Imputation HLA et analyse génomique de la coinfection VIH/VHC

Marc Jeanmougin

le cnam

École doctorale Sciences des Métiers de l'Ingénieur (Paris)

Génomique, Bioinformatique et Applications

THÈSE DE DOCTORAT

présentée par : **Marc JEANMOUGIN**

soutenue le : **21 décembre 2017**

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

Discipline : **Biochimie et biologie moléculaire** *Spécialité* : **Bioinformatique**

Imputation HLA et Analyse génomique de coinfection VIH/VHC

THÈSE dirigée par

Pr. ZAGURY Jean-François — *Professeur du CNAM, Paris*

RAPPORTEURS

Dr. DELANEAU Olivier — *Faculté de Médecine, Université de Genève*

Dr. GUINOT Christiane — *Laboratoire d'informatique, Université de Tours*

PRÉSIDENT

Pr. BAR-HEN Avner — *Professeur du CNAM, Paris*

EXAMINATRICES

Dr. COBAT Aurélie — *Génétique Humaine des maladies infectieuses (UMR 1163, INSERM)*

Dr. DOMINGUEZ Stéphanie — *Service Immunologie Clinique et Maladies Infectieuses, Hôpital Henri-Mondor*

Résumé

La génomique d'association cherche à mettre en évidence des liens entre le génome et des traits ou phénotypes, notamment dans le contexte de maladies. Aujourd'hui, les études les plus courantes en génomique d'association sont les études dites « génome entier », qui analysent autant de variants du génomes (SNPs) que possible, sans préjuger de leur fonction biologique. Cependant, les méthodes de génotypage utilisées pour ces études ne permettent pas toujours d'obtenir des informations précises dans une région hypervariable comme la région HLA (Antigènes Leukocytes Humains), qui joue un rôle crucial dans l'immunité. Les variants génétiques de ces régions sont alors souvent prédits par des approches bioinformatiques.

J'ai développé au cours de ma thèse un nouvel outil, HLA-Check, permettant d'évaluer, à partir de génotypes obtenus par puce de génotypage, la plausibilité de données d'allèles HLA d'un même individu, et démontré que cette technique permettait d'identifier plus précisément les individus dont les allèles HLA avaient été mal caractérisés afin de les retyper ou de les écarter de l'étude. Un article documentant cet outil a été publié dans BMC Bioinformatics (BioMed Central).

J'ai également effectué une étude d'association génome entier sur le déclenchement de la cirrhose chez les patients co-infectés par le VIH (virus de l'immunodéficience humaine) et le VHC (virus de l'hépatite C). L'infection conjointe par ces deux virus est fréquente en raison de modes d'infection similaires, et en outre, l'infection par le VIH stimule l'activité du VHC et accélère la fibrose du foie puis sa cirrhose, causant la mort des patients co-infectés.

Mon étude porte sur 306 patients co-infectés issus de la cohorte ANRS CO-13

HEPAVIH. J'ai pu mettre en évidence trois signaux associés avec le déclenchement de la cirrhose, dont deux qui ont un lien pertinent avec les maladies hépatiques (*CTNND2* et *MIR7-3HG*). L'identification de ces nouveaux variants devrait permettre une meilleure compréhension des mécanismes moléculaires de la cirrhose, et contribuer au développement de nouvelles stratégies diagnostiques ou thérapeutiques. L'article documentant cette étude est en cours de publication.

Mots-clés : *GWAS, SNP, VIH/VHC, cirrhose, HLA, CMH, Imputation*

Abstract in English

Association genomics aims at finding links between the genome and some traits or illnesses. Today, the most frequent studies in this field are genome wide association studies (GWAS), which analyze as many genome variants (mainly Single Nucleotide Polymorphisms) as possible, without any a priori on their biological function. However, genotyping methods used in these studies may be insufficient to get reliable information in higly variable regions such as the HLA which plays a crucial role in immunity, and the genetic variants of such regions are often predicted using bioinformatics approaches.

During my PhD, I have created a new tool, HLA-Check, that allows to rate the plausibility of HLA (Human Leukocyte Antigen) alleles from the genotypes obtained from genotyping chips. I also assessed its performances and showed that it was able to point out individuals with a wrong HLA typing, in order to retype them or remove them from the study. An article documenting this tool was published in BMC Bioinformatics (BioMed Central).

I have also performed a genome-wide association study on cirrhosis outbreak in individuals coinfected with HIV (human immunodeficiency virus) and HCV (hepatitis C virus). Because of similar infection routes (blood-related), co-infection with those two viruses are frequent, and the infection by HIV enhances HCV activity and increases liver fibrosis leading to cirrhosis and death of co-infected patients. Our study has dealt with 306 co-infected patients from the ANRS CO-13 HEPAVIH cohort. I could point out three statistically significant signals, two of them being highly relevant for their involvement in liver diseases (*CTNND2* and *MIR7-3HG*). The identification of these new variants should lead to a better understanding of the molecular mechanisms involved in cirrhosis, and should contribute to the rational developement of new diagnostic or therapuetic

strategies. A publication is under way.

Keywords : *GWAS, SNP, HIV/HCV, cirrhosis, HLA, MHC, Imputation*

Remerciements

Mes premiers remerciements vont évidemment au Pr. Jean-François Zagury, mon directeur de thèse. Pour m'avoir initié à la bioinformatique, pour m'avoir accueilli au sein de son laboratoire et m'avoir encadré pendant plus de trois ans en me laissant une grande liberté, merci.

Un grand merci également au professeur Bar-Hen pour avoir accepté de présider mon jury, aux docteurs Guinot et Delaneau pour avoir bien voulu endosser le rôle de rapporteur du manuscrit en dépit de délais très courts, ainsi qu'aux docteurs Dominguez et Cobat pour avoir accepté de faire partie du jury.

Un remerciement spécial à Sigrid, Cédric et Josselin, pour leurs conseils et enseignements sur la biologie, la bioinformatique, ou les statistiques, sans lesquels rien n'aurait été possible. Ou pas grand-chose.

Merci encore à l'ensemble de mes collègues, grâce à qui l'ambiance est bien plus qu'une simple ambiance de travail. Par ordre approximatif d'ancienneté au Cnam, Christiane [1], Matthieu [2], Toufik [3], Hervé[2], Nathalie [4], Vincent [5], Damien, Charly, Anita, Florent, Manon, Daniela, Jérémy, Vincent [6], Célia, Chloé, Vincent [7] [8], Benjamin, Max, Solène, Émilie, et les collègues de Cochin.

1. qui gère avec brio le côté administratif du laboratoire
2. La révolution ne devrait plus tarder, camarade
3. un peu le sage du laboratoire
4. Qui gagne toujours les *blind tests*
5. le statisticien
6. le stagiaire
7. le matheux
8. sérieusement, combien il y a de Vincent ?

Et enfin, un immense merci à ma famille, toujours là pour moi [9], et à tou·te·s mes ami·e·s, les Toulousains [10], les 09 [12], les grotas [13], #ulminfo [15], les inclassables [16], les auteurs de webcomics [18], les producteurs de chocolat noir [20] et de manière générale toutes les personnes sympathiques avec lesquelles j'ai eu l'occasion d'interagir et qui ont rendu ces dernières années "plutôt cool dans l'ensemble" [21].

9. Même quand je donne beaucoup trop rarement de nouvelles
10. que je vois beaucoup trop peu, Arthur, Thibault, Vincent [11], Rémi ...
11. disons « de Toulouse »
12. Yoann, Ludovic, Antoine, Hang, Fabrice, Floriane, Pablo, *et al.*
13. Ted, Pablo (le même), Paul, Béné, Amiel, Marine, Aurélien [14], Maud, Odile, My-Hai, Nicolas, Valentin, Léa, Louis, Marion, Maxime, *et al.*
14. « Aurélien » ?
15. iXce, olasd, les conscrits...
16. Tsampa, Sabrina, ... [17]
17. L'avantage d'avoir une catégorie « Autres, préciser », c'est que si vous ne vous reconnaissez pas dans les catégories précédentes, et que vous pensez que je vous ai oublié, vous pouvez vous rajouter dans cette catégorie en vous comptant dans les "...". Oh, sauf si vous préférez vous placer dans la catégorie finale, c'est vous qui voyez.
18. XKCD, SMBC, PhD Comics, Existential Comics, Dilbert, et quelques autres [19]
19. Oui, j'ai passé un temps infini à lire des webcomics. Non, je ne regrette rien.
20. Lindt en particulier pour son rapport qualité/prix compatible avec un budget étudiant
42. Je me demande si les gens lisent toutes les footnotes dans l'ordre, ou si celle-ci ne sera jamais lue par personne
21. Je sais, ça fait beaucoup de notes de bas de page. J'aime bien les notes de bas de page [22].
22. Mais qui n'aime pas les notes de bas de page [23] ?
23. Ne vous inquiétez pas, il y en a moins dans la suite du manuscrit.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP Analyse en composantes principales

ADN Acide DésoxyriboNucléique

ANRS France REcherche Nord&Sud Sida-hiv Hépatites

ARN Acide RiboNucléique

ARNm ARN messager

Cnam Conservatoire National des Arts et Métiers

CNV Copy Number Variations

GBA Laboratoire Génomique, Bioinformatique et Applications

HLA Human Leukocyte Antigen

Indel Insertions et délétions

LD Déséquilibre de liaison

MAF Fréquence de l'Allèle Mineur

MHC Complexe Majeur d'Histocompatibilité

NGS Next Generation Sequencing

ONU Organisation des Nations Unies

PCR Polymerase chain reaction

SIDA Syndrome d'ImmunoDéficience Acquise

SNP Single Nucleotide Polymorphism

SSO Sequence-specific oligonucleotide

SSP Sequence-specific primer

VHC Virus de l'Hépatite C

VIH Virus de l'Immunodéficience Humaine

Première partie

Introduction

Chapitre 1

Génétique et maladies

La biologie moléculaire est une branche de la biologie qui a pour but de comprendre le fonctionnement des organismes au travers de l'action de ses molécules, notamment l'ADN et les protéines. Au-delà de cette compréhension, de nombreuses techniques expérimentales ont été développées dans ce contexte pour travailler directement sur ces molécules, notamment toutes les techniques de génie génétique.

La génétique, plus précisément, est la science qui étudie l'hérédité et les gènes, en d'autres termes la partie *héritée* du fonctionnement de l'organisme. On parle de caractères héréditaires, transmis d'un être vivant à ses descendants, habituellement par le biais de la reproduction.

Le support moléculaire de l'hérédité est la molécule d'ADN, que l'on retrouve dans le noyau de toutes les cellules qui composent un organisme.

1.1 Le dogme central en biologie

Le dogme central postule la manière dont l'information génétique se transmet et se transfère, et peut se résumer ainsi : « L'ADN dirige sa propre réplication en ADN identique, ainsi que sa transcription en ARN, pouvant ou non être traduit en protéines. »

Si cette théorie a ses limites et ne permet pas d'expliquer tous les processus en jeu dans la production d'une protéine, elle permet néanmoins de poser un premier cadre explicatif sur les processus observés dans les organismes vivants.

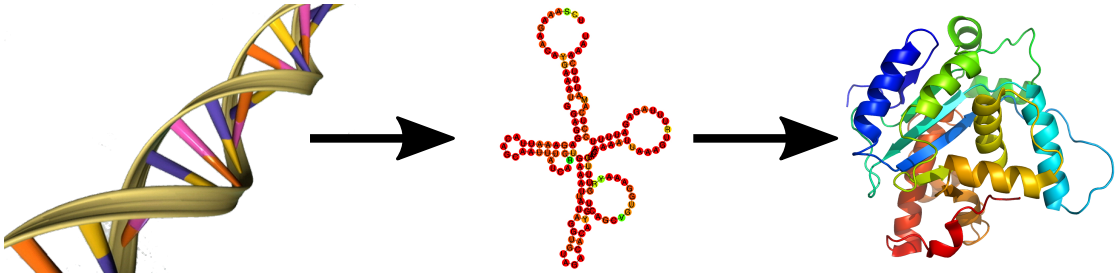

FIGURE 1.1 – Le dogme central
L'ADN est transcrit en ARN, qui est traduit en protéine.

1.1.1 L'ADN, base de la vie cellulaire

L'acide désoxyribonucléique, ou ADN, est une macromolécule à double brin qui permet de stocker l'information génétique de manière stable. Sa molécule se compose d'une succession de nucléotides, formés d'oses (ou sucres), de groupes phosphate, et d'une de quatre bases azotées possibles : Adénine, Thymine, Guanine et Cytosine, abrégées A,T,G,C. Les chromosomes qui contiennent l'information génétique des organismes sont faits de deux brins d'ADN antiparallèles : les bases de chaque brin sont en en vis-à-vis, l'adénine se situe toujours face à la thymine et la guanine face à la cytosine. Le tout forme une double hélice[WC+53] qui confère au complexe formé une grande stabilité. Lors de la division cellulaire, on appelle *réplication* l'opération qui consiste à dupliquer l'ADN.

Le génome humain comporte 23 paires de molécules d'ADN, organisés en structures dénommées *chromosomes*. L'une de ces paires est composée des chromosomes *sexuels* : XX ou XY. Les 22 autres paires sont toujours composées similairement avec notamment les mêmes gènes aux mêmes positions sur le chromosome. Les paires de chromosomes non sexuels (aussi appelés autosomes) sont numérotés par ordre de taille décroissant.

1.1.2 De l'ADN à la protéine

Transcription

L'ARN polymérase est une enzyme responsable de la transcription de parties de l'ADN, appelées *gènes* et représentant environ 1.5% de l'ADN, en acide ribonucléique (ou ARN). Schématiquement, on peut expliquer ce processus de la manière suivante : l'ARN

FIGURE 1.2 – Caryotype humain
observé au microscope, après identification et classement des chromosomes ([Wik17])

polymérase se fixe sur une séquence particulière de l'ADN en amont du gène, appelée le *promoteur*, puis parcourt la séquence d'ADN en construisant incrémentalement la séquence complémentaire au brin lu. Une différence notable entre ADN et ARN est que la thymine est remplacée dans l'ARN par une base structuralement très proche, l'uracile.

L'ARN ainsi formé subit ensuite des modifications de structure importantes, notamment l'ajout de la coiffe (*capping*), l'épisssage (*splicing*) et de polyadénylation. La phase d'*épissage* consiste à séparer les fragments d'ARN entre les brins qui seront effectivement traduits en protéines provenant des séquences d'ADN qu'on qualifiera d'*exons* (pour les parties *ex*primées), des autres, qualifiés d'*introns*. Pour certains gènes, il peut exister des épissages alternatifs, c'est-à-dire plusieurs possibilités de découpe du gène en exons et introns, ce qui résulte en des ARN différents après épissage.

Traduction

Après export du noyau des cellules vers leur cytoplasme, les ARN dits *messagers* (ARNm) sont traduits en protéines au niveau de structures moléculaires complexes, faites d'ARN dit ribosomique, et de protéines. Chaque triplet de nucléotides (ou codon) de l'ARNm code pour un acide aminé (voir table 1.1 le *code génétique*), et l'assemblage des acides aminés forme la séquence d'une protéine, aussi appelée structure primaire.

La protéine peut ensuite se replier ou se combiner avec d'autres protéines pour jouer son rôle fonctionnel dans la cellule ou l'organisme. On recense à ce jour chez l'humain

TABLE 1.1 – Le code génétique

1e base	2e base				3e base
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

plus de 20 000 gènes codant pour une protéine[Con+04].

Il existe d'autres types d'ARN, dénommés ARN *non codant*, qui ne sont pas destinés à être traduits en protéines : par exemple, l'ARN ribosomiques qui forme en grande partie les ribosomes[Whi+90], l'ARN de transfert[LE97] responsable de la traduction du code génétique, ou encore les micro-ARN[Gri+06] qui régulent l'expression de certains gènes.

1.2 Les polymorphismes génétiques, source de diversité

1.2.1 Causes de la diversité génétique

La réplication de l'ADN, est une opération à très faible taux d'erreur. Cependant, ce taux étant non-nul, et que la réplication concerne l'ensemble des trois milliards de paires de bases sur deux chromosomes à chaque division cellulaire (il y a environ 2 trillions de divisions cellulaires dans un corps humain par jour), on observe l'existence de variations génétiques (aussi appelées *polymorphismes*).

Des facteurs extérieurs, comme les radiations[RRK58] ou l'infection par un rétrovirus (virus retranscrivant son ARN dans le génome de la cellule infectée), peuvent également causer de telles variations génétiques.

Lorsqu'ils concernent les cellules germinales, ces polymorphismes sont susceptibles de se transmettre à la descendance, et on obtient ainsi une *variabilité génétique* au sein des espèces. Ces variations génétiques sont souvent sans impact sur les traits et phénotypes, mais l'*héritabilité* des caractères résulte de certains de ces polymorphismes, en particulier lorsqu'elle affecte des gènes : en effet, comme on peut le voir dans la table 1.1, une différence d'un nucléotide peut engendrer une différence sur la protéine produite.

L'ensemble des mutations aléatoires se produisant sur le génome des individus et qui engendre de la diversité au sein des populations est appelé la *dérive génétique*. Cette évolution qui produit des effets phénotypiques profitables ou non à l'espèce est complétée par la *sélection naturelle* : les mutations qui apportent un caractère profitable à l'espèce auront davantage de chances d'être transmises d'une génération à l'autre et de se répandre dans la population, que celles qui désavantagent l'individu dans son environnement.

Méiose et *cross-overs*

La méiose est le processus de divisions cellulaires qui conduit à la formation des gamètes (cellules sexuées) à partir des cellules dites *germinales*, contenant l'information génétique qui sera transmise à la descendance.

Sa particularité est de comporter deux phases de divisions successives du noyau. Les chromosomes ne sont dupliqués que lors de la première phase, ce qui aboutit à la création de cellules ne possédant que la moitié de l'information génétique, avec un unique chromosome issu de chaque paire d'autosomes d'origine paternelle et maternelle.

Les étapes de la méiose peuvent notamment comporter un brassage *intrachromosomique* de cross-over au sein des chromosomes paternels et maternels avant le brassage *interchromosomique* qui répartira les paires de chromosomes dans les gamètes (voir figure 1.3).

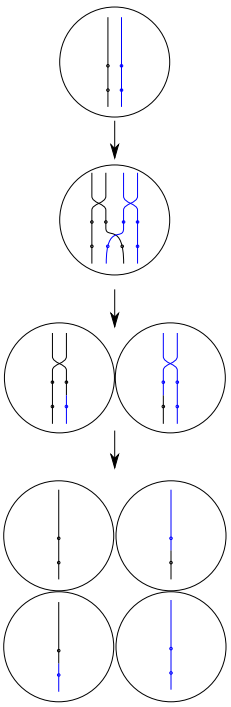

FIGURE 1.3 – Méiose avec brassage intrachromosomique
Des *cross-over* peuvent se former après la duplication de l'ADN, puis les deux phases de division du noyau répartissent l'information dans les gamètes.

Nomenclature

On appellera *allèles* les polymorphismes existants d'un même *locus* ou gène au sein de la population, *génotype* la combinaison des deux allèles d'un locus ou gène chez un individu en particulier (combinaison qui correspond aux deux chromosomes d'origine paternelle et maternelle), et on dira qu'un individu est *hétérozygote* s'il possède deux allèles différents pour un locus, ou *homozygote* sinon.

1.2.2 Les polymorphismes génétiques

En 1959, quelques années après la découverte de la structure de la molécule d'ADN, Lejeune découvre, en observant les caryotypes liés à la trisomie 21[LTG59b], que la structure de celle-ci est susceptible de connaître des variations, en l'occurence ici avec la présence d'un troisième chromosome. Depuis, de nombreux autres types de polymorphismes ont été découverts, et classifiés en plusieurs types.

Les principaux polymorphismes rencontrés dans le génome et étudiés en génétique sont les SNPs, les insertions, délétions, *Copy Number Variations* (CNV), et les satellites.

SNP

Un SNP, ou *Single Nucléotide Polymorphism*, est une variation d'un unique nucléotide. Ces variations représentent environ 90% de la variabilité observée, et on en dénombre plus de 170 millions dans la base de données de référence dbSNP[She+01].

```
Allele1    CGG AAC ATG CGC GGC TAC TAC
Allele2    CGG AAC CTG CGC GGC TAC TAC
```

FIGURE 1.4 – Exemple de SNP
tiré de [Hea+17]

Indels

Les indels sont les insertions ou suppressions (en anglais *deletions*) de portions du génome (selon l'allèle qui est pris pour référence, on peut considérer de manière équivalente que son alternative est une insertion ou une délétion). Lorsqu'ils concernent des gènes, ces polymorphismes peuvent amener à un changement radical de la protéine produite (qui aboutit souvent à des pertes de fonction de la protéine), étant donné que la lecture des codons se décale.

```
Allele1    CGG AAC  .TG CGC GGC TAC TAC
Allele2    CGG AAC CTG CGC GGC TAC TAC
```

FIGURE 1.5 – Exemple d'Indel

(micro)Satellites

Les satellites (ou séquences répétées en tandem) sont des répétitions de séquences du génome. L'unité de répétition (le motif) peut être de longueur variable, allant du *microsatellite* de quelques bases répétées des centaines de fois, au motif α-satellite de

171 paires de bases présent dans les centromères, et pouvant être répété sur plusieurs mégabases. On estime que la moitié du génome environ est composée de satellites[RKD08].

CNV

Les Copy Number Variations (CNV) sont de grands fragments du génomes, d'une taille supérieure à 1000 paires de bases. Ils sont similaires aux satellites en ce qu'ils consistent également en des copies de portions du génome, mais là où les satellites se répètent en un même point du génome, les CNV peuvent avoir des copies disséminées à des endroits variés du génome, y compris sur plusieurs chromosomes différents. Les CNV englobant parfois des gènes entiers, peuvent aboutir à des modifications des niveaux d'expression de ceux-ci, avec des effets phénotypiques.

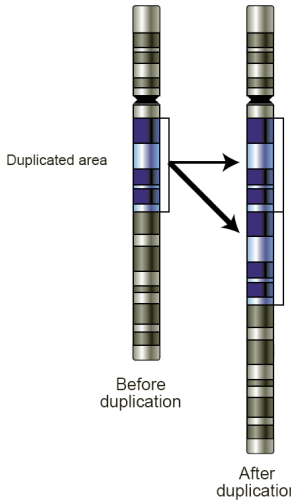

FIGURE 1.6 – Exemple de CNV

1.3 Notions de génétique des populations

La génétique des populations découle des travaux de Gregor Mendel (XIX^e), premier scientifique à théoriser la manière dont les gènes se transmettent entre générations (lois de Mendel). À la suite de ses travaux, la discipline s'est intéressée à l'étude de l'évolution de la présence d'un allèle dans une ou plusieurs populations. Elle utilise des modèles mathématiques et statistiques pour prédire les changements de fréquence des allèles, en

TABLE 1.2 – Équilibre de Hardy-Weinberg pour la distribution des génotypes. La fréquence de l'allèle A est p, celle de l'allèle a est q.

Allèles	A	a
A	AA, $freq = p^2$	Aa, $freq = pq$
a	aA, $freq = qp$	aa, $freq = q^2$

présence d'éventuelles pressions évolutives.

Ces notions sont particulièrement importantes, car les hypothèses à leur origine sont à la base de tous les résultats obtenus en statistique génétique, puis en bioinformatique.

1.3.1 Équilibre d'Hardy-Weinberg

Dans une population, pour un polymorphisme biallélique, on note p la fréquence d'un allèle (A) et q la fréquence de l'autre (a). On considère une absence de sélection sur les allèles concernés, et on a $(p + q) = 1$ donc $(p + q)^2 = p^2 + q^2 + 2pq = 1$.

La loi de Hardy-Weinberg établit que

— dans une population isolée suffisamment large, en l'absence de mutations et de sélection, les fréquences alléliques restent constantes d'une génération à l'autre (Chaque allèle de la génération $n + 1$ ayant une probabilité p d'être issu d'un allèle A et q d'être a, la loi des grands nombres assure le résultat).

— Si les couples et les gamètes se forment au hasard, les fréquences génotypiques sont directement liées aux fréquences alléliques : p^2 sera la fréquence du génotype AA, $2pq$ celle de Aa, et q^2 celle de aa (voir table 1.2).

Une déviation de l'équilibre de Hardy-Weinberg indique en général une falsification des hypothèses d'Hardy-Weinberg : soit que la mutation n'est pas neutre en termes de sélection, soit qu'on est en présence de consanguinité, par exemple.

1.3.2 Déséquilibres de liaison

Lorsque deux polymorphismes sont suffisamment éloignés dans le génome, on considère qu'ils sont transmis de manière relativement indépendante : au niveau populationnel, après plusieurs générations (et donc plusieurs recombinaisons interchromosomiques ou

intrachromosomiques), les probabilités d'observer ces polymorphismes sont effectivement indépendantes.

En revanche, la présence de deux polymorphismes proches sur un même chromosome, par mutation par exemple, introduit une *liaison* entre ces polymorphismes. Ils seront notamment dépendants l'un de l'autre lors des transmissions à la génération suivante : les recombinaisons intrachromosomiques étant relativement rares, elles se produisent peu souvent entre deux polymorphismes proches, donc si l'un est transmis, c'est que son chromosome est présent dans le gamète, et par conséquent l'autre variant également [1], ce qui entraînera un *déséquilibre de liaison.*

Les allèles ne sont donc pas indépendants, et cela se traduit par une différence entre les fréquences des combinaisons des allèles observés, et les produits des fréquences alléliques des deux SNPs (fréquences attendues des combinaisons). D'autres facteurs peuvent également contribuer à une interdépendance, comme par exemple la pression de sélection (si par exemple la présence conjointe de deux mutations n'est pas viable), aussi parle-t-on de manière plus générale de *déséquilibre gamétique.*

On observera alors, sur des allèles A/a et B/b $\begin{cases} f_{AB} = f_A f_B + D \\ f_{Ab} = f_A f_b - D \\ f_{aB} = f_a f_B - D \\ f_{ab} = f_a f_b + D \end{cases}$

Lorsque $D = 0$, il n'y a pas de déséquilibre gamétique observable.

Pour pouvoir comparer des déséquilibres entre polymorphismes ayant des fréquences différentes, on normalise la valeur D en une valeur $D' = \dfrac{D}{D_{max}}$ où $D_{max} = \begin{cases} min(f_{AB}, f_{BB}) si D > 0) \\ min(f_{Ab}, f_{aB}) si D > 0 \end{cases}$
Cette mesure possède l'avantage d'expliciter (lorsque $|D'| = 1$) l'absence d'un génotype possible.

Cependant, la mesure la plus utilisée est le r^2, défini comme

$$r^2 = \frac{D^2}{f_a f_A f_b f_B}$$

Ici, $r^2 = 1$ indique un déséquilibre total entre les polymorphismes : lorsque les allèles sont totalement corrélés et co-transmis. La connaissance de l'un impliquera alors la

1. Cependant, après de nombreuses générations, cette faible probabilité de cross-over finit par se réaliser, et en l'absence d'autres facteurs (sélection notamment), le déséquilibre observé aura tendance à diminuer, et les polymorphismes –bien que proches– à se répartir de manière plus homogène dans la population.

connaissance de l'autre.

1.3.3 Stratification géographique

En raisons des migrations de populations au cours du temps, mais aussi notamment de la dérive génétique, évolution causée aléatoirement, les mutations apparaissant dans une population se sont structurées en fonction des migrations originelles qui la composent, et donc représentent en partie l'évolution géographique de celles-ci.

Pour prendre en compte cette stratification géographique (autrement dit la "spécificité génétique" d'une population) lors d'études génétiques, où l'on s'intéresse aux facteurs causaux en essayant d'ignorer les effets de population, on réalise une analyse par composantes principales de la population testée. Cette analyse permet de résumer au mieux les informations contenues dans les données, en distinguant les caractéristiques pertinentes du bruit.

Lors des tests d'association génétique, [Nov+08] a mis en évidence (Figure 1.7) que l'une des principales causes de la variance des données entre individus provient de la répartition des origines géographiques des individus, en superposant les origines géographiques des individus et leur répartition suivant les deux premiers axes de l'ACP. Pour éviter de détecter des effets dûs à cette seule répartition et donc se concentrer sur des signaux génétiques causaux, on prendra les coordonnées des individus suivant ces deux premiers axes de l'analyse en composante principale (que l'on a déterminé comme fortement corrélés à l'origine géographique, cf. Figure 1.8 où j'ai reconfirmé cette corrélation sur les populations utilisées dans les études du laboratoire) comme covariables de l'analyse. À l'avenir, la mondialisation, et en particulier l'accroissement des échanges internationaux cumulé avec les progrès des transports modernes, va certainement bousculer la relative stabilité génétique des populations observée jusqu'au siècle dernier.

