoire National des Arts et Métiers; AIDS Cancer Vaccine Development Foundation; Neovacs SA; Vaxconsulting. S.L. benefits from a fellowship from the French Ministry of Education, Technology, and Research, and S.L.C. benefits from a fellowship from the ANRS.

[a] S.L.C. and S.L. contributed equally to this work.

[b] The ANRS Genomic Group oversees the AIDS genomic projects of the ANRS; study group members are listed at the end of the text.

genes involved in the control of viral load. Indeed, these 3 GWASs identified the *HCP5* rs2395029 polymorphism, which is in linkage disequilibrium with HLA-B*57 and other immunity genes, such as *MICB*, *TNF*, *LTB*, *BAT1*, and *HLA-C*. An association with the HIV DNA level (reservoir) was also depicted by the PRIMO GWAS [2] for *SDC2* (chromosome 8), whose encoded protein is required for Tat internalization. Finally, the *ZNRD1* locus of chromosome 6 was associated with the control of disease progression but not of viral load in both the Euro-CHAVI (Center for HIV/AIDS Vaccine ImmunologyCenter for HIV/AIDS Vaccine Immunology) [3] and the nonprogressor Genomics of Resistance to Immunodeficiency Virus (GRIV) [4] GWASs.

To identify genetic loci predisposing a person to rapid progression of AIDS rather than disease control, we undertook a case-control GWAS involving a unique cohort of human immunodeficiency virus type 1 (HIV-1)–infected patients who experienced rapid progression, using Illumina HumanHap300 BeadChips. In a manner symmetric to the published nonprogressor GRIV GWAS [4], the use of this extreme phenotype should lead to an enrichment of our knowledge of the genetic factors involved in rapid disease progression. The power of this extreme design has indeed been demonstrated by previous candidate gene studies [5–7].

METHODS

The GRIV cohort. The GRIV cohort was established in France in 1995 to generate a large collection of DNA for genetic studies to identify host genes associated with either rapid progression or nonprogression to AIDS [5, 7]. Only white people of European descent living in France were eligible for enrollment to reduce confounding by population substructure. These criteria limit the influence of ethnic and environmental factors (all subjects live in a similar environment and are infected by B strains) and emphasize how the genetic makeup of each individual determines the various patterns of progression. Rapid progressors were included on the basis of the main clinical outcomes, CD4 T cell count and time to disease progression, and were defined as those who had 2 or more CD4 T cell counts below 300 cells/mm^3 within 3 years after the last seronegative test result. DNA was obtained from fresh peripheral blood mononuclear cells or from Epstein-Barr virus–transformed cell lines. The rapid progression group ($n = 85$) was composed of 73 men and 12 women aged 21–55 years (median, 32 years) at inclusion. At inclusion, the median CD4 T cell count was 230 cells/mm^3 (minimum and maximum values, 20 and 297 cells/mm^3). All patients provided written informed consent before enrollment in the GRIV GWAS.

The seropositive control population. To discriminate between positive signals corresponding to either an association with rapid progression or an association with HIV-1 infection, we needed a group of seropositive control subjects who were not rapid progressors. For that, we used 275 white French subjects who qualified as nonprogressors to AIDS (ie, those who had an asymptomatic HIV-1 infection for >8 years, no receipt of treatment, and a CD4 T cell count consistently remaining at >500 cells/mm^3). This control group was composed of 201 men and 74 women aged 19–62 years (median, 35 years) at inclusion. The median CD4 T cell count of this seropositive control population was 706 cells/mm^3 (minimum and maximum values, 501 and 2298 cells/mm^3).

The SU.VI.MAX seronegative control group. The SU.VI.MAX (Supplémentation en Vitamines et Minéraux Antioxydants) study was a randomized, double-blind, placebo-controlled, primary-prevention trial designed to test the efficacy of daily supplements of antioxidant vitamins and minerals at nutrition-level doses in reducing several major health problems in industrialized countries, especially the main causes of premature death, cancers and cardiovascular diseases. This cohort study was started in 1994 in France and included 12,735 subjects [8]. The control group genotyped in the present study comprised 1352 representative SU.VI.MAX participants, all of them white persons living in France who were HIV-1 seronegative. This control cohort was composed of 525 men and 827 women, with a mean age of 53.1 and 48.5 years, respectively.

The second seronegative control group. The D.E.S.I.R. (Data from an Epidemiological Study on Insulin Resistance Syndrome) program was a 9-year follow-up study designed to clarify the development of insulin resistance syndrome. Subjects were recruited from 1994 to 1996 from volunteers insured by the French social security system, which offers periodic health examinations free of charge [9]. This second control group comprised 697 participants who were both not obese and normoglycemic from the D.E.S.I.R. trial, all French and HIV-1 seronegative. It was composed of 281 men and 416 women aged 30–64 years.

Genotyping method. The GRIV cohort and the 2 seronegative groups were genotyped using Illumina Infinium II HumanHap300 BeadChips (Illumina). Genomic DNA (750 ng) was whole-genome amplified, fragmented, denatured, and hybridized on prepared HumanHap300 BeadChips for a minimum of 16 h at 48°C. Non–specifically hybridized fragments were removed by washing, and the remaining specifically hybridized DNA was fluorescently labeled by a single base extension reaction and detected using a BeadArray scanner (Illumina). Normalized bead-intensity data obtained for each sample were loaded into BeadStudio software (version 3.1; Illumina), which converted fluorescence intensities into single-nucleotide polymorphism (SNP) genotypes.

Quality control. Using the BeadStudio software, we analyzed the crude genotyping data, and SNPs were filtered according to the following parameters. First, samples with a call

Table 1. Best Results Obtained for the Comparison between Rapid Progressors and Control Subjects

SNP	Gene	Chr	Chr position	A1	A2	Allelic frequency (A1), %				OR (95% CI)	Fisher *P* Value, RP-$CTR_{SU.VI.MAX}$	FDR *q* value
						RP	$CTR_{SU.VI.MAX}$	$CTR_{D.E.S.I.R.}$	SCP			
rs4118325	Intergenic[a]	1	107379355	A	G	5.2	19.0	18.7	15.8	0.24 (0.12–0.46)	6.09×10^{-7}	0.17
rs1522232	*SOX5*	12	24285639	T	C	29.1	47.7	48.2	50.7	0.45 (0.32–0.63)	1.80×10^{-6}	0.20
rs1360517	Intergenic	9	12997129	A	G	16.3	5.9	7.0	5.8	3.09 (2.00–4.78)	3.27×10^{-6}	0.20
rs3108919	Intergenic	8	101910722	C	T	43.6	26.6	27.8	27.5	2.13 (1.56–2.91)	3.86×10^{-6}	0.20
rs10800098	*RXRG*	1	163675719	A	G	14.5	4.9	4.7	5.6	3.29 (2.08–5.20)	3.86×10^{-6}	0.20
rs10494056	Intergenic[b]	1	107349442	A	C	5.8	18.6	17.9	16.1	0.27 (0.14–0.52)	4.29×10^{-6}	0.20
rs12351740	Intergenic[c]	9	13010010	T	C	12.8	4.1	4.7	3.8	3.46 (2.12–5.62)	6.63×10^{-6}	0.25
rs1020064	*TGFBRAP1*[d]	2	105264172	T	G	9.3	23.2	25.6	25.2	0.34 (0.20–0.57)	7.04×10^{-6}	0.25

NOTE. *P* values were computed by the Fisher exact test in the allelic frequency mode and are presented with their corresponding allelic frequencies in the various populations (rapid progressors [RP], seronegative control subjects [$CTR_{SU.VI.MAX}$], and the seropositive control population [SCP]), chromosome (Chr) positions, odds ratios (ORs) with 95% confidence intervals (95% CIs), and false-discovery rate (FDR) *q* values. The frequencies in the D.E.S.I.R. second seronegative control group ($CTR_{D.E.S.I.R.}$) are also indicated and are similar to those in the SU.VI.MAX cohort. SNP, single-nucleotide polymorphism.

[a] This intergenic SNP is in linkage disequilibrium with *PRMT6* and intergenic SNPs.

[b] This intergenic SNP is in linkage disequilibrium with *PRMT6* and intergenic SNPs ($r^2 = 0.92$ with rs4118325).

[c] This intergenic SNP is in partial linkage disequilibrium with the intergenic SNP rs1360517 ($r^2 = 0.68$).

[d] This *TGFBRAP1* SNP is in linkage disequilibrium with an intergenic SNP.

rate (percentage of SNPs genotyped by sample) <95% in the Illumina clusters were deleted. Second, the SNPs with a call frequency (percentage of samples genotyped by SNP) <99% were reclustered. Third, after reclustering, samples with a call rate below 97% were deleted. The clustering step can create SNP genotyping errors, which can be prevented by following the Illumina procedure (see http://www.illumina.com/downloads/GTDataAnalysis_TechNote.pdf). This method evaluates the quality of the newly created clusters according to several criteria, which can be manually checked and corrected as necessary. In total, 1300 SNPs were excluded by this Illumina quality control procedure. Finally, after all the quality control steps were performed, the 15,731 SNPs with a call frequency <98% (>2% of missing data) were excluded. This procedure ensures reliable genotyping data with little missing data.

Hardy-Weinberg equilibrium analysis was performed for each SNP in each group by using an exact statistical test implemented in PLINK software ([10]; available at: http://pngu.mgh.harvard.edu/~purcell/plink/). Deviation from Hardy-Weinberg equilibrium in a group of patients suggests that the SNP has a biological effect, while deviation in the control group or all groups suggests a systematic error in genotyping. The 1475 SNPs that were not in the Hardy-Weinberg equilibrium in the SU.VI.MAX control group ($P < 1.0 \times 10^{-3}$) were rejected in this way.

In total, 235 SNPs with low minor allelic frequency (<1%) in the global population were also filtered.

Haplotype inference. Haplotype inference was obtained using the rapid and accurate Shape-IT algorithm [11].

Linkage disequilibrium. For each SNP exhibiting a significant association, we looked for the other SNPs in linkage disequilibrium ($r^2 \geq 0.8$) in the HapMap population of Western European ancestry (CEU, HapMap data Release 21a/phase II, January 2007, on NCBI B35 assembly, dbSNP125; available at http://www.hapmap.org) to identify the genes possibly concerned by the associations. A SNP was assigned to a gene if it was located in the gene or in the 2 kb flanking regions (potential regulatory sequence); otherwise, it was considered intergenic.

Statistical analysis. For each SNP, we performed a standard case-control analysis by using the Fisher exact test (with PLINK software) to compare allelic distributions between the rapid progressors and the control subjects.

To take into account the multiple tests while controlling for the risk of false discovery, we computed for each *P* value a false-discovery rate (FDR) under the form of a *q* value: the *q* value is an estimate of the proportion of false-positive signals below a threshold *P* value [12]. The FDR computation is more complex but more powerful than the standard Bonferroni corrections [13–15], because it allows the identification of more true-positive signals. For polyfactorial diseases in which several genes are at stake, it thus provides a more adapted outlook on the GWAS results than the "all or nothing" Bonferroni cutoff. We thus used an FDR approach called local base estimating, with a 25% threshold for our case-control study.

For all the SNPs meeting this statistical threshold (Table 1), the quality of genotyping was individually rechecked with the BeadStudio software. We also checked that the allelic frequen-

This figure is available in its entirety in the online version of the *Journal of Infectious Diseases.*

Figure 1. Number of independent ($r^2 < 0.5$) single-nucleotide polymorphisms (SNPs) meeting the false-discovery rate threshold of 25%.

This figure is available in its entirety in the online version of the *Journal of Infectious Diseases*.

Figure 2. Quantile-quantile plot for expected (*red*) versus observed (*black*) *P* values from the comparison of rapid progressors with control subjects.

cies in the seropositive control population were similar to those in the seronegative SU.VI.MAX control population for those SNPs of interest, confirming that the observed associations were indeed linked to rapid progression.

Statistical reliability of the rapid progression results. The GRIV rapid progression cohort is unique, and no other independent cohort was readily available for replication. To check in an independent fashion the statistical relevance of the associations with rapid progression, we decided to evaluate the odds of obtaining just as many associations with the same statistical method (ie, the Fisher exact test and FDR cutoff) by comparing the genotypes of 1000 control subgroups (randomly extracted from the D.E.S.I.R. control cohort and composed of 85 subjects each) with the genotypes of the entire SU.VI.MAX control group. For each simulation, we counted the number of independent ($r^2 < 0.5$) SNPs meeting the statistical threshold of an FDR of 25% (Figure 1).

Identification of population stratification. To correct for possible population stratification at the intercontinental level, case and control genotypes were analyzed using STRUCTURE software (version 2.2; see [16] and http://pritch.bsd.uchicago.edu/software.html). We selected a set of 328 SNPs informative for ancestral origin (*F* statistics fixation index, >0.2) based on the Perlegen data set and separated by 5 Mb to avoid linkage disequilibrium. We also included genotypes obtained from unrelated individuals representing the 3 populations studied by the HapMap project to better separate our rapid progressors and control individuals according to their continent of origin. While nearly all case and control subjects were within the range of the white individuals from HapMap, one individual was outside the white subjects cluster in the rapid progressors (decreasing the rapid progression group from 86 to 85 subjects).

To avoid spurious associations resulting from possible population stratification or genotyping errors, a quantile-quantile plot was also produced by plotting the ranked values of the test statistics against the approximated expected order statistic (Figure 2). We also computed the genomic inflation factor λ [17]. The result (λ = 1.038), along with the quantile-quantile plot, suggested little overall effect of stratification.

RESULTS

After the quality control tests, a total of 291,119 autosomal SNPs were tested for association with rapid progression. Figure 3 depicts the distribution of the *P* values along the chromosomes, and Table 1 presents the best signals (for an extended list, see Table 2). Eight associations with FDRs ⩽25% were identified, corresponding to 6 independent ($r^2 < 0.5$) SNPs. Since no replication was readily available, we performed simulations on independent control populations to evaluate the reliability of these results (see Methods and Figure 1). Overall, less than 1% of these tests yielded 6 or more independent signals with a FDR ⩽25% (mean ± standard deviation = 0.4 ± 1.29), suggesting that most of the rapid progression associations found in this study are likely to be true positives. This result underscores the statistical relevance of the rapid progression associations, even with a modest sample size.

The best rapid progression signals were observed for SNPs on chromosome 1 (rs4118325 [$P = 6.09 \times 10^{-7}$] and rs10494056 [$P = 4.29 \times 10^{-6}$]) (Figure 3), which are in linkage disequilibrium themselves ($r^2 = 0.92$) and are also in linkage disequilibrium with SNPs of the *PRMT6* gene. The rs4118325-A allele

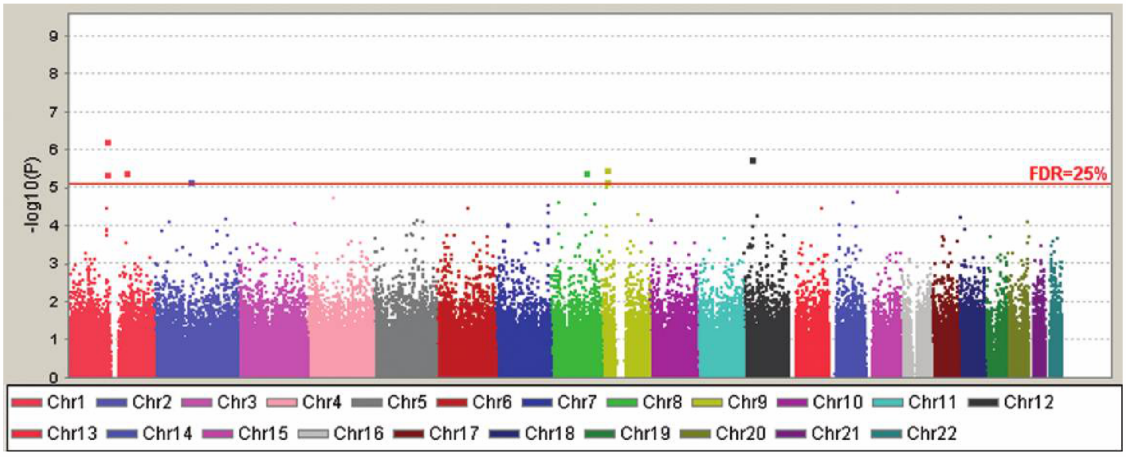


Figure 3. Distributions along the human autosomes of $-\log_{10}$(*P* values) obtained for the comparison of rapid progressors with control subjects. The red line marks the false-discovery rate threshold of 25%. Chr, chromosome.

Table 2. Fifty Best *P* Values Obtained for the Comparison between Rapid Progressors and Control Subjects

This table is available in its entirety in the online version of the *Journal of Infectious Diseases*.

and/or the alleles in linkage disequilibrium are associated with prevention of rapid progression, with an odds ratio (OR) of 0.24 [95% confidence interval [CI], 0.12–0.46] (Table 1). Another association was identified on chromosome 1 in *RXRG* gene ($P = 3.86 \times 10^{-6}$, OR, 3.29 [95% CI, 2.08–5.20]). Finally, SNPs modulating rapid progression were also found in *SOX5* ($P = 1.80 \times 10^{-6}$; OR, 0.45 [95% CI, 0.32–0.63]) and *TGFBRAP1* ($P = 7.04 \times 10^{-6}$; OR, 0.34 [95% CI, 0.20–0.57]). Two other independent SNPs, rs1360517 and rs3108919, met the FDR threshold of 25% (Table 1) but were not close to any known gene (distance, <20 kb).

To complete our analysis, we explored the influence of covariables—such as *CCR5*-Δ32 and *CCR5-P1* haplotypes [7], the HIV-1 infection mode (mucosal or parenteral), age at seroconversion, and sex—on all the associations presented in Table 1. None of them was found to affect the observed signals.

To deepen our genomic analysis, we also tried to combine our results with previously published data from the EuroCHAVI cohort, which assessed the viral load end point [3]. We combined their *P* values with ours using the classical Fisher method. Unfortunately, no combined *P* values met the FDR threshold of 25% (data not shown; best combined *P* value, 6.23×10^{-5}). The lack of significant common signals between the 2 studies may likely stem from the difference in inclusion criteria, in particular the use of viral load versus the use of

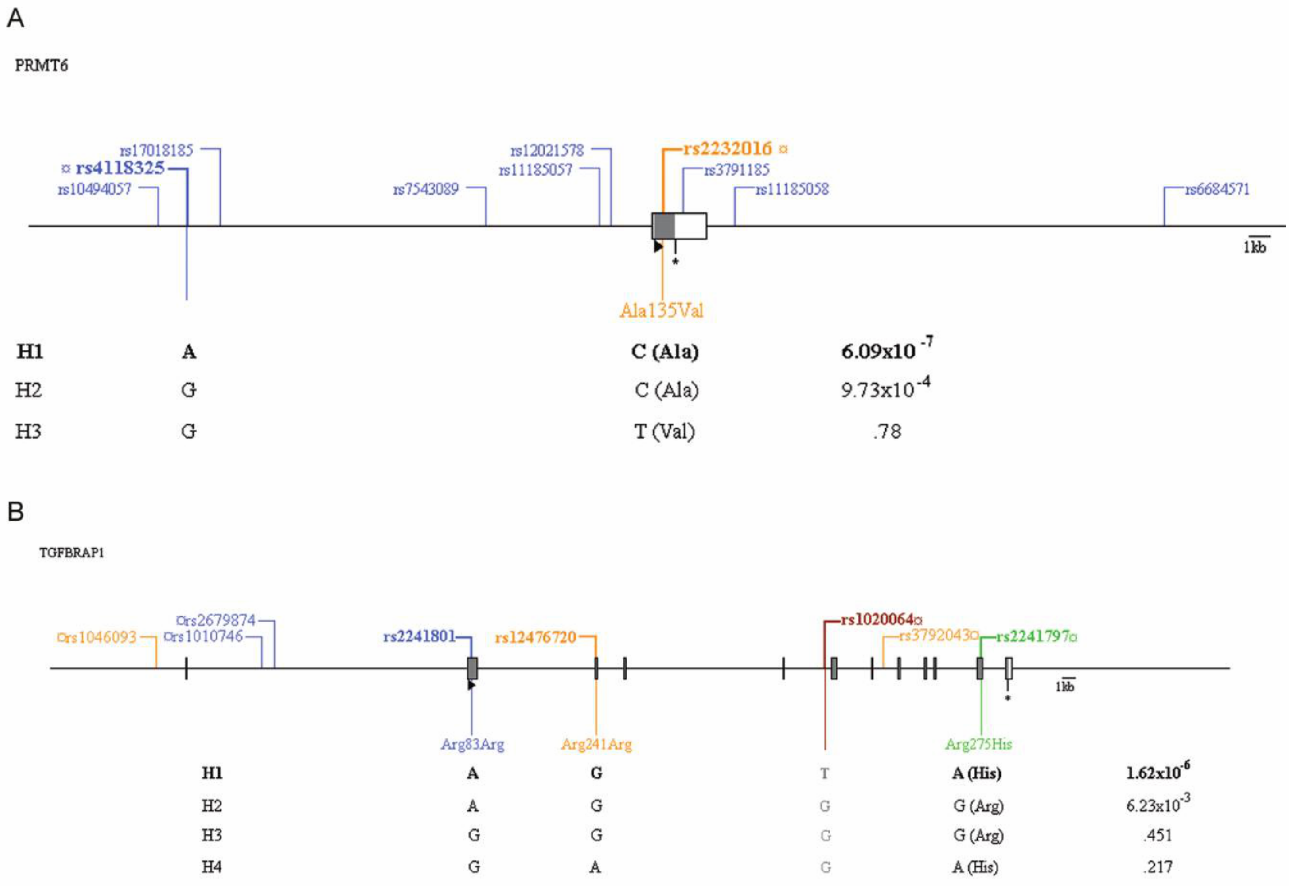


Figure 4. Haplotype maps for *PRMT6* (*A*) and *TGFBRAP1* (*B*). Exons and untranslated regions are symbolized by shaded and unshaded boxes, respectively. The positions of the ATG and stop codons are indicated by a triangle (▶) and by an asterisk (*), respectively. Single-nucleotide polymorphisms (SNPs) in high linkage disequilibrium ($r^2 > 0.9$) are represented with the same color, and the genotyped tagSNPs are marked with the symbol ¤. For greater clarity, only the exonic polymorphisms in linkage disequilibrium with the genotyped tagSNPs ($r^2 > 0.9$) are specified in panel B. The numbers in the right column correspond to the *P* values obtained when comparing the allelic frequency of each haplotype between rapid progressors and control subjects.