1.4 Maladies génétiques et hérédité

Le caractère héritable de certains traits au cours des générations a été observée de longue date : chez l'homme par exemple, des caractères visibles comme la couleur de

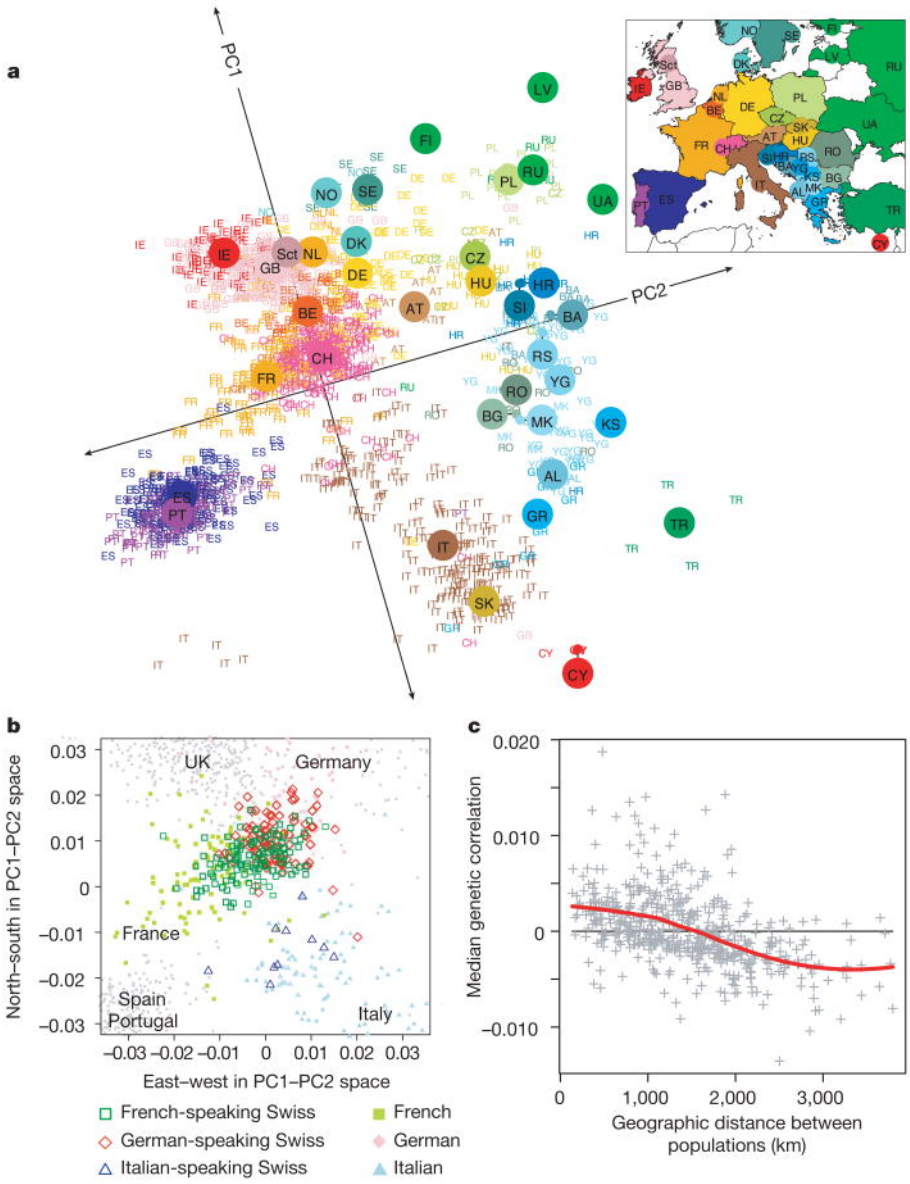

FIGURE 1.7 – Structure de la population en Europe
Cette figure tirée de [Nov+08] met en évidence la corrélation entre la répartition géographique des individus, et les coordonnées obtenues en extrayant les deux premiers axes d'une analyse en composantes principales des génomes des individus : il est ainsi notamment possible, à partir du génome d'un individu, de situer grossièrement son origine géographique.

peau, des yeux, ou des cheveux, sont hérités.

De plus, l'observation chez les animaux ou les plantes de l'héritabilité de certains caractères tels que le rendement en lait ou en grain, a été mis à profit très tôt pour

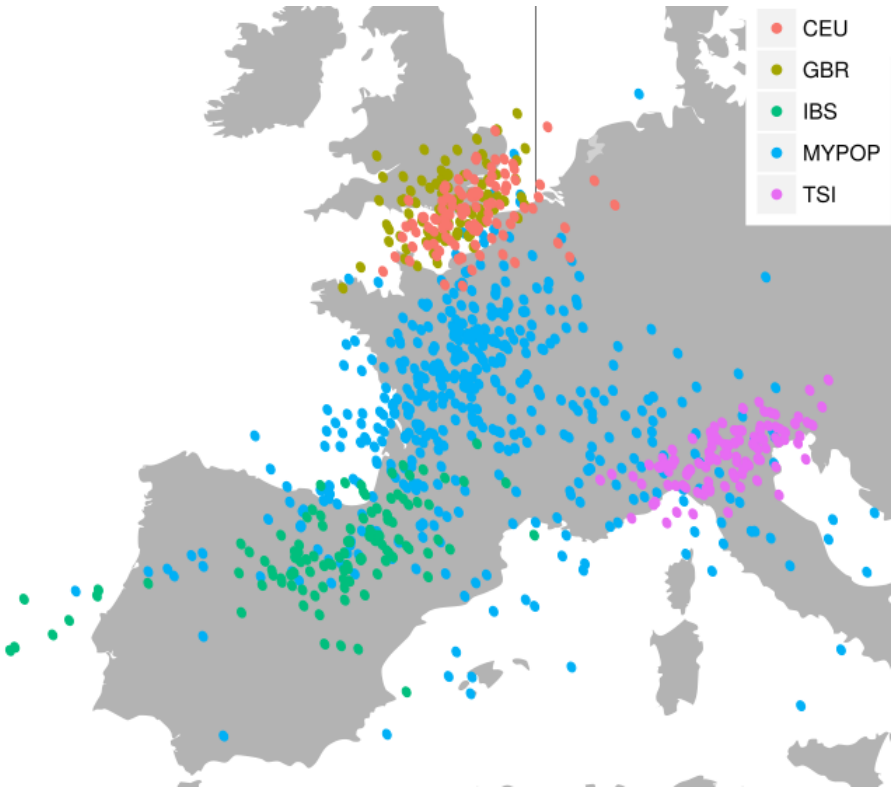

FIGURE 1.8 – structure de la population : HEPAVIH(en bleu) et 1000 Genomes
Nous avons effectué à nouveau les analyses de la figure 1.7 avec les données à notre disposition : les données de 1000 Genomes (d'origines homogènes anglaises, espagnoles et italiennes) et de HEPAVIH (en bleu), majoritairement français. Nous arrivons également à corréler la position suivant les premiers axes de l'analyse en composante principale, et la répartition géographique des populations d'origine.

orienter leur sélection en agriculture. Toutefois, ce n'est qu'à la découverte de l'ADN par Watson et Crick (1953) que l'on a pu identifier le mécanisme biologique à l'origine de l'hérédité.

Dès l'apparition de la possibilité d'observation au microscope des caryotypes, il a été possible de faire le lien entre des maladies, comme le syndrome de Down (trisomie 21), et l'ADN (Lejeune [LTG59a]). Depuis, l'amélioration constante des techniques de biologie moléculaire s'est accompagnée d'une meilleure compréhension des liens entre génome, gènes, et maladies.

De manière générale, une conséquence du dogme central de la biologie moléculaire est qu'un changement de phénotype dû à la génétique (et donc potentiellement héritable)

peut avoir deux principales causes : les mutations peuvent modifier la séquence codante d'une protéine, et dans ce cas la protéine peut être déformée et ne plus remplir son rôle, ou les mutations peuvent influencer sur l'expression, c'est-à-dire la quantité d'une protéine produite par la cellule, contrôlée par divers facteurs. L'exemple le plus flagrant est celui de la trisomie 21 : la présence d'un chromosome 21 supplémentaire entraîne une augmentation de 50% de l'expression d'un unique (petit) chromosome et a des conséquences phénotypiques importantes[Kor+94]. De manière générale, la plupart des signaux trouvés dans les études d'associations (présents dans le GWAS Catalog[Mac+17]) correspondent à des variations d'expression.

1.4.1 Maladies monogéniques

Les polymorphismes génétiques, en particulier lorsqu'ils sont localisés dans des gènes, peuvent engendrer des modifications de la structure des protéines, et potentiellement avoir des effets sur la fonction de cette protéine. Ils peuvent également entraîner des variations dans les niveaux d'expression des gènes, c'est-à-dire de la quantité d'ARNm et/ou de protéines produits à partir d'un gène. Lorsque des modifications se produisent, elles peuvent conduire au mauvais fonctionnement d'un gène en particulier, causant alors une *maladie monogénique*.

La plupart des effets des polymorphismes bialléliques (à deux allèles) peuvent être qualifiés de *dominants, récessifs* ou *additifs* :

— Un effet est dit dominant s'il suffit de la présence de l'allèle associé sur l'un des deux chromosomes qui portent le gène pour obtenir le phénotype,

— récessif s'il faut avoir l'allèle associé sur les deux chromosomes qui portent le gène pour observer son effet,

— et additif si, dans le cas d'un phénotype continu (ne possédant pas deux catégories clairement définies), le phénotype sera d'autant plus présent ou observable que l'allèle associé est présent (c'est-à-dire qu'on observe le caractère dans une plus faible mesure en présence d'une copie de l'allèle, qu'en sa présence sur les deux chromosomes portant le gène).

Les effets récessifs ou dominants sont duaux : si on a deux allèles pour deux effets

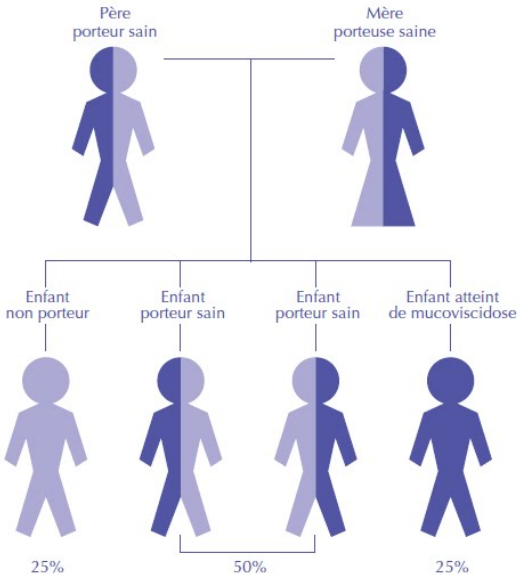

FIGURE 1.9 – Transmission d'une maladie récessive par des parents sains
Deux parents hétérozygotes porteurs d'un allèle d'une maladie monogénique récessive ont une chance sur quatre de transmettre tous deux l'allèle muté, et ainsi d'avoir un enfant atteint de la maladie (ici, la mucoviscidose).

(par exemple dans les groupes sanguins, A (ou B) est dominant, tandis que O est récessif), un des phénotypes sera récessif là où l'autre sera dominant.

Les maladies génétiques monogéniques les plus connues et les plus anciennement observées sont les maladies récessives, comme la drépanocytose (gène *HBB*[Aki+11]), l'hémophilie (gènes *F8* ou *F9* [LV86]), ou l'albinisme(*TYR* pour le type 1[SBR98]) : Dans cette famille de maladies, c'est la présence de la mutation sur chacun des deux chromosomes qui engendre la maladie.

Ces maladies ont notamment fait l'objet d'*études familiales*[HBE05] : pour comprendre les mécanismes de l'hérédité, on observe la transmission du phénotype au sein de familles, ce qui permet d'en déduire la transmission du polymorphisme associé.

Au niveau génétique, on a commencé par étudier les séquences Alu[MMR97] : elles sont relativement simples à observer par fluorescence et nombreuses chez tous les primates (environ 11% du génome humain). De plus, la diversité de leur présence ou absence et de leur longueur a permis d'identifier individuellement les chromosomes (et donc de pouvoir différencier un chromosome paternel d'un chromosome maternel lors des études familiales)

et leurs régions, permettant de visualiser les variations importantes[Ben+08].

À un niveau plus fin, les marqueurs RFLP (Polymorphisme de longueur des fragments de restriction) ont beaucoup été développés à partir des années 1980, grâce aux progrès de la biologie moléculaire et la disponibilité des enzymes de restriction[Rob+09]. Ces enzymes coupent le génome à des sites précis, et la présence ou l'absence d'un site de restriction, en raison de la présence d'un polymorphisme, produira des fragments découpés plus ou moins longs, qui seront donc facilement identifiables. Cette méthode permet donc d'établir une première liste de sites de découpe homozygotes (si tous les fragments découpés par l'enzyme sont de même longueur) ou hétérozygotes.

Certaines études se concentrent également sur les familles avec des jumeaux monozygotes pour mettre en évidence les phénomènes liés à l'environnement et ceux liés à la génétique. Ce type de maladie peut en effet se détecter notamment par des sauts de génération observables : si deux grands-parents possèdent deux copies de l'allèle associé à la maladie, les parents peuvent n'en posséder qu'un, et chacun le transmettre à un enfant, qui sera donc porteur des deux allèles, et donc de la maladie.

Détecter les gènes impliqués revient donc à chercher quelles sont les mutations portées de manière homozygote par les membres malades, et de manière hétérozygotes par les membres sains de la famille.

Les myopathies, par exemple, sont des dégénérescences musculaires. Plus d'une centaine de gènes sont cruciaux pour le développement et le fonctionnement des tissus musculaires, aussi une déficience sur l'un ou l'autre des gènes impliqués amènera une myopathie différente[Cha11].

On recense à ce jour plusieurs milliers de maladies génétiques rares[Rat+12].

1.4.2 Maladies multifactorielles

En revanche, la plupart des maladies communes ne sont pas causées par une unique mutation, mais par un ensemble d'éléments pouvant causer (par exemple, l'exposition au virus ou à la bactérie) ou favoriser (ou au contraire défavoriser, voire empêcher) l'apparition de la maladie. Ces maladies, initialement appelées multigéniques, ont été ensuite appelées

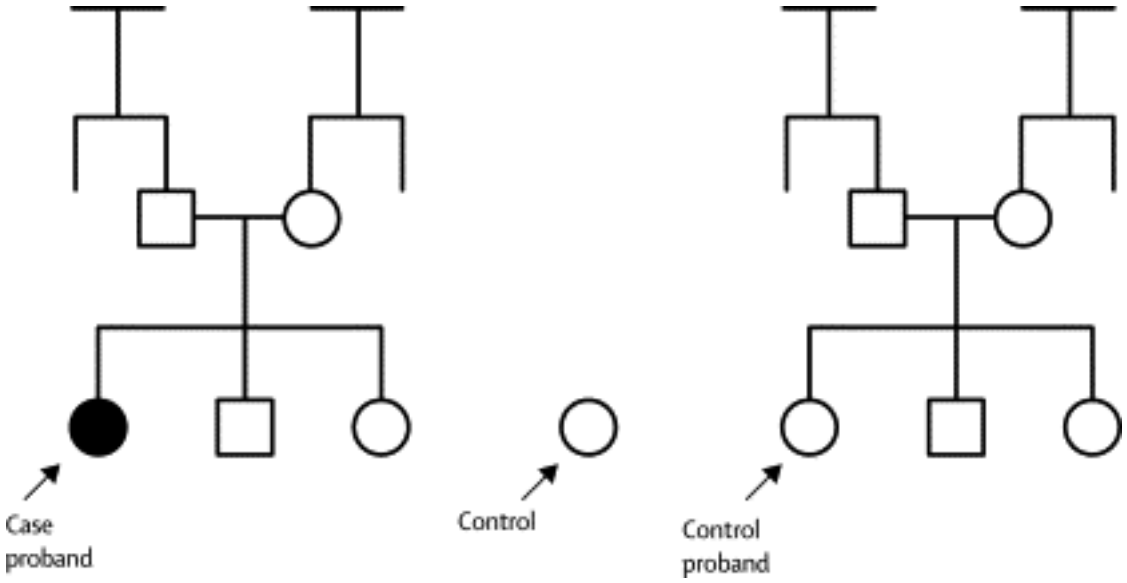

FIGURE 1.10 – Exemple d'une étude familiale
On cherche à déterminer à partir de familles analysées les mécanismes de transmission, et les membres porteurs d'allèles mutés (tiré de [HBE05]).

multifactorielles, pour intégrer le rôle généralement important de l'environnement dans la susceptibilité à ces maladies.

Facteurs de risques

Pour ces maladies, les gènes ne seront la plupart du temps que des *facteurs de risque*, associés plus ou moins fortement avec la maladie.

Un allèle ou un autre d'un gène pourra influencer le départ ou la cinétique de la maladie, et leur étude peut notamment aider à comprendre les mécanismes de fonctionnement de celle-ci et ses voies de signalisation. Au-delà des seuls facteurs génétiques, les autres principaux facteurs de risques sont liés à l'environnement (au sens large, qui prend aussi en compte les actions du patient, ou encore son alimentation) : si par exemple certains gènes peuvent être impliqués dans le développement ou l'évolution d'un cancer du poumon, par exemple, la cigarette et la pollution restent des facteurs importants de déclenchement de la maladie[Hec99].

Dans le cas du VIH, une mutation ($CCR5\Delta32$) a été découverte qui empêche l'infection par le virus, donnant un exemple de facteur génétique ayant une forte incidence sur

l'effet d'un virus (facteur extérieur), et amenant à des recherches sur l'utilisation de cet effet à des fins thérapeutiques[Hüt+09 ; Sam+96 ; Dea+96 ; Rap+97 ; Qui+98]. L'allèle HLA*B35 a également été corrélé au SIDA[Car+99], et l'allèle HLA*B57 :01 a été associé à la fois à un arrêt de la progression de l'infection par le VIH [Mig+00 ; Kas+96 ; Hen+99 ; Fel+07 ; Lim+09] et à l'hypersensibilité à l'abacavir, un traitement du VIH[Het+02].

Pour quantifier l'effet d'un facteur, on peut calculer le risque relatif rapproché associé (ou OR, odds ratio).

Gènes impliqués

Étant donné que l'effet recherché n'est plus ici systématiquement lié à une unique mutation, il est nécessaire de chercher des *associations* entre le génome et le phénotype : à partir d'un grand nombre d'individus, on cherche à trouver des variations du génome statistiquement associées à la distribution des phénotypes de maladies. Comparativement aux études sur les maladies monogéniques, l'implication de plusieurs gènes entraîne des effets individuellement moins forts, et une étude nécessitera davantage d'individus pour isoler un signal.

Ce type d'étude sur de larges population est assez récent : dans les années 1990, elles ont débuté par des approches dites « gènes candidats » s'appuyant sur des tests RFLP ou de séquençage ciblant des gènes connus, puis sont passées à des approches génome entier depuis 2006. Les techniques employées pour ces dernières sont présentées en section 2.1.

Si la recherche en génétique avait initialement l'espoir de pouvoir expliquer la plupart des traits hérités avec un faible nombre de gènes (par exemple, une étude sur l'autisme en 1999 écrivait « Il y a peut-être plus d'une quinzaine de gènes impliqués »[Ris+99]), les évolutions de la recherche ces dernières années ont conduit à se rendre à l'évidence : pour la plupart des traits phénotypiques, il y a probablement un grand nombres de gènes impliqués, la plupart d'entre eux exerçant un effet minime sur le phénotype.

On recense dans le GWAS Catalog plusieurs centaines d'études démontrant des associations avec les gènes HLA, pour plus d'un millier d'associations.

Les méta-analyses, qui combinent différentes études, et possédent une puissance

statistique plus forte, permettent souvent de mettre en évidence davantage de signaux. Par exemple, une méta-analyse (agrégation d'analyses) sur le diabète de type 1 a pu mettre en évidence plus de 40 loci affectant de manière significative le risque de contracter la maladie[Bar+09].

Héritabilité manquante et hypothèse omnigénique

Les études génétiques se succédant ont identifié des centaines de variants génétiques associés avec des maladies multifactorielles. Cependant, les variants identifiés ne sont souvent associés qu'à des augmentations mineures de risques, n'expliquant que partiellement la transmission observée des susceptibilités à ces maladies. On qualifie ce manque par le terme d'*héritabilité manquante*[Man+09]. Plusieurs explications ont été proposées pour tenter d'expliquer ce phénomène, comme la présence de très nombreux variants ayant un faible effet[BEV16], l'implication de variants rares non encore détectés, ou encore une prise en compte incomplète des effets de l'environnement.

Par exemple, le modèle *omnigénique*[BLP17] considère que pour les traits complexes, *tous* les gènes ont, à un certain degré, une influence : les gènes directement reliés au phénotype ont l'influence la plus forte, mais les gènes codant pour des protéines interagissant avec les protéines reliées au phénotypes ont une influence, bien que moins forte, et les gènes codant pour des protéines interagissant avec celles-ci également, etc., et par propagation, tous les gènes (Figure 1.11). Ce type d'hypothèse, contrastant avec les considérations initiales, laisse penser qu'il sera difficile (voire impossible) de jamais obtenir et caractériser toute l'héritabilité d'un phénotype.

1.4.3 HLA et maladies

Le complexe majeur d'histocompatibilité (CMH) est un système impliqué dans les réponses immunitaires. Il a été découvert par Jean Dausset à partir de 1952 en étudiant les réactions des globules blancs aux transfusions[Dau58], ce qui lui vaudra le prix Nobel de médecine en 1980, partagé avec George Snell[Sne12] et Baruj Benacerraf[Ben81].

Le CMH est composé d'un groupe de protéines de surface situées sur les cellules de l'organisme, dont le rôle est de présenter des antigènes (petits fragments de proteines du

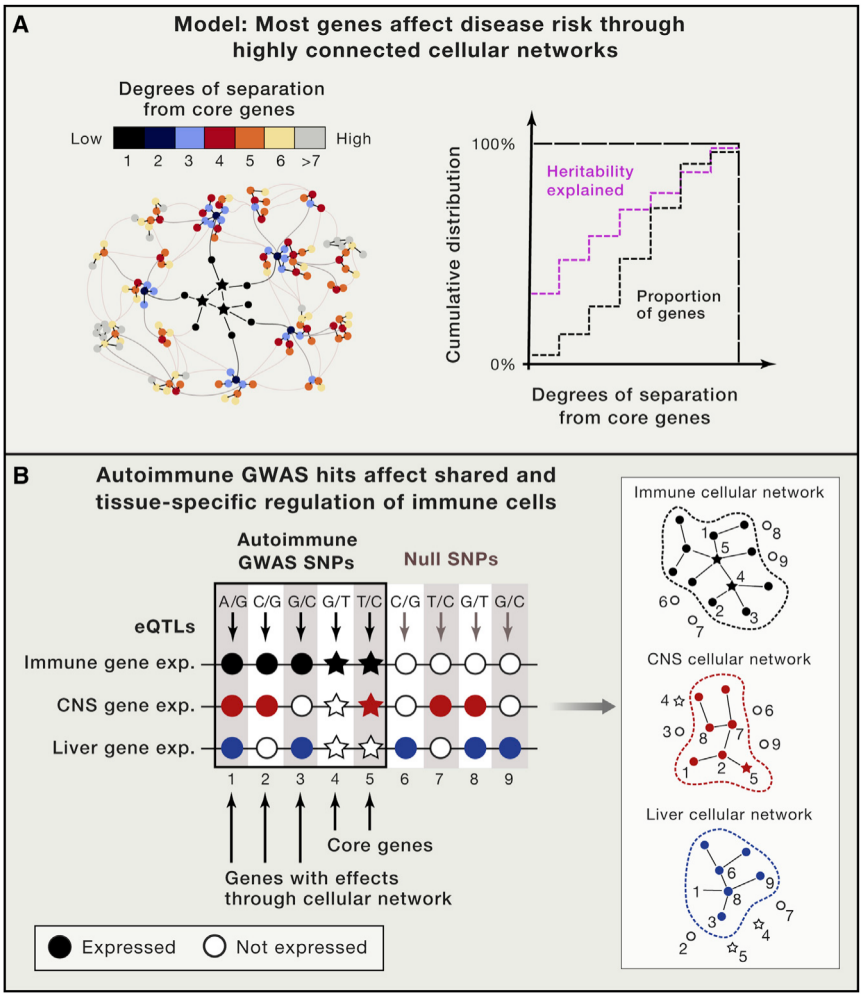

FIGURE 1.11 – Modèle omnigénique
(Tiré de [BLP17]) Pour un phénotype de maladie donné, un nombre restreint de gènes ont un effet direct sur la maladie. En revanche, par les effets des réseaux d'interactions, la plupart des gènes seront à faible distance de ces gènes centraux, et pourront avoir un effet non nul sur la maladie. La plupart de l'héritabilité reste expliquée par le petit nombre de gènes au cœur de l'action.
Pour les traits qui sont reliés à différents tissus, l'effet d'un SNP sera une moyenne pondérée de ses effets sur les divers types de cellules.

soi ou du non soi) aux cellules du système immunitaire, ce qui déclenche si nécessaire une réponse immunitaire.

On distingue le CMH de classe I (CMH I) et le CMH de classe II. Le premier est un ensemble de protéines de surface présentes sur toutes les cellules de l'organisme. Les proteines du CMH I présentent des peptides propres à la cellule (des molécules qui

nécessitent la machinerie cellulaire, des protéines du soi ou des protéines virales par exemple) appelés "antigènes endogènes". Le CMH de classe II (CMH II) est un ensemble de protéines de surface présentes principalement sur des cellules du système immunitaire (notamment les cellules présentatrices d'antigènes « professionelles », ou CPA, telles que les cellules dendritiques). Les proteines du CMH II présentent majoritairement des antigènes d'origine extracellulaire dits "antigènes exogènes" (fragments de bactéries, parasites, cellules infectées...).

Le CMH humain est dénommé HLA (Human Leukocyte Antigen) et est codé par un ensemble de gènes localisés dans une région du chromosome 6. Il existe dans l'espèce humaine une grande variabilité génétique de la région du HLA, avec jusqu'à 4000 allèles connus pour certains gènes. Les différences génétiques entre les molécules du HLA des individus se traduisent en physiologie par des différences d'aptitude à présenter des antigènes spécifiques et déclencher une réponse efficace, et en pathologie par des différences de susceptibilité à de nombreuses maladies comme les maladies infectieuses, les maladies autoimmunes, ou encore dans le succès des greffes.

L'importance des conséquences de la variabilité de la région HLA, dans les maladies notamment, explique qu'un grand nombre d'études et un effort de recherche particulier soit mené sur le complexe majeur d'histocompatibilité.

Par exemple, le HLA a notamment été associé avec :

— HLA-B*27 avec la spondylarthite ankylosante[Fel+03]
— HLA-DRB1*15 avec la sclérose en plaques[Pat+13]
— HLA-B*35, HLA-A*29, HLA-B*22, HLA-B*14, HLA-C*8 et HLA-B*57 avec le VIH[Car+99 ; Lim+09 ; Hen+99]
— HLA-DR avec la polyarthrite rhumatoïde[Kla+06]
— HLA-DR avec le lupus erythémateux[JSG82]
— HLA-A*2 et HLA-B*37 avec la grippe[CP89]
— HLA-DR, HLA-DQ, HLA-B*38 avec le diabète de type 1[Erl+08 ; Nej+07]
— HLA-DRB avec la longévité [Iva+98]
— etc.

La partie II de cette thèse sera notamment dédiée à la recherche d'informations sur

le HLA à partir de diverses sources de données.

Chapitre 2

Génomique d'association

L'acquisition et le traitement de données biologiques de tous types et leur profusion en particulier depuis la fin du XX^e siècle, a donné lieu à l'émergence d'une nouvelle discipline prolongeant jusqu'au *big data* la biostatistique existant depuis le milieu du XX^e siècle, associant entre autres des biologistes, des mathématiciens (en particulier statisticiens), et des informaticiens : la *bioinformatique*.

Cette nouvelle discipline vise à résoudre des problèmes d'ordre biologique liés aux progrès des techniques de biologie moléculaire avec l'aide des outils modernes d'analyse et des progrès de l'informatique sur cette période, et est aujourd'hui très centrée sur la résolution des problèmes biologiques liés à la masse de données générées par les nouvelles biotechnologies (omics), en s'appuyant sur les progrès informatiques et statistiques.

La bioinformatique se subdivise en deux principales branches : d'une part, l'étude des séquences s'intéresse à l'ADN, à l'ARN, et aux processus amenant à la production des protéines ; et d'autre part la bioinformatique structurale qui s'intéresse aux protéines, leurs configurations 3D, et leurs fonctions et interactions avec les tissus et molécules de leur environnement(notamment grâce à la cristallographie). Beaucoup de travaux s'orientent aussi sur l'étude des réseaux, c'est-à-dire des interactions de manière générale (entre gènes, protéines, cellules...) avec des problématiques de modélisation.

Mon travail a porté pour sa part sur la *génomique*, qui est la branche de la bioinformatique visant à étudier le génome (ici, le génome humain particulièrement), pour notamment aider à comprendre les fonctions associées aux gènes et les effets des polymorphismes

génétiques, et leur impact sur les maladies.