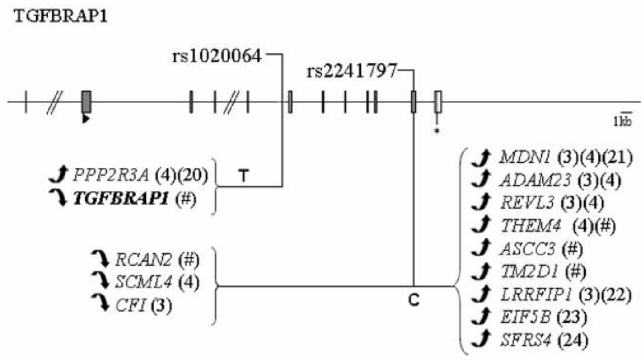


Figure 5. Correlation between some *TGFBRAP1* polymorphisms and differential gene expression according to the Genevar database [18] and the Dixon database [19]. Exons and untranslated regions are symbolized by shaded and unshaded boxes, respectively. The positions of the ATG and stop codons are indicated by a triangle (▶) and by an asterisk (*), respectively. The modulation of gene expression is indicated by the arrow direction: increased (↑) or decreased (↓) expression. The numbers correspond to the bibliographic references of previous works linking these genes to AIDS. The present genomewide association study of rapid progression is indicated by a pound sign (#).

CD4 T cell count, and also from the elimination of very rapid progressors in the Euro-CHAVI study, since such patients could not exhibit a sufficiently prolonged viral load set point [3].

In the past, we have often observed that haplotypes may be more informative than individual SNPs; this was notably shown for *CCR5* [7], *CXCR1* [6], and *HLA* [5]. We thus decided to explore the haplotype patterns for the genes exhibiting the best signals in our rapid progression GWAS, limiting the investigation to exonic and promoter SNPs (Figure 4). For *PRMT6*, we demonstrated that the effect of the *PRMT6* SNP rs4118325 could be tracking a haplotype (composed of rs4118325 and rs2232016 [Ala135Val]) through linkage disequilibrium ($r^2 = 0.12$, $D' = 1$) (Figure 4*A*). No differential messenger RNA (mRNA) expression could be significantly associated with either of these SNPs using the Genevar database [18] or the Dixon database [19].

For *TGFBRAP1*, the rs1020064 SNP was in high linkage disequilibrium ($r^2 = 0.97$) with a haplotype containing 3 exonic SNPs (rs2241801 [Arg83Arg], rs12476720 [Arg241Arg], and rs2241797 [Arg275His]) (Figure 4*B*). The SNP rs2241797 (Arg275His) appeared to be essential, since the signal disappeared when that SNP was removed from the haplotype ($P > 5.0 \times 10^{-2}$) and remained identical when either of the 2 synonymous SNPs were removed ($P = 1.62 \times 10^{-6}$). Interestingly, rs1020064 has been associated with differential expression of *TGFBRAP1* in the Genevar database and of *PPP2R3A* [20] in the Dixon database (Figure 5). More strikingly, in the Dixon database the nonsynonymous SNP rs2241797 was significantly associated with the differential expression of several proteins (Figure 5), among which 4 have been independently described to interact with HIV-1 (*MDN1* [21], *LRRFIP1* [22], *EIF5B* [23], and *SFRS4* [24]), and several have exhibited a positive association in an AIDS GWAS. Of note, no differential expression could be found for the 2 synonymous SNPs rs2241801 and rs12476720. For the *SOX5* rs1522232 and *RXRG* rs10800098 SNPs, no haplotype involving exonic or promoter SNPs and no differential mRNA expression could be found.

DISCUSSION

For the first time, the extreme HIV-1 rapid progression phenotype was specifically examined in a GWAS. The power of using the GRIV extreme design has been demonstrated by previous candidate gene studies [5–7, 25] and by the previous GWAS of nonprogressors [4]. Here we have identified 6 novel associations with rapid progression with ORs as high as 4, emphasizing the power of this extreme design in spite of a relatively modest sample size. As a comparison, there was only 1 independent ($r^2 < 0.5$) signal passing the FDR threshold of 25% in the GWAS of the nonprogressor GRIV cohort [4]. No replication was readily available, since this specific rapid progression design is rather unique in the world. The biological relevance of the rapid progression associations and their statistical validation through simulations with a second control group (<1% odds of finding as many independent signals) are nevertheless compelling. The use of an extreme population may provide a strong contrast and indeed help unravel new genetic factors, behaving as a magnifying glass [25, 26].

Four of the 6 SNPs associated with rapid progression to AIDS were clearly linked to a gene (distance, <2 kb). *SOX5* (OR, 0.45) encodes a transcription regulator notably expressed in lymph nodes and lymphoid tissues [27] and is also known to be involved in the transforming growth factor β (TGF-β)/SMAD chondrogenesis signaling pathway [28, 29]. There is no other experimental evidence linking this protein to the pathogenesis of HIV-1 infection. *RXRG* (OR, 3.29) encodes a retinoic acid nuclear receptor mediating the antiproliferative effects of retinoic acid and is known to repress HIV-1 transcription and replication [30, 31]. Several studies have also associated vitamin A deficiency with a bad prognosis in patients with AIDS, but the benefit of vitamin A supplementation in AIDS patients is still controversial [32, 33]. The genetic association for *PRMT6* (OR, 0.24) points toward a direct interaction between the host and the virus. Indeed, *PRMT6* codes for an arginine *N*-methyltransferase previously reported to methylate HIV-1 Tat and Rev, impairing their function [34, 35]. The product of *PRMT6* also methylates the high mobility group protein HMGA1, thereby modifying HMGA1 interactions with DNA [36], which could alter HIV-1 integration into the human genome [37]. The modulation of *PRMT6* transcription could

thus affect HIV-1 integration or replication and prevent rapid progression. Finally, an association with rapid progression was discovered for *TGFBRAP1* (OR, 0.34). *TGFBRAP1* is expressed in most lymphoid tissues and is involved in the TGF-β signaling pathway [38]. TGF-β is a pleiotropic immunosuppressive cytokine involved in immune homeostasis and in the differentiation of and balance between type 17 T helper (Th17) cells (T cells protecting the mucosal barrier integrity [39]) and regulatory T (T_{reg}) cells (T cells essential for immune suppression [40]). Interestingly, recent works have also supported a combined role for TGF-β and retinoic acid for T_{reg} cell differentiation [41, 42]. T_{reg} cells are rapidly induced after simian immunodeficiency virus (SIV) and HIV-1 infection [43, 44] and may have a deleterious effect during the chronic phase of infection [45, 46]. Moreover, an impairment in the balance between Th17 and T_{reg} cells during HIV-1 infection was recently described [40, 47, 48]. Finally, TGF-β can stimulate HIV-1 Tat transcription [49] and, reciprocally, be induced by Tat early during infection [50], supporting a key role for the intermingled effects of TGF-β and Tat in the early development of HIV-1 infection. Overall, our results support the central role played by the TGF-β pathway and the balance between Th17 and T_{reg} cells in AIDS progression. Interestingly, *PRMT6* and *TGFBRAP1* haplotypes involving promoter and exonic SNPs were also associated with disease progression; moreover, some of the haplotype SNPs could be associated with downstream differential mRNA expression (Figure 5).

In conclusion, as for all genetic studies, our results—including SNPs with higher *P* values, which were not discussed—will need to be confirmed by replication in other cohorts and by other investigations, such as fine gene mapping and biological experiments.

ANRS GENOMIC GROUP

The ANRS Genomic Group is composed of Prof Jean-François Delfraissy (Agence Nationale de Recherche sur le SIDA, Paris), Dr Laurence Meyer (Hôpital Kremin-Bicêtre, France), Prof Philippe Broët (Hôpital Kremlin-Bicêtre, France), Dr Cyril Dalmasso (Hôpital Kremlin-Bicêtre, France), Prof Patrice Debré (Hôpital La Salpétrière, Paris), Dr Ioannis Théodorou (Hôpital La Salpétrière, Paris), Prof Christine Rouzioux (Hôpital Necker, Paris), Cédric Coulonges (Conservatoire National des Arts et Métiers, Paris), Sigrid Le Clerc (Conservatoire National des Arts et Métiers, Paris), Sophie Limou (Conservatoire National des Arts et Métiers, Paris), and Prof Jean-François Zagury (Conservatoire National des Arts et Métiers, Paris).

Acknowledgments

We are grateful to all the patients and medical staff who have kindly collaborated with the GRIV project.

References

1. Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD. Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov **2008**; 7:221–30.
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. 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.
4. 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.
5. 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.
6. Vasilescu A, Terashima Y, Enomoto M, et al. A haplotype of the human CXCR1 gene protective against rapid disease progression in HIV-1+ patients. Proc Natl Acad Sci U S A **2007**; 104:3354–9.
7. 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.
8. Hercberg S, Galan P, Preziosi P, 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 Mineraux AntioXydants Study. Int J Vitam Nutr Res **1998**; 68:3–20.
9. 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.
10. 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.
11. Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics **2008**; 9:540.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B **1995**; 57:289–300.
13. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stati Med **1990**; 9:811–8.
14. Perneger TV. What's wrong with Bonferroni adjustments. BMJ **1998**; 316:1236–8.
15. Storey JD, Tibshirani R. Statistical significance for genomewide studies. PNAS **2003**; 100:9440–5.
16. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics **2000**; 155:945–59.
17. Devlin B, Roeder K. Genomic control for association studies. Biometrics **1999**; 55:997–1004.
18. Ge D, Zhang K, Need AC, et al. WGAViewer: software for genomic annotation of whole genome association studies. Genome Res **2008**; 18:640–3.
19. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. Nat Genet **2007**; 39:1202–7.
20. Ammosova T, Washington K, Debebe Z, Brady J, Nekhai S. Dephosphorylation of CDK9 by protein phosphatase 2A and protein phosphatase-1 in Tat-activated HIV-1 transcription. Retrovirology **2005**; 2:47.
21. Brass AL, Dykxhoorn DM, Benita Y, et al. Identification of host proteins required for HIV infection through a functional genomic screen. Science **2008**; 319:921–6.
22. Wilson SA, Brown EC, Kingsman AJ, Kingsman SM. TRIP: a novel double stranded RNA binding protein which interacts with the leucine rich repeat of flightless I. Nucleic Acids Res **1998**; 26:3460–7.
23. Wilson SA, Sieiro-Vazquez C, Edwards NJ, et al. Cloning and characterization of hIF2, a human homologue of bacterial translation initi-

ation factor 2, and its interaction with HIV-1 matrix. Biochem J **1999**; 342:97–103.
24. Exline CM, Feng Z, Stoltzfus CM. Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type 1 vif gene expression. J Virol **2008**; 82:3921–31.
25. 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.
26. Froguel P, Blakemore AI. The power of the extreme in elucidating obesity. N Engl J Med **2008**; 359:891–3.
27. Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A **2002**; 99:4465–70.
28. Furumatsu T, Tsuda M, Taniguchi N, Tajima Y, Asahara H. Smad3 induces chondrogenesis through the activation of SOX9 via CREB-binding protein/p300 recruitment. J Biol Chem **2005**; 280:8343–50.
29. Ikeda T, Kawaguchi H, Kamekura S, et al. Distinct roles of Sox5, Sox6, and Sox9 in different stages of chondrogenic differentiation. J Bone Miner Metab **2005**; 23:337–40.
30. Kiefer HL, Hanley TM, Marcello JE, Karthik AG, Viglianti GA. Retinoic acid inhibition of chromatin remodeling at the human immunodeficiency virus type 1 promoter: uncoupling of histone acetylation and chromatin remodeling. J Biol Chem **2004**; 279:43604–13.
31. Maeda Y, Yamaguchi T, Hijikata Y, et al. All-trans retinoic acid attacks reverse transcriptase resulting in inhibition of HIV-1 replication. Hematology **2007**; 12:263–6.
32. Austin J, Singhal N, Voigt R, et al. A community randomized controlled clinical trial of mixed carotenoids and micronutrient supplementation of patients with acquired immunodeficiency syndrome. Eur J Clin Nutr **2006**; 60:1266–76.
33. Mehta S, Fawzi W. Effects of vitamins, including vitamin A, on HIV/AIDS patients. Vitam Horm **2007**; 75:355–83.
34. Invernizzi CF, Xie B, Richard S, Wainberg MA. PRMT6 diminishes HIV-1 Rev binding to and export of viral RNA. Retrovirology **2006**; 3:93.
35. Xie B, Invernizzi CF, Richard S, Wainberg MA. Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat interactions with both cyclin T1 and the Tat transactivation region. J Virol **2007**; 81:4226–34.
36. Sgarra R, Lee J, Tessari MA, et al. The AT-hook of the chromatin architectural transcription factor high mobility group A1a is arginine-methylated by protein arginine methyltransferase 6. J Biol Chem **2006**; 281: 3764–72.
37. Li L, Yoder K, Hansen MS, Olvera J, Miller MD, Bushman FD. Retroviral cDNA integration: stimulation by HMG I family proteins. J Virol **2000**; 74:10965–74.
38. Charng MJ, Zhang D, Kinnunen P, Schneider MD. A novel protein distinguishes between quiescent and activated forms of the type I transforming growth factor beta receptor. J Biol Chem **1998**; 273:9365–8.
39. Manel N, Unutmaz D, Littman DR. The differentiation of human T_H-17 cells requires transforming growth factor-beta and induction of the nuclear receptor RORgammat. Nat Immunol **2008**; 9:641–9.
40. de St Groth BF, Landay AL. Regulatory T cells in HIV infection: pathogenic or protective participants in the immune response? AIDS **2008**; 22:671–83.
41. Benson MJ, Pino-Lagos K, Rosemblatt M, Noelle RJ. All-trans retinoic acid mediates enhanced T reg cell growth, differentiation, and gut homing in the face of high levels of co-stimulation. J Exp Med **2007**; 204:1765–74.
42. Mucida D, Park Y, Kim G, et al. Reciprocal TH17 and regulatory T cell differentiation mediated by retinoic acid. Science **2007**; 317:256–60.
43. Estes JD, Wietgrefe S, Schacker T, et al. Simian immunodeficiency virus–induced lymphatic tissue fibrosis is mediated by transforming growth factor beta 1-positive regulatory T cells and begins in early infection. J Infect Dis **2007**; 195:551–61.
44. Kekow J, Wachsman W, McCutchan JA, Cronin M, Carson DA, Lotz M. Transforming growth factor beta and noncytopathic mechanisms of immunodeficiency in human immunodeficiency virus infection. Proc Natl Acad Sci U S A **1990**; 87:8321–5.
45. Kared H, Lelievre JD, Donkova-Petrini V, et al. HIV-specific regulatory T cells are associated with higher CD4 cell counts in primary infection. AIDS **2008**; 22:2451–60.
46. Weiss L, Donkova-Petrini V, Caccavelli L, Balbo M, Carbonneil C, Levy Y. Human immunodeficiency virus-driven expansion of CD4+CD25+ regulatory T cells, which suppress HIV-specific CD4 T-cell responses in HIV-infected patients. Blood **2004**; 104:3249–56.
47. Brenchley JM, Paiardini M, Knox KS, et al. Differential Th17 CD4 T-cell depletion in pathogenic and nonpathogenic lentiviral infections. Blood **2008**; 112:2826–35.
48. Favre D, Lederer S, Kanwar B, et al. Critical loss of the balance between Th17 and T regulatory cell populations in pathogenic SIV infection. PLoS Pathog **2009**; 5:e1000295.
49. Li JM, Shen X, Hu PP, Wang XF. Transforming growth factor beta stimulates the human immunodeficiency virus 1 enhancer and requires NF-kappaB activity. Mol Cell Biol **1998**; 18:110–21.
50. Zocchi MR, Contini P, Alfano M, Poggi A. Pertussis toxin (PTX) B subunit and the nontoxic PTX mutant PT9K/129G inhibit Tat-induced TGF-beta production by NK cells and TGF-beta–mediated NK cell apoptosis. J Immunol **2005**; 174:6054–61.

Genomewide Association Study on a Rapid Progression Cohort Identifies New Susceptibility Alleles for AIDS

Sigrid Le Clerc[1,2,3*], Sophie Limou[1,2,3,4*], Cédric Coulonges[1,3], Wassila Carpentier[3], Christian Dina[5], Lieng Taing[1], Olivier Delaneau[1], Taoufik Labib[1,2], Rob Sladek[6], ANRS Genomic Group[3], Christiane Deveau[3], Hélène Guillemain[1], Rojo Ratsimandresy[1], Matthieu Montes[1], Jean-Louis Spadoni[1], Amu Therwath[7], François Schächter[1], Fumihiko Matsuda[8], Ivo Gut[4], Jean-Daniel Lelièvre[2], Yves Lévy[2], Philippe Froguel[5,9], Jean-François Delfraissy[3], Serge Hercberg[10], Jean-François Zagury[1,2,3#]

SUPPLEMENTARY ONLINE CONTENT

Figure 1: To check the relevance of the rapid progression associations in an independent fashion, we compared the genotypes of 1,000 subgroups of size 85 randomly extracted from the D.E.S.I.R. control cohort (697 individuals), with the genotypes of the entire SU.VI.MAX control group. For each test, we counted the number N of independent (r^2<0.5) SNPs respecting the 25% FDR threshold: N varied between 0 and 8. The histogram presents the number of tests for each value of N. Overall, the odds of getting 6 or more independent SNPs is less than 5 out of 1000.

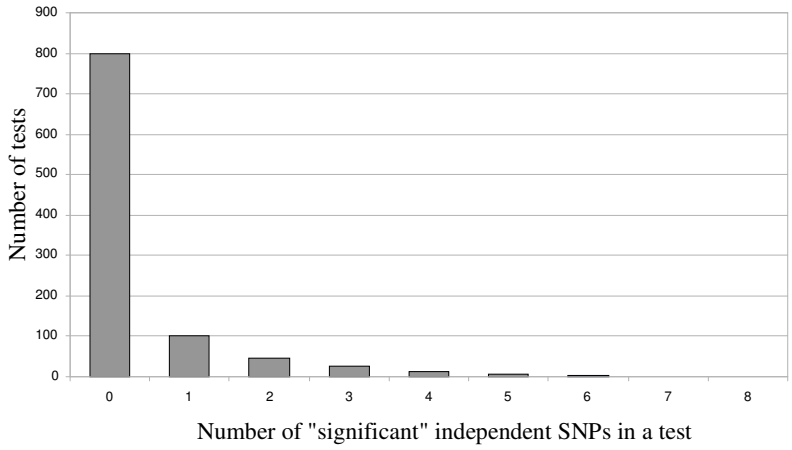


Figure 2: Quantile-quantile plot for expected (red) *vs* observed (black) *p-values* from the comparison of rapid progressors with control subjects. X axis: -$\log_{10}$(observed *p-values*); Y axis: -$\log_{10}$(expected *p-values* under the null hypothesis).