2.1 Génotypage à l'échelle du génome entier : techniques de biologie moléculaire

Le *génotypage* dénote l'ensemble des techniques qui visent à déterminer les allèles des polymorphismes génétiques. En 2001, [Lan+01] et [Ven+01] ont publié une première version du séquençage du génome humain, qui a servi de référence de base, et a rapidement permis des progrès en la matière, et notamment la détermination des parties conservées du génome sur la population, et les mutations fréquentes (présentes dans au moins 5% d'une population)

2.1.1 Puces de génotypage

Les puces de génotypage sont des puces permettant d'obtenir le génotype (ou *génotyper*) jusqu'à plusieurs millions de SNPs (polymorphismes sur un unique nucléotide) du génome simplement et rapidement, en se basant sur une séquence connue comme conservée en amont du polymorphisme, servant d'amorce. Il suffit ensuite d'ajouter la base complémentaire avec un marqueur de couleur (en général rouge pour le G, vert pour le A) par simple polymérisation d'une base, et selon la couleur (dans ce cas, rouge pour GG, verte pour AA, orange pour AG) on peut déterminer le génotype de la personne au niveau du polymorphisme testé.

Avantages et Limites

Cette méthode a l'avantage d'être relativement rapide et peu chère, et de pouvoir tester en parallèle plusieurs millions de SNPs.

En revanche, elle a deux principales limites : D'une part, il faut connaître en avance le polymorphisme visé : les mutations rares ne sont pas forcément répertoriées (et, si elles ont une importance phénotypique, ne seront pas détectées) ; et d'autre part la méthode demande à ce qu'il existe une zone conservée sans polymorphismes connus, d'un côté ou de l'autre du polymorphisme. Dans les zones hautement polymorphiques, cette

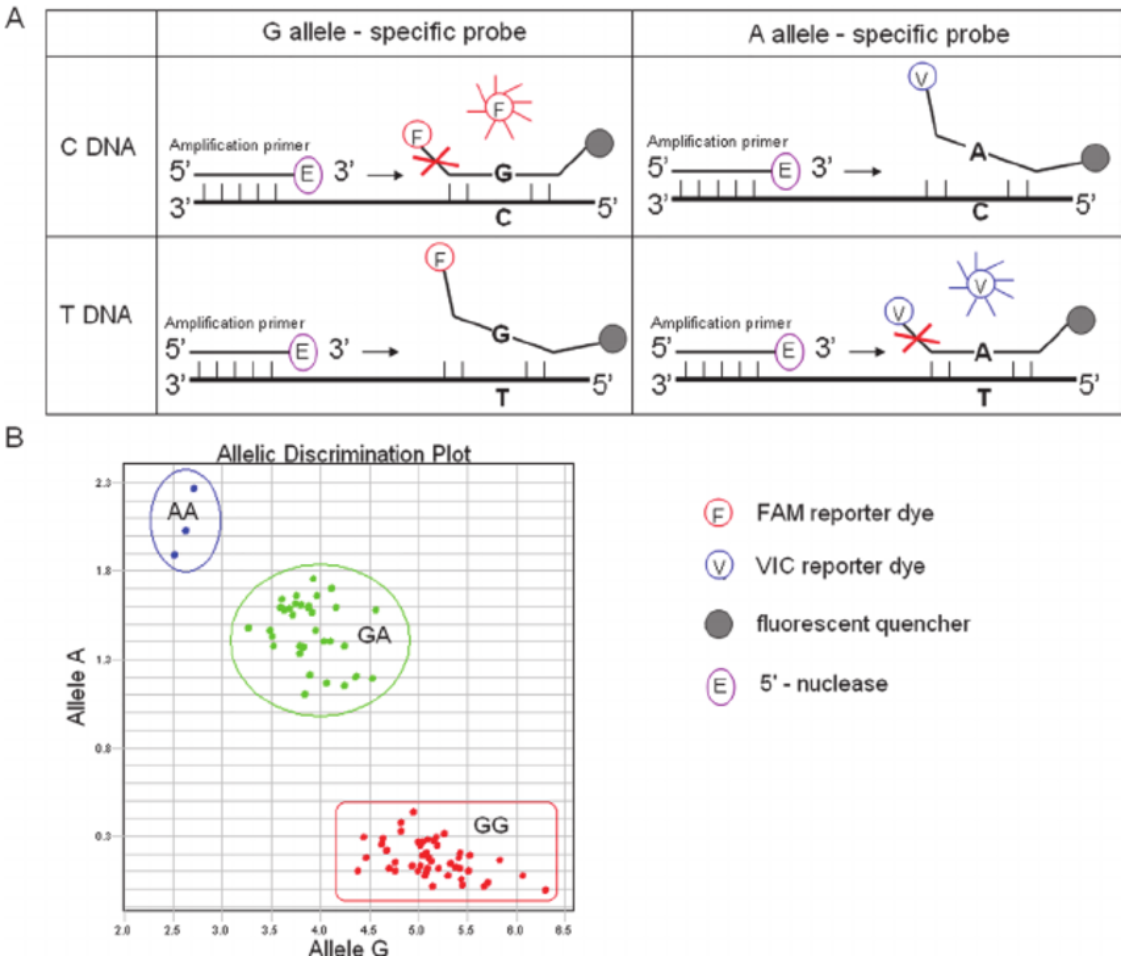

FIGURE 2.1 – Amorce de génotypage
[Zha+15] Une amorce se fixe sur la partie connue, assez longue pour être unique dans le génome, et produira une luminescence différente suivant quelle base se fixera sur le locus visé.

hypothèse est loin d'être toujours vérifiée. Par exemple, dans certains gènes du HLA, on peut observer plusieurs dizaines de SNPs à la suite, empêchant d'utiliser cette méthode pour les déterminer.

2.1.2 NGS (Séquençage Nouvelle Génération)

Le séquençage haut débit, aussi appelé séquençage nouvelle génération (Next Generation Sequencing ou High-throughput sequencing en anglais) est un ensemble de technologies relativement récentes pour séquencer (c'est-à-dire lire, linéairement) le génome de manière rapide, avec des lectures pouvant contenir plusieurs centaines de bases,

pouvant souvent fonctionner à partir de peu d'ADN et ce, de manière hautement parallèle.

La relative démocratisation ces dernières années du séquençage haut débit permet de réduire significativement les coûts du séquençage : D'après les estimations du NIH[NIH16], si le premier séquençage a probablement coûté des millions de dollars, en 2016 le coût d'un séquençage du génome entier est de l'ordre de grandeur de 1000€. En 2017, Illumina (une des principales entreprises du domaine du génotypage) a récemment annoncé que le coût du séquençage pourrait rapidement tomber sous la barre des 100$[Ill].

Le génotypage du HLA se fait depuis deux ans essentiellement par séquençage, qui a atteint une qualité et un prix sans concurrence pour cette région hautement polymorphe. Afin de tenir compte des erreurs, on peut par ailleurs effectuer jusqu'à plusieurs centaines de lectures de la région visée pour aligner les fragments au mieux.

2.2 Outils et ressources

La multiplication des moyens d'acquérir des données biologiques de tous types et la diversité des analyses possibles de celles-ci ont donné lieu à des milliers d'utilisations de l'outil informatique, que ce soit pour stocker de l'information (bases de données spécialisées, ontologies[Smi+07], standardisation d'échange de données...), pour l'annoter, la traiter ou pour l'analyser. Nous ne nous intéresserons ici qu'aux outils principaux utilisés pour ma thèse.

2.2.1 Ressources principales en génomique

La première nécessité quand on considère la masse de données produite en génétique consiste à les stocker, classer et rendre accessibles. Dans ce secteur, le laboratoire GBA a par exemple utilisé les données du projet *HapMap*[Gib+03], puis du projet *1000 Genomes*[BS15]. Cette base de donnée vise plus particulièrement à créer le catalogue le plus étendu possible de génotypes humains, et d'identifier les variants présents dans au moins 1% d'une population.

Les bases de données de RefSeq [PTM06] et Ensembl[Yat+15] ne fournissent pas de génotypes entiers, mais des séquences annotées, qui sont des commentaires sur les

fonctions connues des gènes, allèles, ou polymorphismes.

À un niveau plus agrégé, les bases de données du National Center for Biotechnology Information (NCBI) [Whe+07] mettent à disposition les informations sur tous les polymorphismes connus. dbSNP[She+01], par exemple, recense tous les SNPs connus pour leur donner un identifiant unique, en effectuant le suivi des fusions d'identifiants de SNPs (deux SNPs initialement pensés comme différents s'avérant finalement n'en être qu'un), leur statut de validation (dans quelles bases de données ont-ils été identifiés ?), etc.

D'autres projets, tels que Gene Ontology[Ash+00 ; Con+15b] ou le GWAS Catalog[Wel+13], visent à structurer la manière de partager et de communiquer les informations se rapportant aux données biologiques, afin de faciliter leur traitement automatisé.

Il existe également de très nombreuses bases de données contenant des données génomiques de cohortes d'étude, qu'elles soient privées (chez des sociétés comme 23 and me, par exemple) ou créées par des organismes publics. Pour des raisons évidentes de vie privée, les données qu'elles contiennent sont très rarement publiques. J'ai pour ma part utilisé des données génomiques publiques de 1000Genomes[Con+15a], d'une cohorte d'étude du diabète[Hil+10 ; Ric+09], d'une cohorte générale anglaise[Lap+15], et d'une cohorte d'étude de la coinfection SIDA-Hépatite de l'ANRS[Lok+10].

2.2.2 Phasage

Un problème majeur des puces de génotypages telles que décrites plus haut est que si elles fournissent bien la liste des polymorphismes de type SNPs, elles ne permettent pas d'en déduire l'information *phasée* : quels sont les allèles (sur les points d'hétérozygotie) présents sur le même brin chromosomique ? Pour répondre à cette question, on utilisera les *haplotypes* : un haplotype correspond simplement à une combinaison de n SNPs présents sur un même chromosome. Ainsi, la connaissance des deux haplotypes donne toute l'information génétique.

Le problème auquel on se retrouve confronté est combinatoire : pour un génotype à n SNPs hétérozygotes, il y a 2^n manières de le décomposer en haplotypes. Tirant parti des déséquilibres de liaison entre SNPs, et des taux de recombinaisons connus suite aux

cross-overs se produisant lors de la méiose, les logiciels de phasage visent à établir quels SNPs sont présents sur quel chromosome (paternel ou maternel), en général à partir d'haplotypes déjà connus au préalable.

Les principaux logiciels de phasage (Shape-IT[DMZ12 ; DZM13] et beagle[BB07]) se basent sur l'utilisation de modèles de Markov cachés (HMM) avec des heuristiques de parcimonie.

2.2.3 Imputation

Une fois obtenus les haplotypes, l'utilisation de bases de données d'haplotypes connus avec autant de SNPs que possible (toujours en utilisant des modèles de Markov) permet d'utiliser le déséquilibre de liaison pour *imputer* des SNPs, c'est-à-dire obtenir statistiquement des informations sur les SNPs non génotypés, pour peu qu'ils soient suffisamment liés aux SNPs déjà obtenus.

En raison de leur fonctionnalités communes (notamment, l'imputation a besoin du phasage), les logiciels de phasage sont associés à des outils d'imputation : beagle[BB16] possède son propre module d'imputation, tandis que les formats de fichiers de Shape-IT et d'IMPUTE2[HDM09] sont compatibles entre eux.

2.2.4 Gestion de données

La nécessité de pouvoir partager les données entre chercheurs a amené à la création de formats de fichiers standardisés, afin de permettre le développement d'outils basés sur ces formats. Les deux formats de fichier les plus courants pour de simples données génomiques sont les standards de *plink*[Pur+07], et le format vcf[Dan+11]. Bien que la version 4.0 (la plus récente) du format vcf soit plus versatile que le format de fichier de plink (elle permet notamment d'agréger des données typées avec des données imputées, alors que plink est limité aux données sans probabilités), la plupart des outils et protocoles utilisés au laboratoire sont adaptés à plink.

En particulier, le logiciel plink est adapté pour effectuer simplement les nombreux contrôles qualité usuels sur les données génomiques (effectuer des calculs de masse sur les données manquantes, vérifier l'équilibre de Hardy-Weinberg, etc.)

```
ID7631 ID7631 0 0 1 1 A A G G A A A A T T C C C C T T C T A A
ID3502 ID3502 0 0 1 1 G A A A A A G A C T T T C C T T T T G G
ID7647 ID7647 0 0 1 1 G A A G A A G A T T T C C C T T T T G A
ID7284 ID7284 0 0 1 1 A A G G G A A A C T T C A C T T T T A A
ID7247 ID7247 0 0 1 1 A A A G A A A A C T T C C C T T T T G A
ID7287 ID7287 0 0 1 1 A A A G A A A A A T T C C C C T T C T A A
ID7608 ID7608 0 0 1 1 A A A G A A A A C C T T A C G T C T G G
ID3079 ID3079 0 0 1 1 A A G G A A A A A T T C C C C G T T T A A
ID3302 ID3302 0 0 1 1 A A A G A A A A A T T T C C C G T C T G A
ID7406 ID7406 0 0 1 1 G A A G A A G A T T T C C C T T C T G A

6       rs9392299       0       203909
6       rs1418708       0       205610
6       rs1418707       0       205800
6       rs1418706       0       205878
6       rs11961507      0       206313
6       rs9502959       0       206599
6       rs17806289      0       208822
6       rs17133789      0       209462
6       rs17133787      0       209980
6       rs7756332       0       211941
```

FIGURE 2.2 – exemple de fichier PLINK comportant un fichier ped et un fichier map Un des fichiers comporte les informations des SNPs (chromosome, nom, position), tandis que l'autre comporte les informations des individus (ID de la famille, ID de l'individu ...) et leur génotype.

2.3 Études d'association sur cohorte

Les études d'association cherchent à corréler des loci du génome à des phénotypes, discrets ou continus.

À cette fin, on va comparer les génotypes des membres d'une cohorte et les comparer, soit entre cas (affectés par une maladie, par exemple) et contrôles (non-affectés), soit essayer de corréler le génotype avec l'intensité d'un phénotype. Historiquement, les principaux types d'études génétiques ont été sur gènes candidats, puis sur génome entier par puce, et enfin sur exome ou génome entier par séquençage.

Pour chaque polymorphisme (allèle) étudié, une étude donnera une $p - valeur$. Cette valeur représente la probabilité d'obtenir une association au moins aussi forte que celle observée sous l'hypothèse nulle (par hasard). On considère en général qu'un *signal* est statistiquement significatif s'il a moins de 5% de chance d'être présent "par hasard", et plus la valeur p est faible (soit par une meilleure association, soit par le test d'une plus grande population), plus le test est concluant. Lorsque l'on teste des hypothèses multiples (comme plusieurs gènes, ou plusieurs polymorphismes), ce seuil de 5% nécessite d'être ajusté.

2.3.1 Gènes candidats

Une étude *gènes candidats* porte sur un nombre restreints de gènes dont on pense a priori qu'ils sont impliqués dans le phénotype. Elle étudie ensuite les polymorphismes dans ces gènes et leur corrélation au phénotype.

2.3.2 Génome entier

Une étude *génome entier* par puce[Man09], ou GWAS en anglais (pour genome-wide association study) teste indistinctement tous les polymorphismes disponibles dans le génome sur la puce, qui couvre en général le génome entier avec des SNPs répertoriés dans 1000Genomes. Cette approche a l'avantage de ne pas privilégier une fonction ou une localisation génétique *a priori*, mais la conséquence du très grand nombre de variants testés est un seuil statistique attendu très fort. Le seuil de significativité pour les études

génomes entiers est une p-valeur de 5.10^{-8} pour prendre en compte les tests multiples mais aussi les déséquilibres de liaison existant entre les SNPs[Hog+08], ce seuil étant considéré comme strict[PIP11].

La section 2.4 détaille les étapes classiques d'une GWAS.[Bal06]

2.3.3 Exome

Le dernier type classique d'études d'association est l'étude sur l'exome : Ici, au lieu de tester tous les polymorphismes du génome, on se limite aux gènes pour ne considérer que les zones exoniques, ou dans les versions plus récentes les régions régulatrices en amont du gène , en supposant que les variants impliqués ont des effets directs sur la structure de protéines. Cette méthode est d'autant plus populaire que les méthodes de séquençage modernes proposent la possibilité de séquencer l'exome seulement pour une fraction du prix du séquençage complet du génome (puisque l'exome couvre environ 1.5% du génome[Ng+08]), en permettant de détecter les variants rares de l'exome.

La justification de l'exome par rapport à l'approche génome entier par puce est justement que des variants rares peuvent jouer un rôle important dans les phénotypes étudiés. L'exploitation des données génomes entier par NGS reste limitée aux variants fréquents par puces et aux variants rares dont seule une fraction est détectée par l'approche d'exome. Néanmoins, l'approche génome entier par NGS va se populariser dans la mesure où le prix du séquençage devrait encore beaucoup diminuer.

2.4 GWAS

Cette section détaille les matériels et méthodes couramment utilisés lors d'études d'association, notamment dans le laboratoire GBA.

2.4.1 Contrôles qualité

La première étape, à partir des données de génotypes, consiste à en effectuer un contrôle qualité, pour éliminer tout problème pouvant avoir été causé par un défaut de génotypage. Par exemple, un taux élevé de données manquantes sur un SNP ou un

patient peut indiquer un défaut de la puce du patient ou des sondes utilisées sur ce SNP ; et un grand taux d'hétérozygotie sur un patient indiquera une contamination probable.

2.4.2 Effets de population et ACP

Pour éviter d'isoler des associations liées à la structure de la population (voir la section 1.3.3), il est fondamental d'utiliser les données de génotypage de la cohorte afin de détecter si certains individus sont issus d'une population différente de celle de la majorité, afin de les écarter de l'analyse.

À cette fin, on réalise une analyse en composantes principales (ACP) avec le logiciel EIGENSTRAT[Pri+06], qui déterminera les combinaisons indépendantes de SNPs contribuant le plus à la variabilité génétique observée. Dans le système d'axes constitué par les deux premières composantes principales de l'analyse, les individus sont représentés par un nuage de points, idéalement ne comportant pas d'éléments extrêmes.

Une fois écartés les individus extrêmes, l'analyse en composante principale est refaite, et on utilise alors les deux premières composantes comme covariables dans les analyses[MD11].

Deuxième partie

L'imputation du HLA

Chapitre 3

Le HLA

3.1 Antigènes HLA

Le HLA est un ensemble de protéines de surface des cellules, qui jouent un rôle essentiel dans la régulation du système immunitaire humain. Les gènes correspondants, codant pour ces protéines, sont situés dans la région $6p21$ sur le chromosome 6, qui est l'une des régions les plus polymorphiques du génome. Le gène du HLA se subdivise en HLA de classe I (avec les gènes *HLA-A*, *HLA-B*, et *HLA-C*) et de classe II (principalement *HLA-DP, HLA-DQ, HLA-DR*).

3.2 Fonctions biologiques

Le système HLA un rôle central dans l'immunité. Les protéines du HLA de classe I sont présentes à la surface de quasiment toutes les cellules de l'organisme, et servent en premier lieu à présenter aux cellules du système immunitaire des morceaux de protéines (peptides, antigènes), nécessitant la machinerie cellulaire, notamment lorsqu'il s'agit de protéines étrangères, comme c'est le cas lors d'infections virale. Leur présence et leur expression permet donc de différencier les cellules saines appartenant au corps, de cellules à détruire.

Les protéines du HLA de classe II ne sont la plupart du temps présentes qu'à la surface de cellules du système immunitaire telles que les cellules présentatrices d'antigènes. Elles présentent des peptides extracellulaires aux lymphocytes T CD4+, et si le peptide est

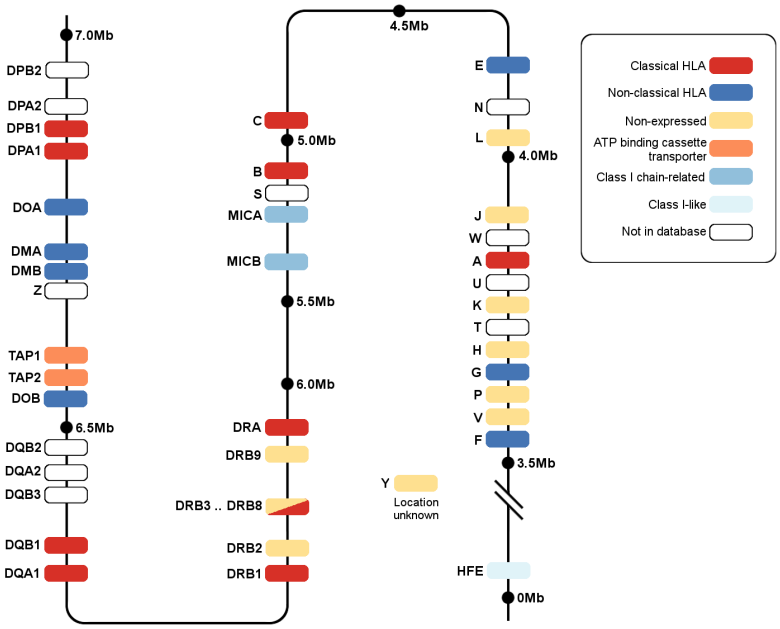

FIGURE 3.1 – Gènes du HLA
Cette figure présente les gènes HLA du complexe majeur d'histocompatibilité, dans la région 6p21 du chromosome 6 (tiré de http://hla.alleles.org/alleles/).

reconnu comme une menace (d'origine bactérienne, par exemple), une réponse immunitaire spécifique sera mise en place.

Il existe dans l'espèce humaine une grande variabilité génétique de la région du HLA, avec jusqu'à 4000 allèles connus pour certains gènes. Les différences génétiques entre les molécules du HLA des individus se traduisent en physiologie par des différences d'aptitude à présenter des antigènes spécifiques et déclencher une réponse efficace, et en pathologie par des différences de susceptibilité à de nombreuses maladies

Dans le cas des maladies dites *autoimmunes*, le corps déclenche une réaction immunitaire contre lui-même (dans des maladies telles que la sclérose en plaques, le diabète de type 1 ou encore la polyarthrite rhumatoïde). [Vin12]

Le HLA joue encore un rôle majeur dans les greffes d'organes ou de moelle osseuse[Mor+02] : Pour qu'une greffe soit possible et ne mène pas à un rejet, le système immunitaire du donneur et de l'accepteur doivent être compatibles, et en particulier, les protéines HLA du donneur et du receveur être suffisamment similaires (par exemple,

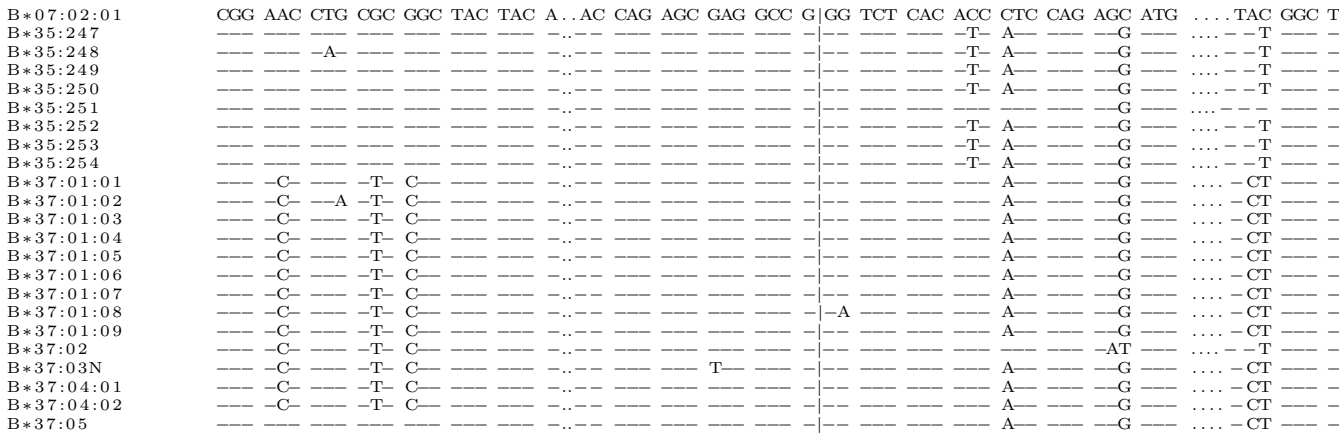

FIGURE 3.2 – Exemple d'alignement d'allèles HLA
Alignements d'allèles du HLA B, à 2 ou 3 champs (tiré de http://hla.alleles.org/alleles/).

présenter les mêmes peptides à la surface).

3.3 Typage HLA

Le HLA, de par son importance, a été énormément analysé dans des études génomiques. Le principal problème se posant alors est que son haut degré de polymorphisme (le *HLA-B*, par exemple, possède plus de 4828 allèles[Rob+14]) implique une grande difficulté à l'utilisation de la plupart des méthodes simples de génotypage (les puces de génotypage, par exemple) et force l'utilisation de méthodes plus complexes, plus itératives, et plus coûteuses.

La classification des allèles du HLA est hiérarchisée en différents champs :

— Deux allèles HLA identiques au premier champ ont un site de liaison similaire

— Deux allèles HLA identiques également au deuxième champ correspondent à la même protéine

— Deux allèles HLA qui diffèrent seulement au-delà du troisième champ ont des mutations synonymes, ou présentes dans les introns du gène.

3.3.1 Méthodes

Historiquement (jusqu'au début du XXIe siècle), le typage HLA se faisait par une analyse sérologique[Cla+85] (donc sans utiliser d'ADN), jusqu'à l'apparition des méthodes basées sur une amplification PCR (Polymerase chain reaction), puis plus récemment, des méthodes de séquençage direct.

SSP & SSO

Les méthodes SSP et SSO (pour *sequence-specific primer* et *sequence-specific oligonucleotide*)[Dun12 ; Ric12] amplifient la quantité d'ADN avec une PCR puis tentent d'accrocher des amorces connues, chacune de ces deux méthodes ayant ses avantages et inconvénients (par exemple, la méthode SSP est beaucoup plus rapide, mais ne permet pas simplement de paralléliser le typage d'un grand nombre d'individus)

Séquençage

L'essor du séquençage cette dernière décennie s'est étendu au typage HLA[Say+01], avec une fiabilité croissante et la possibilité de détecter plus simplement de nouveaux polymorphismes. En général, il est suffisant de se restreindre à l'exon 2 des gènes étudiés (et l'exon 3 pour les gènes de classe I) pour identifier correctement le site de liaison du gène.

3.3.2 Difficultés

Polymorphismes

La présence d'un très grand nombre de polymorphismes dans la région du HLA rend la tâche de son typage particulièrement difficile, tout particulièrement pour les méthodes à bases d'amorces : la présence de mutations dans la séquence sur laquelle doit se greffer l'amorce rend par exemple le typage difficile. La figure 3.2 présente un exemple de polymorphismes connus dans une petite région du *HLA-B* avec une grande densité de polymorphismes sur une poignée d'allèles proches (il y a environ 5000 allèles connus)

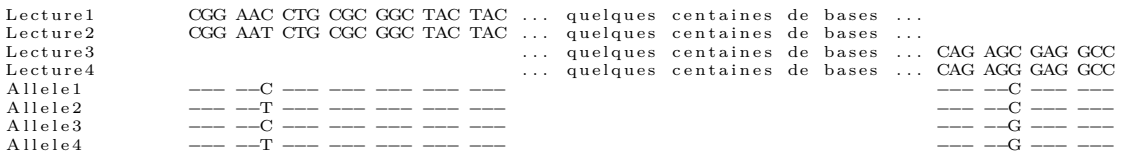

FIGURE 3.3 – Exemple d'ambiguïté inhérente
Il est impossible ici de déterminer précisément si on est en présence de (Allele1, Allele4) ou de (Allele2, Allele3)

Ambiguïtés

De plus, y compris quand la lecture de tout l'ADN de la région HLA se fait correctement, le typage HLA peut présenter des ambiguïtés inhérentes à la lecture par morceaux de l'ADN, qui sont en fait des ambiguïtés de phasage.

En effet, les lectures de l'ADN, quelles que soient les méthodes employées, sont relativement courtes à l'échelle du génome, et deux variations différentes éloignées de plus de quelques centaines de bases, si elles ne présentent pas de variation hétérozygotes régulièrement réparties entre, ne seront pas phasables. Hors, il existe des paires de paires d'allèles HLA reproduisant cette configuration (Exemple figure 3.3).

3.4 L'imputation HLA

Au début de ma thèse, le typage direct du HLA présentant des difficultés techniques ou un coût élevé, l'idée a été émise de pouvoir *imputer* les allèles HLA, comme on impute des SNPs, en exploitant des déséquilibres de liaison connus entre les allèles du HLA et les SNPs de la région, et plusieurs logiciels avaient été proposés pour répondre à cette problématique[LDM08].