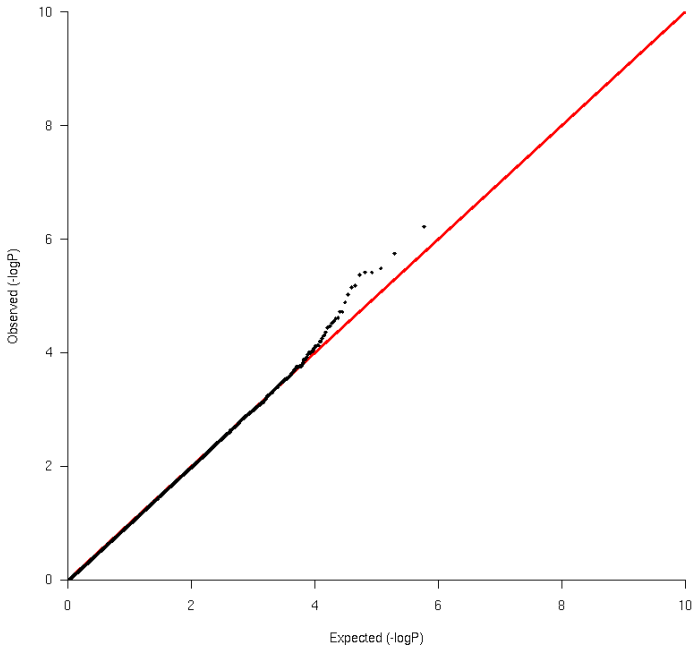


Table 2: Fifty best *p-values* obtained for the comparison between rapid progressors and control subjects.

				Allelic Frequency (A1), %					Fisher	FDR
SNP	Chr	A1	A2	RP	$CTR_{SU.VI.MAX}$	$CTR_{D.E.S.I.R.}$	SCP	OR (95% CI)	$P_{RP\text{-}CTRSuvimax}$	qvalue
rs4118325	1	A	G	5.2	19.0	18.7	15.8	0.24 (0.12-0.46)	6.09E-07	0.17
rs1522232	12	T	C	29.1	47.7	48.2	50.7	0.45 (0.32-0.63)	1.80E-06	0.20
rs1360517	9	A	G	16.3	5.9	7.0	5.8	3.09 (2.00-4.78)	3.27E-06	0.20
rs3108919	8	C	T	43.6	26.6	27.8	27.5	2.13 (1.56-2.91)	3.86E-06	0.20
rs10800098	1	A	G	14.5	4.9	4.7	5.6	3.29 (2.08-5.20)	3.86E-06	0.20
rs10494056	1	A	C	5.8	18.6	17.9	16.1	0.27 (0.14-0.52)	4.29E-06	0.20
rs12351740	9	T	C	12.8	4.1	4.7	3.8	3.46 (2.12-5.62)	6.63E-06	0.25
rs1020064	2	T	G	9.3	23.2	25.6	25.2	0.34 (0.20-0.57)	7.04E-06	0.25
rs1556032	9	C	T	66.3	48.9	50.1	48.7	2.05 (1.48-2.84)	9.47E-06	0.31
rs4489981	15	T	C	65.2	47.8	49.1	43.8	2.04 (1.47-2.81)	1.30E-05	0.38
rs12642133	4	A	G	19.2	34.7	33.0	29.6	0.45 (0.30-0.66)	1.91E-05	0.46
rs11097821	4	A	C	19.2	34.8	33.0	29.6	0.44 (0.30-0.66)	1.92E-05	0.46
rs11629182	14	A	G	13.4	27.4	26.0	28.7	0.41 (0.26-0.64)	2.44E-05	0.52
rs12549196	8	G	A	22.7	11.1	11.3	9.8	2.36 (1.62-3.44)	2.53E-05	0.52
rs6998986	8	A	G	56.5	39.8	38.4	40.7	1.95 (1.43-2.67)	2.78E-05	0.54
rs7810799	7	A	G	7.0	1.5	2.7	2.9	5.11 (2.62-9.95)	3.00E-05	0.54
rs1762495	1	T	C	7.0	18.5	19.6	15.8	0.33 (0.18-0.60)	3.40E-05	0.55
rs7985331	13	G	A	22.4	10.9	11.7	13.5	2.34 (1.60-3.43)	3.54E-05	0.55
rs1408700	6	A	G	13.4	5.0	5.7	4.2	2.96 (1.85-4.75)	3.61E-05	0.55
rs10277069	7	A	G	7.0	1.5	2.7	2.9	4.87 (2.51-9.45)	4.41E-05	0.64
rs3105453	8	C	T	59.9	43.7	46.6	43.6	1.92 (1.40-2.63)	4.89E-05	0.67
rs1986364	9	G	A	9.9	22.2	23.7	22.1	0.38 (0.23-0.64)	5.11E-05	0.67

rs2138010	12	A	C	10.5	3.4	2.7	3.6	3.36 (1.97-5.71)	5.51E-05	0.69
rs2238536	18	A	G	14.0	5.4	5.7	5.6	2.86 (1.80-4.54)	6.26E-05	0.74
rs7608404	2	A	G	39.5	25.0	23.0	25.2	1.95 (1.42-2.68)	6.41E-05	0.74
rs1751274	10	T	C	1.2	8.6	8.3	7.7	0.12 (0.03-0.51)	7.39E-05	0.76
rs11744511	5	C	A	19.2	33.3	32.5	32.2	0.47 (0.32-0.70)	7.42E-05	0.76
rs6027527	20	T	C	6.5	17.1	17.7	19.5	0.33 (0.18-0.62)	7.57E-05	0.76
rs10495913	2	G	A	23.0	37.8	35.8	36.0	0.49 (0.34-0.71)	7.59E-05	0.76
rs10044036	5	T	C	5.9	16.2	17.4	14.6	0.32 (0.17-0.61)	8.14E-05	0.79
rs26602	5	A	G	10.5	22.7	21.7	24.0	0.40 (0.24-0.65)	8.50E-05	0.79
rs1449004	3	G	A	45.4	30.6	31.8	33.9	1.88 (1.38-2.57)	8.98E-05	0.79
rs1882070	7	G	A	23.9	12.5	12.4	13.1	2.20 (1.52-3.18)	9.57E-05	0.79
rs4983306	14	C	T	15.2	28.4	28.9	24.6	0.45 (0.29-0.69)	9.66E-05	0.79
rs4723646	7	C	T	23.9	12.5	12.3	12.7	2.18 (1.51-3.16)	9.96E-05	0.79
rs9788561	14	T	C	53.0	37.8	37.3	38.2	1.85 (1.36-2.52)	1.01E-04	0.79
rs10267245	7	C	T	6.5	1.4	N/A	2.9	4.79 (2.40-9.55)	1.01E-04	0.79
rs1412304	9	C	A	17.5	8.0	8.8	7.8	2.43 (1.60-3.69)	1.04E-04	0.79
rs949650	12	A	G	26.2	41.0	42.0	41.1	0.51 (0.36-0.72)	1.07E-04	0.80
rs1788823	18	T	C	47.7	33.0	33.1	36.7	1.85 (1.36-2.52)	1.22E-04	0.85
rs1819548	1	G	A	7.6	18.5	19.7	14.9	0.36 (0.20-0.64)	1.23E-04	0.85
rs12743478	1	T	C	5.3	15.1	15.1	14.0	0.31 (0.16-0.61)	1.23E-04	0.85
rs7595132	2	A	G	3.5	12.5	10.9	10.7	0.25 (0.11-0.58)	1.33E-04	0.87
rs9435337	1	C	T	8.8	20.0	18.7	15.6	0.38 (0.22-0.65)	1.33E-04	0.87
rs333985	2	A	C	16.9	7.7	9.2	9.3	2.45 (1.60-3.74)	1.35E-04	0.87
rs6996185	8	C	T	36.7	23.3	23.1	20.9	1.90 (1.38-2.62)	1.49E-04	0.88
rs11135856	8	A	G	22.1	11.6	12.7	10.4	2.17 (1.48-3.16)	1.60E-04	0.88
rs662627	5	T	C	37.3	23.9	25.7	24.0	1.89 (1.37-2.60)	1.67E-04	0.88

rs2682826	12	T	C	41.9	28.0	28.6	27.5	1.85 (1.35-2.54)	1.73E-04	0.88
rs7001509	8	G	A	13.4	5.5	6.9	5.3	2.65 (1.66-4.23)	1.73E-04	0.88

NOTE. The *p-values* were computed with Fisher's exact tests in the allelic frequency mode and presented with their corresponding allelic frequency in the different populations (rapid progressors [RP], control subjects [$CTR_{SUVIMAX}$], second control group [CTR_{DESIR}] and seropositive control population [SCP]), chromosome (Chr), Odds-Ratios (OR), 95% confidence interval (95% CI) and FDR *q-value*.

3. Criblage des SNPs de faible fréquence issus d'études d'association 'génome entier' réalisées chez les patients infectés par le VIH-1

ARTTICLE 3

Screening Low Frequency SNPs from Genome Wide Association Study Reveals a New Risk Allele for Progression to AIDS

Sigrid Le Clerc, Cédric Coulonges, Olivier Delaneau, Danielle Van Manen, Joshua T. Herbeck, Sophie Limou, Ping An, Jeremy J. Martinson, Jean-Louis Spadoni, Amu Therwath, Jan H. Veldink, Leonard H. van den Berg, Lieng Taing, Taoufik Labib, Safa Mellak, Matthieu Montes, Jean-François Delfraissy, François Schächter, Cheryl Winkler, Philippe Froguel, James I. Mullins, Hanneke Schuitemaker, Jean-François Zagury

JAIDS; (sous presse ~ 2010)

RESUME

Pas moins de 7 études 'génome entier' ont été publiées concernant le SIDA et seulement des associations dans le locus du *HLA* sur le chromosome 6 et *CXCR6* avaient passé le seuil de significativité 'génome entier' de Bonferroni.

Nous avons exploité les données issues de trois études 'génome entier' réalisées précédemment, ciblant particulièrement les SNPs de fréquence faible (fréquence de l'allèle mineur (MAF) <5%). Deux groupes composés de 365 non-progresseurs à long terme et de 47 progresseurs rapides ont été comparés à un groupe de 1394 contrôles séronégatifs pour le VIH.

Sur les 8584 SNPs avec des MAF< 5% dans les cas et dans les contrôles (seuil de Bonferroni= $5,89x10^{-6}$), quatre SNPs sont associés au phénotype de non-progression. Le meilleur résultat est pour le SNP rs2395029 localisé dans le gène HCP5 (p=$8,54x10^{-15}$, OR : 3,41) dans le chromosome 6. Deux autres SNPs en déséquilibre de liaison partiel avec le SNP rs2395029 sont significatifs : *C6orf48* (p=$3,03x10^{-10}$, OR=2,9) et *NOTCH4* (p=$9,08x10^{-07}$, OR : 2,32). La quatrième association correspond au rs2072255, localisé dans le gène *RICH2* (p=$3,30x10^{-06}$, OR : 0,43) sur le chromosome 17. Lorsque l'on utilise le SNP de *HCP5* (rs2395029) comme covariable, les signaux de *C6orf48* et *NOTCH4* disparaissent, mais le signal localisé dans *RICH2* reste significatif.

L'analyse des fréquences faibles a mis un avant une nouvelle association dans le gène *RICH2*. De manière intéressante RICH2 interagit avec BST-2 connue pour être un facteur majeur de restriction de l'infection par le VIH-1. Notre étude a ainsi identifié un nouveau gène candidat pour l'étiologie moléculaire du SIDA.

SCREENING LOW FREQUENCY SNPS FROM GENOME WIDE ASSOCIATION STUDY REVEALS A NEW RISK ALLELE FOR PROGRESSION TO AIDS

Sigrid Le Clerc[1,2,3], Cédric Coulonges[1,3], Olivier Delaneau[1], Danielle Van Manen[5], Joshua T. Herbeck[6], Sophie Limou[1,2,3,4], Ping An[7], Jeremy J. Martinson[8], Jean-Louis Spadoni[1], Amu Therwath[9], Jan H. Veldink[10], Leonard H. van den Berg[10], Lieng Taing[1], Taoufik Labib[1], Safa Mellak[1], Matthieu Montes[1], Jean-François Delfraissy[3], François Schächter[1], Cheryl Winkler[7], Philippe Froguel[11,12], James I. Mullins[6], Hanneke Schuitemaker[5], Jean-François Zagury[1,2,3#]

1: Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, Paris, France.

2: Université Paris 12, INSERM U955, Créteil, France.

3: ANRS Genomic Group (French Agency for Research on AIDS and Hepatitis), Paris, France.

4: CEA/Institut de Génomique, Centre National de Génotypage, Evry, France.

5: Department of Experimental Immunology, Sanquin Research, Landsteiner Laboratory, Center for Infectious Diseases and Immunity Amsterdam (CINIMA) Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands.

6: University of Washington School of Medicine, Department of Microbiology, Seattle, WA, USA.

7: Laboratory of Genomic Diversity, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, MD, USA

8: Department of Human Genetics University of Pittsburgh, Pittsburgh, PA, USA

9: Laboratoire d'Oncologie Moléculaire, Université Paris 7, Paris, France.

10: Rudolf Magnus Institute of Neuroscience, Department of Neurology, University Medical Center Utrecht, 3584 CX, Utrecht, The Netherlands

11: UMR CNRS 8090, Institut Pasteur de Lille, Lille, France.

12: Genomic Medicine, Hammersmith Hospital, Imperial College London, London, UK.

#: Corresponding Author: Jean-François Zagury, 292 rue Saint Martin, 75003 Paris, France, tel: 33 1 58 80 88 20, fax: 33 1 58 80 87 86, e-mail: zagury@cnam.fr

Sources of support: This work was supported by Agence Nationale de Recherche sur le SIDA (ANRS), Sidaction, Fondation de France, Innovation 2007 program of Conservatoire National des Arts et Métiers (CNAM), AIDS Cancer Vaccine Development Foundation, Neovacs SA, Vaxconsulting. Sophie Limou benefits from a fellowship from the French Ministry of Education, Technology and Research and Sigrid Le Clerc benefits from a fellowship of ANRS. 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 financially supported by the Netherlands National Institute for Public Health and the Environment. The authors acknowledge funding from the Netherlands Organization for Scientific Research (TOP, registration number 9120.6046). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute. UO1-AI-35042, UL1-RR025005 (GCRC), UO1-AI-35043, UO1-AI-35039, UO1-AI-35040, UO1-AI-35041. This project has been funded in part with federal funds from the National Institutes of Health, under contract HHSN261200800001E. 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 U.S. Government. This Research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Running head: Low frequency SNP associations in AIDS GWAS

Abstract

Background: Seven genome-wide association studies (GWAS) have been published in AIDS and only associations in the HLA region on chromosome 6 and *CXCR6* have passed genome-wide significance.

Methods: We reanalyzed the data from three previously published GWAS, targeting specifically low frequency SNPs (minor allele frequency (MAF)<5%). Two groups composed of 365 slow progressors (SP) and 147 rapid progressors (RP) from Europe and the US were compared with a control group of 1394 seronegative individuals using Eigenstrat corrections.

Results: Of the 8584 SNPs with MAF<5% in cases and controls (Bonferroni threshold=5.8×10^{-6}), four SNPs showed statistical evidence of association with the SP phenotype. The best result was for *HCP5* rs2395029 ($p=8.54 \times 10^{-15}$, OR=3.41) in the HLA locus, in partial linkage disequilibrium with two additional chromosome 6 associations in *C6orf48* ($p=3.03 \times 10^{-10}$, OR=2.9) and *NOTCH4* (9.08×10^{-07}, OR=2.32). The fourth association corresponded to rs2072255 located in *RICH2* ($p=3.30 \times 10^{-06}$, OR=0.43) in chromosome 17. Using *HCP5* rs2395029 as a covariate, the *C6orf48* and *NOTCH4* signals disappeared, but the *RICH2* signal still remained significant.

Conclusion: Besides the already known chromosome 6 associations, the analysis of low frequency SNPs brought up a new association in the *RICH2* gene. Interestingly, RICH2 interacts with BST-2 known to be a major restriction factor for HIV-1 infection. Our study has thus identified a new candidate gene for AIDS molecular etiology and confirms the interest of singling out low frequency SNPs in order to exploit GWAS data.

Keywords: AIDS; HIV-1; genome-wide association study; SNP; disease progression; *RICH2*.

Introduction

In recent years, the development of new DNA technologies has allowed the successful completion of genome-wide association studies (GWAS) and multiple genetic associations were identified in several diseases such as Celiac disease[1], Schizophrenia[2], or Type 2 diabetes[3]. In AIDS, several GWAS have been published since 2007[4-10] and associations passing genome-wide significance have been found solely in chromosome 6 in the region of HLA (*HCP5*[4-6], *HLA-C*[4, 5, 11]) and in the *CXCR6* gene[12]. The design of genotyping chips tends to rely mainly on common variants with minor allele frequencies (MAF) > 5%. Moreover, the power to detect low allele frequency SNP associations in AIDS is weakened for two complementary reasons : 1. lower frequency means less individuals at stake and thus weaker p values, and 2. AIDS-related genomic cohorts have enrolled fewer patients compared to other pathologies such as chronic kidney disease[13] or Type 2 diabetes[3]. However, low frequency SNPs are predicted to have the potential for greater functional consequences than common alleles and may contribute strongly to genetic susceptibility to common diseases[14-16]; thus, they constitute very good candidates for genetic association studies. We therefore decided to re-analyze the genome-wide data obtained from the GRIV cohort on slow progression [6] and on rapid progression[7], by focusing specifically on low frequency SNPs (MAF < 5 %). In order to increase the power of the study, we also included in our analysis rapid and slow progressors from the Dutch ACS cohort[11] and from the American MACS156 group[8].

Methods

The case groups.

Slow Progressors and Rapid Progressors were gathered from 3 HIV cohorts:

The GRIV cohort.

The GRIV (Genomics of Resistance to Immunodeficiency Virus) cohort, established in 1995 in France, is a collection of DNA samples to identify host genes associated with slow-progression and with rapid progression to AIDS[17-20]. Only Caucasians of European descent living in France were eligible for enrolment to reduce confounding by population substructure. These criteria limit the influence of the ethnic and environmental factors (all subjects live in a similar environment and are infected by HIV-1 subtype B strains) and put an emphasis on the genetic make-up of each individual in determination of rapid (RP) or slow progression (SP) to AIDS. The RP and the SP were included on the basis of the main clinical outcomes, CD4 T-cell count and time to disease progression. SP were defined as asymptomatic HIV-1 infected individuals for more than 8 years, no treatment and a CD4 T-cell count above 500/mm^3. The SP group (n=270) was composed of 200 males and 70 females aged at inclusion from 19 to 62 (mean=35). Rapid progressors (RP) were stringently defined as having a two or more CD4 T-cell counts below 300/mm^3 less than 3 years after the last seronegative testing. The RP group (n=84) was composed of 72 males and 12 females aged at inclusion from 21 to 55 (median=32). DNA was obtained from fresh peripheral blood mononuclear cells or from EBV-transformed cell lines. All patients provided written informed consent before enrolment in the GRIV genetic association study.

The ACS cohort.

The ACS (Amsterdam Cohort Studies) cohort is composed of 316 HIV-1 homosexual men and 100 HIV-1 drug users. This cohort was established to follow the course of HIV-1 infection using various endpoints related to HIV-1 infection and AIDS. The ACS participants were described in detail previously[11].

Since CD4 T-cell count was assessed during routine clinical follow-up, we could extract the ACS SP and RP patients respecting the GRIV criteria (SP: n=36, RP: n=41). The SP and RP status was easily determined among seroconverter subjects since the date of seroconversion was known. We could also extract SP from seroprevalent subjects when the time of seropositivity was known to be higher than 8 years.

The MACS156 group.

The MACS156 study comprises a subset of 156 HIV-1 homosexual men enrolled in the MACS (MultiCenter AIDS Cohort Study) cohort, a prospective cohort originally established to investigate the natural history of HIV infection[21]. This subset of MACS European American participants was chosen to be enriched with the extremes AIDS progression phenotypes [8]. The MACS156 participants were described in detail previously[8].

Since CD4 T-cell count was assessed during routine clinical follow-up, we could extract the MACS SP and PR respecting the GRIV definition (SP: n=59, RP: n=22). As with the ACS cohort, the SP and the RP were selected from seroconverter subjects. SP were also selected from seroprevalent subjects.

Control groups.

Three Caucasian control groups from France, The Netherlands, and the USA were merged and used as a control group.

The D.E.S.I.R. control group

The Data from an Epidemiological Study on Insulin Resistance Syndrome (D.E.S.I.R) program was a 9-year follow-up study designed to clarify the development of the insulin resistance syndrome. Subjects were recruited from 1994 to 1996 from volunteers insured by

the French social security system, which offers periodic health examinations free of charge[22]. This control group comprised 694 participants both non obese and normoglycemic of the D.E.S.I.R. trial, all French and HIV-1 seronegative. It was composed of 281 males and 413 females aged from 30 to 64 years.

Dutch controls (CTR-ACS).

This control group corresponds to 376 Dutch subjects genotyped with HumanHap300 BeadChips[23].

Illumina control group

This control group corresponds to Caucasian subjects genotyped with HumanHap300 BeadChips from the Illumina Genotyping Control Database (www.illumina.com). There were 324 individuals.

Genotyping method, quality control. The GRIV cohort, the ACS cohort, and the control groups were genotyped using the Illumina Infinium II HumanHap300 BeadChips (Illumina, San Diego, CA, USA). The genotyping quality was assessed for each group using the BeadStudio software (version 3.1, Illumina). Missing data >2%, minor allele frequency (<1%) and deviations from Hardy-Weinberg equilibrium in the control groups ($p<1.0x10^{-3}$) were removed during these quality control steps[6, 7, 11]. The MACS156 group genotype data were obtained through the Affymetrix GeneChip Human Mapping 500K Array (Affymetrix, Santa Clara, CA, USA). Different quality control filters were applied to ensure reliable genotyping data[8]. For all the cohorts, we removed outliers exhibiting non-Caucasian ancestry, cohort by cohort, using the Eigenstrat method[24].

SNP selection. We considered two pooled case groups (SP from GRIV, ACS, and MACS156 group on the one hand, RP from GRIV, ACS, and MACS156 on the other hand) and the pooled control groups (D.E.S.I.R., CTR-ACS, Illumina-CTR). We retained the 8584 SNPs exhibiting a MAF < 5% for the SP-CTR comparison (Bonferroni $5.8x10^{-6}$), and 10295 SNPs exhibiting a MAF < 5% for the RP-CTR comparison (Bonferroni $4.8x10^{-6}$). It was important to choose SNPs with low frequency in either groups since we were looking for factors either promoting progression (low MAF in SP compared to CTR or low MAF in CTR compared to RP) or preventing progression (low MAF in RP compared to CTR or low MAF in CTR compared to SP). The choice to screen specifically low frequency SNPs stems from two complementary reasons: **1.** a biological reason: HIV-1 infection is a multi-factorial disease with several genetic factors impacting disease. Several groups have pointed out that most signals involved in complex diseases should deal with the low frequency variants spectrum [14-16]. Indeed, this observation was confirmed in AIDS since the main signal found up to now was in *HCP5* with a low frequency variant [4-6, 9]. **2.** a statistical fact : in our case-control configuration, a low frequency either in the CTR or in the CASE group means systematically a weaker p value for a given Odds Ratio (OR which measures the real biological impact of the SNP). For instance, for an OR of 0.5 in the dominant mode, the p values obtained are: 0.02 with a SP MAF of 2% and a CTR MAF=3.9%, $1.5x10^{-5}$ with a SP MAF of 5% and a CTR MAF=9.5%, $9.98x10^{-8}$ with a SP MAF of 10% and CTR MAF=18.2%. Moreover, in the Illumina genotyping chips used, SNPs with low MAF are under-represented in genotyping chips compared to SNPs with higher MAF (data not shown). Overall for a biological effect measured by a given OR, SNP associations are thus more difficult to identify in the low frequency spectrum since they exhibit weaker p values by essence and since they are under-represented and thus artificially penalized through global Bonferroni corrections.

Statistical analysis. We performed a case-control analysis by comparing either the SP group (n=365) or the RP group (n=147) consisting in GRIV, ACS, and MACS patients with the control group consisting of D.E.S.I.R, CTR-ACS, and Illumina control individuals (n=1394). The statistical analysis was performed by a logistic regression (with SNPtest software[25]) in the dominant mode, taking into account stratification by adding the 2 first Eigenstrat PC axes as covariates using EIGENSOFT. Testing for association under the dominant model was appropriate since we lacked power to test for associations under the recessive model and additionnaly, in this context, the dominant model is identical to the additive mode. For each SNP passing the Bonferroni threshold, we recomputed the regression by adding the *HCP5* SNP (rs2395029) as a covariate to check for non-independence due to linkage disequilibrium (with SNPtest software[25]).

SNP imputation. Using SNPtest Impute software [25] it was possible to impute untyped SNPs of the MACS156 study subjects, absent from the the Affymetrix GeneChip Human Mapping 500K Array (Affymetrix, Santa Clara, CA, USA) and present in the Illumina HumanHap300 BeadChips (Illumina, San Diego, CA, USA). They were imputed using the HapMap release 21 phased data for the European population (CEU) as panel of reference (http://www.hapmap.org). Only the SNPs imputed with high reliability (imputation quality score[25] P>0.9) were retained.

Internal replication. After comparing the total SP group (combined from GRIV, ACS, and MACS 156) with the total control group (combined from D.E.S.I.R, CTR-ACS, and CTR_Illumina), we also performed an individual analysis of each group GRIV, ACS, and MACS156 for the four SNPs passing the Bonferroni threshold. For GRIV, we checked the result obtained in our previous GWAS[6]. For ACS, we tested the association between the four

SNPs and times to time to AIDS93 by linear regression. For MACS156 group, we also tested the association between the SNPs and time to AIDS93 by linear regression.

Linkage disequilibrium (LD). For each SNP exhibiting a significant association, we looked for the other SNPs in linkage disequilibrium ($r^2 \geq 0.9$) in the HapMap population of Western European ancestry (CEU, HapMap data Release 21a/phase II January 2007, on NCBI B35 assembly, dbSNP125, http://www.hapmap.org) in order to identify the genes possibly tracked by the SNP associations. A SNP was assigned to a gene if it was located within the gene or in the 2kb flanking regions (potential regulatory sequence), otherwise it was considered intergenic.

Bioinformatics exploration. To further explore the associations observed, we tried to identify putative modifications in mRNA expression as shown in *Genevar*[26] and Dixon[27] databases, in splicing (FastSNP[28], http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp/), in polyadenylation (polyAH, http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter and polyApred, http://www.imtech.res.in/raghava/polyapred/submission.html), or in transcription factor binding sites (SignalScan, http://www-bimas.cit.nih.gov/molbio/signal/, TESS, http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME, and TFSearch, http://www.cbrc.jp/research/db/TFSEARCH.html, derived from TRANSFAC database). We used the Genecards database to look for the tissues and organs expressing the proteins (GeneCards[29], http://www.genecards.org).

Results

For the 8584 SNPs with a MAF < 5 %, we tested a total of 365 SP patients and compared them with a control cohort of 1394 seronegative individuals (from France, the Netherlands, and the US, see methods). Four signals passed the Bonferroni threshold in the dominant mode, three in chromosome 6 and one in chromosome 17 (Table 1). Not unexpectently, the best result was obtained for the well-replicated *HCP5* rs2395029 ($p=8,54x10^{-15}$). The two other associations found in chromosome 6 were for rs9368699 in *C6orf48* ($p= 3,03x10^{-10}$) and rs8192591 in *NOTCH4* ($p=9,08x10^{-07}$). These genes are in partial linkage disequilibrium with HCP5-rs2395029 (resp. r^2=0.68, r^2=0.57). The fourth association corresponded to the chromosome 17 SNP rs2072255 located in the *RICH2* gene ($p=3.30x10^{-6}$). This SNP is in full LD (r^2=1) with rs2072254 corresponding to a synonymous change (Gly188Gly) of *RICH2* (Figure 1 and see map, Supplemental Digital Content 1). The fours SNPs corresponded to the higher end of the MAF distribution of the SNP studied (see histogram, Supplemental Digital Content 2), which is logical since larger numbers of subjects lead to stronger p values.

In order to evaluate the role of LD in these associations, we recomputed the p-values using the *HCP5* SNP (rs2395029) as covariate. The two chromosome 6 SNPs were not significant in the adjusted analysis (p=0.78 for rs9368699 in *C6orf48*, p=0.31 for rs8192591 in *NOTCH4)*, but the association remained statistically robust for the chromosome 17 SNP after controlling for the *HCP5* SNP, $p=1.82x10^{-6}$ for rs2072255 in *RICH2*. In line with this computation, we found that the rs2072255 frequency was not significantly different between the SP elite controllers[12] and among the SP non elite controllers [12] (p=0.7).

The subjects carrying the rs2072255-A allele were 9.04% in the SP group (Table 1), 18.87% in the control group, and 17.1% in the RP group (this excludes the hypothesis of an effect on infection). Interestingly these frequencies were consistent within each of the three SP groups as well as within each of the three control groups (Figure 2). Moreover, the positive signal for association of rs2072255 was confirmed in the individual cohorts GRIV, ACS, and

MACS156 study: the comparison of the NP with D.E.S.I.R controls in GRIV led to p= $8.1x10^{-5}$, the analysis of ACS by linear regression led to p=0.05, and the analysis of the MACS156 group by linear regression led also to p=0.05. A table summarizing the results in the different cohorts is provided in supplemental digital content 3. Finally, the rs2072254 *RICH2* exonic SNP (in LD with the rs2072255) is located in a splicing site according to FastSNP[28] (Figure 1).

When comparing the 10295 selected SNPs between the 147 RP patients with the 1394 seronegative control group, no signal passed the Bonferroni threshold.

Discussion

We decided to reanalyze previous GWAS data on AIDS cohorts by focusing specifically on low frequency SNPs (MAF<5%). For that, we combined rapid and slow progressors from three international cohorts from France (GRIV), Netherlands (ACS), and US (MACS156 study) totalling 365 SP and 147 RP, who were compared with 1394 controls (seronegative individuals). No association was found when comparing the RP group with the CTR group. This was not a surprise since the RP group comprised only 147 individuals and for a MAF of 5% in the CTR, one needed to get a quite strong biological effect (OR>2.8) to pass the Bonferroni threshold. Four SNPs passed the Bonferroni threshold when comparing the SP group with the CTR group. Among them, three are in chromosome 6 and were previously found significant by several studies : rs2395029 in *HCP5*[4-6, 8, 9, 11], rs9368699 in *C6orf48*[5, 6], rs8192591 in *NOTCH4*[5]. *NOTCH4* is an interesting candidate gene due to its role in immune regulation and the *NOTCH4* rs8192591 corresponds to a non-synonymous Gly534Ser protein variant. This association was found independent from the *HCP5* rs2395029 by Fellay et al.[5], however the signal disappeared in our own study when using *HCP5* rs2395029 as covariate.

A possible explanation for this discrepancy could be the use of viral load as an endpoint in the study by Fellay et al.[5] while we used here a progression phenotype.

The fourth signal identified in the present study is new and corresponds to rs2072255 in the chromosome 17 *RICH2* gene. The *RICH2* gene encodes a Rho-type GTPase activator composed of 818 amino acid (89,247 kDa) with an intracellular localisation. It is expressed highly in the brain, and at a basal level in several tissues notably in the lymph nodes[29]. A recent study has shown that RICH2 is a part of the physical link between BST-2 and the actin cytoskeleton and prevents the internalisation of BST-2[30]. RICH2 could thus contribute to the externalisation of BST-2 which prevents HIV-1 virion budding and release[31]. This is a counteraction of HIV-1 Vpu known to favor internalization and degradation of BST-2[31]. The rs2072255-A *RICH2* allele favors progression to AIDS since 18.87% of the CTR carry the variant while only 9.04% of the SP carry it (Table 1, Figure 2). Interestingly, rs2072255 is in total LD with the SNP rs2072254 located in a splicing site of *RICH2* as suggested by FastSNP (Figure 2). If the rs2072254-G minor allele alters mRNA splicing, it could lead to a down-modulation of *RICH2* and thus explain a diminished effect of BST-2 against HIV-1 production.

The identification of three chromosome 6 signals already confirmed by several studies shows the relevance of targeting specifically low frequency SNPs in GWAS. The genetic and biological data regarding the *RICH2* signal are also quite compelling and provide a new relevant candidate gene to explore the molecular etiology of HIV-1 pathogenesis. Further genetic and experimental studies will be needed to confirm and understand the effect of *RICH2* in AIDS pathogenesis.

Acknowledgements

The authors are grateful to all the patients and medical staff who have kindly collaborated with the GRIV project. Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS)

with centers (Principal Investigators) at The Johns Hopkins Bloomberg School of Public Health (Joseph B. Margolick, Lisa P. Jacobson), Howard Brown Health Center, Feinberg School of Medicine, Northwestern University, and Cook County Bureau of Health Services (John P. Phair, Steven M. Wolinsky), University of California, Los Angeles (Roger Detels), and University of Pittsburgh (Charles R. Rinaldo). Website located at http://www.statepi.jhsph.edu/macs/macs.html. 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 U.S. Government.

References

1. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet.* Apr 2010;42(4):295-302.

2. O'Donovan MC, Craddock N, Norton N, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet.* Sep 2008;40(9):1053-1055.

3. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet.* May 2008;40(5):638-645.

4. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. *Science.* Aug 17 2007;317(5840):944-947.

5. Fellay J, Ge D, Shianna KV, et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* Dec 2009;5(12):e1000791.

6. 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.* Feb 1 2009;199(3):419-426.

7. 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.* Oct 15 2009;200(8):1194-1201.

8. Herbeck JT, Gottlieb GS, Winkler CA, 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.* Feb 15 2010;201(4):618-626.

9. 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(12):e3907.

10. Pelak K, Goldstein DB, Walley NM, et al. Host determinants of HIV-1 control in African Americans. *J Infect Dis.* Apr 15 2010;201(8):1141-1149.

11. 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.* Jan 2 2009;23(1):19-28.

12. Limou S, Coulonges C, Herbeck JT, et al. Multi-Cohort Genetic Association Study Reveals CXCR6 as a New Chemokine Receptor Involved in AIDS Long-Term Non-Progression. *The Journal of infectious diseases.* 2010;in press.

13. Kottgen A, Pattaro C, Boger CA, et al. New loci associated with kidney function and chronic kidney disease. *Nat Genet.* May 2010;42(5):376-384.

14. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet.* Apr 2007;80(4):727-739.

15. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet.* Jan 2008;82(1):100-112.

16. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* Jun 2008;40(6):695-701.

17. Flores-Villanueva PO, Hendel H, Caillat-Zucman S, et al. Associations of MHC ancestral haplotypes with resistance/susceptibility to AIDS disease development. *J Immunol.* Feb 15 2003;170(4):1925-1929.

18. 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.* Jun 1 1999;162(11):6942-6946.

19. 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.* Mar 29 1997;349(9056):922-923.

20. Winkler CA, Hendel H, Carrington M, et al. Dominant effects of CCR2-CCR5 haplotypes in HIV-1 disease progression. *J Acquir Immune Defic Syndr.* Dec 1 2004;37(4):1534-1538.

21. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR, Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol.* Aug 1987;126(2):310-318.

22. 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.* Aug 1996;44(4):373-375.

23. van Es MA, Veldink JH, Saris CG, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat Genet.* Oct 2009;41(10):1083-1087.

24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* Aug 2006;38(8):904-909.

25. 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.* Jul 2007;39(7):906-913.

26. Ge D, Zhang K, Need AC, et al. WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.* Apr 2008;18(4):640-643.

27. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet.* Oct 2007;39(10):1202-1207.

28. Yuan HY, Chiou JJ, Tseng WH, et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* Jul 1 2006;34(Web Server issue):W635-641.

29. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* Apr 1997;13(4):163.

30. Rollason R, Korolchuk V, Hamilton C, Jepson M, Banting G. A CD317/tetherin-RICH2 complex plays a critical role in the organization of the subapical actin cytoskeleton in polarized epithelial cells. *J Cell Biol.* Mar 9 2009;184(5):721-736.

31. Tokarev A, Skasko M, Fitzpatrick K, Guatelli J. Antiviral activity of the interferon-induced cellular protein BST-2/tetherin. *AIDS Res Hum Retroviruses.* Dec 2009;25(12):1197-1210.

Legends

Table 1: Four SNPs passed the Bonferroni threshold after comparing slow progressors against controls on low frequency SNPs (MAF < 5%). For each SNP, the chromosome, the minor allele A1, the minor allele frequency (MAF), the frequency of the group in the dominant mode (used for association tests), the *p value* in the dominant mode, the odds ratio (OR), and the genes concerned (SNP and/or SNP(s) in LD with $r^2 \geq 0.9$ in Caucasians HapMap data) are provided. A SNP was assigned to a gene if it was located in the gene or within the 2kb flanking regions.

Figure 1: Genetic map of the *RICH2* gene region. 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 by an asterisk (*), respectively. The 2 SNPs rs2072255 (intronic) and rs2072254 (exonic) are represented ($r^2=1$) and a zoom with the genetic sequences is provided.

Figure 2: Frequency of individuals carrying the rs2072255-A variant (dominant mode) in the various groups : SP GRIV (n=270), SP ACS (n=31), SP MACS (n=59), the D.E.S.I.R. control group (n=694). CTR-ACS (n=376), the Illumina CTR (n=324), and the RP group (n=147). The frequencies within each of the GRIV RP, ACS RP, and MACS156 study RP groups was similar.