3.4.1 SNP2HLA[Jia+13]

SNP2HLA fonctionne sur un principe simple : chacun des allèles possibles de chaque gène HLA est encodé dans le génotype comme un SNP binaire dénotant la présence ou l'absence de l'allèle. Le génotype se retrouve ainsi augmenté de quelques centaines de pseudo-allèles. Le logiciel fait ensuite appel au logiciel beagle[BY09] pour phaser les

génotypes, puis pour imputer à partir d'un panel de référence donné.

3.4.2 HLA*IMP2[Dil+13]

HLA*IMP2 fonctionne sur le même principe que SNP2HLA, mais n'est disponible que sous la forme d'une interface web (qui pose donc des problèmes de confidentialité des données), et sans précisions particulières sur les données de référence utilisées (ce qui pose des problèmes pour comparer le logiciel sur un pied d'égalité avec les alternatives), et une version précédente du logiciel, HLA*IMP[Dil+11] a été vendue à Affimetrix.

3.4.3 HIBAG[Zhe+14]

HIBAG (pour *HLA Imputation using attribute BAGging*) est un package R pour utiliser une méthode statistique de classification dite "attribute bagging"[BGQ03]. D'un panel de référence, il permet de construire itérativement un ensemble d'arbres de décisions utilisés pour *voter* pour la classification de la paire d'allèles HLA d'un sujet en fonction des SNPs connus pour la puce utilisée. Si la construction, itérative, d'un ensemble d'arbres de décisions le meilleur possible est un processus algorithmiquement beaucoup plus lent que le phasage utilisé dans les méthodes précédentes, le grand avantage de cette méthode est que le typage à partir d'un classificateur déjà construit pour une population donnée et un ensemble de SNP disponibles donné, est quasi instantané.

3.4.4 HLA*PRG[Dil+16]

Contrairement aux précédents, HLA*PRG détecte simplement le typage HLA avec des données de séquençage génome entier.

3.4.5 Outils du laboratoire

Le laboratoire GBA ayant développé un outil de phasage similaire à beagle, Shape-IT, j'ai entrepris de reproduire le pipeline de SNP2HLA avec Shape-IT et Impute2.

Logiciel	A	C	B	DRB1	DQB1
SNP2HLA (papier)	97.2	96.1	94.7	89.3	95.5
beagle+beagle	97.4	96.2	96.1	91.8	94.4
ShapeIT+Impute	97.6	95.8	96	87.9	94.8
beagle+Impute	98.0	95.6	96.1	92.3	96.8
HIBAG	97.4	96.0	96.7	90.6	97.0

TABLE 3.1 – Comparaison de différentes méthodes d'imputation du HLA (% d'allèles correctement imputés)

3.4.6 Comparaison

En étude préliminaire, j'ai comparé les différentes méthodes d'imputation HLA disponibles (Table 3.1) avec différents logiciels de phasage et d'imputation (beagle+beagle correspond à SNP2HLA classique). Le tableau 3.1 montre que Shape-IT, pourtant connu pour être meilleur que beagle[DZM13] n'améliore pas le niveau d'imputation des allèles HLA qui reste élevé, mais très imparfait pour une éventuelle utilisation en diagnostic. Le chapitre suivant revient plus en détail sur ces résultats.

À défaut d'améliorer la qualité de l'imputation du HLA directement, j'ai porté mon travail sur la fiabilité de l'utilisation de ces logiciels, que je présente dans le chapitre suivant.

Chapitre 4

HLA-Check

4.1 Problématique

Étant donné le caractère important de l'étude du HLA, et notamment des allèles des gènes HLA, il est intéressant de disposer d'outils d'imputation HLA pour par exemple associer dans des GWAS certains allèles du HLA avec des phénotypes donnés. Toutefois, nous remarquons que leur fiabilité est limitée (cf table 3.1), et cette incertitude sur les types HLA dans les données imputées à étudier se traduit en une incertitude sur les associations trouvées entre les allèles HLA et le phénotype. Afin d'obtenir des résultats d'association les plus pertinents possibles, nous nous intéressons donc à la question suivante :

Comment améliorer la fiabilité de l'imputation HLA ?

4.2 Pistes envisagées

4.2.1 Problème d'imputation

La première piste envisagée a été de partir de l'approche de SNP2HLA, relativement simple, puis tenter d'améliorer la qualité d'imputation des pseudo-SNPs. Pour cela, nous nous sommes appuyés sur les logiciels Shape-IT[DMZ12] et IMPUTE2[HDM09], plus versatiles et performants que la version de beagle utilisée par SNP2HLA (SNP2HLA utilise beagle3[BY09 ; BB07 ; Bro06], alors que beagle4[BB16] apporte des améliorations substantielles et utilise des formats de fichier plus communs).

Bien que l'implémentation nous aie permis de valider l'approche utilisant ces logiciels en lieu et place de beagle, nous n'avons pas pu constater d'amélioration des performances d'imputation, les deux approches ayant des résultats semblables. A posteriori, nous expliquons ces résultats par la grande similitude des algorithmes employés, et le fait que Shape-IT ne se distingue de beagle3 qu'en présence de déséquilibres de longue distance (plusieurs centaines de MB), tandis que nous imputons ici des SNPs dans le MHC qui est une région relativement petite (de l'ordre de 5Mbases)

4.2.2 Apprentissage machine

Une autre option envisagée a été d'utiliser les machines à support de vecteur[BGV92], un outil d'apprentissage machine ayant été utilisé avec succès dans de nombreuses applications, et notamment en classification de textes, en particulier avec l'utilisation d'un noyau non linéaire[SS02].

Toutefois, cette option n'a pas permis d'obtenir de classification satisfaisante, probablement en raison du grand nombre de paires de HLA impliqués. Par ailleurs, l'étude du fonctionnement de HIBAG[Zhe+14], utilisant une méthode d'apprentissage machine radicalement différente des autres méthodes testées, mais produisant des résultats du même ordre de grandeur, nous amène à penser que la quantité d'information extractible du dataset d'entraînement ne permettra pas d'améliorer significativement l'imputation.

4.2.3 Identification des allèles mal imputés

À défaut d'améliorer directement la fiabilité de l'imputation, l'axe de recherche que nous avons poursuivi s'est ensuite focalisé sur la question « Peut-on identifier quels sont les allèles ou individus imputés avec plus de certitude ? ».

Pour "ajouter de l'information" à celle présente dans le panel de référence, nous avons donc décidé d'utiliser le panel de référence de 1000 génomes, qui, s'il ne contient pas directement d'information HLA, permet d'imputer des SNPs présents dans la région, et plus particulièrement dans les exons du HLA.

Connaissant la séquence des allèles possibles des gènes HLA, une manière possible de valider la qualité d'une imputation HLA sera donc de voir dans quelle mesure l'imputation

des SNPs dans ceux-ci est compatible avec le HLA donné [1]. Pour cela, nous allons comparer toutes les combinaisons d'allèles connus sur le gène HLA étudié, avec les variants du gène prédits par imputation génétique à partir des données de génotypage par puce. Les allèles HLA connus, et tout particulièrement les séquences au niveau des exons 2 et 3, sont tous répertoriés dans une base de données mise à jour tous les trois mois sur `http://hla.alleles.org`. Il ne restera plus qu'à observer la compatibilité existant entre une paire d'allèles et la séquence prévue par imputation pour un individu.

4.3 Implémentation

La section suivante présente HLA-Check[Jea+17], un outil pour évaluer la pertinence d'informations HLA en la comparant à l'imputation des SNPs dans les exons du HLA.

4.4 HLA-Check : evaluating HLA data from SNP information

La section suivante (DOI:10.1186/s12859-017-1746-1)présente HLA-Check.

4.4.1 Résumé en français

Contexte

Le complexe majeur d'histocompatibilité, région du génome humain, et plus spécifiquement les gènes HLA (antigènes leukocytes humains), ont un rôle prépondérant dans de nombreuses maladies humaines. Grâce aux progrès récents des méthodes de séquençage (comme le Séquençage Nouvelle Génération, NGS), le génotypage précis de cette région est devenu possible, mais reste relativement onéreux. Afin d'obtenir les informations d'llèles HLA pour les millions d'échantillons déjà génotypés ces dix dernières années, des outils bioinformatiques efficaces tels que SNP2HLA ou HIBAG ont été développés, qui extraient l'information HLA du déséquilibre de liaison entre les allèles HLA et les marqueurs SNP de la région CMH.

1. On pourrait même naïvement imaginer convertir cette validation en une méthode d'imputation, mais les données SNP de référence sont très parcellaires sur le HLA (notamment en raison de leur difficulté à imputer) et sont loin de permettre de différencier tous les allèles.

Résultats

Dans notre étude, nous avons utilisé ShapeIT et Impute2 pour développer une méthode d'imputation similaire à celle de SNP2HLA, et obtenu une qualité comparable d'imputation sur un jeu de données européen. Nous avons ensuite développé un nouvel outil, HLA-Check, qui permet la détection de typages HLA peu plausibles vis à vis des génotypes SNP dans la région, et montré que combiner cet outil à un outil d'imputation HLA permet d'augmenter sa fiabilité, en particulier pour les gènes HLA de classe I.

Conclusion

Notre outil, HLA-Check, a permis d'identifier un petit nombre de typages HLA (moins de 10%) dans un jeu de données, et les individus identifiés peuvent ensuite être retirés ou retypés par NGS pour une étude d'association HLA.

Jeanmougin *et al. BMC Bioinformatics* (2017) 18:334
DOI 10.1186/s12859-017-1746-1

BMC Bioinformatics

SOFTWARE Open Access

HLA-check: evaluating HLA data from SNP information

Marc Jeanmougin, Josselin Noirel, Cédric Coulonges and Jean-François Zagury*

Abstract

Background: The major histocompatibility complex (MHC) region of the human genome, and specifically the human leukocyte antigen (HLA) genes, play a major role in numerous human diseases. With the recent progress of sequencing methods (eg, Next-Generation Sequencing, NGS), the accurate genotyping of this region has become possible but remains relatively costly. In order to obtain the HLA information for the millions of samples already genotyped by chips in the past ten years, efficient bioinformatics tools, such as SNP2HLA or HIBAG, have been developed that infer HLA information from the linkage disequilibrium existing between HLA alleles and SNP markers in the MHC region.

Results: In this study, we first used ShapeIT and Impute2 to implement an imputation method akin to SNP2HLA and found a comparable quality of imputation on a European dataset. More importantly, we developed a new tool, HLA-check, that allows for the detection of aberrant HLA allele calling with regard to the SNP genotypes in the region. Adding this tool to the HLA imputation software increases dramatically their accuracy, especially for HLA class I genes.

Conclusion: Overall, HLA-check was able to identify a limited number of implausible HLA typings (less than 10%) in a population, and these samples can then either be removed or be retyped by NGS for HLA association analysis.

Keywords: Human leukocyte antigen, Major histocompatibility complex, Imputation

Background

Human Leukocyte Antigen (HLA) genes are coding for cell surface antigen proteins responsible for a major function of the immune system, the detection of foreign or abnormal antigens [1]. These genes are located on the short arm of chromosome 6, in a region known as the major histocompatibility complex (MHC). They play a ubiquitous role in medicine, most notably in autoimmune diseases [2, 3], infectious diseases [4, 5], and transplant medicine [6].

The MHC is among the most polymorphic regions in the human genome, with up to 4000 known alleles for each class I gene, and up to 2000 alleles for class II HLA genes (case of *HLA-DRB1*) [7]. Furthermore, there is a strong impact of natural selection in the evolutionary history of the MHC that creates long-range linkage disequilibrium observed between many if not most variants in this region [8], that further complicates the task of widely-used genomic tools such as imputation algorithms. Imputation algorithms typically use a reference panel to infer statistical patterns from linkage disequilibrium, that allows them to impute missing data in other datasets, usually using Hidden Markov Models on haplotypes [9].

The HLA typing technologies have evolved in the past few years from Sequence-Specific Primers (SSP) and Sequence-Specific Oligonucleotide Probes (SSOP) to Next-Generation Sequencing (NGS) [10]. SSP and SSOP were until fairly recently the best way to detect variations in the MHC but required known constant primers which could fail in the HLA region since some genes can have almost all of their nucleotides display polymorphisms (Single Nucleotide Polymorphism, hereafter SNP) [11]. These old methods also focused mostly on exons 2 and 3 for class I HLA genes (which code for the binding site), or just exon 2 in class II HLA genes, so many recorded HLA alleles are only known from these exons. NGS methods have now become robust enough to be used routinely [12], but are still too expensive for many research groups to afford: the order of magnitude of an HLA typing by sequencing is today 15 euros per allele typing, so

*Correspondence: zagury@cnam.fr
Laboratoire Génomique, Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint-Martin, 75003 Paris, France

120 euros per individual for all 8 class I and class II loci. For panels consisting of thousands of people, this amounts to hundred of thousand euros for a typing by sequencing, while imputation methods and HLA-check allow to use already generated SNP data at no additional cost.

When typing HLA, the level of precision is usually called one-field (previously "2-digit"), two-field (or "4-digit"), or more. The first field indicates the serological antigen carried by an allotype, and the next ones the unique protein sequence. The next fields (not used in this study) indicate synonymous genetic polymorphisms. We'll for instance denote an allele of the *HLA-A* gene at the one-field level as *HLA-A**02 and at the two-field level as *HLA-A**02:01.

Thanks to the availability of large reference panels being genotyped both by genotyping chips (Illumina, Affymetrix, other) and by NGS in the HLA region, several imputation methods have been developed in the past few years: SNP2HLA [13] (modeling HLA alleles as binary SNPs when running imputation software beagle), HIBAG [14] (R package using attribute bagging), or HLA*IMP [15] (Web service now discontinued). They exhibit a fairly good imputation accuracy level in the tests performed, ranging from 90 to 97% according to the HLA gene at stake [13]. Of course, this performance may greatly vary with the reference panel provided, as some studies have shown for instance that using a European panel for a Finnish population may lead to poor results [16]. It is also worth noting that these range of results do not allow any use of these methods in clinical settings, where the costs of HLA typing outlined above are minor compared to the medical consequences of a mistyping.

As discussed in previously published works, there are two important limitations for imputation methods: first, the diversity of the reference panel is crucial for the quality of imputation, and the possibility of errors in the reference panels due to failures in gold-standard typing methods may limit the imputation accuracy.

In the present work, we have developed a new tool which aims at limiting these sources of errors by evaluating the plausibility of the HLA alleles attributed to an individual given his SNP genotypes in the HLA region. With this tool, we could at the same time find errors in reference panels, and also evaluate the soundness of imputed HLA types obtained by any imputation tool. We show that we manage to drastically improve imputation, reaching 99% accuracy for some HLA genes while only eliminating a few individuals, and discuss the possible consequences of these observations.

Implementation

Data material

We primarily used the T1DGC (Type 1 Diabetes Genetics Consortium) cohort as our reference panel [17, 18] of 5225 European unrelated individuals. Genotype data included 7135 SNPs within the MHC region obtained with the Illumina Immunochip platform, and classical HLA allele typing for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, *HLA-DPB1* and *HLA-DRB1* at a two-field resolution. The T1DGC reference panel can be obtained from the NIDDK repository at https://www.niddkrepository.org/niddk/home.do. This panel was the reference also used in previous studies [13], and it will allow for an easier comparison with state-of-the-art tools. We used this panel as provided originally with the SNP2HLA package.

As a testing panel for our imputation method, we used the British 1958 Birth Cohort (1958BC) [19] composed of 2434 individuals genotyped on Illumina HumanHap550 and also typed by gold-standard methods at two-field or one-field levels for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1*. Access to this data was obtained through the Wellcome Trust Case Control Consortium Data Access Committee and could be done from the European Genome-phenome Archive (EGA) at https://www.ebi.ac.uk/ega/. 1958BC was also used as a testing panel in previous studies [13].

These panels cover a variety of existing common alleles in European population: for 1958BC and T1DGC respectively, we have 25 and 51 alleles for *HLA-A*, 44 and 98 for *HLA-B*, 21 and 34 for *HLA-C*, 34 and 52 for *HLA-DRB1*, and 18 and 19 for *HLA-DQB1*. The method used in these panels to attribute HLA alleles has since their publication been shown to cause some systematic errors, for instance *HLA-DRB1*14:54* vs *DRB1*14:01:01* in [20].

We have also used the panel of 5008 haplotypes from various origins [21], assembled by the 1000 Genomes project in which the SNP/indels were phased thanks to the ShapeIT software [22]. This panel was required to extend (by imputation with IMPUTE2 [23]) the SNP coverage of genotypes obtained by chips into the HLA exonic regions, since this step is needed for HLA-check.

We also used the HLA reference database, called IPD-IMGT/HLA Database (http://hla.alleles.org/), version 3.22 (Oct 2015). This database defines at all levels (protein, cDNA, and gDNA) all the known HLA alleles.

HLA typing by imputation

To impute HLA from SNP data, we included all HLA alleles, each of which was represented as a biallelic SNP marker: present or absent. We used the ShapeIT software to phase haplotypes, and IMPUTE2 software to impute the HLA alleles. Then, we kept only alleles with a post-probability dosage of more than 0.5, thus defining the individuals for which HLA was "called". In 99.9% of the cases this indeed gave us exactly two alleles.

Measure of the accuracy of HLA imputation

HLA-check checks if an HLA allele attribution is compatible with the given SNP genotype of an individual. To assess the accuracy of HLA-check and to measure its efficiency, we imputed the HLA of 2434 individuals in the testing panel (1958BC) with ShapeIT and Impute2 as described above. We measured accuracy as the fraction of correctly assigned HLA allele over all called alleles (i.e., discarding alleles with post-probability dosage less than 0.5). We used such a measure keeping in mind the potential applications: indeed, if we impute HLA alleles for an individual, we will be first interested in seeing if there is a called result, second in knowing if this result is actually correct.

HLA-check: the approach

The principle of the method is to compare the SNP genotypes of an individual obtained by any experimental (i.e. chip) or computational method (i.e. imputation) with the SNP genotypes defined by all the known combinations of HLA alleles from the IPD-IMGT/HLA Database (http://hla.alleles.org). We then check if the attributed HLA allele pair is among the best matches by computing a discrepancy measure *D*.

Our approach is straightforward: we first try to impute as many SNPs as possible in the HLA exonic regions, using the best reference set at our disposal. As of today, the 1000 genomes project has the best coverage of SNPs with a large panel of already phased genotypes from various populations all over the genome, but similar results will likely be obtained with the Haplotype Reference Consortium [24] in the near future. This SNP extension phase needs to be done as precisely as possible, in order to get a coverage as precise and as complete as possible in the exonic sequences of the HLA genes.

For an individual, for each SNP in the exonic segments of a given HLA gene, we then compare the post-probabilities genotype obtained after its imputation with the genotypes from all possible allele pairs derived from the HLA reference genome, obtained through the available exonic HLA cDNA sequences of IPD-IMGT/HLA Database (http://hla.alleles.org). We do a per-SNP evaluation of the discrepancy measure where *D*, for a pair of HLA alleles, at a given SNP marker, is the probability of this HLA pair to be incorrect. For instance, take rs41541913. If in the HLA definition, *HLA-A**01:01 has a guanine (G nucleotide) and *HLA-A**80:01 has an cytosine (C nucleotide), and the posterior probabilities for the tested individual are 10%CC, 25%GG and 65%CG, we consider that the imputation post-probabilities indicate $s = 0.35$ of disagreement with the (*HLA-A**01:01, *HLA-A**80:01) HLA allele pair.

Then, *D* obtained for an HLA allele combination for a given individual is simply obtained by summing it for all SNP markers available in the gene[1]:

$$D(genome, HLA_1, HLA_2) = \sum_{rs} D(rs)$$

After computing *D* for all the pairs of HLA alleles of the IPD-IMGT/HLA Database, we compare the discrepancy measure of the attributed HLA of an individual with the best one obtained among all HLA pairs tested: the final *D* we compute is simply the difference between those. A difference of 0 indicates that we reached the least possible discrepancy for the attributed HLA allele pair, making us highly confident in the validity of this HLA typing, while a high discrepancy indicates a mismatch between the best possible HLA fitting with the genotype and the attributed HLA, suggesting that these HLA alleles are unlikely to be well attributed.

Interestingly, our approach depends mainly on the definitions of the HLA alleles from the IPD-IMGT/HLA Database. All HLA allele combinations are tested, and only the quality of imputation in the HLA exons and the SNP coverage will have an impact.

Results

Replication of SNP2HLA results with ShapeIT and IMPUTE2

We first replicated the SNP2HLA method for imputation of HLA alleles, by phasing the same reference panel with the ShapeIT [22] software developed by our group and by imputing SNPs and HLA in the region with the IMPUTE2 software. The results were quite similar to those obtained by SNP2HLA for the quality of imputation (Table 1).

HLA-check: detection of spurious alleles

We also developed a tool, HLA-check, to detect spurious attributed HLA alleles, as described in Material and Methods. This tool relies both on the precise SNP imputation in the exonic parts of HLA genes currently using the 1000 genome reference panel and on the genetic description in the exonic regions of all known HLA allele pairs using the HLA reference database (see pipeline Fig. 1).

We first evaluated the soundness of HLA-check on the T1DGC reference panel that contains 5225 individuals typed at a two-field resolution, who were

Table 1 Comparison of imputation scores using vanilla SNP2HLA vs our method: the results are quite similar

HLA (two-field)	Test population	*D* (ShapeIT/Impute2)	*D* (SNP2HLA)
HLA-A	865	97.6	97.1
HLA-B	1495	96	94.9
HLA-C	813	95.8	95.9
HLA-DRB1	800	87.9	87.8
HLA-DQB1	974	94.8	96.4

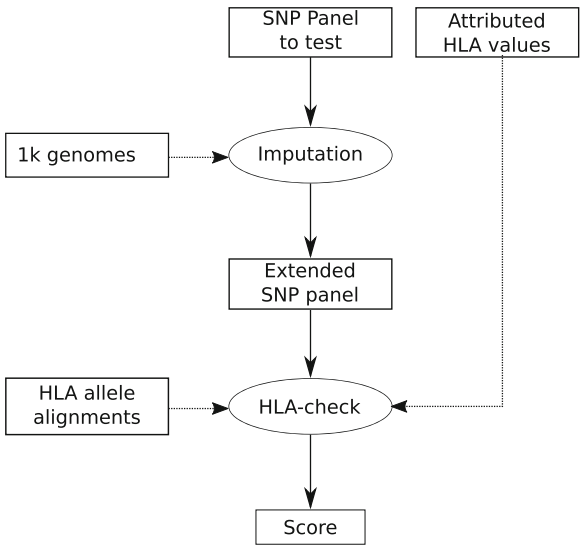

Fig. 1 HLA-check pipeline: We start by augmenting the panel to test to get more SNP using an imputation phase with the 1000genomes data, then we compare those SNP alleles with their theoretical values for all possible HLA allele pairs

also genotyped using an Illumina chip. For that, we compared the discrepancy measures given by HLA-check for the HLA alleles attributed to the T1DGC individuals with those obtained in randomized tests in which the HLA assignments were shuffled between individuals. Figure 2 presents the curves obtained for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* in the original and in the shuffled T1DGC population. Several randomizations of the T1DGC population were tested with identical results.

As expected, we observed a clear-cut difference between the distribution of D for randomized HLA types and the distribution of D for the real population at almost all loci, except for *DRB1*. In *DRB1*, the genotyping by chip is likely of poorer quality since there are several paralogous sequences in the neighborhood (genes *DRB3*, *DRB4*, *DRB5* and pseudo-genes *DRB2*, *DRB6*, *DRB7*, *DRB8*, *DRB9*) and there are also fewer SNPs in the exonic parts of the gene (3 times less than for HLA class 1 genes for instance, cf Table 2). This has also been observed in other works, for instance in HLA*PRG [25] that takes into account paralogous sequences to reach a decent HLA typing rate (from NGS sequences) while observing high similarity between *HLA-DRB1,DRB3,DRB4* and *DRB5*, or in [26].

This precludes the use of HLA-check for *HLA-DRB1* and explains the unsatisfactory overlap observed in D distributions. In *HLA-DQA1*, the small numbers of bars is due to the very small number of alleles. To choose the cutoff value, we modeled the distribution of D attributed to random HLA alleles as a normal distribution and discarded those inferior to $\mu + 2\sigma$ where μ is the mean value and σ the standard deviation. The chosen values for the cutoff were then rounded to 2 for *HLA-A* and *HLA-C*, and to 7 for *HLA-B*. This result shows it is possible to discriminate discrepancy measures likely corresponding to a plausible HLA allele attribution from the ones corresponding to an aberrant one.

Impact of the cleaning of the reference panel on imputation quality

We first thought it was possible to identify aberrant HLA attribution from reference panels and delete them for future HLA imputation studies. We indeed tried this approach on 1958BC using T1DGC as the reference panel, but observed no measurable improvement of the HLA imputation accuracy, likely due to the small numbers of removed individuals at stake. Alternatively, using 1958BC as a reference panel, we see that only in the case of *HLA-C*, which had around 5% of dubious HLA typings (a much higher rate than for other HLA genes), we were able to significantly improve the imputation rate of HLA imputation in the T1DGC panel.

Trading cohort size for precision in HLA imputation

We then used HLA-check to evaluate the credibility of imputed HLA alleles, our goal being to detect and eliminate subjects whose imputed HLA alleles were deemed unlikely, hence improving the accuracy of the HLA imputation.

The first step was to remove from the test population (1958BC) the few individuals for whom D on the real (typed) HLA allele was deemed too high, thus eliminating potential badly typed people from our study. These individuals could have shown up as not imputed properly in the following steps. The number of individuals removed following this initial step are provided for the *HLA-A*, *HLA-B*, and *HLA-C* typed at one-field and two-field (Table 3). *HLA-B* is known to be more difficult to impute and type due to its greater polymorphism and heterogeneity, and we also mirror this observation here by having worse results with it than with other class I loci. We also provide our results for *HLA-DQB1*, even if it does not compare well with the class I HLA genes. We have seen that our approach is not relevant for *HLA-DRB1*.

We then imputed the HLA alleles at one-field and two-field of the 1958BC population using the reference cohort (T1DGC) two-field HLA alleles, and computed D obtained on the imputed HLA. People with too high discrepancy were suspected of having a wrong HLA obtained by the imputation, and were labeled as such. This group was indeed very enriched in wrongly imputed

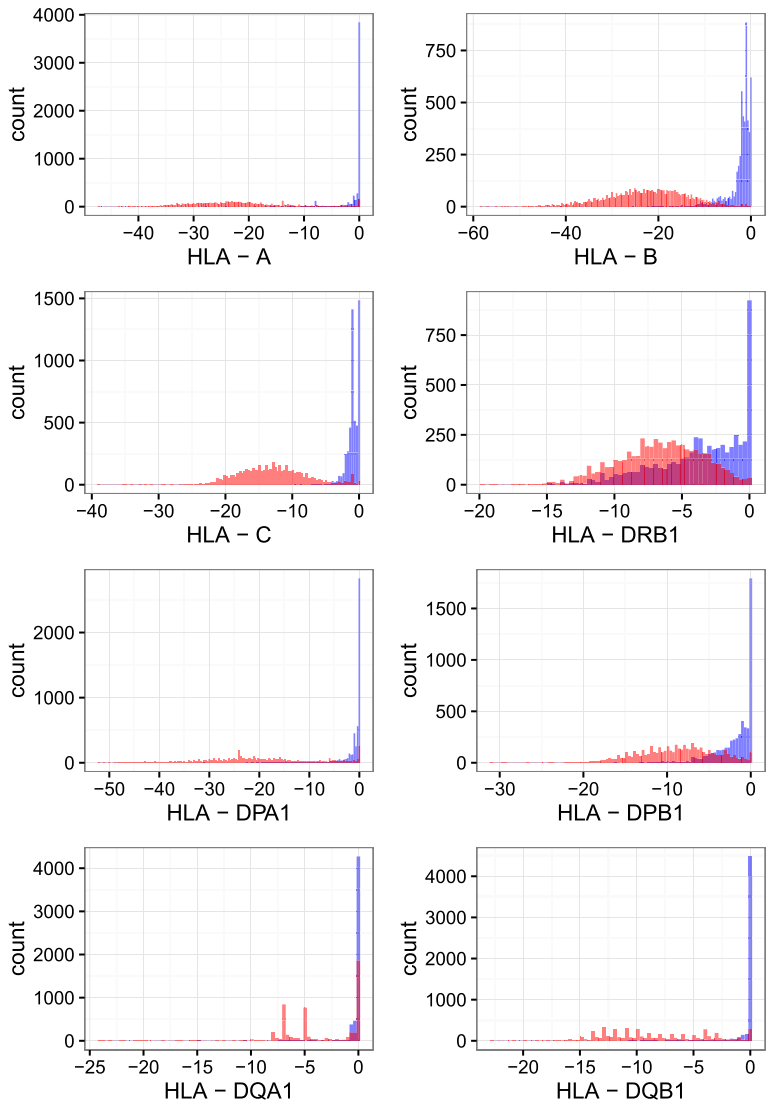

Fig. 2 Histogram of HLA-check distance (*D*) distribution obtained for each HLA locus (*A, B, C, DRB, DPA, DPB, DQA, DQB*) comparing the actual T1DGC HLA types (*blue*) with a randomized set of HLA types from T1DGC (*red*). In the x-axis, the *D* value, in the y-axis, the number of subjects (genotypes) obtaining this value. For each SNP genotype, *D* was computed as described in the Material and Methods. One can see the clear-cut difference between the two distributions for class I HLA genes

HLA alleles (as compared with their HLA attributed by genotyping), and by categorizing them "dubious" we were able to greatly increase the success rate on the remaining test subjects (Table 4). In that table we also give the error rate (rightly typed individuals filtered out by our method).