Tables and Figures

Table 1

SNP	chr	A1	A2	MAF_{SP}	MAF_{CTR}	Frequency dominant mode : A1A1 & A1A2		P_{SP-CTR}	OR	Gene(s) / LD
						SP_{ALL}	CTR_{ALL}			
rs2395029	6	C	A	10.19%	3.41%	20.38%	6.81%	$8.54x\ 10^{-15}$	**3.41** CI 95%: 2.4-4.8	*HCP5, intergenic, MICB*
rs9368699	6	G	A	9.20%	3.48%	18.13%	6.96%	$3.03x10^{-10}$	**2.9** CI 95%: 2.04-4.16	*C6orf48, RDBP, TNXB*
rs8192591	6	T	C	6.83%	3.15%	13.63%	6.30%	$9.08x10^{-7}$	**2.32** CI 95%: 1.56- 3.41	*NOTCH4 GPSM3*
rs2072255	17	A	G	4.79%	10.04%	9.04%	18.87%	$3.30x10^{-6}$	**0.43** CI 95%: 0.28-0.63	*RICH2*

Figure 1

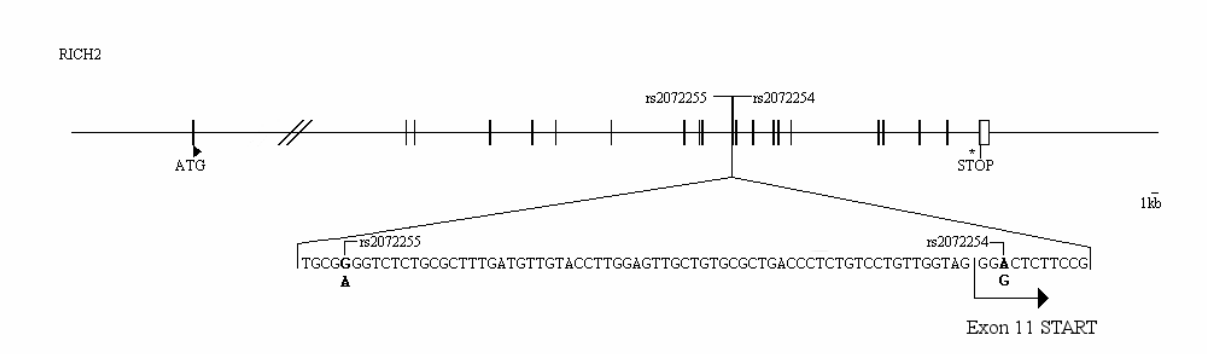
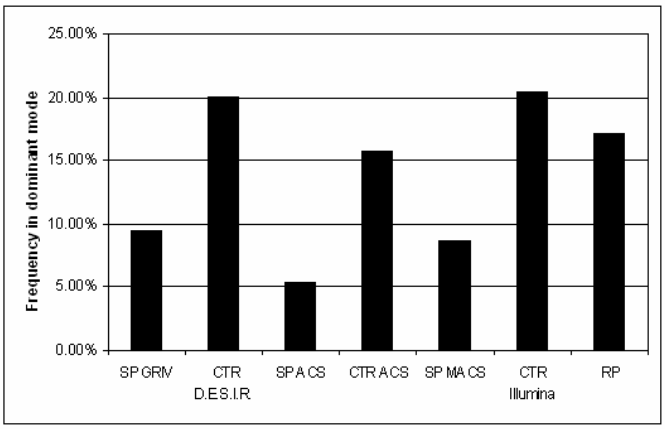
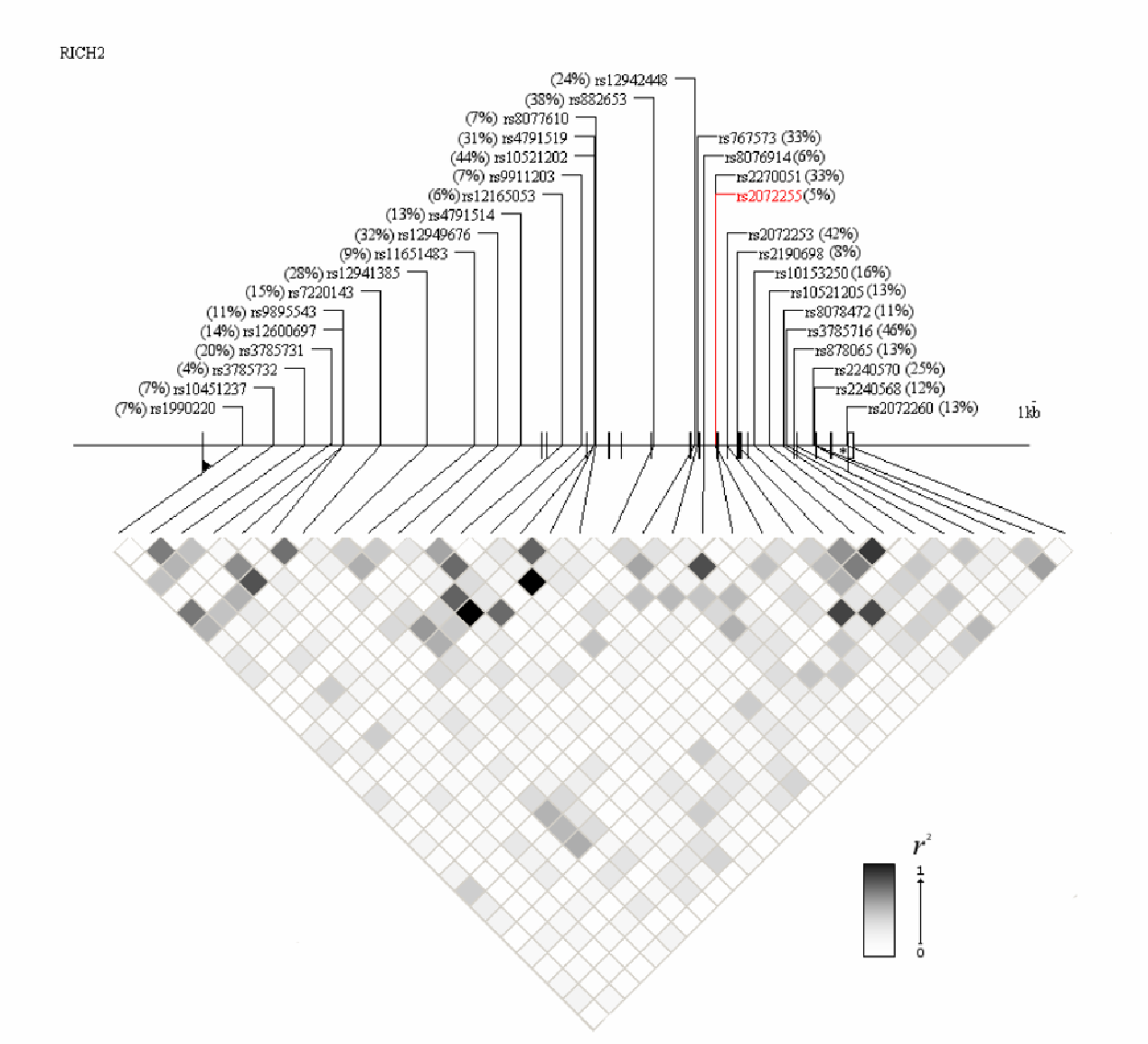


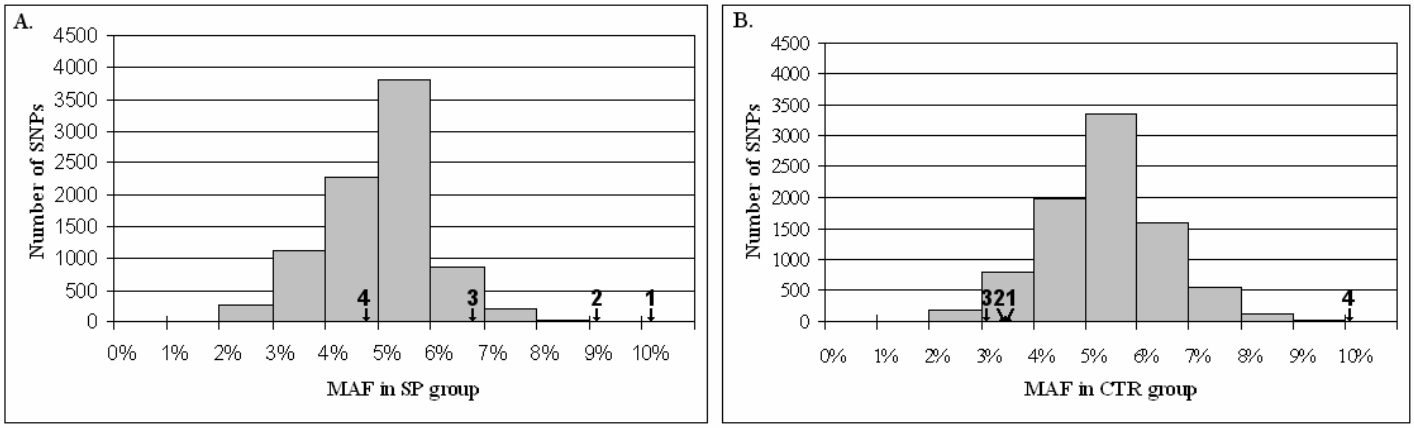
Figure 2



Supplemental Digital Content 1: Genetic map and Linkage Disequilibrium (LD, r^2) map presenting *RICH2* SNPs genotyped by the Illumina Infinium II HumanHap300 beadChips (Illumina, San Diego, CA, USA). 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 by an asterisk (*), respectively. In parenthesis next to the SNP name, the lowest MAF either in the SP or in the CTR group is provided. There were 2 SNPs with MAF <5% that were selected in our study: the significant SNP rs2072255 which was reliably imputed (imputation quality score P=0.99), and a second SNP rs3785732 (imputation quality score P=0.78) which was excluded from the analysis. The significant SNP is marked in red.



Supplemental Digital Content 2: Distribution of the MAF of the SNPs studied in this work in the SP group (left panel) and in the CTR group (right panel). The MAF of the SNPs associated with SP are shown by the numbers 1, 2, 3 and 4 corresponding respectively to SNPs *HCP5* rs2395029, *C6orf48* rs9368699, *NOTCH4* rs8192591, *RICH2* rs2072255.



Supplemental Digital Content 3: Summary of the results obtained in the various cohorts. The first column shows the comparison the SP subset (GRIV, ACS, MACS156 group) with the CTR subjects (D.E.S.I.R., CTR ACS, CTR illumina) with the minor allele frequency in the dominant mode and the p value. The second and third columns show the internal replication in the individuals cohorts ACS and MACS156 study : minor allele frequency in the dominant mode, p-value obtained by linear regression measuring time to AIDS93 according to the genotype in the dominant mode.

	SP_{ALL} N=365	**CTR_{ALL} N=1394**	**P_{SP-CTR}**	**ACS N=416**	**P_{AIDS93}**	**MACS156 N=156**	**P_{AIDS93}**
rs2395029-C	20.38%	6.81%	$8.5x\ 10^{-15}$	7.4%	$3x10^{-4}$	6.6%	$5.46x10^{-4}$
rs9368699-G	18.13%	6.96%	$3x10^{-10}$	6.9%	$2.9x10^{-4}$	12.4%	$9.9x10^{-3}$
rs8192591-T	13.63%	6.30%	$9x10^{-7}$	5.9%	1	10.4%	$5.56x10^{-1}$
rs2072255-A	9.04%	18.87%	$3.3x10^{-6}$	15.2%	$4.2x10^{-2}$	14.7%	$4.82x10^{-2}$

Quatrième partie :

Discussion

1. Bilan des études d'association réalisées sur la cohorte GRIV

1.1. Travaux sur 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 [176].

La principale nouveauté de cette analyse 'génome entier' réside dans l'étude de patients séropositifs VIH-1, caractérisés par le phénotype extrême de non-progression vers le SIDA. Nous avons répliqué le signal majeur du SNP rs2395029 localisé dans le gène *HCP5* (p=6,79x10^{-10}), déjà mis en lumière lors de l'analyse sur la cohorte Euro-CHAVI [137]. Cette association peut être expliquée par le déséquilibre de liaison entre le rs2395029 et des SNPs localisés dans des gènes majeurs de la région *HLA* tels que *HLA-B, MICB, TNF, LTB* et *BAT1*. L'hypothèse que le gène *HCP5* pourrait interagir avec le VIH via un mécanisme d'ARN antisens avait été émise, cependant, il n'existe pas, à ce jour, de preuves en faveur de cette hypothèse. L'allèle *HLA-B*57* est associé au contrôle de la réplication du VIH-1 et de la progression vers le SIDA (Introduction, chapitre 3.2.2) [177]. MICB est un ligand pour les cellules T $CD8^+$ et NK, cellules clés pour la réponse anti-VIH (Introduction, chapitre 2.2.2). BAT1 est un composant essentiel de la machinerie d'épissage et d'export de l'ARN [178] et est également un régulateur négatif de cytokines pro-inflammatoires [179]. LTB est un modulateur essentiel de l'inflammation [180]. Enfin, le TNF est une cytokine pro-inflammatoire, ayant été largement explorée dans l'étude de l'infection VIH [181]. D'un point de vue biologique, tous ces gènes constituent donc de très bons candidats qui pourraient intervenir dans la pathogenèse du SIDA. Cependant, le fort déséquilibre de liaison dans cette région rend difficile la discrimination du ou des gènes causaux, il sera donc important d'utiliser des stratégies différentes telles que les explorations fonctionnelles pour pouvoir approfondir l'analyse.

Les données de la première étude 'génome entier' sur la cohorte Euro-CHAVI [137] étant disponibles, nous avons réalisés une méta-analyse entre nos deux études. Cette approche a permis de mettre en évidence essentiellement des polymorphismes de la **région *HLA*** (gènes

TNXB, *BAT2-3*, *RDBP*, *HLA-C*, *PSORS1C1*), dont la majorité est liée au signal *HCP5*. Afin d'affiner notre analyse, nous avons entrepris une étude stratifiée par les génotypes du SNP rs2395029, la plupart des signaux précédents disparaissent. Les premiers signaux indépendants du SNP *HCP5* impliquent des polymorphismes du locus ***ZNRD1/RNF39***, également localisé dans la région *HLA,* suggérant un rôle indépendant dans le phénotype de progression de la maladie. D'un point de vue fonctionnel, le gène *RNF39* est peu caractérisé. L'expression du gène *ZNRD1*, codant pour une sous-unité de l'ARN polymérase I, est modulée par plusieurs allèles du locus *ZNRD1/RNF39* d'après la base de données d'expression *Genevar* [172, 173].

L'étude réalisée par notre équipe constituait la première GWAS ciblant un phénotype extrême de progression et a confirmé le rôle majeur de la région *HLA* pour la non-progression à long terme et pour le contrôle de la charge virale.

1.2. Travaux sur la progression rapide

Nous avons réalisé une étude d'association 'génome entier' à l'aide de puces Illumina HumanHap300 basée sur la comparaison de 85 progresseurs rapides avec 1352 contrôles séronégatifs [182].

Comme dans le cas de l'analyse sur la non-progression à long terme, l'étude du phénotype extrême de progression rapide constituait une nouveauté. La puissance de cette cohorte de progresseurs rapides avait déjà été démontré à travers différentes analyses du type 'gène candidat' [114]. Cependant lors de l'analyse 'génome entier', aucun signal n'a atteint le seuil de significativité statistique de Bonferroni. Pour faire face à cette difficulté, nous avons réalisé une approche statistique basée sur la méthode du False Discovery Rate (FDR), permettant également de prendre en compte le problème des tests multiples [163]. Afin de valider la pertinence de cette méthodologie, nous avons évalué par simulations les chances d'obtenir autant de signaux positifs sur des populations contrôles de même taille choisies au hasard : nous avons ainsi comparé les génotypes de 1000 sous-groupes de 85 individus extraits aléatoirement d'une population contrôle indépendante avec les génotypes de nos 1352 contrôles. Pour chacune de ces simulations, nous avons compté le nombre d'associations indépendantes atteignant le seuil de FDR à 25%. Moins de 1% des tests a permis d'obtenir au moins 6 signaux indépendants avec une telle méthodologie (moyenne=0,4 ± 1,29). Ce résultat souligne la pertinence de l'approche FDR utilisée, malgré

la taille modeste de notre population, et confirme la puissance des cohortes à phénotype extrême pour la découverte de nouveaux facteurs génétiques de risque.

Cette approche nous a permis de mettre en évidence 6 nouvelles associations avec des OR (Odd Ratio) allant au delà de 4, suggérant des effets biologiques très forts. Le meilleur signal a été observé pour les SNPs rs4118325 (p=6,09x10^{-7}, OR=0,24) et rs1044056 (p=4,29x10^{-6}, OR=0,27) en déséquilibre de liaison (r^2=0,92). Ces deux SNPs sont eux aussi en déséquilibre de liaison avec des SNPs du gène *PRMT6.* D'autres signaux avec des FDR inférieur à 25% et localisés dans des gènes ont pu être identifiés : ***SOX5*** (rs1522232, p=4,29x10^{-6}, OR=0,45), ***RXRG*** (rs10800098, p=4,29x10^{-6}, OR=3,29), ***TGFBRAP1*** (rs1020064, p=4,29x10^{-6}, OR=0,34). Enfin, deux autres SNP relativement éloignés de tous gènes (±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 [183, 184]. RXRG code pour un récepteur nucléaire de l'acide rétinoïque, impliqué dans la répression de la transcription du VIH-1 [185, 186]. PRMT6 est une arginine N-méthyltransférase pouvant méthyler les protéines Tat et Rev du VIH-1 [187, 188], et la molécule HMGA1 [189], 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 [190]. TGFBRAP1 est impliquée dans la voie de signalisation du TGFβ [191], cytokine immunosuppressive pléiotropique. Le TGFβ est notamment impliqué dans la différenciation des cellules Th17 et Treg, importantes pour le maintien de l'intégrité mucosale et le contrôle de l'inflammation [192-195]. 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.

Nous avons exploité les haplotypes des régions mises en lumière dans cette étude 'génome entier', en nous limitant au SNPs des exons et du promoteur. Pour *PRMT6* nous avons montré que le signal pourrait être lié à l'haplotype impliquant le SNP rs4118325 et un SNP exonique non synonyme (rs2232016, [Ala135Val]). Nous avons également démontré que le SNP rs1020064 du gène *TGFBRAP1*, est en déséquilibre avec un haplotype de 3 SNPs. exoniques (rs2241801 [Arg83Arg], rs12476720 [Arg241Arg], et rs2241797 [Arg275His]). Le polymorphisme non-synonyme rs2241797 semble essentiel pour l'association, puisque le signal disparaît lorsque ce SNP est retiré de l'haplotype, contrairement aux deux SNPs

synonymes Arg83Arg et Arg241Arg. De plus, d'après les bases de données d'expression *Genevar* [172, 173] et Dixon [174], le SNP rs1020064 est corrélé à l'expression différentielle de *TGFBRAP1* et *PPP2R3A*, et rs2241797 à celle de plusieurs gènes dont certains ont été décrits comme interagissant avec le VIH-1.

Il est important de préciser ici que, depuis la publication de l'article, ces signaux de progression rapide ont pu être répliqués partiellement dans des cohortes indépendantes telles que ACS et Euro-CHAVI, à l'exception du signal *TGFBRAP1* (données non publiées).

1.3. Travaux sur les SNPs de faible fréquence

Nous avons analysé les données issues d'étude 'génome entier', en nous focalisant sur les SNPs de faible fréquence, obtenant ainsi un seuil de Bonferroni de $5{,}8 \times 10^{-6}$. Nous avons combiné les progresseurs rapides et les non-progresseurs à long terme de 3 cohortes (GRIV, ACS et l'analyse MACS156) totalisant 365 NP et 147 PR que nous avons comparés à 1394 contrôles.

Aucune association n'a été découverte lors de la comparaison des progresseurs rapides avec les contrôles. Nous n'avons pas été surpris par ce résultat, car en tenant compte des effectifs du groupe PR et du spectre de fréquences, il aurait fallu un odds ratio de 2,8, c'est à dire un effet biologique très fort pour pouvoir passer le seuil de Bonferroni. Dans le cas des non-progresseurs à long terme 4 SNPs ont passé le seuil statistique, parmi ces résultats 3 SNPs sont situés sur le chromosome 6 et ont été identifiés dans d'autres études : rs2395029 dans le gène *HCP5* (p=$8{,}54 \times 10^{-15}$) [137, 141, 142, 145, 150, 176], rs9368699 dans *C6orf48* (p= $3{,}03 \times 10^{-10}$) [137, 176], rs8192591 dans *NOTCH4* (p=$9{,}08 \times 10^{-07}$) [141]. *NOTCH4* constitue un candidat intéressant du fait de son rôle dans la régulation de l'immunité. L'analyse précédemment faite sur la cohorte Euro-CHAVI avait indiqué une indépendance entre cette association et le SNP rs2395029, cependant ce n'est pas le cas dans notre étude. Une explication possible pour cette différence est l'utilisation de deux phénotypes distincts : la charge virale ou la progression vers le SIDA.

Le quatrième signal identifié correspond au SNP rs2072255 (p=$3{,}30 \times 10^{-6}$) dans le gène *RICH2* localisé sur le chromosome 17. L'allèle rs2072255-A de *RICH2* favorise la progression vers le SIDA avec une fréquence de 9,04% chez les NP contre 18.97% chez les

contrôles. Le gène *RICH2* s'avère être un candidat intéressant, en effet, le produit de ce gène pourrait empêcher l'internalisation de BST-2 [196], qui est connue pour prévenir le bourgeonnement et la libération virale du VIH-1 [197]. De manière intéressante, le SNP rs2072255 est en déséquilibre de liaison total avec le SNP rs2072254 qui se situerait dans un site d'épissage d'après FastSNP http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp/). Si ce polymorphisme altère l'épissage de l'ARNm, cela pourrait mener à une diminution de l'expression de *RICH2* et expliquer l'action réduite de BST-2.

L'identification de signaux déjà confirmés souligne la pertinence de travailler sur les SNPs de fréquence faible. Les données biologiques et génétiques concernant *RICH2,* font de lui un nouveau gène candidat très intéressant. Des études expérimentales et génétiques supplémentaires sont nécessaires pour confirmer et comprendre le rôle de RICH2 dans la pathogenèse du SIDA.

1.4. Comparaison 'génome entier' entre LTNP et PR

En parallèle des trois études 'génome entier' présentées ci-dessus, nous avons également réalisé la comparaison des LTNP (n=275) avec les PR (n=85), mais cette étude n'a pas permis d'identifier de signaux supplémentaires. La comparaison systématique entre la population LTNP et la population de PR se révèle moins pertinente que la comparaison entre l'une de ces populations 'cas' et une population 'contrôle' pour plusieurs raisons : (i) perte en puissance statistique, car les effectifs des populations LTNP et PR sont plus petits que celui du groupe contrôle ; (ii) notre expérience passée nous a montré que la plupart des signaux étaient soient associés à la non progression, soit à la progression rapide, rarement aux deux (iii) la détection d'un signal par ce type d'analyse ne permet pas, sans une population contrôle, de déterminer si le polymorphisme est associé à la non-progression à long terme ou à la progression rapide ; (iv) enfin, la comparaison LTNP-PR ne permet pas de dévoiler les signaux liés à la susceptibilité et à l'acquisition du VIH-1.