Discussion

We have developed a simple method to detect aberrant HLA attribution in individuals knowing their SNP genotype in the MHC region. This method is useful for experimentally typed HLA as well as imputed data. In the experimentally typed cohorts, we found very few obvious

Table 2 Number of SNP markers used for each HLA gene (exonic SNPs that can be imputed from 1000genomes)

HLA	*HLA-A*	*HLA-B*	*HLA-C*	*HLA-DRB1*	*HLA-DPA1*	*HLA-DPB1*	*HLA-DQA1*	*HLA-DQB1*
#snps	118	118	110	41	46	42	71	74

Table 3 Test subjects eliminated a priori from the 1958BC test dataset

HLA	*HLA-A*	*HLA-B*	*HLA-C*	*HLA-DQB1*
1958BC (two-field)	18/865	60/1495	77/813	28/974
1958BC (one-field)	35/1669	61/1562	99/1291	103/1701

errors of allele attribution for all HLA genes in T1DGC (less than 1%), significantly more in 1958BC (Table 3). This difference may be explained by the improvement of typing methods between those two cohorts (T1DGC is much more recent). However, after HLA imputation, we could detect several individuals with falsely attributed HLA, and when removing them thanks to HLA-check, we achieved a score of accuracy around 99% for *HLA-A*, *HLA-B*, and *HLA-C* at two-field on the test group 1958BC. These results are quite satisfactory for class I gene alleles since we gain more than 2.5 points of accuracy compared to the use of SNP2HLA alone (Table 4). HLA-check did not yield as good results for class II HLA genes with only a small improvement of accuracy for *HLA-DQB1* (from 94% to 96%) and no improvement at all for *HLA-DRB1*. This latter gene is known to be difficult for genotyping. *HLA-DPA1* and *HLA-DPB1* data were not available for testing (no data in the testing panel). As expected, the results for 1-field typing are even higher, with at least 99.2% accuracy.

Even if we are able to categorize a group of people containing a higher proportion of wrongly imputed, and found possible to identify precisely extreme individuals with clearly false typing, it is still difficult to detect in the people we remove those with false typing from the others (correct people outlined in Table 4), and thus to provide a general model for the detected discrepancies.

There are other tools for HLA imputation such as HIBAG which exhibited quite good performances [14]. We also tried our method on HIBAG for two-field class I HLA genes, and it gave similar results and improvements (Table 5). Note that this is not directly comparable to the other imputation method since we used a pre-built reference file (this tool gives HLA imputation results much faster than other methods, but the reference data for a given chip needs to be computed and is a very time-consuming process), so we could not control the reference panel or use the same as with SNP2HLA.

Conclusion

In the last decade, millions of individuals have been typed through genotyping chips for genetic association studies and the current accuracy of imputation software such as SNP2HLA may limit the statistical power for finding new associations on HLA genes. The use of HLA-check would certainly remove a small proportion of individuals, but could allow a higher accuracy in association detection justifying its use for research purposes. Moreover, these removed individuals could be individually retyped if needed (they are about 5% of typings). To this end, HLA-check can be downloaded for its local use. HLA-check performs very quickly (on a personal computer): only a few seconds per tested individual are needed to obtain the final comparison value for a given HLA. This method should not have a direct impact on HLA typing for medical purposes since current sequencing methods already reach 100% of accuracy at G group [27] level (exons 2 and 3 for class I and exon 2 for class II).

Availability and requirements

HLA-check is available under the MIT license at url https://github.com/mclegrand/HLA-check/. This license expressly allows for any use or modification for one's own needs. It is available as a platform-independent C++11 source code, and can be compiled with openMP to enable threading.

Table 4 Imputation accuracy without any processing, then with the filtering applied with our scoring method

Gene	Test population	Base imputation	People removed	Sub-population imputation
HLA (two-field)				
HLA-A	865	97.6	30 (40% correct)	99.6
HLA-B	1495	96	112 (65% correct)	98.5
HLA-C	813	95.8	81 (60% correct)	99.5
HLA-DQB1	974	94.8	40 (60% correct)	96.1
HLA (one-field)				
HLA-A	1669	97.8	62 (40% correct)	99.9
HLA-B	1562	97.1	119 (70% correct)	99.5
HLA-C	1291	95.8	125 (60% correct)	99.9
HLA-DQB1	1701	98.2	175 (90% correct)	99.2

T1DGC was used as our reference panel and 1958BC as our test panel. We also precise the percentage of *correct* imputations that were removed

Table 5 Imputation accuracy for HIBAG. Unlike SNP2HLA, we did not use T1DGC as a reference panel but a precomputed model due to the way HIBAG works. Nevertheless, our results are very similar to those obtained with SNP2HLA

HLA (two-field)	Test population	Base imputation	People removed	Sub-population imputation
HLA-A	865	97	36	99.5
HLA-B	1495	95	84	97.3
HLA-C	813	95	83	99.7

Endnote

[1] We also tried to compute the sum the log of (1-s) to obtain the combined probability of all markers, but this approach gave a considerable importance to imputation errors and ended up with more imprecise results than the simple sum.

Abbreviations
1958BC: British 1958 Birth cohort; EGA: European Genome-phenome archive; HLA: Human leukocyte antigen; IPD-IMGT: Immuno Polymorphism Database - International ImMunoGeneTics information system; MHC: Major histocompatibility complex; NIDDK: National [USA] Institute of diabetes and digestive and kidney diseases; NGS: Next generation sequencing; SNP: Single nucleotide polymorphism; SSOP: Sequence-specific oligonucleotide probes; SSP: Sequence-specific primers; T1DGC: Type 1 diabetes genetics consortium

Acknowledgments
The authors thank Olivier Delaneau for helpful discussions, and the various sources of data used in this study.

Funding
MJ benefits from a PhD fellowship from the French ministry of education. This funding did not play any role in the study.

Availability of data and materials
T1DGC
The datasets analyzed during the current study were freely available on the SNP2HLA website at the time we got them but were retracted later, so you might be able to obtain them by contacting SNP2HLA authors.
1958BC
The data that support the findings of this study are available from the European Genome-Phenome Archive but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Wellcome Trust Case Control Consortium Data Access Committee.
1000 genomes haplotypes
The datasets analyzed during the current study are available in the impute2 repository, https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
HLA alleles alignments
The datasets analysed during the current study are available via the http://hla.alleles.org/ website.
HLA-check
HLA-check is available under the MIT license at url https://github.com/mclegrand/HLA-check/.

Authors' contributions
MJ wrote the software and performed the analyses. JFZ conceived the study. MJ and JFZ analyzed the results. JN and CC made intellectual contributions for the progress of the study. MJ and JFZ wrote the paper with inputs and corrections from JN and CC. All authors read and approved the final manuscript.

Ethic approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 28 September 2016 Accepted: 26 June 2017
Published online: 11 July 2017

References

1. Janeway CA, Travers P, Walport M, Shlomchik MJ. Immunobiology: the immune system in health and disease. Curr Biol. 1997;1:11.
2. Simmonds MJ, Gough SCL. The HLA region and autoimmune disease: associations and mechanisms of action. Curr Genomics. 2007;8(7):453–65.
3. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Rioux JD. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet. 2008;4(4):e1000024.
4. Hill AV. The immunogenetics of human infectious diseases. Annu Rev Immunol. 1998;16(1):593–617.
5. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. Clin Microbiol Rev. 2009;22(2):370–85.
6. Petersdorf EW. Genetics of graft-versus-host disease: the major histocompatibility complex. Blood Rev. 2013;27(1):1–12.
7. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43(D1):D423-31.
8. Ahmad T, Neville M, Marshall SE, Armuzzi A, Mulcahy-Hawes K, Crawshaw J, Walton R. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Hum Mol Genet. 2003;12(6):647–56.
9. Delaneau O, Zagury JF. Haplotype inference. Data Production Anal Population Genomics: Methods Protoc. 2012;177–96. ISBN 978-1-61779-870-2. http://www.springer.com/in/book/9781617798696.
10. Erlich H. HLA DNA typing: past, present, and future. Tissue antigens. 2012;80(1):1–11.
11. de Bakker PI, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. Hum Mol Genet. 2012;21(R1): R29–36.
12. Cereb N, Kim HR, Ryu J, Yang SY. Advances in DNA sequencing technologies for high resolution HLA typing. Human Immunol. 2015;76(12):923–27.
13. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, de Bakker PI. Imputing amino acid polymorphisms in human leukocyte antigens. PloS ONE. 2013;8(6):e64683.
14. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, Weir BS. HIBAG-HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014;14(2):192.
15. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA* IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011;27(7):968–72.
16. Vlachopoulou E, Lahtela E, Wennerström A, Havulinna AS, Salo P, Perola M, Lokki ML. Evaluation of HLA–DRB1 imputation using a Finnish dataset. Tissue Antigens. 2014;83(5):350–5.
17. Hilner JE, Perdue LH, Sides EG, Pierce JJ, Wägner AM, Aldrich A, Akolkar B. Designing and implementing sample and data collection for an

international genetics study: the Type 1 Diabetes Genetics Consortium (T1DGC). Clinical Trials. 2010;7(1 suppl):S5–32.
18. Rich SS. Special Issue: Fine mapping of the MHC Region for Type 1 diabetes genes. Diabetes Obes Metab 11.s1. 2009. https://www.cabdirect.org/cabdirect/abstract/20093046189.
19. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Saunders G, Laurent T. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015;47(7):692–5.
20. Xiao Y, Lazaro AM, Masaberg C, Haagenson M, Vierra–Green C, Spellman S, Hurley CK. Evaluating the potential impact of mismatches outside the antigen recognition site in unrelated hematopoietic stem cell transplantation: HLA–DRB1* 1454 and DRB1* 140101. Tissue Antigens. 2009;73(6):595–8.
21. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
22. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–81.
23. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955–95.
24. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
25. Dilthey AT, Gourraud PA, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-Accuracy HLA type inference from whole-genome sequencing data using population reference graphs. PLoS Comput Biol. 2016;12(10):e1005151.
26. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, Zheng X. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J. 2017. http://dx.doi.org/10.1038/tpj.2017.7. https://www.ncbi.nlm.nih.gov/pubmed/28440342.
27. Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Vina M, Geraghty DE, Holdsworth R, Hurley CK, Lau M, Lee KW, Mach B, Mayr WR, Maiers M, Müller CR, Parham P, Petersdorf EW, Sasazuki T, Strominger JL, Svejgaard A, Terasaki PI, Tiercy JM, Trowsdale J. Nomenclature for factors of the HLA system. Tissue Antigens. 2010;75:291–455.

Troisième partie

Analyses génomiques de coinfection VIH/VHC

Chapitre 5

VIH, VHC, et coinfection

5.1 Le VIH

5.1.1 La maladie

Le VIH, pour Virus de l'Immunodéficience Humaine, est un rétrovirus affectant le système immunitaire humain, et est responsable du syndrome d'immunodéficience acquise (SIDA). Il a été découvert en 1983[Bar+83] et l'infection cible en priorité les lymphocytes T CD4[Kla+84 ; Zag+86], principaux coordinateurs de l'immunité. En conséquence, l'infection aboutit à la destruction du système immunitaire, perturbé par la colonisation virale[Cof+95]. L'épidémie de SIDA, depuis le début des années 80, s'est étendue au monde entier, et reste un des problèmes majeurs de santé publique dans le monde, avec notamment un comité de l'ONU dédié au SIDA (ONUSIDA[HIV16]) et recensant l'épidémie[UNA09].

Les trois méthodes de transmissions principales du VIH sont la voie sexuelle (principale voie de contamination), la voie sanguine, et la voie verticale (mère-enfant).

5.1.2 Évolution de l'infection

La progression de l'infection est caractérisée par deux facteurs : la charge virale (quantité de virus), et la quantité de lymphocytes T CD4+ dans le sang. Trois phases composent l'infection (Figure 5.1) :

— la primo-infection, généralement sans effet caractéristique particulier, avec une chute brutale de la quantité de lymphocytes et une forte augmentation de la

charge virale ;

— la phase de latence est une phase où l'activité du virus et la réponse immunitaire se compensent : il n'y a pas d'effets phénotypiques, et la diminution du taux de CD4+ comme l'augmentation de la charge virale sont très faibles. Cette phase peut durer plusieurs années sans traitement.

— la phase de SIDA est caractérisée par l'arrivée de maladies opportunistes, prenant avantage de la chute du taux de lymphocytes T CD4+.

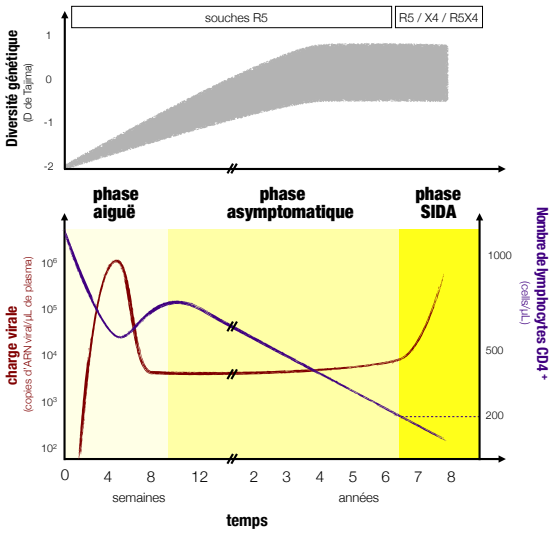

FIGURE 5.1 – Phases du VIH
La phase aiguë correspond à un pic de charge virale et une chute du taux de CD4+, et est suivie d'une longue phase asymptomatique durant laquelle le taux de CD4+ diminue lentement, avant l'entrée en phase SIDA, où la charge virale progresse à nouveau et où le malade devient vulnérable aux maladies opportunistes(tiré de [AM12]).

5.1.3 Physiopathogenèse de l'infection par le VIH-1

Cycle de réplication

Pour ralentir ou contrer l'infection par le VIH, ou pour prolonger la phase de latence (asymptomatique) avant la phase SIDA, les recherches se sont concentrées sur les mécanismes d'action et de réplication du virus[Fre01]. La figure 5.2 détaille les étapes du cycle de réplication du virus avec les protéines humaines impliquées à chaque étape.

Hypothèses de physiopathogenèse

L'infection par le VIH entraîne une un déficit de l'immunité cellulaire causé en grande partie par la baisse de la population de lymphocytes T CD4+, mais également par d'autres mécanismes pas encore élucidés.

De manière directe, on peut expliquer en partie la baisse du taux de CD4+ par une lyse (destruction de la cellule causée par l'attaque de sa membrane) directe des cellules infectées. Cependant, même chez des sujets affectés par le SIDA, seuls environ 10% des CD4+ exprimant le génome viral, ce qui ne suffit pas à expliquer la déplétion observée[BAV91].

Indirectement, plusieurs hypothèses existent pour expliquer la baisse du taux de CD4+ :

- Une réponse spécifique contre le VIH pourrait affecter les CD4+ adsorbant les antigènes des cellules infectées[Wal+87]
- Des superantigènes viraux pourraient activer les lymphocytes T conduisant à l'apoptose (autodestruction) des CD4+[Est+94]
- Certaines protéines virales diminueraient la présence des HLA de classe I à la surface des cellules infectées, les protégeant de la détection par le système immunitaire[GIS98]
- La présence du virus pourrait déséquilibrer la proportion des CD4+ ayant les fonctions Th1 et Th2, cet équilibre étant important pour une réponse immunitaire adaptée[CS93]
- L'attaque spécifique du VIH dans les muqueuses gastro-intestinales peut déclencher la présence dans le sang de bactéries capables de maintenir le système immunitaire en hyperactivité chronique pendant la durée de l'infection[MTS13]

Ces hypothèses sont non-exclusives, et font partie des pistes envisagées pour élucider les mécanismes d'action du VIH.

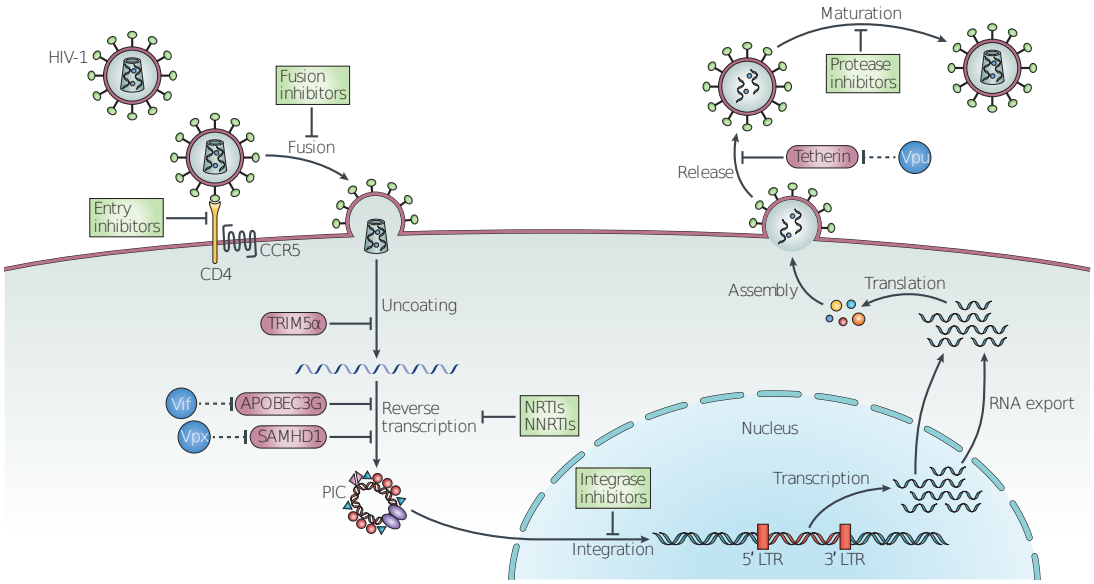

FIGURE 5.2 – Cycle de réplication du VIH
(Tiré de [BRD13])
Cette figure illustre les principales étapes du cycle de réplication du VIH-1 : la liaison aux récepteurs et co-récepteurs du CD4, la fusion avec la membrane, la libération de l'ARN, la transcription inverse de l'ARN du VIH dans l'ADN, puis la translocation dans le noyau. Une fois l'ADN viral intégré à l'ADN hôte, la transcription permet de recréer de l'ARN viral et des protéines, qui viendront se placer à la surface de la cellule pour former des virions immatures. Enfin, ces virions sont relâchés, maturent, et iront infecter d'autres cellules.
Le diagramme indique également les voies d'action des grandes familles d'antirétroviraux (en vert), et les facteurs clés de restriction du VIH avec leurs antagonistes viraux.

5.1.4 Études d'association génétique dans le SIDA

Un grand nombre d'études génétiques ont été menées sur la réponse de l'hôte à l'infection par le VIH, pour aider à comprendre les mécanismes moléculaires mis en jeu dans l'interaction virus-hôte, notamment la déplétion observée du système immunitaire, des études "gènes candidats"[AW10 ; OH13] aux études "génome entier" depuis 2007[Fel+07].

Plusieurs études génome entier ont été menées, entre autres dans le laboratoire GBA, par analyse de la cohorte GRIV composée de patients à profils de progression extrêmes vers le SIDA (comparant des patients progressant très lentement dans la phase de latence ou « non progresseurs », contre des patients à la phase de latence très courte, ou « progresseurs rapides »)[Hen+96 ; Hen+01]. Ces études génome entier se sont appuyées

sur des approches statistiques mais aussi bioinformatiques (MAF faibles, eQTL, etc.) et ont révélé plusieurs gènes potentiellement impliqués dans le développement de la maladie, comme HLA*B57, déjà connu, mais aussi *CXCR6*, *RICH2*, *PRMT6*, *SOX5*, ou *TGFBRAP1* [Lim+09 ; Le +09 ; Lim+10 ; Le +11 ; Spa+15].

5.2 Le VHC

5.2.1 L'hépatite C

L'hépatite C est une maladie du foie causée par un virus. Le VHC (virus de l'hépatite C) est un virus à ARN du genre Hepacivirus de la famille des Flaviviridae, et a été découvert en 1989 après plusieurs années de recherche pour un virus jusque là qualifié d'hépatite "non A non B". Du fait de l'absence de contrôles spécifiques, ce virus a notamment contaminé jusqu'à 90% des hémophiles transfusés pré-1990, et on estime dans le monde à plus de 170 millions le nombre de personnes infectées, avec d'importantes variations suivant les pays[Moh+13], et différents génotypes du virus ont été observés, à des proportions variant également géographiquement (Figure 5.3).

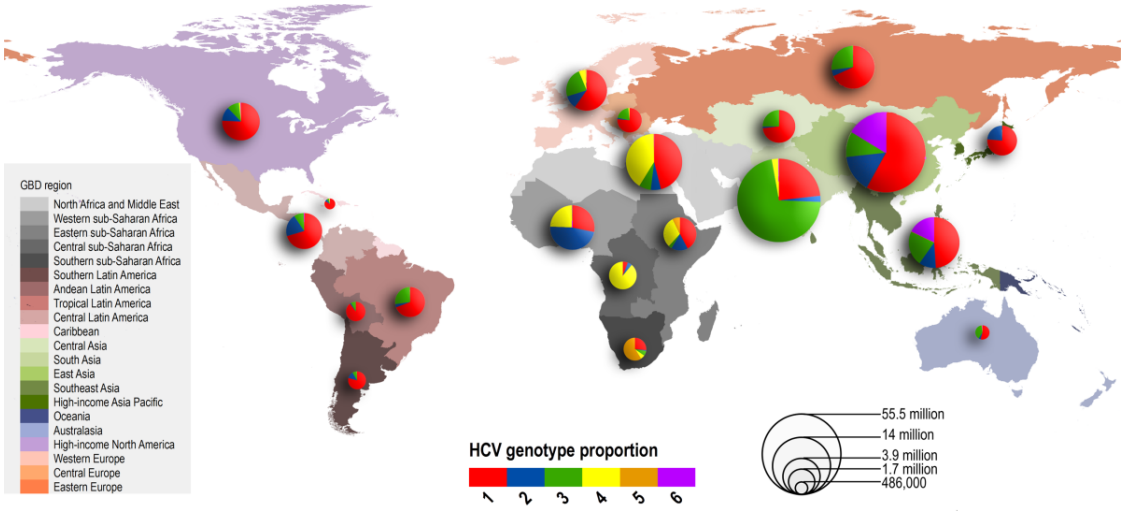

FIGURE 5.3 – Prévalence relative des différents génotypes du VHC par région. La taille des diagrammes est proportionnelle à la prévalence du virus dans la région. (D'après [Mes+15])

Le virus se transmet essentiellement par le sang, que ce soit par transfusion ou injections (par exemple de drogues) avec des seringues contaminées : On estime de 80%

des toxicomanes sont affectés par le virus[Van+90].

5.2.2 Évolution de l'infection

Le virus du VHC infecte principalement les cellules hépatiques, où l'on peut retrouver plusieurs millions de copies d'ARN viral par gramme de tissu infecté, alors qu'il est peu détecté dans le reste du corps. Le virus est responsable à la fois de l'infection aiguë et de l'infection chronique. La forme aiguë de l'infection est dans la majorité des cas asymptomatique. Dans 20 à 50% des cas, les patients vont éliminer le virus, et dans 50 à 80% des cas, les individus avec une hépatite aiguë vont évoluer vers une hépatite chronique. L'infection chronique est caractérisée par la présence en continu de l'ARN du VHC dans le foie. L'infection chronique va entraîner des lésions au niveau du foie qui vont mener à la fibrose des tissus hépatiques, qui pourra mener dans un tiers des cas à la cirrhose, qui peut à son tour évoluer vers un hépatocarcinome cellulaire (HCC). L'évolution de la fibrose est aussi généralement asymptomatique et la détection d'une cirrhose peut être trop tardive et ne pas permettre de traiter les patients. Pour ces raisons, le suivi de ceux-ci est très important, et le développement d'outils de prédiction génétique pourrait largement contribuer à un meilleure évaluation des patients et à un meilleur suivi.

Cette évolution de la fibrose est influencée par un nombre de facteurs importants, comme le sexe, l'âge, ou la surconsommation d'alcool.

5.2.3 Physiopathogenèse de l'infection par le VHC

Cycle de réplication

La figure 5.4 présente les principales étapes du cycle du VHC.

Hypothèses de physiopathogenèse

Les mécanismes d'action du VHC sont mieux connus que ceux du VIH, et principalement indirects : En effet, le virus du VHC n'entraînant pas la destruction des cellules du foie, un des principaux facteurs de la pathogenèse de l'hépatite C chronique est la réponse immunitaire de l'hôte. On sait notamment que l'infection chronique des cellules

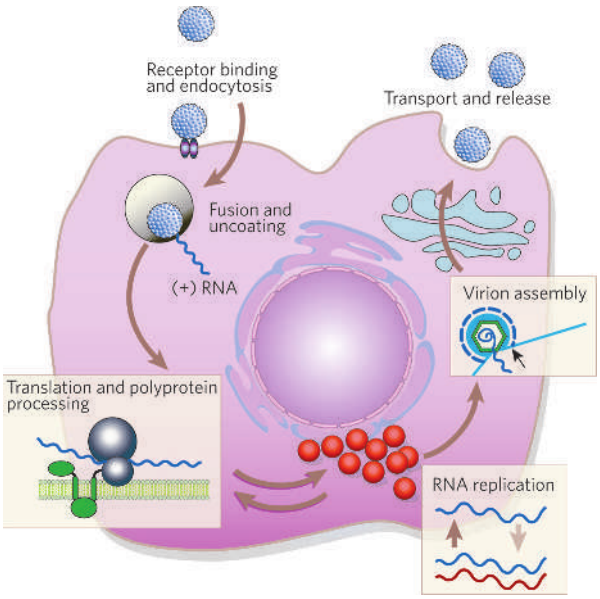

FIGURE 5.4 – Cycle du VHC
Après son entrée dans la cellule et décapsidation, le VHC traduit ses protéines, se réplique, puis reforme de nouveaux virions. [LR05]

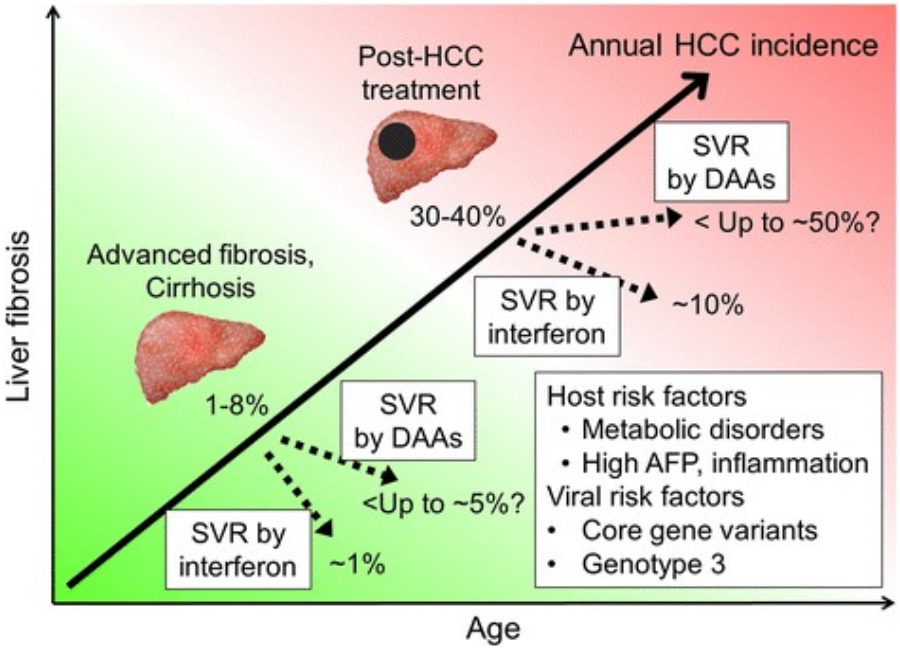

FIGURE 5.5 – Évolution de l'état du foie infecté par le VHC vers l'hépatocarcinome
Tiré de [Bau+17]

hépatiques par le VHC va créer un déséquilibre des messagers de l'immunité (cytokines) qui pourrait aboutir à une réponse immunitaire qualitativement inadaptée à l'élimination

du virus, et être à l'origine d'une réaction inflammatoire excessive favorisant l'évolution de la fibrose au cours de l'infection.[Fie10 ; Sob+01].

5.2.4 Études génomiques sur le VHC

Plusieurs GWAS se sont intéréssé à étudier les variants génétiques associés à la progression du VHC : [Rau+10] pour *IFNL3* et *IFNL4*, [Mik+11] pour *DEPDC5*, [Kum+11] pour *MICA*, [Pat+12] pour *RNF7*, *MERTK*, et *TULP1*, [Ura+13] pour *C6orf10* et *BTNL2*, [Mik+13] pour *HLA-DQ* et [Dug+13] retrouve *IFNL4* et *HLA-DQ*. La figure 5.6 résume les étapes de la progression associées à chaque gène.

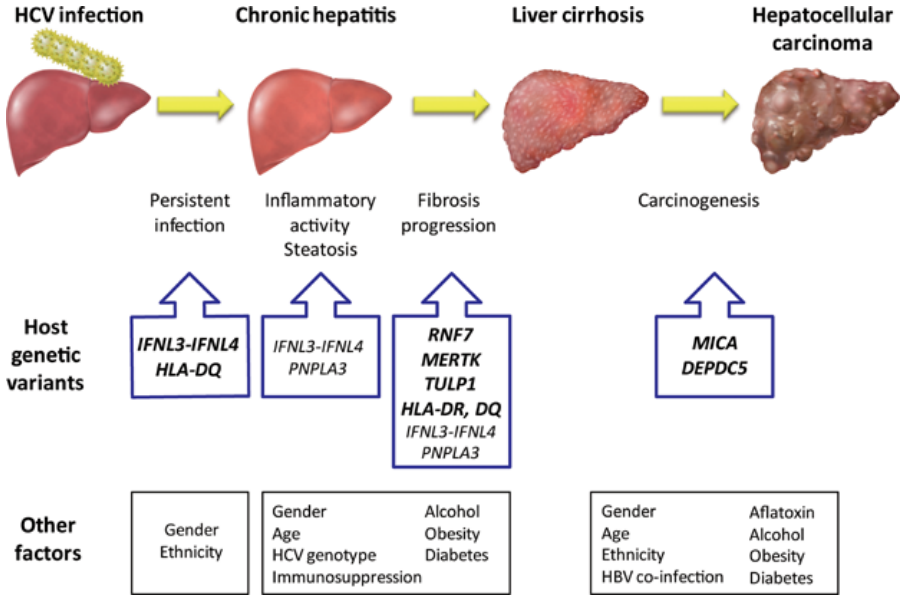

FIGURE 5.6 – Gènes impliqués dans la progression clinique du VHC
Les gènes en gras ont été identifiés par GWAS, les autres par approche gènes candidats.
Tiré de [MT16]

Chapitre 6

Étude METAVIR

6.1 Coinfection VIH/VHC

En raison des modes d'infection similaires pour le VIH et VHC (en particulier par le sang), on retrouve fréquemment une infection VHC par les individus infectés par le VIH. Le facteur de risque le plus important est notamment l'injection de drogue par intraveineuse, pouvant fréquemment transmettre les deux virus simultanément ([Ted+03] note une valeur-p $< 10^{-4}$).

La coinfection avec le VIH a des effets significatifs sur l'action du VHC : L'immunodépression causée par le VIH, ou une interaction encore mal comprise entre les deux virus, provoque une augmentation de la charge virale du VHC[Bel+98]. De plus, certains traitements du VIH par HAART entraînent également une augmentation des niveaux d'ARN du VHC[Mil+05].

Cette augmentation de la charge virale entraîne une accélération de ses effets sur le foie et de la progression de l'hépatite, et reste, avec le SIDA, une des causes principales de décès chez les patients co-infectés VIH/VHC[Lew+08].

Réciproquement, en revanche, l'infection par le VHC ne semble pas avoir d'impact mesuré sur l'évolution de l'infection par le VIH ou pour la phase SIDA.

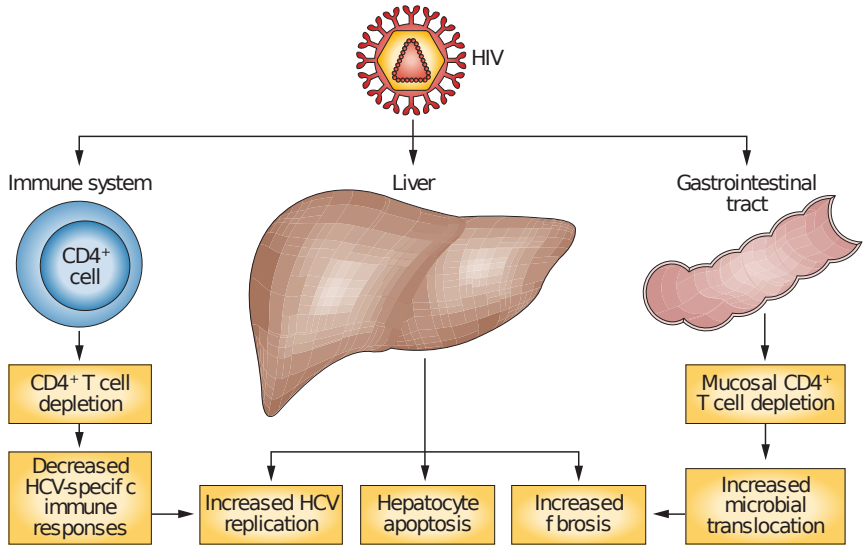

FIGURE 6.1 – Facteurs impactant l'état du foie dans la coinfection VIH-VHC
L'infection VIH entraîne une réaction immunitaire affaiblie contre le VHC, donc une augmentation de la réplication du VHC, une inflammation hépatique, et enfin une fibrose accrue. Tiré de [CFC14]

6.2 La cohorte ANRS CO13–HEPAVIH

En France, estimant à 24% des infectés du VIH les co-infectés par le VHC, l'ANRS (Agence Nationale de Recherche SIDA/hépatites) a créé en 2006 une cohorte d'étude sur la coinfection : HEPAVIH, dont le but est de mieux comprendre les interactions entre les deux virus et leurs traitements, et caractériser l'histoire naturelle de la co-infection VIH-VHC en termes de morbidité et de mortalité et ses déterminants.

La cohorte contient plus d'un millier de patients dont une minorité sont génotypés, suivis dans le temps avec une grande quantité de variables sur l'évolution de leur situation (par exemple, la pression du foie ou le score METAVIR, ou le taux de lymphocytes T CD4+) et leur traitement (prise de bithérapie/trithérapie, molécules utilisées...).

6.2.1 Étude génétique sur la coinfection[Ulv+16]

La première étude génétique sur la coinfection VIH-VHC portant sur la cohorte ANRS HEPAVIH a identifié plusieurs associations génétiques influençant le score METAVIR, mesurant la progression de la fibrose, et a souligné la qualité de la cohorte HEPAVIH.

Nous avons poursuivi les investigations en nous focalisant sur le passage à la cirrhose, en comparant les patient ayant déclenché une cirrhose et les patients à faible fibrose.

Ce faisant, nous avons identifié de nouvelles associations significatives, impliquant les gènes *CTNND2* et *MIR7-3HG*. La publication inclue ci-après détaille nos travaux et est en cours de soumission.

6.2.2 Résumé en français de l'article en soumission

Contexte

Il y a des indices convergents pour indiquer une implication de variants génétiques dans la fibrose hépatique chez les patients infectés par le virus de l'hépatite C (VHC), mais cet aspect a été peu étudié chez les patients co-infectés avec le VHC et le virus de l'immunodéficience humaine (VIH). Nous avons effectué une approche cas-contrôle, comparant les patients avec une fibrose faible et modérée (F0-F2) à sévère (F4), dans une étude d'association génome entier (GWAS), chez les patients coinfectés VIH/VHC de la cohorte française ANRS-CO13 HEPAVIH. La fibrose hépatique a été évaluée en terme de score METAVIR, un score qualitatif, à partir du FibroScan©, FibroTest©, ou biopsie du foie. Après contrôle qualité, l'étude génome entier a été effectuée sur 280 patients caucasiens, comparant 205 patients avec un score METAVIR faible (<F3) avec 75 patients au score METAVIR élevé (F4), sur 8 426 897 SNPs génotypés (Illumina Omni2.5 BeadChip) ou correctement imputés.

Résultats

Trois signaux significatifs au niveau du génome entier ont été obtenus (p-value$< 5 \times 10^{-8}$). Le premier, dans la région 1q31, correspondant à rs72727113 (p-valeur$= 8.18 \times 10^{-9}$), est un SNP isolé sans lien apparent avec un gène. Le deuxième signal, obtenu sur le SNP rs3258099 (p-valeur$= 6.05 \times 10^{-9}$) dans la région 5p15, comprend 5 SNPs introniques du gène *CTNND2*, qui semble impliqué dans les mécanismes biologiques de l'hépatocarcinome. Le troisième signal, obtenu sur le SNP rs74554920 (p-valeur$= 4.79 \times 10^{-8}$), dans la région 19p13, comprend 2 SNPs sur le gène *MIR7-3HG*.

6.3 GWAS study reveals *CTNND2* and *MIR7-3HG* gene polymorphisms associations in liver fibrosis severity and hepatocellular carcinoma in HIV/HCV coinfected patients

Marc Jeanmougin, Damien Ulveling, Sigrid Le Clerc, Taoufik Labib, Josselin Noirel, Vincent Laville, Cédric Coulonges, Wassila Carpentier, Dominique Salmon, François Dabis, Yves Lévy, Stéphanie Dominguez and Jean-François Zagury and the HEPAVIH ANRS CO13 cohort

Keywords : HIV/HCV, Liver Disease, GWAS, METAVIR, *CTNND2*

Abstract

Background : There is growing evidence that human genetic variants contribute to liver fibrosis in subjects with hepatitis C virus (HCV) monoinfection, but this aspect has been little investigated in patients coinfected with HCV and human immunodeficiency virus (HIV). We performed a case control approach comparing mild to moderate fibrosis (F0-F2) to severe fibrosis (F4) using genome-wide association study (GWAS) in HIV-HCV coinfected patients included in the French ANRS CO13 HEPAVIH cohort. Liver fibrosis was evaluated in METAVIR score from FibroScan®, Fibrotest® or liver biopsy, providing a qualitative fibrosis score. After quality control, GWAS was conducted on 280 Caucasian patients, by comparing 205 patients with low METAVIR score (< F3) with 75 patients with high METAVIR score (F4), for a total of 8,426,597 genotyped (Illumina Omni2.5 BeadChip) or reliably imputed SNPs.

Results : Three signals of genome-wide significance (p-value $< 5 \times 10^{-8}$) were obtained. The first, on chromosome 1q31 and corresponding to rs72727113 ($p - value = 8.18 \times 10^{-9}$) a single imputed SNP with no obvious link to a gene. The second signal, obtained through rs2158099 ($p - value = 6.05 \times 10^{-9}$) on chromosome region 5p15, includes 5 intronic variants of the *CTNND2* gene which seems involved in HCV-induced hepatocellular carcinoma mechanisms. The third signal, obtained through rs74554920 ($p - value = 4.79 \times 10^{-8}$) on chromosome region 19p13, includes 2 variants on the

MIR7-3HG gene.

6.3.1 Introduction

One third of human immunodeficiency virus (HIV) infected patients are coinfected with chonic viral hepatitis C (HCV) and dual infection worsen fibrosis progression and liver related diseases compared to HCV monoinfection[Kon+14 ; Sul+07]. Since the introduction of highly active antiretroviral therapy (HAART), the survival of HIV patients has improved while HCV has emerged as a major comorbid disease causing progressive liver disease, potentially leading to end-stage liver disease, hepatocellular carcinoma, or death[Bic+01 ; Ros+03]. Risk factors such as active HCV RNA replication, HIV or HBV coinfection, alcohol consumption, age, and obesity are well know to impact liver fibrosis severity but account for only a small proportion of the variability in liver fibrosis development[Ben+01 ; Rüe+14]. Otherwise access to new direct active drugs against HCV (DAA) with high rate of virological success up to 95% slow down the rate of fibrosis progression and even allow fibrosis regression and prevent the occurrence of cirrhosis complications. However a small percentage of individuals even cured will develop hepatocellular carcinoma. The persistence of comorbidities such as steatohepatitis or alcohol consumption may contribute to maintaining this risk, but host factors must also probably be involved. Indeed there is growing evidence to suggest that host genetic factors are involved in rapid fibrosis progression and hepatocellular carcinoma pathogenesis[Lut+11 ; RBG12].

A high number of genome wide association studies have been performed revealing new factors that influenced treatment efficacy or clinical course in HCV infection. The main association in HCV related phenotypes was a nucleotide polymorphisms (SNPs) near interferon-l4 (IFNL4) locus, known as interleukin-28B (IL28B), with the response to pegylated interferon (PEG-IFN) plus ribavirin (RBV) therapy[Ge+09 ; Sup+09 ; Tan+11], as well as with the spontaneous clearance of HCV[Rau+10 ; Tho+09]. By candidate gene approach, polymorphisms of the IL28B gene have also been associated with liver inflammation and hepatic fibrosis in HCV monoinfection[Abe+10 ; Boc+12]. Three GWAS have focused on the HCV-related liver fibrosis outcome in patients[Pat+12 ; Ura+13 ;

Ulv+16]. The first, in a Caucasian population of HCV monoinfected, identified four susceptibility loci, three of which were linked to genes involved in apoptosis[Pat+12]. The second, conducted in Japan in HCV monoinfected subjects, detected variants within the HLA region[Ura+13]. Finally, our group identified a 3p25 locus associated with liver stiffness evaluated by FibroScan in the HIV/HCV coinfected HEPAVIH ANRS C013 cohort replicated in HCV monoinfection, involving genes linked to cell structure (CAV3) and HCV replication (RAD18)[Ulv+16]. GWAS identified DEPDC5 and MICA as susceptibility genes for HCV-induced hepatocarcinoma in Japanese populations[Kum+11 ; Mik+11]. Regarding the new treatment (DAA) success rate, a group has investigated the development of hepatocellular carcinoma after clearance of HCV infection and they highlighted one SNP located in the gene TLL1.

Regarding the lack of specific investigation of the HIV/HCV coinfected patients and the need of biological markers of liver disease progression, we have investigated the genetic factors the extreme phenotype of the ANRS CO13 HEPAVIH cohort in HIV/HCV coinfected subjects. In the first approach our team was interested in the rate of progression of fibrosis[Ulv+16], in the present study we were interested only in extreme phenotypes. We therefore carried out this GWAS to identify new genetic markers associated with liver fibrosis by comparing subjects with a low METAVIR score (F0F1F2 group) and with subjects with a high METAVIR score (F4 group), in the prospective French HEPAVIH ANRS CO13 cohort of HIV/HCV coinfected patients.

6.3.2 Materials and methods

Study design and population

This study was based on a population sample of 351 patients with HIV/HCV coinfection who had provided written informed genetic consent and were enrolled in the French National Agency for Research on AIDS and Viral Hepatitis (ANRS) CO13 HEPAVIH cohort[Lok+10]. The enrollment criteria were as follows : patients aged 18 years or over, with chronic HIV/HCV coinfection confirmed by a positive test for HIV-1 antibody and by an HCV RNA assay (regardless of clinical stage, sex or transmission group). The dual infection in this cohort was well-characterized because patients had full follow-up medical

visits every six months in cases of cirrhosis and annually for patients without cirrhosis (median follow-up for the cohort : 4 years). A self-administered questionnaire was used to record socio-demographic characteristics and data concerning past and current smoking habits, drug use and alcohol consumption.

351 patients were genotyped, however, after quality control and validation of the dataset information (availability of covariates) ; only 306 patients remained included for the genetic analyses (see below). All these patients are treated with HAART and two-thirds received a pegylated interferon +/- ribavirin treatment for HCV infection (n=200). 27 patients have achieved a sustained virological response (SVR) after treatment before the inclusion in the cohort but continued the monitoring. Because of their low number and in order to assess the fibrosis evolution after SVR, we kept those patients in the analysis. We also assessed the effect of SVR on GWAS results by analyzing the most significant signals by removing of these patients.

Liver histology and staging of liver fibrosis for the binary comparison of low versus high METAVIR score groups

The stage of liver fibrosis in patients coinfected was determined by two concordant evaluations among a liver biopsy, a Fibrotest®, or a FibroScan® , leading to the standardized attribution of the METAVIR score in a five-point scale from F0 to F4[Cas+05]. The distribution of individuals according to their METAVIR score is presented in Supplemental Table A.1 and Supplemental Figure A.1.

Genotyping and quality control procedures

The Illumina Omni2.5 BeadChip (Illumina, San Diego, USA), which contains 2,391,739 markers, was used to genotype all patients. Raw data were first analyzed with Genome Studio software (version 1.6.3 ; Illumina) to obtain genotype calls. 17 samples with a call rate (percentage of SNPs genotyped per sample) $< 95\%$ were removed. Two pairs of individuals displayed high levels of relatedness on analysis with PLINK software[Pur+07].

One patient from each of these pairs was retained for the analysis. One patient was exhibiting an anomalous heterozygous rate and was also excluded. We subsequently

excluded 260,600 SNPs with a low call rate ($< 98\%$), 246 SNPs not in Hardy-Weinberg equilibrium (HWE, $p - value < 1 \times 10^{-6}$) and 751,337 SNPs with a low minor allele frequency (MAF $< 1\%$). In total, 1,372,673 SNPs were retained for imputation procedures and analysis.

Population stratification analysis

The genotypes were analyzed by principal component analysis (PCA), with the SMARTPCA utility of the EIGENSOFT package version 4.2[Pri+06], to correct for possible population stratification. We carried out two rounds of SMARTPCA, which identified 25 outliers. These outliers were excluded from subsequent analyses (Supplemental Figure A.2), which were carried out on a final sample of 306 patients used for the GWAS. A third analysis without outliers was performed to determine the eigenvectors (Supplemental Figure A.2). In the statistical analysis, we used the two first eigenvectors as covariates to correct for population substructure in the association analyses.

Imputation procedures

The genotype data were phased with SHAPEIT2, and the phased data were then imputed with IMPUTE2[DZM13 ; DMZ12]. Haplotype data for 1094 European individuals from the phase 3 integrated variant set of the 1000 Genomes project released in in october 2014[Con+10] were used as a reference panel. After all procedures had been carried out, we retained 8,895,889 SNPs with a high imputation quality score, as measured by the information metric (> 0.9), and with a MAF $> 5\%$.

Covariates used for the association analysis

We focused more specifically on the genetic factors affecting liver fibrosis, by adjusting the regression analyses for possible confounding factors defined at the enrollment date : sex, age (in years), alcohol consumption (never, previous and current), HCV genotype (1 versus other), duration of HIV/HCV infection (in years) and effective time on antiretroviral treatment for HIV infection (in years). The full set of covariates was available of 306 patients. The top two eigenvectors were also used to correct for population substructure.

Statistical tests

Genomic associations tests were performed by the SNPTEST 2.4.1 software[Mar+07], with a complete set of the relevant covariates, in the additive and dominant models. The classical genome-wide significance threshold ($p - values \leq 5 \times 10^{-8}$) was used[BM12 ; Mar+07]. We retained for the analyses all SNPs with an imputation and association score greater than 0.9 and a MAF greater than 5%. In the additive model, we were left with 6,374,602 SNPs, and 6,309,788 in the dominant model. To ensure that the presence of patients with SVR did not impact the results, we also did the analyses without the patients in SVR.

Replication

Replication studies were performed in two cohorts of patients infected with HCV alone. The 22 SNPs with $FDR < 0.15$ in the primary scan were tested for replication in the primary cohort studied in a previous GWAS of liver fibrosis progression[Pat+12]. This replication sample combined data from two cohorts of adult patients of European descent from France (the ANRS Genoscan study group) and Switzerland (the Swiss Hepatitis C Cohort Study) with chronic HCV infection. In these patients, liver fibrosis stage was determined by examining a liver biopsy specimen obtained before treatment METAVIR scores from F0 to F4 were obtained. Genotyping was performed with a combination of Illumina arrays (HumanCNV370-Duo and Human1M-Duo), and imputation was performed with the IMPUTE2 software package and the 1000 Genomes Project haplotypes released on June 16, 2014, in the phase 1 integrated variant set release. FPR was available for 1,064 individuals and was used to test for replication of the 22 SNPs with $FDR < 0.15$.

6.3.3 Results

Sample description

Baseline demographic description of the 306 patients retained for the genetic association study (see Methods) is described in Supplemental Table A.1. The data shown were recorded at the enrollment date of patients. About 75% of patients were coinfected

via intravenous drug injection and the median age at infection was about 20 years. The median duration of coinfection was 23 years. Men accounted for 70% of the patients in the cohort and most patients were infected with HCV genotype 1. Median CD4+ cell count was 454 cells/mm3 and median HIV viral load was 3.7 log viral copies/ml. The median duration of antiretroviral treatment for HIV infection was about 10.3 years.

The phenotype F3 being widely considered as less precisely detected than others (some patients classified as F3 could be in an early cirrhotic state, while others could still regress to F2), we decided to only consider extreme and clearly defined phenotypes, dropping the F3 category (26 patients). Thus, patients were split between F0, F1, F2 group, vs. F4. Sex ratio and mean age did not differ significantly between the two groups. Median CD4+ cell counts were slightly lower for the F4 group than for the F0F1F2 group (p=0.08), whereas there was no significant difference in HIV viral load between these two groups (Supplemental Table A.1). We found that dual infection time was significantly correlated with the F4 phenotype (p=0.001). To ensure that the presence of patients with SVR did not impact the results, we also did the analyses without the 27 patients in SVR, and confirmed that the results did not change significantly.

Associations following the comparison of the F0F1F2 and F4 groups

After the quality-control analysis and imputation procedures, 280 patients remained included in the study for a total of 8 895 889 genotyped or reliably imputed variants. The 205 individuals exhibiting a low METAVIR score (< F3) were compared with the 75 individuals exhibiting a high METAVIR score (F4) in the dominant and in the additive genetic models. The QQ-plot is shown in Supplemental Figure A.3, and the p-value distribution is shown in the Manhattan plot in Supplemental Figure A.4. The inflation factor is close to one ($\lambda = 1.06$ for dominant, and $\lambda = 1.07$ for additive), as expected.

In the additive model, the GWAS identified three variants in two distinct loci showing a genome-wide significance level ($p - values < 5 \times 10^{-8}$), for which information is provided in Table 6.1. The first significant signal, on chromosome 19, corresponded to a cluster of two variants (rs11669331 SNP and rs11669473 indel, r2= 0.91), in high linkage disequilibrium (r2 > 0.8) with 12 SNPs and notably two located in intronic region of

SNP	Chr	Allele	MAF F0F1F2	MAF F4	OR	p-value
rs72727113	1	A	3%	15%	11	1.63×10^{-8}
rs556921174	19	G	4%	16%	8.4	4.79×10^{-8}
rs74554920	19	G	4%	16%	8.4	4.79×10^{-8}

TABLE 6.1 – Associations from with the binary METAVIR F0F1F2/F4 phenotype in additive mode such that $p - value < 5 \times 10^{-8}$, p-values adjusted with covariates (see Methods)

SNP	Chr	Allele	MAF F0F1F2	MAF F4	OR	p-value
rs72727113	1	A	3%	15%	14	7.76×10^{-9}
rs35268064	5	T	53 %	29 %	0.18	1.45×10^{-8}
rs11367210	5	G	55 %	31 %	0.17	9.88×10^{-9}
rs1423493	5	T	55 %	31 %	0.17	9.95×10^{-9}
rs2074260	5	T	54 %	30 %	0.18	3.52×10^{-8}
rs2158099	5	G	55 %	31 %	0.17	9.95×10^{-9}

TABLE 6.2 – Associations from with the binary METAVIR F0F1F2/F4 phenotype in dominant mode such that $p - value < 5 \times 10^{-8}$, p-values adjusted with covariates (see Methods)

the *MIR7-3HG* gene (cf Supplemental Tables A.2 and A.3). The rs11669331-G allele favored a high score of METAVIR with 16% in the F4 group vs. 4.36% in the F0, F1, F2 group (Figure 1a). The second signal is related to the 1q31 chromosome region and corresponded to a single imputed SNP rs72727113 within an intergenic region (beta = 2.4, SE = 0.43, $p - value = 1.6 \times 10^{-8}$). Multivariate analysis confirmed that these association results within the cluster were supported by a single signal.

In the dominant model, the GWAS identified six variants in two distinct loci showing a genome-wide significance level ($p - values < 5 \times 10^{-8}$), for which information is provided in Table 6.2. The first significant signal, for chromosome region 5p15, corresponded to a cluster of five variants (three SNPs and two indels) in linkage disequilibrium (r2 > 0.8) with four SNPs (cf Supplemental Table A.2). All these SNPs are located in intronic part of *CTNND2* (cf Supplemental Table A.3). The rs2158099 SNP G allele promoted a high METAVIR score (Figure 1b). The second signal of the 1q31 was rs72727113 SNP and was already associated in the additive model, here at p-value 7.76×10^{-9}.

We have performed an approach with more extreme phenotypes by removing the F2

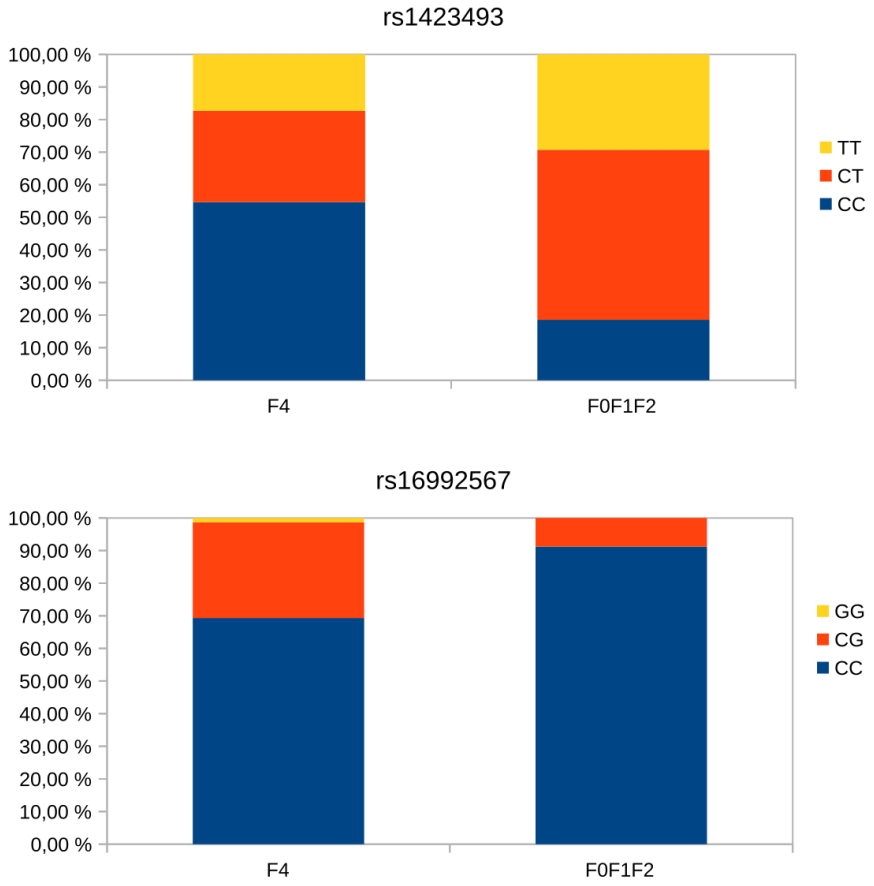

FIGURE 6.2 – Distribution by genotype for the SNP (a) rs1423493 and (b) rs16992567 according to the F0F1F2 and F4 METAVIR score groups.
For each genotype on the x-axis, the METAVIR scores group is indicated on the y-axis, together with the percentage of the subjects in the corresponding group. Cohort size is 205 for F012 group and 75 for the F4 group.

group. In this scenario, only 224 patients remain (75 in the F4 group and 149 in the F0F1 group). The only genome wide significant model is the signal of the 5p15 region in the *CTNND2* gene, in the dominant model.

Investigation of association with a FDR < 0.15

The SNPs with a FDR < 0.15 (supplemental Table 4 and 5), corresponding to 8 clusters, were tested for replication in additive and dominant model in a previous Franco-Swiss GWAS of liver fibrosis in patients infected with HIV alone. The associated clusters and their related genes are display in the supplemental Table 6.

The rs138336017 SNP on chromosome five is the only one replicated in the GENOSCAN cohort and it was imputed in the two cohorts (p = 0.03).

Replication of other study signals

Two previous GWAS were interested in similar phenotypes by comparing METAVIR group[Pat+12 ; Ura+13]. We tried to replicate the main results of these two studies ($p - value < 10^{-6}$), which are summarized in Supplemental Table A.7. We replicated two signals in the 6p21 region, the first one rs3817963 SNP in Patin et al.[Pat+12] and the second one rs9380516 SNP in Urabe et al.[Ura+13] (respectively p-value 0.018 and 0.038, see supplementary Table A.7).