2. Comparaison des signaux obtenus avec ceux des autres études génomiques sur le SIDA

2.1. Comparaison avec les approches 'gène candidat'

Les approches 'gène candidat' 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, chapitre 3.2.2.). Ces méthodes ont confirmé ou infirmé l'implication des gènes étudiés, permettant ainsi d'éclaicir certains des mécanismes de pathogenèse. Lorsque l'on compare les résultats obtenus par ce type d'approche à ceux des études d'association 'génome entier' sur la cohorte GRIV, 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.

Cependant il faut souligner que des signaux comme le CCR5-Δ32, reconnus par tous (voir Introduction, chapitre 3.2.2.), n'ont pas été retrouvés par l'approche 'génome entier'. Cela s'explique simplement parce que le variant CCR5-Δ32 n'est pas représenté sur les puces. D'autres gènes ne seront pas retrouvés pour des raisons biologiques et statistiques : par exemple, un effet biologique faible conduisant à une valeur *p-value* trop faible par rapport aux corrections de Bonferroni utilisées dans l'exploitation des données 'génome entier'.

L'approche 'génome entier' sur les progresseurs rapides a mis en lumière 6 signaux dans des loci peu soupçonnés jusqu'à présent (*PRMT6*, p=6,09x10^{-7} ; *SOX5*, p=4,29x10^{-6} ; *RXRG*, p=4,29x10^{-6} ; *TGFBRAP1*, p=4,29x10^{-6}) [182]. Ces résultats démontrent l'intéret des études 'génome entier' dans la découverte de nouveaux signaux associés à la progresssion vers le SIDA en s'affranchissant des a priori biologiques.

2.2. Comparaison avec les autres études 'génome entier' publiées

Nous avons évoqué dans ce manuscrit l'ensemble des études 'génome entier' réalisées dans le cadre du SIDA (voir Introduction, chapitre 3.2.3.). Ces résultats ont tous confirmé

l'importance de la région *HLA*. Certaines études ont aussi souligné l'importance 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, 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 [150, 198, 199]. 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 études 'génome entier' abordées jusqu'ici, ont été entreprises sur des cohortes présentant différentes caractéristiques. L'hétérogénéité observée entre ces groupes réside dans les critères de choix lors du recrutement (charge virale, cellules T $CD4^+$, trithérapie, séroconvertis, séroprévalents, profils extrêmes, hommes/femmes, mode d'infection homosexuel, consommateur de drogues en intraveineuse), les phénotype étudiés (qualitatifs : LTNP, PR, *'elite controllers'* ; quantitatifs : charge virale, cellules T $CD4^+$, temps jusqu'au développement du SIDA...), le nombre d'individus recrutés (45 à 2554 individus), l'origine populations variée (européenne : Angleterre, Australie, France, USA... ; afro-américaine), et le génotypage sur différentes plateformes (Illumina, Affymetrix, à l'aide de puces ciblant de 300K à 1M de SNPs).

Toutes ces différences, peuvent être en grande partie à l'origine de l'hétérogénéité observée dans les résultats des ces analyses et de la difficulté de réplication entre les cohortes. Malgré ces aspects négatifs, les différences entre ces cohortes font leur complémentarité. En s'intéressant à divers phénotypes, les associations détectées peuvent être liées à des phénomènes se déroulant à différents moments, dans différents compartiments ou contextes de l'infection VIH. Il est donc important d'une part de continuer à étudier et comparer les cohortes existantes et d'autre part de développer de nouvelles cohortes bien définies, ciblant de nouvelles populations (africaines, asiatiques...), de nouveaux phénotypes...

Dans le cas de deux études réalisées sur la cohorte GRIV, sur les SNPs de faible fréquence et sur les non progresseurs à long terme non 'élites', les résultats du rs2072255 (*RICH2*) et du rs2234358 (*CXCR6*) ont pu être répliqués dans la cohorte ACS et le groupe MACS156 malgré les différences entre les cohortes, permettant même dans le cas de *CXCR6* de passer le seuil statistique 'genome entier'. De plus, le résultat sur le SNP de *HCP5* a été répliqué [137, 142, 145, 176] dans un grand nombre de cohortes VIH^+. Ces trois résultats

démontrent que la réplication est posssible malgré l'hétérogénéité entre les groupes et renforcent l'idée de complémentarité entre ces différentes cohortes.

3. Critique des études 'génome entier'

3.1. Aspects positifs

L'atout majeur des études 'génome entier' repose sur une approche exhaustive, sans aucun a priori en criblant l'ensemble du génome. Ainsi, de nouvelles régions non soupçonnées précédemment peuvent être découvertes. Un autre aspect non négligeable réside dans l'obtention d'un grand nombre de données génétiques pour un coût raisonnable.

Comme nous l'avons vu, ce type d'étude dans le cadre du SIDA a permis l'identification de nouveaux SNPs qui n'étaient précédemment pas associés au VIH-1. Certains signaux tels que *HCP5/HLA-B*57*, ont pu être confimés.

Pour conclure, ces analyses ont généré une grande quantité de données dans divers domaines, pouvant mener à une meilleure compréhension des mécanismes moléculaires sous-jacents de chacun des phénotypes étudiés.

3.2. Aspects négatifs

L'utilisation de méthodes statistiques standardisées et stringentes liées aux tests multiples assure la robustesse des résultats (détection de 'vrais positifs'), mais sont également à l'origine de nombreux résultats 'faux négatifs' d'effets plus modestes, qu'il est impossible de discriminer des signaux 'vrais négatifs' par les méthodes classiques d'exploitation 'génome entier'. Ces corrections statistiques ne permettent pas une exploitation optimale des données générées.

Un autre aspect négatif des études 'génome entier' réside dans l'hypothèse de départ qui associe un variant commun à une maladie commune [96]. Ainsi les SNPs avec des MAF<5% ont été largement négligés lors de la conception des technologies de génotypage. Depuis quelques années l'implication des SNPs avec un spectre de fréquences varié est soutenue par un grand nombre d'équipes de recherche [102, 103, 104]. Actuellement de nouveaux outils de génotypage sont disponibles et permettent de mieux cibler ces SNPs (voir Introduction, chapitre 3.1.4). Cependant comme évoqué dans le troisième article de cette

thèse, les seuils statistiques ‘génome entier’ rendent difficile l’émergence de SNPs avec de faibles fréquences sur des cohortes standard.

Au sein des variations communes (MAF> 5%), même si la couverture de la diversité du génome par les puces de génotypage est très efficace, elles ne ciblent cependant pas forcément tous les marqueurs fréquents des populations. Par exemple, les puces Illumina HumanHap300 capturent tous les SNPs fréquents de la population européenne de la Phase I de HapMap avec un seuil $r^2 \geq 0,8$, cependant, elles ne couvrent que 77% des SNPs fréquents de la population européenne de la Phase II de HapMap avec un seuil $r^2 \geq 0,8$[99]. Ainsi, en fonction de la technologie utilisée, certains polymorphismes fréquents impliqués dans la maladie peuvent être manqués par les études d'association ‘génome entier’.

Les puces de génotypage ciblent des SNPs marqueurs ou tagSNPs qui constituent une bonne couverture du génome. Les SNPs identifiés dans les études d'association ‘genome entier’ sont donc des marqueurs d'une région et non nécessairement des variants causaux (*e.g.* rs2395029 illustre le profil complexe de déséquilibre de liaison de la région *HLA* et reflète de nombreux allèles et gènes candidats tels que *HLA-B*57*, *TNF*, *MICB*...). Cet aspect implique une analyse plus fine sur les régions mises en lumière.

Tous ces facteurs expliquent le fait que les variants identifiés dans les études d'association 'génome entier' ne sont généralement responsables que d'une fraction des phénotypes observés dans les maladies.

4. Perspectives dans le cadre du SIDA

4.1. Approfondir les signaux obtenus

Les tagSNPs étudiés avec les puces de génotypages permettent la mise en évidence d'une région plutôt que du variant causal. Des investigations supplémentaires génétiques ou fonctionnelles seront nécessaires, afin d'identifier le ou les variants réellement responsables. Une meilleure connaissance des régions génétiques associées devrait permettre d'établir comment un variant identifié peut influencer les mécanismes de pathogenèse (haplotypage, interaction). Des expérimentations biologiques pourront confirmer l'effet d'une variation et montrer qu'elle est problement causale (par exemple sur-expression du gène). Des banques de données bioinformatiques comme Genevar [172, 173] ou Dixon [174] existent ainsi et peuvent forunir directement de telles informations.

4.2. Développement des cohortes et méta-analyses

4.2.1. Développement des cohortes

L'augmentation de la taille des cohortes VIH existantes devrait permettre l'accroissement de la puissance statistique de détection des signaux. Dans le cadre de la cohorte GRIV, une phase de recrutement de progresseurs rapides est en cours.

Le développement de nouvelles cohortes ciblant de nouvelles populations (africaines, asiatiques...), et/ou de nouveaux phénotypes (HEPS, acquisition du VIH-1, réponse aux traitements HAART, maladies non-SIDA apparaissant sous HAART -*e.g.* maladies rénales, cardiovasculaires, diabètes, ...) permettra aussi d'explorer d'autres phénotypes et d'identifier de nouveaux facteurs génétiques associés à l'infection VIH-1.

4.2.2. Analyses combinées de cohortes (méta-analyses)

Les méta-analyses, combinant les données génomiques de plusieurs études, ont déjà démontré leur efficacité dans le cas de nombreuses pathologies en révélant de nombreux nouveaux signaux. En ce qui concerne le SIDA, une collaboration entre les équipes GRIV [176,

[182], ACS [150], MACS156 [145] et USA HIV-1 [151], a donné naissance à un projet de méta-analyse entre les différentes études, qui est actuellement en cours. D'autre part, un effort international a permis la mise en commun des données génomiques issues de toutes les cohortes VIH, afin d'explorer les facteurs génétiques impliqués dans la transmission et l'acquisition de l'infection VIH, qui n'avaient jusqu'ici pas été étudiés (projet IHAC, *International HIV-1 Acquisition Consortium*). Ce projet est actuellement en cours, il nécessite une intégration des données phénotypiques, une évaluation et une correction minutieuse de la stratification des populations, et une prise en charge bioinformatique conséquente (gestion de la masse de données et imputation des données issues de différentes plateformes) afin de révéler de nouveaux facteurs génétiques associés à l'infection VIH.

4.3. Développement de nouvelles heuristiques statistiques

Lors des analyses 'génome entier', l'utilisation de tests statistiques stringents tels que les corrections de Bonferroni sont nécessaires pour assurer la détection de signaux 'vrais positifs'. Il existe aujourd'hui des approches différentes telles que le FDR, pour aider à contrôler le taux de faux positifs. Ainsi cette méthode a permis, dans le cadre d'analyse 'génome entier', de découvrir plusieurs associations liées à la maladie d'Alzheimer [200], et à la progression rapide vers le SIDA [182]. Dans le but d'affiner la détection des signaux, l'équipe GRIV développe actuellement un nouveau logiciel de correction statistique évaluant le nombre de tests indépendants dans une région génétique.

Comme nous l'avons vu dans le troisième article de cette thèse, des études priorisées peuvent permettre d'identifier des associations supplémentaires. Nous avons ainsi choisi d'analyser les SNPs de faible fréquence en leur appliquant les critères statistiques usuels, mais il est aussi possible de baser les analyses sur des connaissances biologiques plus spécifiques de la pathologie étudiée. Ce type d'approche correspond à une approche 'gène candidat' élargie, avec des a priori moins limités qui favoriseront la découverte possible de signaux insoupçonnés.

4.4. Développement de nouvelles approches bioinformatiques

4.4.1. Les données 'multi-marqueurs'

Les données génomiques 'multi-marqueurs' ont été peu exploitées jusqu'à aujourd'hui. Il existe des logiciels qui permettent de générer et d'analyser les haplotypes sur l'ensemble du

génome [171]. L'analyse des haplotypes présente un grand intêret d'un point de vue biologique, en effet, les haplotypes correspondent à des combinaisons d'allèles transmises de générations en générations, 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 [201]. De plus, leur complexité les rend certainement plus adaptés pour se corréler et décrire la diversité des mécanismes biologiques. Une analyse sur les patients de la cohorte GRIV est actuellement en cours.

D'autre part, les interactions épistatiques nécessitent des recherches supplémentaires, en effet la plupart des études d'association 'génome entier' ne montrent pas de preuve de ce type d'interaction. Elles existent probablement, mais elles sont indectables car exlues ou minimisées par les méthodes d'analyse bioinformatiques [202].

4.4.2. Les données relatives aux études fonctionnelles haut-débit

Au cours de ces dernières années, des études fonctionnelles à haut-débit ont été réalisées *in vitro* dans le cadre du SIDA et ont identifié des gènes impliqués dans le cycle viral. Trois études 'génome entier' utilisant des siARN ont ainsi permis d'inactiver un à un les gènes du génome dans des modèles cellulaires humains, et d'évaluer ensuite la capacité de réplication du virus [175 , 203 , 204]. De façon similaire, une étude de criblage 'génome entier' à l'aide de shARN (*small hairpin* ARN) a été menée [205]. Des études protéomiques ont identifié des protéines exprimées différentiellement à différentes étapes de l'infection VIH (phases précoce/tardive, infection bystander...) [206 , 207 , 208 , 209]. D'autres études haut débit basées sur le criblage fonctionnel d'ADNc [210], de microARN [211] et sur le profil de transcription [212] ont également été conduites. Bien que les modèles expérimentaux mis en place dans toutes ces études soient à chaque fois des systèmes différents et non physiologiques, ces études représentent une source indiscutable de cibles candidates pour l'exploration de l'infection VIH. Ces études sont peu recouvrantes en termes de facteurs identifiés, mais ont cependant mis en lumière des voies de signalisation communes.

Les facteurs identifiés lors de ces études à haut débit *in vitro* représentent des gènes candidats prioritaires à comparer avec les données génétiques *in vivo* issues des GWAS. La combinaison de l'ensemble de ces données dans un système bioinformatique de priorisation devrait permettre de dresser une image cohérente de la réplication virale dans la cellule hôte.

4.5. Nouvelles technologies

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 SNPs, notamment les SNPs rares avec des MAF inférieures à 1%. Ainsi, il est possible aujourd'hui 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 d'autres polymorphismes comme les CNV ou les indels, offrant ainsi une couverture plus complète de la diversité entre les individus et de nouvelles informations sur la structure du génome.

Il existe 3 aspects qui représentent un enjeu pour le développement et la réussite du séquençage haut débit : statistique, informatique, et financier.

D'un point de vue **statistique** ce type d'approche démultiplie le nombre de tests, et requiert donc de nouvelles procédures de priorisation des très nombreux polymorphismes identifiés ou le développement de nouveaux outils statistiques et bioinformatiques [213].

Les ressources **informatiques** afin de gérer les données de séquençage 'genome 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.

Enfin le **coût** des technologies de séquençage reste très élevé. Les deux produits les plus connus qui permettent aujourd'hui de séquencer la totalité du génome ainsi que l'exome humain sont Illumina/Solexa's GA_{II} et Life/APG's SOLiD 3. Elles présentent toutes les deux des prix similaires avec des coûts d'environ 10 à 20 000 $ pour le génome et de 3 à 5 000 $ pour l'exome. Même si ces deux techniques de séquençage sont bien différentes, les résultats obtenus dans les deux cas sont de qualité égale (voir la revue [214]). Le développement de ces technologies représente un enjeu majeur actuel, et un grand nombre d'entreprises travaillent sur la diminution des coûts et l'amélioration de la qualité des séquences générées. On peut imaginer que le coût du séquençage du génome entier sera à terme de 1000 à 2000 $.

A l'heure actuelle le choix du séquençage de l'exome peut permettre de réduire les coûts en passant de 3,2 Gb à 38 Mb à séquencer. Cette approche présente l'avantage de permettre un meilleur séquençage de chaque région ciblée (plus de passes) et donc beaucoup

plus de précision [215]. De plus d'un point de vue pratique, il semble plus raisonnable de cibler les régions des exomes qui sont les seules facilement interprétables à ce jour. Il serait intéressant de développer une approche incluant d'autres zones fonctionnelles telles que les régions promotrices des gènes.

Les puces de génotypages 2,5M actuelles, même si elles ne tiennent pas encore compte de toutes les données du projet 1000 génomes, peuvent apporter une solution moins coûteuse et moins fastidieuse que le séquençage. En effet, leur coût est de l'ordre de 500 Euros et les outils bioinformatiques d'imputation devraient permettre de retrouver des variants assez rares. Encore une fois tout dépend de la taille de la population testée, mais si l'on considère qu'à coût égal on peut génotyper 6 fois plus de sujets avec les puces de génotypage, elles peuvent fournir un avantage statistique non négligeable, même pour les variants assez rares.

L'utilisation de populations extrêmes pourrait constituer elle aussi une solution alternative, en recherchant à mettre en valeur l'incidence de variations rares sur certains gènes. En effet, un fort contraste génétique (effet 'loupe') favorise la découverte de nouveaux facteurs : à effectif équivalent, ces cohortes sont plus puissantes statistiquement que les cohortes constituées de patients à tous stades de la maladie [216, 217, 218] et elles peuvent constituer des sources d'identification de gènes importants. L'étude de la cohorte GRIV a ainsi confirmé la puissance des phénotypes extrêmes, y compris avec des populations de taille réduite et pourrait être très utile dans ce type d'approche. Aujourd'hui, si 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.

Dans le cadre du SIDA, 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 [102, 103, 104]. Les '*elite controllers*' du VIH présentent un phénotype très rare dans la population (moins de 1% des personnes infectées). Des effets très forts de la région *HLA* ont été retrouvés dans la cohorte GRIV pour la population '*elite controller*' [176] et une publication récente a exactement confirmé ce résultat [219]. Ainsi les *'elite controllers'* non HLA-B57 pourraient constituer une population de choix pour ce type d'approche.

Conclusion

Les travaux exposés dans cette thèse correspondent à une approche moderne de la biologie, rendue possible depuis peu grâce aux progrès de la génétique moléculaire. Au cours de cette thèse, j'ai participé aux études d'association 'génome entier' réalisées sur les patients de la cohorte GRIV et au développement de stratégies pour exploiter les données de génotypage.

La première étude a confirmé 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.

La seconde étude a révélé de nouveaux gènes associés à la progression rapide vers le SIDA (*PRMT6*, *SOX5*, *RXRG*, et *TGFBRAP1*), soulignant l'importance du contrôle de la réplication virale et de l'inflammation.

La troisième étude était focalisée sur l'analyse des SNPs de faible fréquence issus de trois études 'génome entier' précédemment réalisées dans le cadre du SIDA et a permis de dévoiler un nouveau candidat pertinent, le gène *RICH2*.

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, ou encore permettre de développer des outils diagnostiques qui aideront les médecins à mieux prédire les évolutions de leurs patients.

Pour conclure, ce travail aura constitué pour moi un apprentissage particulièrement enrichissant en abordant un sujet multidisciplinaire à l'interface entre biologie, génétique, statistique et bioinformatique.

Bibliographie

1. Gottlieb MS. Pneumocystis pneumonia--Los Angeles. . *MMWR Morb Mortal Wkly Rep.* Jun 1981;96(6):980-981; discussion 982-983.

2. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR Morb Mortal Wkly Rep.* Jul 3 1981;30(25):305-308.

3. Follow-up on Kaposi's sarcoma and Pneumocystis pneumonia. *MMWR Morb Mortal Wkly Rep.* Aug 28 1981;30(33):409-410.

4. Jaffe HW, Bregman DJ, Selik RM. Acquired immune deficiency syndrome in the United States: the first 1,000 cases. *J Infect Dis.* Aug 1983;148(2):339-345.

5. Barre-Sinoussi F, Chermann JC, Rey F, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science.* May 20 1983;220(4599):868-871.

6. Gallo RC, Salahuddin SZ, Popovic M, et al. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science.* May 4 1984;224(4648):500-503.