6.3.4 Discussion

We performed a genome wide association study on liver fibrosis using the subjects with extreme METAVIR score and this study has highlighted three major signals never described to date. Two of them have a biological relevance with implications in the pathogenesis of liver fibrosis and hepatocellelar carcinoma. We have detected three significant signals associated with severe fibrosis defined by METAVIR scores in patients with HIV/HCV coinfection. These signals were persistent with or without patients who achieved a SVR after treatment, suggesting mechanisms independent from a sustained virological response. This underscores the importance of monitoring HIV HCV co-infected patients even after achieving a sustained virological response because they may develop

cancer despite HCV cure. We could unfortunately not analyze this population who develop a cancer or a complication of cirrhosis because these patients could not be included in the study because of lost to follow up or death. For this reason we compared the most extreme patients of our cohort with a METAVIR score F4 against the patients with METAVIR score inferior to F3.

The most significant associations found in this GWAS are located in the *CTNND2* gene and include five intronic variants. The *CTNND2* gene encodes for a δ-catenin, an adhesive junction associated protein of the armadillo/β-catenin superfamily. The δ-catenin is important in brain and eye development[Isr+04] but also in cancer formation[Lu10]. Several splice variants were described, and an isoform and dose feedback of the gene expression is important in cellular morphogenesis, apoptosis, and cancer[Zha+10]. Indeed, this gene is involved in the mechanisms of maintenance of adherence junctions which are responsible for cell-cell contact[Rey07]. he δ-catenin affects the localization and stability of p120-catenin (CTNND1) by competitively interacting with E-cadherin that plays an essential role in the formation of adherence junctions and in the function of epithelial cells[Yan+10]. It has been reported that abnormal expression of E-cadherin and p120-catenin in intra-hepatic cholangio-carcinoma is correlated with tumor differentiation, intra-hepatic metastasis, and survival of patients [Zha+08]. Additionally, *CTNND2* is expressed in vascular endothelium and the simple lack of one gene copy is sufficient to impair endothelial cell motility and vascular assembly in vitro and pathological angiogenesis in vivo[DeB+10]. *CTNND2* has also been reported as deregulated in hepatocellular carcinoma[Che+14 ; Wan+13]. Furthermore, δ-catenin has been reported to interact with β-catenin[Lu+99]. This interaction is of particular interest in HCV induced hepatocellular carcinoma since β-catenin together with the E-cadherin play a significant role in liver physiology and pathology through the Wnt/Fzd/β-catenin signaling pathway[RKS15] and they have been placed on the list of seric markers of liver carcinogenesis[RKS15].

Regarding the second signal, LD investigation of the 1q31 region (Supplemental Table A.2) did not reveal any obvious link with a gene or biological function.

Two SNPs located in the *MIR7-3HG* gene were associated significantly with severe fibrosis. *MIR7-3HG* is a RNA gene and it contains the MIR7-3 microRNA. This host

gene and the micro RNA have been involved in the hepatocellular carcninoma[Wan+16 ; Yan+13] Regarding, the success of DAA treatment, the need of biomarker to detect SVR patient who are susceptible to develop a cancer and the recent emergence of miRNA as a possible candidate for treatment or biomarker and a crucial role of them in the progression of the HCV infection (e.g. miR-122)[Shr+15], a deeper investigation of these two genes seems very relevant.

We investigated the SNPs with FDR inferior to 0.15 and we performed a replication analysis in an independent cohort of HCV monoinfected patients. Interestingly we obtained one SNP with a significant p-value. The rs138336017 SNP is located on the chromosome five.

The investigation of the GWAS focusing on similar phenotypes highlighted two SNPs located in the region 6p21, the first one rs3817963 SNP in Patin et al.[Pat+12] and the second one rs9380516 SNP in Urabe et al.[Ura+13] (respectively p-value 0.018 and 0.038, see supplementary Table 7). These results underlined the importance of the 6p21 region for the extreme phenotypes

Our GWAS identified several new loci associated with evolution toward extreme METAVIR score in HIV/HCV coinfected patients, on regions 5p15, 1q31 and 19p13 regions. One result was found in *CTNND2*, a gene involved in the Wnt/Fzd/β-catenin signaling pathway which plays a role in liver cancer physiopathology. The second relevant signal is located in the *MIR7-3HG* gene wich has been link with the hepatocellular carcinoma. These findings provide new insight into the molecular mechanisms of liver fibrosis and hepatocarcinoma pathogenesis in patients with HIV/HCV coinfection. These genomics tools might be useful to better patient's characterization to recommend rapid access to DAA treatment and regular liver function check regardless of HCV cure. As for any GWAS, it will be important to try and replicate these signals in other coinfection cohorts. Several HIV/HCV coinfection cohorts have been described[PK15] and these interesting results may trigger new genetic studies.

Bibliography

[Abe+10] Hiromi Abe, Hidenori Ochi, Toshiro Maekawa, C Nelson Hayes, Masataka Tsuge, Daiki Miki, Fukiko Mitsui, Nobuhiko Hiraga, Michio Imamura, Shoichi Takahashi et al. « Common variation of IL28 affects gamma-GTP levels and inflammation of the liver in chronically infected hepatitis C virus patients ». In : *Journal of hepatology* 53.3 (2010), p. 439-443 (cf. p. 71).

[Ben+01] Yves Benhamou, Vincent Di Martino, Marie Bochet, Geneviève Colombet, Vincent Thibault, Amélie Liou, Christine Katlama et Thierry Poynard. « Factors affecting liver fibrosis in human immunodeficiency virus–and hepatitis C virus–coinfected patients : impact of protease inhibitor therapy ». In : *Hepatology* 34.2 (2001), p. 283-287 (cf. p. 71).

[Bic+01] Ioana Bica, Barbara McGovern, Rakesh Dhar, David Stone, Katherine McGowan, Rochelle Scheib et David R Snydman. « Increasing mortality due to end-stage liver disease in patients with human immunodeficiency virus infection ». In : *Clinical infectious diseases* 32.3 (2001), p. 492-497 (cf. p. 71).

[BM12] William S Bush et Jason H Moore. « Genome-wide association studies ». In : *PLoS computational biology* 8.12 (2012), e1002822 (cf. p. 75).

[Boc+12] Pierre-Yves Bochud, Stéphanie Bibert, Zoltán Kutalik, Etienne Patin, Julien Guergnon, Bertrand Nalpas, Nicolas Goossens, Lorenz Kuske, Beat Müllhaupt, Tillman Gerlach et al. « IL28B alleles associated with poor hepatitis C virus (HCV) clearance protect against inflammation and fibrosis in patients infected with non-1 HCV genotypes ». In : *Hepatology* 55.2 (2012), p. 384-394 (cf. p. 71).

[Cas+05] Laurent Castéra, Julien Vergniol, Juliette Foucher, Brigitte Le Bail, Elise Chanteloup, Maud Haaser, Monique Darriet, Patrice Couzigou et Victor de Lédinghen. « Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C ». In : *Gastroenterology* 128.2 (2005), p. 343-350 (cf. p. 73).

[Che+14] Rongxin Chen, Yinying Dong, Xiaoying Xie, Jie Chen, Dongmei Gao, Yinkun Liu, Zhenggang Ren et Jiefeng Cui. « Screening candidate metastasis-associated genes in three-dimensional HCC spheroids with different metastasis potential ». In : *International journal of clinical and experimental pathology* 7.5 (2014), p. 2527 (cf. p. 80).

[Con+10] 1000 Genomes Project Consortium et al. « A map of human genome variation from population scale sequencing ». In : *Nature* 467.7319 (2010), p. 1061 (cf. p. 74).

[DeB+10] Laura M DeBusk, Kimberly Boelte, Yongfen Min et P Charles Lin. « Heterozygous deficiency of δ-catenin impairs pathological angiogenesis ». In : *Journal of Experimental Medicine* (2010), jem-20091097 (cf. p. 80).

[DMZ12] Olivier Delaneau, Jonathan Marchini et Jean-François Zagury. « A linear complexity phasing method for thousands of genomes ». In : *Nature methods* 9.2 (2012), p. 179-181 (cf. p. 74).

[DZM13] Olivier Delaneau, Jean-Francois Zagury et Jonathan Marchini. « Improved whole-chromosome phasing for disease and population genetic studies ». In : *Nature methods* 10.1 (2013), p. 5-6 (cf. p. 74).

[Ge+09] Dongliang Ge, Jacques Fellay, Alexander J Thompson, Jason S Simon, Kevin V Shianna, Thomas J Urban, Erin L Heinzen, Ping Qiu, Arthur H Bertelsen, Andrew J Muir et al. « Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance ». In : *Nature* 461.7262 (2009), p. 399 (cf. p. 71).

[Isr+04] Inbal Israely, Rui M Costa, Cui Wei Xie, Alcino J Silva, Kenneth S Kosik et Xin Liu. « Deletion of the neuron-specific protein delta-catenin leads to severe cognitive and synaptic dysfunction ». In : *Current Biology* 14.18 (2004), p. 1657-1663 (cf. p. 80).

[Kon+14] Monica A Konerman, Shruti H Mehta, Catherine G Sutcliffe, Trang Vu, Yvonne Higgins, Michael S Torbenson, Richard D Moore, David L Thomas et Mark S Sulkowski. « Fibrosis progression in human immunodeficiency virus/hepatitis C virus coinfected adults : prospective analysis of 435 liver biopsy pairs ». In : *Hepatology* 59.3 (2014), p. 767-775 (cf. p. 71).

[Kum+11] Vinod Kumar, Naoya Kato, Yuji Urabe, Atsushi Takahashi, Ryosuke Muroyama, Naoya Hosono, Motoyuki Otsuka, Ryosuke Tateishi, Masao Omata, Hidewaki Nakagawa et al. « Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma ». In : *Nature genetics* 43.5 (2011), p. 455-458 (cf. p. 72).

[Lok+10] Marc-Arthur Loko, Dominique Salmon, Patrizia Carrieri, Maria Winnock, Marion Mora, Laurence Merchadou, Stéphanie Gillet, Elodie Pambrun, Jean Delaune, Marc-Antoine Valantin et al. « The French national prospective cohort of patients co-infected with HIV and HCV (ANRS CO13 HEPAVIH) : early findings, 2006-2010 ». In : *BMC infectious diseases* 10.1 (2010), p. 303 (cf. p. 72).

[Lu+99] Qun Lu, Mercedes Paredes, Miguel Medina, Jianhua Zhou, Robert Cavallo, Mark Peifer, Lisa Orecchio et Kenneth S Kosik. « δ-Catenin, an adhesive junction–associated protein which promotes cell scattering ». In : *The Journal of cell biology* 144.3 (1999), p. 519-532 (cf. p. 80).

[Lu10] Qun Lu. « δ-Catenin dysregulation in cancer : interactions with E-cadherin and beyond ». In : *The Journal of pathology* 222.2 (2010), p. 119-123 (cf. p. 80).

[Lut+11] P Lutz, JC Wasmuth, HD Nischalke, N Vidovic, F Grünhagei, F Lammert, J Oldenburg, JK Rockstroh, T Sauerbruch et U Spcngler. « Progression of liver fibrosis in HIV/HCV genotype 1 co-infected patients is related to the T allele of the rs I2979860 polymorphism of the

IL28B gene ». In : *European journal of medical research* 16.8 (2011), p. 335 (cf. p. 71).

[Mar+07] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean et Peter Donnelly. « A new multipoint method for genome-wide association studies by imputation of genotypes ». In : *Nature genetics* 39.7 (2007), p. 906 (cf. p. 75).

[Mik+11] Daiki Miki, Hidenori Ochi, C Nelson Hayes, Hiromi Abe, Tadahiko Yoshima, Hiroshi Aikata, Kenji Ikeda, Hiromitsu Kumada, Joji Toyota, Takashi Morizono et al. « Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers ». In : *Nature genetics* 43.8 (2011), p. 797-800 (cf. p. 72).

[Pat+12] Etienne Patin, Zoltán Kutalik, Julien Guergnon, Stéphanie Bibert, Bertrand Nalpas, Emmanuelle Jouanguy, Mona Munteanu, Laurence Bousquet, Laurent Argiro, Philippe Halfon et al. « Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection ». In : *Gastroenterology* 143.5 (2012), p. 1244-1252 (cf. p. 71, 72, 75, 79, 81).

[PK15] Lars Peters et Marina B Klein. « Epidemiology of hepatitis C virus in HIV-infected patients ». In : *Current Opinion in HIV and AIDS* 10.5 (2015), p. 297-302 (cf. p. 81).

[Pri+06] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick et David Reich. « Principal components analysis corrects for stratification in genome-wide association studies ». In : *Nature genetics* 38.8 (2006), p. 904 (cf. p. 74).

[Pur+07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly et al. « PLINK : a tool set for whole-genome association and population-based linkage analyses ». In : *The American Journal of Human Genetics* 81.3 (2007), p. 559-575 (cf. p. 73).

[Rau+10] Andri Rauch, Zoltán Kutalik, Patrick Descombes, Tao Cai, Julia Di Iulio, Tobias Mueller, Murielle Bochud, Manuel Battegay, Enos Bernasconi, Jan Borovicka et al. « Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure : a genome-wide association study ». In : *Gastroenterology* 138.4 (2010), p. 1338-1345 (cf. p. 71).

[RBG12] Monika Rau, Katharina Baur et Andreas Geier. « Host genetic variants in the pathogenesis of hepatitis C ». In : *Viruses* 4.12 (2012), p. 3281-3302 (cf. p. 71).

[Rey07] Albert B Reynolds. « p120-catenin : Past and present ». In : *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1773.1 (2007), p. 2-7 (cf. p. 80).

[RKS15] Karol Rogacki, Aldona Kasprzak et Adrian Stępiński. « Alterations of Wnt/β-catenin signaling pathway in hepatocellular carcinomas associated with hepatitis C virus ». In : *Polish Journal of Pathology* 66.1 (2015), p. 9-21 (cf. p. 80).

[Ros+03] Eric Rosenthal, Marilyne Poiree, Christian Pradier, Christian Perronne, Dominique Salmon-Ceron, Loic Geffray, Robert P Myers, Philippe Morlat, Gilles Pialoux, Stanislas Pol et al. « Mortality due to hepatitis C-related liver disease in HIV-infected patients in France (Mortavic 2001 study) ». In : *Aids* 17.12 (2003), p. 1803-1809 (cf. p. 71).

[Rüe+14] S Rüeger, PY Bochud, Jean-François Dufour, B Müllhaupt, D Semela, MH Heim, D Moradpour, A Cerny, R Malinverni, DR Booth et al. « Impact of common risk factors of fibrosis progression in chronic hepatitis C ». In : *Gut* (2014), gutjnl-2014 (cf. p. 71).

[Shr+15] Shubham Shrivastava, Robert Steele, Ranjit Ray et Ratna B Ray. « MicroRNAs : role in hepatitis C virus pathogenesis ». In : *Genes & diseases* 2.1 (2015), p. 35-45 (cf. p. 81).

[Sul+07] Mark S Sulkowski, Shruti H Mehta, Michael S Torbenson, Yvonne Higgins, Sherilyn C Brinkley, Ruben Montes de Oca, Richard D Moore, Nezam H Afdhal et David L Thomas. « Rapid fibrosis progression among HIV/hepatitis C virus-co-infected adults ». In : *Aids* 21.16 (2007), p. 2209-2216 (cf. p. 71).

[Sup+09] Vijayaprakash Suppiah, Max Moldovan, Golo Ahlenstiel, Thomas Berg, Martin Weltman, Maria Lorena Abate, Margaret Bassendine, Ulrich Spengler, Gregory J Dore, Elizabeth Powell et al. « IL28B is associated with response to chronic hepatitis C interferon-α and ribavirin therapy ». In : *Nature genetics* 41.10 (2009), p. 1100-1104 (cf. p. 71).

[Tan+11] Yasuhito Tanaka, Masayuki Kurosaki, Nao Nishida, Masaya Sugiyama, Kentaro Matsuura, Naoya Sakamoto, Nobuyuki Enomoto, Hiroshi Yatsuhashi, Shuhei Nishiguchi, Keisuke Hino et al. « Genome-wide association study identified ITPA/DDRGK1 variants reflecting thrombocytopenia in pegylated interferon and ribavirin therapy for chronic hepatitis C ». In : *Human molecular genetics* 20.17 (2011), p. 3507-3516 (cf. p. 71).

[Tho+09] David L Thomas, Chloe L Thio, Maureen P Martin, Ying Qi, Dongliang Ge, Colm O'hUigin, Judith Kidd, Kenneth Kidd, Salim I Khakoo, Graeme Alexander et al. « Genetic variation in IL28B and spontaneous clearance of hepatitis C virus ». In : *Nature* 461.7265 (2009), p. 798 (cf. p. 71).

[Ulv+16] Damien Ulveling, Sigrid Le Clerc, Aurélie Cobat, Taoufik Labib, Josselin Noirel, Vincent Laville, Cédric Coulonges, Wassila Carpentier, Bertrand Nalpas, Markus H Heim et al. « A new 3p25 locus is associated with liver fibrosis progression in human immunodeficiency virus/hepatitis C virus-coinfected patients ». In : *Hepatology* 64.5 (2016), p. 1462-1472 (cf. p. 72).

[Ura+13] Yuji Urabe, Hidenori Ochi, Naoya Kato, Vinod Kumar, Atsushi Takahashi, Ryosuke Muroyama, Naoya Hosono, Motoyuki Otsuka, Ryosuke Tateishi, Paulisally Hau Yi Lo et al. « A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region ». In : *Journal of hepatology* 58.5 (2013), p. 875-882 (cf. p. 71, 72, 79, 81).

[Wan+13] Hsei-Wei Wang, Tsung-Han Hsieh, SSu-Yi Huang, Gar-Yang Chau, Chien-Yi Tung, Chien-Wei Su et Jaw-Ching Wu. « Forfeited hepatogenesis program and increased embryonic stem cell traits in young hepatocellular carcinoma (HCC) comparing to elderly HCC ». In : *BMC genomics* 14.1 (2013), p. 736 (cf. p. 80).

[Wan+16] Feiran Wang, Yong Qiang, Lirong Zhu, Yasu Jiang, Yinda Wang, Xian Shao, Lei Yin, Jiahui Chen et Zhong Chen. « MicroRNA-7 downregulates the oncogene VDAC1 to influence hepatocellular carcinoma proliferation and metastasis ». In : *Tumor Biology* 37.8 (2016), p. 10235-10246 (cf. p. 81).

[Yan+10] Ilhwan Yang, Ockyoung Chang, Qun Lu et Kwonseop Kim. « δ-Catenin affects the localization and stability of p120-catenin by competitively interacting with E-cadherin ». In : *Molecules and cells* 29.3 (2010), p. 233-237 (cf. p. 80).

[Yan+13] Hui Yang, Yun Zhong, Hui Xie, XiaoBo Lai, Miqing Xu, Yuqiang Nie, Shiming Liu et Yu-Jui Yvonne Wan. « Induction of the liver cancer-downregulated long noncoding RNA uc002mbe. 2 mediates trichostatin-induced apoptosis of liver cancer cells ». In : *Biochemical pharmacology* 85.12 (2013), p. 1761-1769 (cf. p. 81).

[Zha+08] Bo Zhai, He-Xin Yan, Shu-Qin Liu, Lei Chen, Meng-Chao Wu et Hong-Yang Wang. « Reduced expression of P120 catenin in cholangiocarcinoma correlated with tumor clinicopathologic parameters ». In : *World Journal of Gastroenterology : WJG* 14.23 (2008), p. 3739 (cf. p. 80).

[Zha+10] Jiao Zhang, Jian-Ping Lu, David M Suter, Karl-Heinz Krause, M Elizabeth Fini, Baoan Chen et Qun Lu. « Isoform-and dose-sensitive feedback interactions between paired box 6 gene and δ-catenin in cell differentiation and death ». In : *Experimental cell research* 316.6 (2010), p. 1070-1081 (cf. p. 80).

Quatrième partie

Discussion et Conclusion

Chapitre 7

Discussion

Mon travail de thèse a porté sur deux aspects de la génomique :

— D'une part, j'ai créé un outil, HLA-Check, pour tester et améliorer la fiabilité d'outils d'acquisition du HLA

— D'autre part, j'ai découvert de nouveaux signaux impliqués dans l'évolution de la co-infection VIH/VHC en réalisant des analyses génomiques sur la cohorte HEPAVIH à l'aide d'outils et procédures existantes.

7.1 L'imputation du HLA

Dans le contexte du début de ma thèse, l'outil principal de génotypage des génomes était la puce de génotypage, utilisée pour des études d'association génome entier. Les outils de NGS[GMM16], bien qu'en essor, avaient encore un coût élevé qui ne laissait pas présager la brutale chute des coûts observée en 2015 (figure 7.1). En 2017, Illumina a également annoncé une nouvelle plateforme permettant à court terme le séquençage du génome pour moins de 100$ par génome. De plus, la plupart des études sur des cohortes modestes ne permettent d'identifier que des variants communs, qui sont précisément ceux visés par les puces de génotypages.

Ces évolutions technologiques permettent d'envisager des projets de génotypage de plus en plus ambitieux : depuis HAPMAP (environ 200 individus de 4 populations d'origines ethniques différentes) et 1000Genomes (un millier d'individus de multiples origines ethniques), le NHS (National Health Service, système de santé publique anglais)

travaille désormais sur le « 100 000 genomes project » prévoyant le séquençage de cent mille génomes, uniquement à l'échelle du Royaume-Uni.

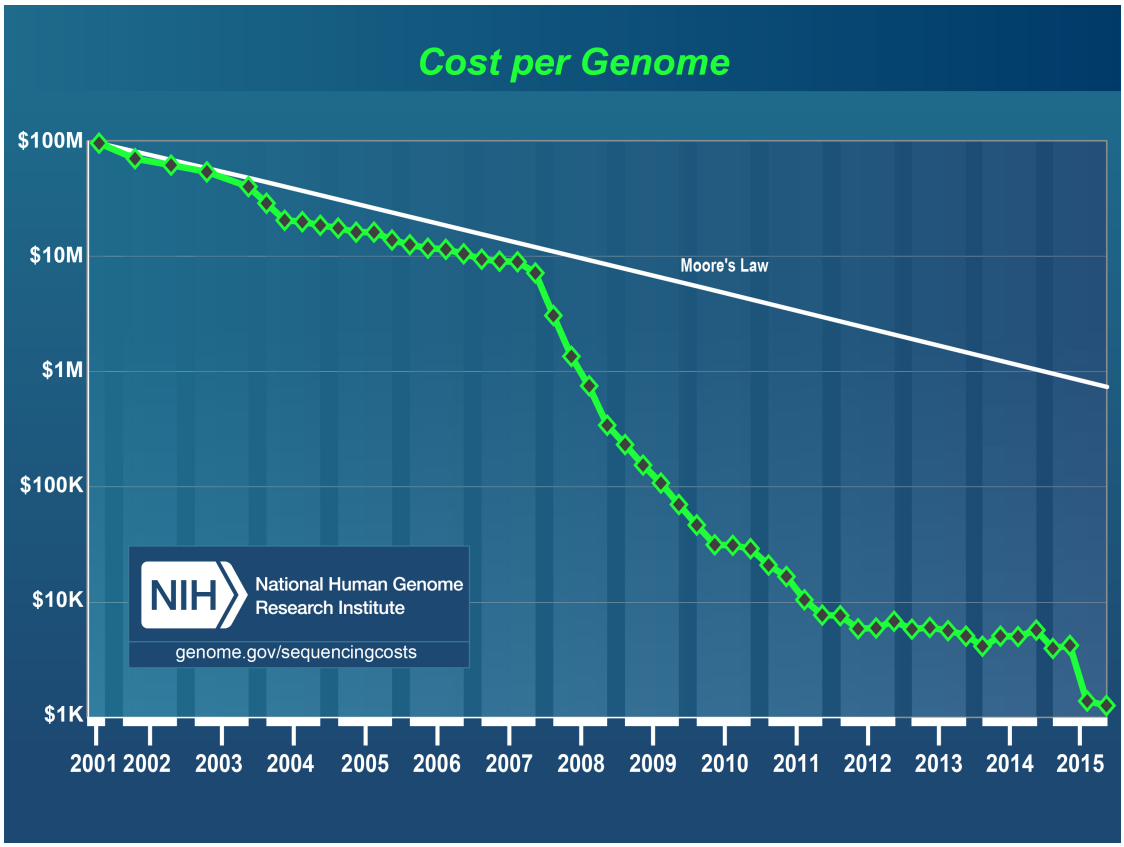

FIGURE 7.1 – Coût du séquençage d'un génome dans le temps

En raison de son polymorphisme élevé, le CMH reste une zone qui bien qu'importante (détectée notamment dans des GWAS du laboratoire GBA sur le VIH), est difficile à génotyper. Les outils d'imputation du HLA à partir de données SNP, tels que SNP2HLA, très récent, sont une piste prometteuse pour la détection des allèles HLA dans les GWAS.

Dans un premier temps, j'ai testé et étudié les différents outils d'imputation du HLA, et essayé d'améliorer l'imputation proposée par SNP2HLA en utilisant les logiciels développés par le laboratoire, supposément plus performants, en l'occurrence Shape-IT[DZM13] et IMPUTE2[How+12] à la place de beagle[BB07]. Les résultats obtenus ont été similaires à ceux d'origine, mais pas significativement supérieurs.

Après une recherche plus approfondie, nous en avons déduit que l'amélioration

apportée par Shape-IT – environ une inversion de mieux toutes les quelques centaines de milliers de paires de bases – correspond à des effets observables sur de grandes régions, tandis que notre étude se concentrait sur quelques dizaines de SNPs (les gènes étudiés ne font que quelques centaines ou milliers de bases) répartis sur une petite région (le MHC ne fait qu'environ 4Mb). Ainsi, cette différence d'échelle explique que l'on n'observe pas de différence significative, les algorithmes employés étant globalement équivalents par ailleurs.

J'ai également étudié le logiciel HIBAG[Zhe+14], qui utilise des méthodes d'*attribute bagging* à la place des modèles de Markov utilisés dans les autres logiciels : il construit un modèle en testant des ensembles de SNPs statistiquement associés aux paires d'allèles HLA dans le panel de test, ne sélectionnant que quelques milliers d'ensembles de SNPs. Pour déterminer le HLA, les ensembles de SNPs sont testés puis calculent à la majorité la paire d'allèles la plus associée au génotype.

Cette méthode rend la construction du modèle très lente (il faut tester de très nombreux ensembles de SNPs, sélectionnés aléatoirement), mais en contrepartie le typage d'un individu est quasi immédiat. De plus, la forme du logiciel (un paquet R) le rend beaucoup plus simple à utiliser que les autres outils. Du point de vue de la précision, HIBAG donne des résultats environ aussi précis que ceux obtenus avec les autres outils.

Observant qu'à partir des mêmes sources de données, les divers outils d'imputation arrivaient à des résultats similaires, j'ai tenté d'améliorer l'imputation en intégrant diverses sources de données, hétérogènes, détaillées dans la table 7.1 : là où un outil d'imputation utilise des SNPs de panels de référence et de test pour déterminer les allèles HLA, HLA-Check utilisera non seulement les SNPs du panel (imputés, donc possédant de l'information du panel de référence d'imputation) et les types HLA à juger, mais aussi les alignements de séquence génomique des allèles HLA connus.

Méthode	Imputation HLA	HLA-Check
Entrée	Panel de référence (SNPs) Panel de test (SNPs)	Typages HLA SNPs des exons (imputés) Alignements HLA
Sortie	HLA imputés	Mesure de confiance dans les HLA

TABLE 7.1 – Imputation et HLA-Check : types de données

J'ai à cet effet utilisé une métrique simple de distance entre une paire d'allèles de gènes et un ensemble de SNPs imputés sur les exons de ces gènes : Pour chaque locus, on prend la probabilité donnée par l'imputation que la paire de nucléotides à ce SNP ne corresponde pas à la paire d'allèles de gène considérée, puis on additionne ces probabilités pour tous les loci disponibles pour obtenir cette distance.

Si d'un point de vue purement mathématique, cette solution n'apparaît pas idéale car elle ne permet pas d'obtenir la probabilité que la paire d'allèles corresponde au génotype, la métrique alternative, l'addition des logarithmes des probabilités (pour avoir la probabilité combinée) ne prend, elle, pas en compte d'une part la non-indépendance des probabilités générées par les outils d'imputation (au contraire, d'ailleurs, la plupart des SNPs proches imputés seraient très dépendants), et d'autre part donnerait un poids énorme à une erreur d'imputation sur un seul locus.

Cette approche a permis d'améliorer la fiabilité de l'imputation HLA en isolant les typages qui, selon notre métrique, apparaissent les moins fiables, et donc de diminuer d'un facteur deux environ la proportion d'individus mal imputés, sans éliminer un nombre rédhibitoire de sujets (moins de 5% en général).