7. Clavel F, Guetard D, Brun-Vezinet F, et al. Isolation of a new human retrovirus from West African patients with AIDS. *Science.* Jul 18 1986;233(4761):343-346.

8. Adjorlolo-Johnson G, De Cock KM, Ekpini E, et al. Prospective comparison of mother-to-child transmission of HIV-1 and HIV-2 in Abidjan, Ivory Coast. *Jama.* Aug 10 1994;272(6):462-466.

9. Marlink R, Kanki P, Thior I, et al. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science.* Sep 9 1994;265(5178):1587-1590.

10. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. *Science.* Jan 28 2000;287(5453):607-614.

11. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science.* Jan 27 1995;267(5197):483-489.

12. Bister K, Jansen HW. Oncogenes in retroviruses and cells: biochemistry and molecular genetics. *Adv Cancer Res.* 1986;47:99-188.

13. Jern P, Coffin JM. Effects of retroviruses on host genome function. *Annu Rev Genet.* 2008;42:709-732.

14. Swain A, Coffin JM. Mechanism of transduction by retroviruses. *Science.* Feb 14 1992;255(5046):841-845.

15. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem.* 1998;67:1-25.

16. Jones KA, Peterlin BM. Control of RNA initiation and elongation at the HIV-1 promoter. *Annu Rev Biochem.* 1994;63:717-743.

17. Paxton W, Connor RI, Landau NR. Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis. *J Virol.* Dec 1993;67(12):7229-7237.

18. Gottlinger HG, Dorfman T, Sodroski JG, Haseltine WA. Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. *Proc Natl Acad Sci U S A.* Apr 15 1991;88(8):3195-3199.

19. Krausslich HG, Facke M, Heuser AM, Konvalinka J, Zentgraf H. The spacer peptide between human immunodeficiency virus capsid and nucleocapsid proteins is essential for ordered assembly and viral infectivity. *J Virol.* Jun 1995;69(6):3407-3419.

20. Pettit SC, Moody MD, Wehbie RS, et al. The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J Virol.* Dec 1994;68(12):8017-8027.

21. Guo X, Liang C. Opposing effects of the M368A point mutation and deletion of the SP1 region on membrane binding of human immunodeficiency virus type 1 Gag. *Virology.* May 10 2005;335(2):232-241.

22. Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, Varmus HE. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature.* Jan 21 1988;331(6153):280-283.
23. Pierson TC, Doms RW, Pohlmann S. Prospects of HIV-1 entry inhibitors as novel therapeutics. *Rev Med Virol.* Jul-Aug 2004;14(4):255-270.
24. Hauber J, Cullen BR. Mutational analysis of the trans-activation-responsive region of the human immunodeficiency virus type I long terminal repeat. *J Virol.* Mar 1988;62(3):673-679.
25. Sodroski J, Patarca R, Rosen C, Wong-Staal F, Haseltine W. Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III. *Science.* Jul 5 1985;229(4708):74-77.
26. Romani B, Engelbrecht S, Glashoff RH. Functions of Tat: the versatile protein of human immunodeficiency virus type 1. *J Gen Virol.* Jan 2010;91(Pt 1):1-12.
27. Malim MH, Hauber J, Le SY, Maizel JV, Cullen BR. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature.* Mar 16 1989;338(6212):254-257.
28. Fritz CC, Zapp ML, Green MR. A human nucleoporin-like protein that specifically interacts with HIV Rev. *Nature.* Aug 10 1995;376(6540):530-533.
29. Ahmad N, Venkatesan S. Nef protein of HIV-1 is a transcriptional repressor of HIV-1 LTR. *Science.* Sep 16 1988;241(4872):1481-1485.
30. Guy B, Kieny MP, Riviere Y, et al. HIV F/3' orf encodes a phosphorylated GTP-binding protein resembling an oncogene product. *Nature.* Nov 19-25 1987;330(6145):266-269.
31. Hammes SR, Dixon EP, Malim MH, Cullen BR, Greene WC. Nef protein of human immunodeficiency virus type 1: evidence against its role as a transcriptional inhibitor. *Proc Natl Acad Sci U S A.* Dec 1989;86(23):9549-9553.
32. Kim S, Ikeuchi K, Byrn R, Groopman J, Baltimore D. Lack of a negative influence on viral growth by the nef gene of human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A.* Dec 1989;86(23):9544-9548.
33. Aiken C, Konner J, Landau NR, Lenburg ME, Trono D. Nef induces CD4 endocytosis: requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. *Cell.* Mar 11 1994;76(5):853-864.
34. Greenberg ME, Iafrate AJ, Skowronski J. The SH3 domain-binding surface and an acidic motif in HIV-1 Nef regulate trafficking of class I MHC complexes. *EMBO J.* May 15 1998;17(10):2777-2789.
35. Herbein G, Gras G, Khan KA, Abbas W. Macrophage signaling in HIV-1 infection. *Retrovirology.* 2010;7:34.
36. Greene WC. The molecular biology of human immunodeficiency virus type 1 infection. *N Engl J Med.* Jan 31 1991;324(5):308-317.
37. Strebel K, Daugherty D, Clouse K, Cohen D, Folks T, Martin MA. The HIV 'A' (sor) gene product is essential for virus infectivity. *Nature.* Aug 20-26 1987;328(6132):728-730.
38. Henriet S, Mercenne G, Bernacchi S, Paillart JC, Marquet R. Tumultuous relationship between the human immunodeficiency virus type 1 viral infectivity factor (Vif) and the human APOBEC-3G and APOBEC-3F restriction factors. *Microbiol Mol Biol Rev.* Jun 2009;73(2):211-232.
39. Guatelli JC. Interactions of viral protein U (Vpu) with cellular factors. *Curr Top Microbiol Immunol.* 2009;339:27-45.
40. Cullen BR. Regulation of human immunodeficiency virus replication. *Annual review of microbiology.* 1991;45:219-250.
41. Freed EO. HIV-1 and the host cell: an intimate association. *Trends Microbiol.* Apr 2004;12(4):170-177.

42. Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature.* Dec 20-1985 Jan 2 1984;312(5996):763-767.

43. Klatzmann D, Champagne E, Chamaret S, et al. T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature.* Dec 20-1985 Jan 2 1984;312(5996):767-768.

44. Rizzuto CD, Wyatt R, Hernandez-Ramos N, et al. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science.* Jun 19 1998;280(5371):1949-1953.

45. Deng H, Liu R, Ellmeier W, et al. Identification of a major co-receptor for primary isolates of HIV-1. *Nature.* Jun 20 1996;381(6584):661-666.

46. Bleul CC, Farzan M, Choe H, et al. The lymphocyte chemoattractant SDF-1 is a ligand for LESTR/fusin and blocks HIV-1 entry. *Nature.* Aug 29 1996;382(6594):829-833.

47. Berger EA, Murphy PM, Farber JM. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annu Rev Immunol.* 1999;17:657-700.

48. Stevenson M. Portals of entry: uncovering HIV nuclear transport pathways. *Trends Cell Biol.* Jan 1996;6(1):9-15.

49. Laughlin MA, Zeichner S, Kolson D, et al. Sodium butyrate treatment of cells latently infected with HIV-1 results in the expression of unspliced viral RNA. *Virology.* Oct 1993;196(2):496-505.

50. Han Y, Lassen K, Monie D, et al. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol.* Jun 2004;78(12):6122-6133.

51. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell.* Aug 23 2002;110(4):521-529.

52. Zagury D, Bernard J, Leonard R, et al. Long-term cultures of HTLV-III--infected T cells: a model of cytopathology of T-cell depletion in AIDS. *Science.* Feb 21 1986;231(4740):850-853.

53. Wiergers K, Rutter G, Kottler H, Tessmer U, Hohenberg H, Krausslich HG. Sequential steps in human immunodeficiency virus particle maturation revealed by alterations of individual Gag polyprotein cleavage sites. *J Virol.* Apr 1998;72(4):2846-2854.

54. Levy JA. *HIV and the pathogenesis of AID*. Washington DC: American Society for Microbiology; 2007.

55. Levy JA. HIV pathogenesis: 25 years of progress and persistent challenges. *AIDS.* Jan 14 2009;23(2):147-160.

56. Hazan U, Romero IA, Cancello R, et al. Human adipose cells express CD4, CXCR4, and CCR5 [corrected] receptors: a new target cell type for the immunodeficiency virus-1? *FASEB J.* Aug 2002;16(10):1254-1256.

57. Munier S, Borjabad A, Lemaire M, Mariot V, Hazan U. In vitro infection of human primary adipose cells with HIV-1: a reassessment. *AIDS.* Nov 21 2003;17(17):2537-2539.

58. Homsy J, Meyer M, Tateno M, Clarkson S, Levy JA. The Fc and not CD4 receptor mediates antibody enhancement of HIV infection in human cells. *Science.* Jun 16 1989;244(4910):1357-1360.

59. Alexaki A, Liu Y, Wigdahl B. Cellular reservoirs of HIV-1 and their role in viral persistence. *Curr HIV Res.* Sep 2008;6(5):388-400.

60. Redel L, Le Douce V, Cherrier T, et al. HIV-1 regulation of latency in the monocyte-macrophage lineage and in CD4+ T lymphocytes. *J Leukoc Biol.* Apr 2010;87(4):575-588.

61. Perelson AS, Essunger P, Cao Y, et al. Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature.* May 8 1997;387(6629):188-191.
62. Finzi D, Blankson J, Siliciano JD, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med.* May 1999;5(5):512-517.
63. Zagury D, Lachgar A, Chams V, et al. Interferon alpha and Tat involvement in the immunosuppression of uninfected T cells and C-C chemokine decline in AIDS. *Proc Natl Acad Sci U S A.* Mar 31 1998;95(7):3851-3856.
64. Fisher AG, Collalti E, Ratner L, Gallo RC, Wong-Staal F. A molecular clone of HTLV-III with biological activity. *Nature.* Jul 18-24 1985;316(6025):262-265.
65. Brinchmann JE, Albert J, Vartdal F. Few infected CD4+ T cells but a high proportion of replication-competent provirus copies in asymptomatic human immunodeficiency virus type 1 infection. *J Virol.* Apr 1991;65(4):2019-2023.
66. Lifson JD, Feinberg MB, Reyes GR, et al. Induction of CD4-dependent cell fusion by the HTLV-III/LAV envelope glycoprotein. *Nature.* Oct 23-29 1986;323(6090):725-728.
67. Sodroski J, Goh WC, Rosen C, Campbell K, Haseltine WA. Role of the HTLV-III/LAV envelope in syncytium formation and cytopathicity. *Nature.* Jul 31-Aug 6 1986;322(6078):470-474.
68. Garg H, Blumenthal R. Role of HIV Gp41 mediated fusion/hemifusion in bystander apoptosis. *Cell Mol Life Sci.* Oct 2008;65(20):3134-3144.
69. Plata F, Autran B, Martins LP, et al. AIDS virus-specific cytotoxic T lymphocytes in lung disorders. *Nature.* Jul 23-29 1987;328(6128):348-351.
70. Walker BD, Chakrabarti S, Moss B, et al. HIV-specific cytotoxic T lymphocytes in seropositive individuals. *Nature.* Jul 23-29 1987;328(6128):345-348.
71. Fauci AS. Host factors and the pathogenesis of HIV-induced disease. *Nature.* Dec 12 1996;384(6609):529-534.
72. Kion TA, Hoffmann GW. Anti-HIV and anti-anti-MHC antibodies in alloimmune and autoimmune mice. *Science.* Sep 6 1991;253(5024):1138-1140.
73. Sadeghi HM, Weiss L, Kazatchkine MD, Haeffner-Cavaillon N. Antiretroviral therapy suppresses the constitutive production of interleukin-1 associated with human immunodeficiency virus infection. *J Infect Dis.* Aug 1995;172(2):547-550.
74. Zagury JF, Bernard J, Achour A, et al. Identification of CD4 and major histocompatibility complex functional peptide sites and their homology with oligopeptides from human immunodeficiency virus type 1 glycoprotein gp120: role in AIDS pathogenesis. *Proc Natl Acad Sci U S A.* Aug 15 1993;90(16):7573-7577.
75. Vieillard V, Strominger JL, Debre P. NK cytotoxicity against CD4+ T cells during HIV-1 infection: a gp41 peptide induces the expression of an NKp44 ligand. *Proc Natl Acad Sci U S A.* Aug 2 2005;102(31):10981-10986.
76. Calabrese LH. Autoimmune manifestations of human immunodeficiency virus (HIV) infection. *Clin Lab Med.* Jun 1988;8(2):269-279.
77. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science.* May 24 2002;296(5572):1439-1443.
78. Ter-Grigorov VS, Krifuks O, Liubashevsky E, Nyska A, Trainin Z, Toder V. A new transmissible AIDS-like disease in mice induced by alloimmune stimuli. *Nat Med.* Jan 1997;3(1):37-41.
79. Dalgleish AG, Wilson S, Gompels M, et al. T-cell receptor variable gene products and early HIV-1 infection. *Lancet.* Apr 4 1992;339(8797):824-828.
80. Laurence J, Hodtsev AS, Posnett DN. Superantigen implicated in dependence of HIV-1 replication in T cells on TCR V beta expression. *Nature.* Jul 16 1992;358(6383):255-259.

81. Chowdhury IH, Munakata T, Koyanagi Y, Kobayashi S, Arai S, Yamamoto N. Mycoplasma can enhance HIV replication in vitro: a possible cofactor responsible for the progression of AIDS. *Biochem Biophys Res Commun.* Aug 16 1990;170(3):1365-1370.
82. Chowdhury MI, Munakata T, Koyanagi Y, Arai S, Yamamoto N. Mycoplasma stimulates HIV-1 expression from acutely- and dormantly-infected promonocyte/monoblastoid cell lines. *Arch Virol.* 1994;139(3-4):431-438.
83. Clerici M, Shearer GM. A TH1-->TH2 switch is a critical step in the etiology of HIV infection. *Immunol Today.* Mar 1993;14(3):107-111.
84. Romagnani S, Del Prete G, Manetti R, et al. Role of TH1/TH2 cytokines in HIV infection. *Immunol Rev.* Aug 1994;140:73-92.
85. Graziosi C, Pantaleo G, Gantt KR, et al. Lack of evidence for the dichotomy of TH1 and TH2 predominance in HIV-infected individuals. *Science.* Jul 8 1994;265(5169):248-252.
86. Ameisen JC, Estaquier J, Idziorek T, De Bels F. The relevance of apoptosis to AIDS pathogenesis. *Trends Cell Biol.* Jan 1995;5(1):27-32.
87. Estaquier J, Idziorek T, de Bels F, et al. Programmed cell death and AIDS: significance of T-cell apoptosis in pathogenic and nonpathogenic primate lentiviral infections. *Proc Natl Acad Sci U S A.* Sep 27 1994;91(20):9431-9435.
88. Gougeon ML, Montagnier L. Programmed cell death as a mechanism of CD4 and CD8 T cell deletion in AIDS. Molecular control and effect of highly active anti-retroviral therapy. *Ann N Y Acad Sci.* 1999;887:199-212.
89. Somma F, Tuosto L, Gilardini Montani MS, Di Somma MM, Cundari E, Piccolella E. Engagement of CD4 before TCR triggering regulates both Bax- and Fas (CD95)-mediated apoptosis. *J Immunol.* May 15 2000;164(10):5078-5087.
90. Tateyama M, Oyaizu N, McCloskey TW, Than S, Pahwa S. CD4 T lymphocytes are primed to express Fas ligand by CD4 cross-linking and to contribute to CD8 T-cell apoptosis via Fas/FasL death signaling pathway. *Blood.* Jul 1 2000;96(1):195-202.
91. Hashimoto F, Oyaizu N, Kalyanaraman VS, Pahwa S. Modulation of Bcl-2 protein by CD4 cross-linking: a possible mechanism for lymphocyte apoptosis in human immunodeficiency virus infection and for rescue of apoptosis by interleukin-2. *Blood.* Jul 15 1997;90(2):745-753.
92. Stevenson M. HIV-1 pathogenesis. *Nat Med.* Jul 2003;9(7):853-860.
93. Shacklett BL. Mucosal immunity to HIV: a review of recent literature. *Curr Opin HIV AIDS.* Sep 2008;3(5):541-547.
94. Herbeuval JP, Shearer GM. HIV-1 immunopathogenesis: how good interferon turns bad. *Clin Immunol.* May 2007;123(2):121-128.
95. Mira MT, Alcais A, di Pietrantonio T, et al. Segregation of HLA/TNF region is linked to leprosy clinical spectrum in families displaying mixed leprosy subtypes. *Genes Immun.* Jan 2003;4(1):67-73.
96. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* Sep 13 1996;273(5281):1516-1517.
97. The International HapMap Project. *Nature.* Dec 18 2003;426(6968):789-796.
98. A haplotype map of the human genome. *Nature.* Oct 27 2005;437(7063):1299-1320.
99. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* Oct 18 2007;449(7164):851-861.
100. Howie BN, Carlson CS, Rieder MJ, Nickerson DA. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum Genet.* Aug 2006;120(1):58-68.
101. Schulze TG, Zhang K, Chen YS, Akula N, Sun F, McMahon FJ. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum Mol Genet.* Feb 1 2004;13(3):335-342.

102. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* Jun 2008;40(6):695-701.
103. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet.* Apr 2007;80(4):727-739.
104. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet.* Jan 2008;82(1):100-112.
105. Goldstein DB. Common genetic variation and human traits. *N Engl J Med.* Apr 23 2009;360(17):1696-1698.
106. 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.* Jan 2003;57(1):25-33.
107. Fellay J. Host genetics influences on HIV type-1 disease. *Antivir Ther.* 2009;14(6):731-738.
108. O'Brien SJ, Nelson GW. Human genes that limit AIDS. *Nat Genet.* Jun 2004;36(6):565-574.
109. 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.* Sep 27 1996;273(5283):1856-1862.
110. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature.* Aug 22 1996;382(6593):722-725.
111. Martin MP, Dean M, Smith MW, et al. Genetic acceleration of AIDS progression by a promoter variant of CCR5. *Science.* Dec 4 1998;282(5395):1907-1911.
112. McDermott DH, Zimmerman PA, Guignard F, Kleeberger CA, Leitman SF, Murphy PM. CCR5 promoter polymorphism and HIV-1 disease progression. Multicenter AIDS Cohort Study (MACS). *Lancet.* Sep 12 1998;352(9131):866-870.
113. Smith MW, Dean M, Carrington M, 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 (SFCC), ALIVE Study. *Science.* Aug 15 1997;277(5328):959-965.
114. Winkler CA, Hendel H, Carrington M, et al. Dominant effects of CCR2-CCR5 haplotypes in HIV-1 disease progression. *J Acquir Immune Defic Syndr.* Dec 1 2004;37(4):1534-1538.
115. Hogan CM, Hammer SM. Host determinants in HIV infection and disease. Part 2: genetic factors and implications for antiretroviral therapeutics. *Ann Intern Med.* May 15 2001;134(10):978-996.
116. Ioannidis JP, Rosenberg PS, Goedert JJ, 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.* Nov 6 2001;135(9):782-795.
117. Hendel H, Henon N, Lebuanec H, et al. Distinctive effects of CCR5, CCR2, and SDF1 genetic polymorphisms in AIDS progression. *J Acquir Immune Defic Syndr Hum Retrovirol.* Dec 1 1998;19(4):381-386.
118. An P, Nelson GW, Wang L, et al. Modulating influence on HIV/AIDS by interacting RANTES gene variants. *Proc Natl Acad Sci U S A.* Jul 23 2002;99(15):10002-10007.
119. Modi WS, Goedert JJ, Strathdee S, et al. MCP-1-MCP-3-Eotaxin gene cluster influences HIV-1 transmission. *AIDS.* Nov 7 2003;17(16):2357-2365.

120. Goldschmidt V, Bleiber G, May M, Martinez R, Ortiz M, Telenti A. Role of common human TRIM5alpha variants in HIV-1 disease progression. *Retrovirology.* 2006;3:54.
121. Bleiber G, May M, Martinez R, et al. Use of a combined ex vivo/in vivo population approach for screening of human genes involved in the human immunodeficiency virus type 1 life cycle for variants influencing disease progression. *J Virol.* Oct 2005;79(20):12674-12680.
122. An P, Bleiber G, Duggal P, et al. APOBEC3G genetic variants and their influence on the progression to AIDS. *J Virol.* Oct 2004;78(20):11070-11076.
123. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature.* Aug 8 2002;418(6898):646-650.
124. Carrington M, Nelson GW, Martin MP, et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science.* Mar 12 1999;283(5408):1748-1752.
125. Magierowska M, Theodorou I, Debre P, et al. Combined genotypes of CCR5, CCR2, SDF1, and HLA genes can predict the long-term nonprogressor status in human immunodeficiency virus-1-infected individuals. *Blood.* Feb 1 1999;93(3):936-941.
126. Kaslow RA, Dorak T, Tang JJ. Influence of host genetic variation on susceptibility to HIV type 1 infection. *J Infect Dis.* Feb 1 2005;191 Suppl 1:S68-77.
127. Carrington M, Martin MP, van Bergen J. KIR-HLA intercourse in HIV disease. *Trends Microbiol.* Dec 2008;16(12):620-627.
128. Martin MP, Gao X, Lee JH, et al. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet.* Aug 2002;31(4):429-434.
129. Jennes W, Verheyden S, Demanet C, et al. Cutting edge: resistance to HIV-1 infection among African female sex workers is associated with inhibitory KIR in the absence of their HLA ligands. *J Immunol.* Nov 15 2006;177(10):6588-6592.
130. Diop G, Hirtzig T, Do H, 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.* Nov 2006;60(9):569-577.
131. Do H, Vasilescu A, Diop G, et al. Associations of the IL2Ralpha, IL4Ralpha, IL10Ralpha, and IFN (gamma) R1 cytokine receptor genes with AIDS progression in a French AIDS cohort. *Immunogenetics.* Apr 2006;58(2-3):89-98.
132. Do H, Vasilescu A, Diop G, et al. Exhaustive genotyping of the CEM15 (APOBEC3G) gene and absence of association with AIDS progression in a French cohort. *J Infect Dis.* Jan 15 2005;191(2):159-163.
133. Flores-Villanueva PO, Yunis EJ, Delgado JC, et al. Control of HIV-1 viremia and protection from AIDS are associated with HLA-Bw4 homozygosity. *Proc Natl Acad Sci U S A.* Apr 24 2001;98(9):5140-5145.
134. 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.* Jun 1 1999;162(11):6942-6946.
135. 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.* Mar 29 1997;349(9056):922-923.
136. Vasilescu A, Heath SC, Ivanova R, et al. Genomic analysis of Th1-Th2 cytokine genes in an AIDS cohort: identification of IL4 and IL10 haplotypes associated with the disease progression. *Genes Immun.* Sep 2003;4(6):441-449.
137. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. *Science.* Aug 17 2007;317(5840):944-947.

138. de Bakker PI, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet.* Oct 2006;38(10):1166-1172.
139. Kulski JK, Dawkins RL. The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16. *Immunogenetics.* May 1999;49(5):404-412.
140. Vernet C, Ribouchon MT, Chimini G, Jouanolle AM, Sidibe I, Pontarotti P. A novel coding sequence belonging to a new multicopy gene family mapping within the human MHC class I region. *Immunogenetics.* 1993;38(1):47-53.
141. Fellay J, Ge D, Shianna KV, et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* Dec 2009;5(12):e1000791.
142. 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(12):e3907.
143. White RA, Pasztor LM, Richardson PM, Zon LI. 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(3-4):272-275.
144. Suske G, Bruford E, Philipsen S. Mammalian SP/KLF transcription factors: bring in the family. *Genomics.* May 2005;85(5):551-556.
145. Herbeck JT, Gottlieb GS, Winkler CA, 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.* Feb 15 2010;201(4):618-626.
146. Wang L, Zhu J, Shan S, et al. Repression of interferon-gamma expression in T cells by Prospero-related homeobox protein. *Cell Res.* Sep 2008;18(9):911-920.
147. Pelak K, Goldstein DB, Walley NM, et al. Host determinants of HIV-1 control in African Americans. *J Infect Dis.* Apr 15 2010;201(8):1141-1149.
148. Shrestha S, Aissani B, Song W, Wilson CM, Kaslow RA, Tang J. Host genetics and HIV-1 viral load set-point in African-Americans. *AIDS.* Mar 27 2009;23(6):673-677.
149. Limou S, Coulonges C, Herbeck JT, et al. Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J Infect Dis.* Sep 15 2010;202(6):908-915.
150. 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.* Jan 2 2009;23(1):19-28.
151. 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.* Jan 26 2007;3(1):e19.
152. Deng HK, Unutmaz D, KewalRamani VN, Littman DR. Expression cloning of new receptors used by simian and human immunodeficiency viruses. *Nature.* Jul 17 1997;388(6639):296-300.
153. 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 Invest.* Mar 2001;107(5):595-601.
154. Landro L, Damas JK, Halvorsen B, et al. CXCL16 in HIV infection - a link between inflammation and viral replication. *Eur J Clin Invest.* Nov 2009;39(11):1017-1024.
155. 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.* Jul 1 2008;181(1):81-91.
156. Hercberg S, Galan P, Preziosi P, 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 Mineraux AntioXydants Study. *Int J Vitam Nutr Res.* 1998;68(1):3-20.

157. 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.* Aug 1996;44(4):373-375.
158. Crow JF. Eighty years ago: the beginnings of population genetics. *Genetics.* Jul 1988;119(3):473-476.
159. Hardy GH. Mendelian Proportions in a Mixed Population. *Science.* Jul 10 1908;28(706):49-50.
160. Weinberg W. On the demonstration of heredity in man. *Naturkunde in Wurttemberg, Stuttgart* 1908;64:368-382.
161. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* May 2005;76(5):887-893.
162. 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.* Sep 2007;81(3):559-575.
163. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy Stat Soc* 1995;Ser.B:289-300.
164. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med.* Jul 1990;9(7):811-818.
165. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ.* Apr 18 1998;316(7139):1236-1238.
166. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* Aug 5 2003;100(16):9440-9445.
167. Fisher R. *Statistical Methods for Research Workers*. Edinburgh; 1932.
168. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* Jun 2000;155(2):945-959.
169. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* Dec 1999;55(4):997-1004.
170. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* Aug 2006;38(8):904-909.
171. Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics.* 2008;9:540.
172. Ge D, Zhang K, Need AC, et al. WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.* Apr 2008;18(4):640-643.
173. Stranger BE, Forrest MS, Clark AG, et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet.* Dec 2005;1(6):e78.
174. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet.* Oct 2007;39(10):1202-1207.
175. Brass AL, Dykxhoorn DM, Benita Y, et al. Identification of host proteins required for HIV infection through a functional genomic screen. *Science.* Feb 15 2008;319(5865):921-926.
176. 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.* Feb 1 2009;199(3):419-426.
177. Stephens HA. HIV-1 diversity versus HLA class I polymorphism. *Trends Immunol.* Jan 2005;26(1):41-47.
178. Reed R, Hurt E. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell.* Feb 22 2002;108(4):523-531.
179. Allcock RJ, Williams JH, Price P. The central MHC gene, BAT1, may encode a protein that down-regulates cytokine production. *Genes Cells.* May 2001;6(5):487-494.

180. Rennert PD, Browning JL, Mebius R, Mackay F, Hochman PS. Surface lymphotoxin alpha/beta complex is required for the development of peripheral lymphoid organs. *J Exp Med.* Nov 1 1996;184(5):1999-2006.
181. Kedzierska K, Crowe SM. Cytokines and HIV-1: interactions and clinical implications. *Antivir Chem Chemother.* May 2001;12(3):133-150.
182. 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.* Oct 15 2009;200(8):1194-1201.
183. Furumatsu T, Tsuda M, Taniguchi N, Tajima Y, Asahara H. Smad3 induces chondrogenesis through the activation of SOX9 via CREB-binding protein/p300 recruitment. *J Biol Chem.* Mar 4 2005;280(9):8343-8350.
184. Ikeda T, Kawaguchi H, Kamekura S, et al. Distinct roles of Sox5, Sox6, and Sox9 in different stages of chondrogenic differentiation. *J Bone Miner Metab.* 2005;23(5):337-340.
185. Kiefer HL, Hanley TM, Marcello JE, Karthik AG, Viglianti GA. Retinoic acid inhibition of chromatin remodeling at the human immunodeficiency virus type 1 promoter. Uncoupling of histone acetylation and chromatin remodeling. *J Biol Chem.* Oct 15 2004;279(42):43604-43613.
186. Maeda Y, Yamaguchi T, Hijikata Y, et al. All-trans retinoic acid attacks reverse transcriptase resulting in inhibition of HIV-1 replication. *Hematology.* Jun 2007;12(3):263-266.
187. Invernizzi CF, Xie B, Richard S, Wainberg MA. PRMT6 diminishes HIV-1 Rev binding to and export of viral RNA. *Retrovirology.* 2006;3:93.
188. Xie B, Invernizzi CF, Richard S, Wainberg MA. Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat Interactions with both cyclin T1 and the Tat transactivation region. *J Virol.* Apr 2007;81(8):4226-4234.
189. Sgarra R, Lee J, Tessari MA, et al. The AT-hook of the chromatin architectural transcription factor high mobility group A1a is arginine-methylated by protein arginine methyltransferase 6. *J Biol Chem.* Feb 17 2006;281(7):3764-3772.
190. Li L, Yoder K, Hansen MS, Olvera J, Miller MD, Bushman FD. Retroviral cDNA integration: stimulation by HMG I family proteins. *J Virol.* Dec 2000;74(23):10965-10974.
191. Charng MJ, Zhang D, Kinnunen P, Schneider MD. A novel protein distinguishes between quiescent and activated forms of the type I transforming growth factor beta receptor. *J Biol Chem.* Apr 17 1998;273(16):9365-9368.
192. Brenchley JM, Paiardini M, Knox KS, et al. Differential Th17 CD4 T-cell depletion in pathogenic and nonpathogenic lentiviral infections. *Blood.* Oct 1 2008;112(7):2826-2835.
193. Favre D, Lederer S, Kanwar B, et al. Critical loss of the balance between Th17 and T regulatory cell populations in pathogenic SIV infection. *PLoS Pathog.* Feb 2009;5(2):e1000295.
194. Fazekas de St Groth B, Landay AL. Regulatory T cells in HIV infection: pathogenic or protective participants in the immune response? *AIDS.* Mar 30 2008;22(6):671-683.
195. Manel N, Unutmaz D, Littman DR. The differentiation of human T(H)-17 cells requires transforming growth factor-beta and induction of the nuclear receptor RORgammat. *Nat Immunol.* Jun 2008;9(6):641-649.
196. Rollason R, Korolchuk V, Hamilton C, Jepson M, Banting G. A CD317/tetherin-RICH2 complex plays a critical role in the organization of the subapical actin cytoskeleton in polarized epithelial cells. *J Cell Biol.* Mar 9 2009;184(5):721-736.

197. Tokarev A, Skasko M, Fitzpatrick K, Guatelli J. Antiviral activity of the interferon-induced cellular protein BST-2/tetherin. *AIDS Res Hum Retroviruses.* Dec 2009;25(12):1197-1210.
198. Catano G, Kulkarni H, He W, 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(11):e3636.
199. 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.* Dec 2009;10(8):673-677.
200. Beecham GW, Martin ER, Li YJ, et al. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. *Am J Hum Genet.* Jan 2009;84(1):35-43.
201. Clark AG. The role of haplotypes in candidate gene studies. *Genet Epidemiol.* Dec 2004;27(4):321-333.
202. Frankel WN, Schork NJ. Who's afraid of epistasis? *Nat Genet.* Dec 1996;14(4):371-373.
203. Konig R, Zhou Y, Elleder D, et al. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell.* Oct 3 2008;135(1):49-60.
204. Zhou H, Xu M, Huang Q, et al. Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe.* Nov 13 2008;4(5):495-504.
205. Yeung ML, Houzet L, Yedavalli VS, Jeang KT. A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J Biol Chem.* Jul 17 2009;284(29):19463-19473.
206. Chan EY, Sutton JN, Jacobs JM, Bondarenko A, Smith RD, Katze MG. Dynamic host energetics and cytoskeletal proteomes in human immunodeficiency virus type 1-infected human primary CD4 cells: analysis by multiplexed label-free mass spectrometry. *J Virol.* Sep 2009;83(18):9283-9295.
207. Coiras M, Camafeita E, Urena T, et al. Modifications in the human T cell proteome induced by intracellular HIV-1 Tat protein expression. *Proteomics.* Apr 2006;6 Suppl 1:S63-73.
208. Molina L, Grimaldi M, Robert-Hebmann V, et al. Proteomic analysis of the cellular responses induced in uninfected immune cells by cell-expressed X4 HIV-1 envelope. *Proteomics.* Sep 2007;7(17):3116-3130.
209. Ringrose JH, Jeeninga RE, Berkhout B, Speijer D. Proteomic studies reveal coordinated changes in T-cell expression patterns upon infection with human immunodeficiency virus type 1. *J Virol.* May 2008;82(9):4320-4330.
210. Nguyen DG, Yin H, Zhou Y, Wolff KC, Kuhen KL, Caldwell JS. Identification of novel therapeutic targets for HIV infection through functional genomic cDNA screening. *Virology.* May 25 2007;362(1):16-25.
211. Nathans R, Chu CY, Serquina AK, Lu CC, Cao H, Rana TM. Cellular microRNA and P bodies modulate host-HIV-1 interactions. *Mol Cell.* Jun 26 2009;34(6):696-709.
212. Imbeault M, Ouellet M, Tremblay MJ. Microarray study reveals that HIV-1 induces rapid type-I interferon-dependent p53 mRNA up-regulation in human primary CD4+ T cells. *Retrovirology.* 2009;6:5.
213. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* Nov 2010;11(11):773-785.
214. Bonetta L. Whole-genome sequencing breaks the cost barrier. *Cell.* Jun 11 2010;141(6):917-919.
215. Bonnefond A, Durand E, Sand O, et al. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS One.* 2010;5(10):e13630.

216. 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.
217. Froguel P, Blakemore AI. The power of the extreme in elucidating obesity. *N Engl J Med.* Aug 28 2008;359(9):891-893.
218. Zhang G, Nebert DW, Chakraborty R, Jin L. Statistical power of association using the extreme discordant phenotype design. *Pharmacogenet Genomics.* Jun 2006;16(6):401-413.
219. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science.* Nov 4 2010.

Liste des publications

Limou S, Coulonges C, Foglio M, Heath S, Diop G, **Le Clerc S**, Hirtzig T, Spadoni JL, Therwath A, Lambeau G, Gut I, Zagury JF. Exploration of associations between phospholipase A2 gene family polymorphisms and AIDS progression using the SNPlex method. *Biomed Pharmacother.* 2008 ; 62(1):31-40.

Dalmasso C, Carpentier W, Meyer L, Rouzioux C, Goujard C, Chaix ML, Lambotte O, Avettand-Fenoel V, **Le Clerc S**, de Senneville LD, Deveau C, Boufassa F, Debre P, Delfraissy JF, Broet P, Theodorou I. 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(12): e3907.

Le Clerc S*, Limou S*, Coulonges C, Carpentier W, Dina C, Delaneau O, Labib T, Taing L, Sladek R, Deveau C, Ratsimandresy R, Montes M, Spadoni JL, Lelièvre JD, Lévy Y, Therwath A, Schächter F, Matsuda F, Gut I, Froguel P, Delfraissy JF, Hercberg S, Zagury JF; ANRS Genomic Group. 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(3): 419-26. (*) equal contribution.

Le Clerc S*, Limou S*, Coulonges C, Carpentier W, Dina C, Taing L, Delaneau O, Labib T, Sladek R; ANRS Genomic Group, Deveau C, Guillemain H, Ratsimandresy R, Montes M, Spadoni JL, Therwath A, Schächter F, Matsuda F, Gut I, Lelièvre JD, Lévy Y, Froguel P, Delfraissy JF, Hercberg S, Zagury JF. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J Infect Dis.* 2009 ; 200(8):1194-201. (*) equal contribution

Limou S, Coulonges C, Herbeck JT, van Manen D, An P, **Le Clerc S**, Delaneau O, Diop G, Taing L, Montes M, van't Wout AB, Gottlieb GS, Therwath A, Rouzioux C, Delfraissy JF, Lelièvre JD, Lévy Y, Hercberg S, Dina C, Phair J, Donfield S, Goedert JJ, Buchbinder S, Estaquier J, Schächter F, Gut I, Froguel P, Mullins JI, Schuitemaker H, Winkler C, Zagury JF. Multi-Cohort Genetic Association Study Reveals CXCR6 as a New Chemokine Receptor Involved in AIDS Long-Term Non-Progression. *J Infect Dis.* 2010 ; 202(6): 908-15.

Le Clerc S, Coulonges C, Delaneau O, Van Manen D, Herbeck JT, Limou S, An P, Martinson JJ, Spadoni JL, Therwath A, Veldink JH, van den Berg LH, Taing L, Labib T, Mellak S, Montes M, Delfraissy JF, Schächter F, Winkler C, Froguel P, Mullins JI, Schuitemaker H, Zagury JF. Screening Low Frequency SNPs from Genome Wide Association Study Reveals a New Risk Allele for Progression to AIDS. *JAIDS*. 2010: in press.

Liste des communications orales

Le Clerc S and Limou S. Large-scale approaches in HIV infection studies. *Genomics and AIDS pathogenesis workshop*. Ermenonville, 2008.

Le Clerc S, Limou S and Zagury JF. Genomewide association study of an AIDS-nonprogression cohort emphasizes the major role of HLA genes in non progression to AIDS. *Conference on Retroviruses and Opportunistic Infections (CROI)*, Montreal, 2009.

Posters

Le Clerc S, Limou S, Coulonges C, Lelièvre JD, Lévy Y, Delfraissy JF, Hercberg S, Zagury JF. Genomewide association study of an AIDS-nonprogression cohort emphasizes the major role of HLA genes in non progression to AIDS. *Conference on Retroviruses and Opportunistic Infections (CROI)*, Montreal, 2009.

Limou S, **Le Clerc S**, van Manen D, Herbeck J, Lévy Y, Estaquier J, Mullins J, Schuitemaker H, Winkler C, Zagury JF. Multi-cohort Identification of a New Chemokine Receptor Risk Allele Involved in AIDS Long-term Non-progression. *Conference on Retroviruses and Opportunistic Infections (CROI)*, San Francisco, 2010.

Sigrid LE CLERC

Analyses ‘génome entier’ de la cohorte GRIV de patients à profil extrême du SIDA

Résumé

Après 25 ans de recherche intensive, aucun vaccin ou traitement définitif contre le SIDA n'existe, et les mécanismes moléculaires de pathogenèse de l'infection VIH-1 ne sont pas clairement élucidés. Les avancées technologiques permettent de comparer des sujets malades avec des sujets contrôles sur tout le génome. Il est ainsi possible d’identifier sans a priori des gènes potentiellement impliqués dans le développement de la maladie avec pour conséquence le développement rationnel de nouvelles stratégies diagnostiques ou thérapeutiques.

Durant ma thèse, j’ai réalisé deux études d’association ‘génome entier’ dans le SIDA, en comparant les 275 non-progresseurs à long terme ou les 85 progresseurs rapides de la cohorte GRIV avec une cohorte de contrôles séronégatifs. J’ai réalisé une troisième analyse en exploitant les données issues de trois études ‘génome entier’ internationales dont la nôtre (France, Pays-Bas, USA), ciblant plus particulièrement les SNPs de fréquence faible (fréquence de l’allèle mineur, MAF<5%).

Ces approches ‘génome entier’ ont réaffirmé le rôle central du HLA dans la progression vers le SIDA, mais aussi dévoilé de nouveaux gènes candidats très pertinents donnant une nouvelle lumière sur les mécanismes moléculaires de la maladie.

Résumé en anglais

After 25 years of intensive research, no vaccine or cure exists against AIDS, and the molecular mechanisms of pathogenesis of HIV-1 infection are not clearly understood. Technological progress has made possible to compare cases versus controls over the whole genome. It is thus possible to identify genes potentially involved in disease development with no a priori, and consequently develop rationally new diagnostic or therapeutic strategies.

During my PhD, I have completed two genome-wide association studies (GWAS) in AIDS, comparing 275 long term non-progressors or the 85 rapid progressors from the GRIV cohort with a cohort of seronegative controls. I have also completed a third analysis exploiting data from three international GWAS including ours (France, Netherlands, USA), targeting particularly low frequency SNPs (minor allele frequency, MAF <5%).

These GWAS approaches have reaffirmed the central role of HLA for progression towards AIDS, but also revealed new relevant candidate genes, shedding a new light on the molecular mechanisms of disease progression.