Perspectives

Notre outil permet d'utiliser des grandes cohortes déjà génotypées par le passé pour effectuer des analyses sur le HLA de manière plus précises en utilisant l'imputation : un typage HLA coûte aujourd'hui de l'ordre de la centaine d'euros, donc éviter le typage d'une cohorte de milliers de patients en utilisant l'imputation peut permettre d'économiser des centaines de milliers d'euros.

D'ores et déjà, une équipe de recherche chinoise nous a contacté pour demander des précisions sur l'utilisation du logiciel dans l'une de leurs études, ainsi que de l'aide pour en interpréter les résultats.

De plus, l'approche utilisée dans HLA-Check n'est pas spécifique au HLA : la méthodologie employée n'a besoin que de SNP annotés, et d'alignements relatifs d'allèles connus d'un gène, aussi elle est applicable à tous les autres gènes très polymorphiques

du génome, par exemples les gènes MICA, ou KIR. Le logiciel lui-même n'aura besoin que de très mineures adaptations pour les prendre en compte. De plus, le logiciel étant open source, n'importe quel acteur de la communauté est libre d'y intégrer de nouveaux gènes à traiter ou SNPs utilisés pour ses propres besoins, ou même d'utiliser une autre métrique de distance entre allèles.

En outre, après une GWAS, l'une des études faites dans les régions présentant un intérêt particulier, ou dans les où l'on a trouvé des associations, est une étude dite sur les *haplotypes*. Dans ces études, on s'intéresse aux effets combinés des SNPs que l'on détermine présents sur le même brin du génome, pour savoir si on peut associer non pas un unique variant d'un SNP au phénotype, mais une combinaison de quelques SNPs, qui produiraient, quand ils sont présents sur le même brin d'ADN, un effet phénotypique fort. Or, lors de ces études, l'imprécision de l'imputation et l'imprécision du phasage provoquent souvent la présence d'haplotypes faux, qui quand on considère par ailleurs le grand nombre d'haplotypes possibles ($2^{n_{\text{SNP}}}$), peut rendre la réalisation des tests statistiques peu fiables.

Dans cette optique, notre approche peut se voir comme une généralisation cette analyse sur les gènes HLA entiers : nous cherchons à étudier les haplotypes HLA dans leur globalité, mais au lieu de considérer toutes les combinaisons possibles de tous les SNPs, on n'examine que les allèles HLA que l'on sait exister au niveau biologique, ce qui simplifie considérablement les configurations à tester, et notre outil permet de rapidement détecter les haplotypes les moins sûrs.

Il pourrait donc être très utile, dans le cadre des analyses de GWAS, d'étendre notre outil à tous les gènes. Le principal obstacle à cette extension est l'inexistence d'une base de données exhaustive des allèles connus des gènes, si possible alignés, tel qu'il existe pour les gènes du HLA.

À plus long terme, en revanche, les méthodes d'acquisition ayant tendance à évoluer vers le séquençage nouvelle génération, les données séquencées devraient être de plus en plus fiables et fournir des haplotypes sûrs de plus en plus longs, et l'utilité de HLA-Check

devrait décroître.

7.2 Analyses génomiques de coinfection VIH-VHC

Les études génome entier sont utilisées depuis environ 2008, et ont connu un grand succès. Les raisons sont multiples : d'une part, la démocratisation des coûts de génotypage a permis à de nombreuses équipes de recherche dans le monde entier de génotyper des population (les puces de génotypage permettent maintenant de génotyper rapidement jusqu'à 5 millions de SNPs), et les GWAS, par leur caractère exhaustif *sans a priori* biologique et la relative simplicité des outils permettant de les effectuer (les protocoles de GWAS sont désormais établis, avec des pipelines d'analyse robustes, parfois automatisés dans les cas les plus simples) comme de leurs fondations statistiques, ont permis de les appliquer dans de nombreux domaines (Figure 7.2).

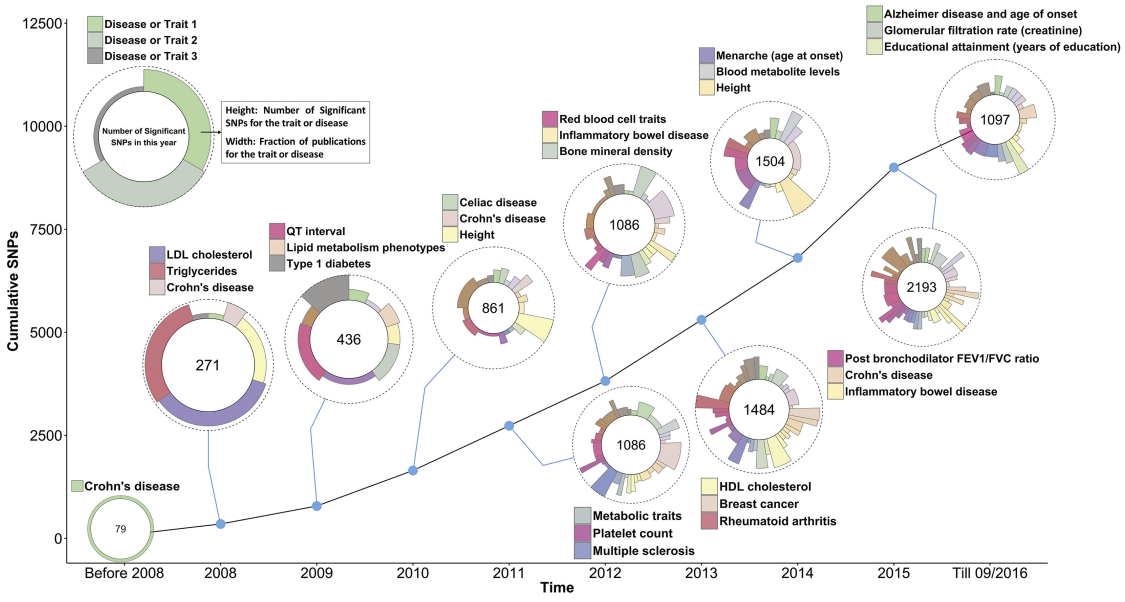

FIGURE 7.2 – Chronologie des découvertes SNP/phénotype par études génome entier Graphique tiré de [Vis+17] : données du GWAS Catalog[Wel+13], comptant uniquement les SNPs à 5.10^{-8}, en enlevant les SNPs en $r^2 > 0.5$

L'étude HEPAVIH

Les mécanismes de la co-infection VIH/VHC étant encore mal compris, et l'équipe GBA ayant une expertise dans les études génomes entier sur les cohortes de maladies, et

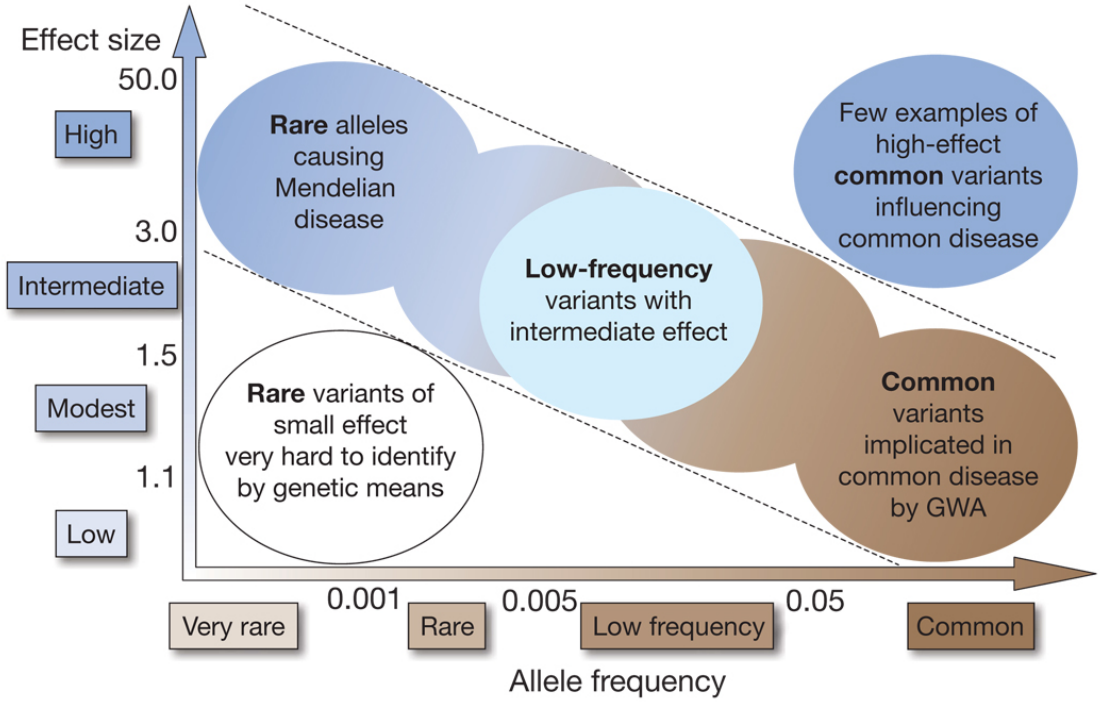

FIGURE 7.3 – Faisabilité de l'identification de variants génétiques par fréquence allélique et force de l'effet génétique (OR)
Plus l'effet est faible, plus la fréquence allélique doit être forte pour détecter l'association, à taille de cohorte égale par ailleurs Graphique tiré de [Man+09]

tout particulièrement sur le VIH, l'équipe a commencé à étudier la cohorte HEPAVIH en collaboration avec l'ANRS dès 2013. Cette collaboration, menée initialement par Damien Ulveling, a donné lieu à une publication[Ulv+16] sur la progression de la fibrose chez les patients co-infectés.

Notre étude s'est davantage orientée sur la fibrose et le passage à la cirrhose : là où [Ulv+16] cherche à étudier la progression tout au long du phénotype, le passage du METAVIR 3(fibrosé, mais non cirrhotique) au METAVIR 4 (cirrhotique), seule étape du processus considérée non réversible, et très variable selon les individus : certains patients passeront rapidement en état cirrhotique, d'autre non.

Nous avons donc procédé à une étude d'association génome entier chez des patients co-infectés VIH/VHC en isolant les patients avec un score METAVIR extrême (une cirrhose), contre ceux qui ont un score METAVIR faible (0 à 2), et l'étude a permis d'isoler trois signaux inédits détaillés ci-après. Deux d'entre eux sont connus pour jouer

a priori un rôle biologique dans le développement de la fibrose ou le cancer du foie. Si nous aurions aimé pouvoir étudier les phénotypes des patients ayant les réponses les plus extrêmes à l'infection comme un cancer ou une cirrhose rapide (en effet, les études sur les phénotypes extrêmes, comme [Le +09] peuvent aider à identifier les allèles les plus impliqués dans un phénotype), leur décès précoce a empêché tout génotypage.

Un des intérêts de la cohorte HEPAVIH est le grand nombre de caractéristiques phénotypiques suivies au cours du temps chez ses membres. Il est donc envisageable d'effectuer d'autres études d'association sur divers phénotypes. En particulier, le laboratoire compte étudier la réponse au traitement du VIH et l'évolution de la charge virale : on pourra ainsi déterminer d'éventuels paramètres génétiques pouvant déterminer pour un individu le traitement le plus susceptible d'être efficace. De plus, nous allons également profiter sous peu d'une mise à jour des informations sur la cohorte permettant d'observer l'évolution des patients sur les dernières années.

Résultats

Le signal le plus important de cette GWAS est situé dans le gène *CTNND2* sur le chromosome 5, avec cinq variants introniques. Ce gène a déjà été associé dans des études sur la formation de cancers[Lu10], la protéine qu'il code (δ-caténine) étant impliquée dans les mécanismes de contact entre cellules[Kos+05]. Notamment, cette protéine interagit directement avec la p120-caténine et la E-cadhérine, deux protéines dont les niveaux d'expression ont été corrélé avec les tumeurs du foie, et la métastase du cancer du foie, ainsi qu'avec le taux de survie des patients.

Le deuxième signal découvert, dans le chromosome 1, semble n'avoir aucun lien avec un gène ni fonction biologique : le gène le plus proche se trouve à plus de 500kb. Les diverses hypothèses pour sa détection sont la possibilité d'un faux positif (la probabilité de faux positif ayant beau être garantie de moins de 5%, elle n'est jamais nulle, et tout signal que l'on trouve dans une étude est toujours susceptible d'être un faux positif) ; mais ce signal peut aussi être révélateur d'un haplotype (si ce SNP est associé à une combinaison de deux allèles de deux gènes alentour, par exemple, sans que l'un ou l'autre, individuellement, ne soient associés au phénotype) ; ou encore correspondre à la possibilité

de nouveaux mécanismes du génome non encore découverts dans cette région.

Le dernier signal est situé dans le gène *MIR7-3HG*, un gène ne codant pas de protéine, mais contenant le micro-ARN MIR-7, qui a déjà été identifié dans des études sur le carcinome hépatocellulaire[Wan+16].

Bien que la région HLA est reliée à certains mécanismes d'action du VIH et du VHC, après avoir tenté d'imputer les allèles HLA, nous n'avons pas pu mettre en évidence d'associations entre un allèle du HLA et le phénotype testé.

Dans une approche exploratoire, cherchant à découvrir des signaux pertinents tout en contrôlant le taux de signaux faux positifs, j'ai également effectué des études de FDR (*false discovery rate*) avec un seuil de 15% (en espérance, on s'attend donc à 15% de faux positifs), avec 8 signaux détectés proches de gènes, dont plusieurs semblent avoir un lien biologique avec le fonctionnement du foie (*MIR7-3HG, KIRREL, FMO11P, CTNND2, STAU2, GYS-1, NRIP-1, LOC105375988*).

Nous avons ensuite contacté une autre cohorte, comportant des patients infectés par le VHC, pour tenter de répliquer ces résultats. L'un de nos signaux, rs138336017, a été répliqué dans la cohorte GENOSCAN avec une p-valeur de 0.03. Ce SNP étant imputé et non génotypé dans nos données, des études du coût de son génotypage ont été entreprises au laboratoire afin de déterminer précisément son degré d'association.

Enfin, nous avons effectué une étude de voies de signalisation, ou *pathways*[Ram+12]. Ces études visent, par exemple à partir des résultats d'une étude génome entier, à déterminer si une voie de signalisation moléculaire est associée au phénotype étudié. On y considère des *ensembles de gènes* avec des points communs, par exemple d'être impliqué dans un processus biologique donné, voire une maladie. On regarde alors si les p-valeurs des SNPs dans cet ensemble de gène diffèrent significativement de celles attendues sous l'hypothèse nulle, pour en extraire la p-valeur de l'ensemble de gènes. Enfin, on détermine si la p-valeur est significative, classiquement (avec une correction statistique si on regarde plusieurs ensembles de gènes, par exemple lorsque l'on teste des bases de données comme KEGG[Kan+09], les pathways de Gene Ontology[Ash+00], ou Reactome[Cro+10]).

J'ai utilisé à cette fin l'outil GSEA[Sub+05]. Les résultats obtenus semblent comporter

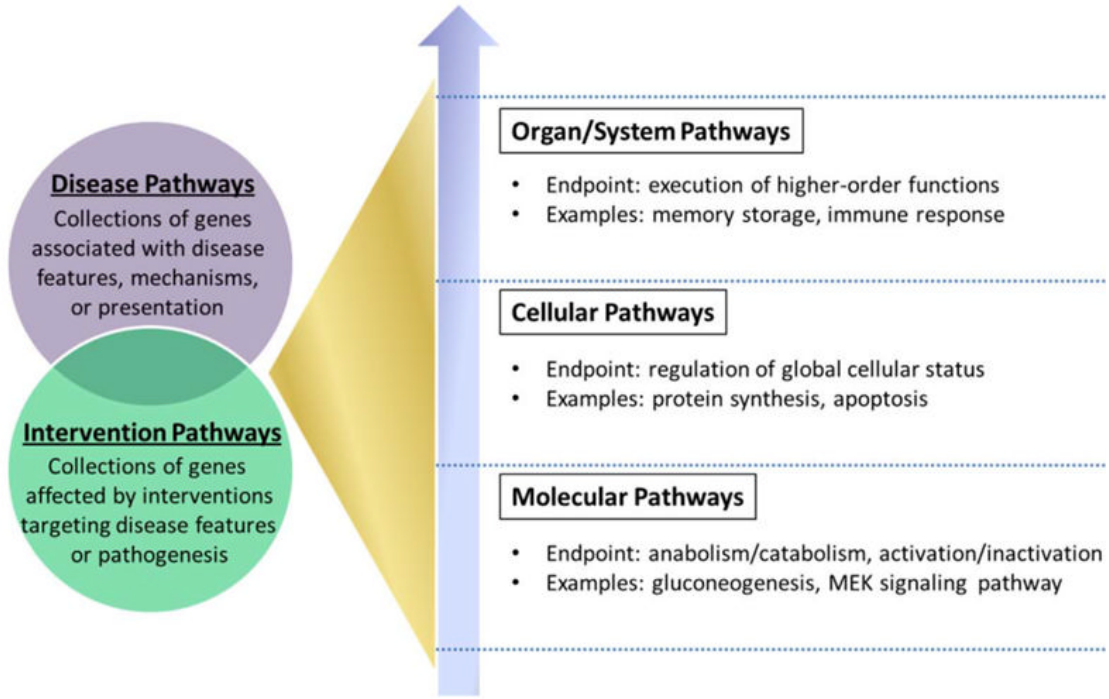

FIGURE 7.4 – Les principaux types de pathways
(Tiré de [Ram+12])

beaucoup de faux positifs, mais il est intéressant de constater qu'un des pathways les plus significatifs est Reactome :Amyloids, un pathway associé à la formation de protéines fibreuses, et donc potentiellement très pertinent pour la fibrose hépatique.

Perspectives

Afin d'établir ou d'infirmer nos signaux, il serait intéressant de chercher à les répliquer dans d'autres cohortes de patients infectés par le VHC. Idéalement, il faudrait tenter de répliquer nos résultats dans d'autres cohortes composées de patients co-infectés par le VIH et VHC, mais en leur absence, après la réplication GENOSCAN, la cohorte utilisée par Urabe[Ura+13] qui s'intéresse également à des phénotypes de fibrose chez des patients monoinfectés, serait un bon candidat pour une seconde réplication.

L'approche par voies de signalisation permet d'identifier des mécanismes biologiques fortement impliqués dans le phénotype : y compris quand les effets des composantes individuelles des éléments de la voie de signalisation sont faibles, une combinaison d'effets

convergents peut se détecter par une telle approche. Nous avons effectué une première analyse, mais il serait intéressant de tester d'autres ensembles de pathways, ainsi que d'autres outils pour les comparer avec le résultat obtenu par GSEA.

Enfin, il est possible que des variants rares soient impliqués dans le déclenchement de la cirrhose. Pour déterminer les variants rares, il serait nécessaire d'effectuer un nouveau génotypage par NGS, mais les coûts actuels liés aux NGS rendent peu probable cette analyse à l'échelle de la cohorte à courte échéance.

Si un traitement pour l'hépatite existe déjà, et les patients sont en général traités et guéris, ces nouvelles études génomiques peuvent aider sur plusieurs points : ils peuvent permettre de trouver des traitements complémentaires, soit plus efficaces (en général, ou seulement sur certaines catégories de patients), soit ayant moins d'effets secondaires, d'analyser d'éventuelles résistances à un traitement, ou encore de trouver des gènes déterminants dans une régression de l'état du patient, et potentiellement importants dans son suivi.

En effet, si nous pouvons détecter les variants qui favorisent le plus une évolution rapide de l'état du foie passant d'une fibrose réversible, à une cirrhose irréversible, il sera d'autant plus simple de pouvoir déterminer pour un patient s'il aura besoin d'un suivi plus particulier au cours de son traitement.

Ces techniques de médecine personnalisée, qui visent à adapter à chacun son suivi et son traitement médical en fonction notamment de ses déterminants génétiques (mais aussi environnementaux), si elles ne sont encore que marginalement utilisées, pourraient être amenées à se généraliser dans le futur en raison de leur intérêt thérapeutique (pour mieux traiter les patients) et financier (pour orienter au mieux le dépistage afin de détecter aussi tôt que possible l'apparition de symptômes chez des patients identifiés comme à risque).

De plus, il existe également une probabilité de régression : certaines personnes en apparence guéries peuvent au bout d'un laps de temps avoir à nouveau une fibrose hépatique, voire une cirrhose. Toujours dans l'optique de se rendre en mesure d'adapter

le suivi des patients (qui peut parfois être contraignant pour ceux-ci) aux risques qu'ils présentent de devoir reprendre un traitement contre le VHC, il serait intéressant d'effectuer des études sur la régression de la fibrose. On pourra ainsi également comparer ces résultats à ceux obtenus pour l'évolution de celle-ci, afin de déterminer la présence d'éventuelles causes communes de ces procédés.

Chapitre 8

Conclusion

Arrivé en 2014 dans une équipe de génomique après un parcours en informatique théorique, j'ai rapidement su m'adapter aux challenges présents dans ce domaine : D'une part, contrairement aux sciences dites « dures », toutes les données manipulées ici sont sujettes à caution et dans la plupart des cas associées à des probabilités. D'autre part, le domaine de la recherche en bioinformatique est extrêmement lié aux technologies pratiques d'acquisition des données dérivées de la biologie moléculaire : l'essor d'une technique de génotypage ou d'une autre qui la supplante (les puces de génotypage versus les NGS, par exemple) conditionne les données, que ce soit leur qualité, leur quantité, et même les outils utilisés pour les manipuler.

Les coûts pratiques d'acquisition ne sont pas non plus à négliger : les données acquises avec des méthodes « passées », tant qu'elles sont traitées avec la rigueur nécessaire, peuvent encore rester utiles, et les rigueurs budgétaires dans le domaine de la recherche contraignent souvent à des économies de moyens entraînant leur réutilisation dans de multiples études.

Ce phénomène est présent dans le cadre des travaux de ma thèse. D'une part, quand je cherche à permettre l'évaluation des imputations du HLA avec les SNPs exoniques disponibles dans HLA-Check, et d'autre part dans l'étude génome entier menée sur HEPAVIH, réutilisant une cohorte déjà utilisée dans une étude précédente[Ulv+16].

Sur cette étude génome entier sur la coinfection VIH/VHC, les résultats obtenus suggèrent de nouveaux gènes potentiellement impliqués dans les mécanismes de la fibrose

pendant la coinfection, qui pourraient se révéler utiles pour la compréhension de ces mécanismes, et *in fine* à l'élaboration de nouvelles pistes diagnostiques ou thérapeutiques.

Dans un domaine où de très nombreux types différents de données cohabitent en étant rarement pris en compte de manière agrégée, j'ai notamment eu l'occasion de démontrer avec HLA-Check que le développement d'outils utilisant des données hétérogènes permet effectivement d'améliorer la pertinence, la précision, ou la confiance dans ses données.

Sur le plan personnel, mon travail de thèse m'a donné la chance de pouvoir mettre à profit certaines de mes compétences informatiques au profit d'un domaine extrêmement dynamique, impliquant à la fois de la génétique, de la statistique, et de la biologie, et ainsi de compléter mon cursus théorique par des applications *Big Data* pratiques, au sein d'un laboratoire dynamique et pluridisciplinaire.

Bibliographie

[Aki+11] Idowu Akinsheye, Abdulrahman Alsultan, Nadia Solovieff, Duyen Ngo, Clinton T Baldwin, Paola Sebastiani, David HK Chui et Martin H Steinberg. « Fetal hemoglobin in sickle cell anemia ». In : *Blood* 118.1 (2011), p. 19-27 (cf. p. 17).

[AM12] Samuel Alizon et Carsten Magnus. « Modelling the course of an HIV infection : insights from ecology and evolution ». In : *Viruses* 4.10 (2012), p. 1984-2013 (cf. p. 60).

[Ash+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig et al. « Gene Ontology : tool for the unification of biology ». In : *Nature genetics* 25.1 (2000), p. 25 (cf. p. 29, 97).

[AW10] Ping An et Cheryl A Winkler. « Host genes associated with HIV/AIDS : advances in gene discovery ». In : *Trends in Genetics* 26.3 (2010), p. 119-131 (cf. p. 62).

[Bal06] David J Balding. « A tutorial on statistical methods for population association studies ». In : *Nature reviews. Genetics* 7.10 (2006), p. 781 (cf. p. 33).

[Bar+09] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras et al. « Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes ». In : *Nature genetics* 41.6 (2009), p. 703-707 (cf. p. 21).

[Bar+83] F Barré-Sinoussi, JC Chermann, F Rey, MT Nugeyre, S Chamaret, J Gruest, C Dauguet, C Axler-Blin, F Vézinet-Brun, C Rouzioux et al. « Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). » In : *Science (New York, NY)* 220.4599 (1983), p. 868 (cf. p. 59).

[Bau+17] Thomas F Baumert, Frank Jühling, Atsushi Ono et Yujin Hoshida. « Hepatitis C-related hepatocellular carcinoma in the era of new generation antivirals ». In : *BMC medicine* 15.1 (2017), p. 52 (cf. p. 65).

[BAV91] JE Brinchmann, J Albert et F Vartdal. « Few infected CD4+ T cells but a high proportion of replication-competent provirus copies in asymptomatic human immunodeficiency virus type 1 infection. » In : *Journal of virology* 65.4 (1991), p. 2019-2023 (cf. p. 61).

[BB07] Sharon R Browning et Brian L Browning. « Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering ». In : *The American Journal of Human Genetics* 81.5 (2007), p. 1084-1097 (cf. p. 30, 45, 90).

[BB16] Brian L Browning et Sharon R Browning. « Genotype imputation with millions of reference samples ». In : *The American Journal of Human Genetics* 98.1 (2016), p. 116-126 (cf. p. 30, 45).

[Bel+98] Marcel Beld, Maarten Penning, Vladimir Lukashov, Martin McMorrow, Marijke Roos, Nadine Pakker, Anneke van den Hoek et Jaap Goudsmit. « Evidence that both HIV and HIV-induced immunodeficiency enhance HCV replication among HCV seroconverters ». In : *Virology* 244.2 (1998), p. 504-512 (cf. p. 67).

[Ben+08] E Andrew Bennett, Heiko Keller, Ryan E Mills, Steffen Schmidt, John V Moran, Oliver Weichenrieder et Scott E Devine. « Active Alu retrotransposons in the human genome ». In : *Genome research* 18.12 (2008), p. 1875-1883 (cf. p. 18).

[Ben81] Baruj Benacerraf. « Role of MHC gene products in immune regulation ». In : *Science* 212.4500 (1981), p. 1229-1238 (cf. p. 21).

[BEV16] Nick H Barton, Alison M Etheridge et Amandine Véber. « The infinitesimal model ». In : *bioRxiv* (2016), p. 039768 (cf. p. 21).

[BGQ03] Robert Bryll, Ricardo Gutierrez-Osuna et Francis Quek. « Attribute bagging : improving accuracy of classifier ensembles by using random feature subsets ». In : *Pattern recognition* 36.6 (2003), p. 1291-1302 (cf. p. 42).

[BGV92] Bernhard E Boser, Isabelle M Guyon et Vladimir N Vapnik. « A training algorithm for optimal margin classifiers ». In : *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, p. 144-152 (cf. p. 46).

[BLP17] Evan A Boyle, Yang I Li et Jonathan K Pritchard. « An Expanded View of Complex Traits : From Polygenic to Omnigenic ». In : *Cell* 169.7 (2017), p. 1177-1186 (cf. p. 21, 22).

[BRD13] Françoise Barré-Sinoussi, Anna Laura Ross et Jean-François Delfraissy. « Past, present and future : 30 years of HIV research ». In : *Nature reviews. Microbiology* 11.12 (2013), p. 877 (cf. p. 62).

[Bro06] Sharon R Browning. « Multilocus association mapping using variable-length Markov chains ». In : *The American Journal of Human Genetics* 78.6 (2006), p. 903-913 (cf. p. 45).

[BS15] Ewan Birney et Nicole Soranzo. « Human genomics : The end of the start for population sequencing ». In : *Nature* 526.7571 (2015), p. 52-53 (cf. p. 28).

[BY09] Brian L Browning et Zhaoxia Yu. « Simultaneous Genotype Calling and Haplotype Phase Inference Improves Genotype Accuracy and Reduces False Positive Associations for Genome-wide Association Studies ». In : *Genetic Epidemiology*. T. 33. 8. WILEY-LISS DIV JOHN WILEY & SONS INC, 111 RIVER ST, HOBOKEN, NJ 07030 USA. 2009, p. 783-783 (cf. p. 41, 45).

[Car+99] Mary Carrington, George W Nelson, Maureen P Martin, Teri Kissner, David Vlahov, James J Goedert, Richard Kaslow, Susan Buchbinder, Keith Hoots et Stephen J O'brien. « HLA and HIV-1 : heterozygote advantage and B* 35-Cw* 04 disadvantage ». In : *Science* 283.5408 (1999), p. 1748-1752 (cf. p. 20, 23).

[CFC14] Jennifer Y Chen, Eoin R Feeney et Raymond T Chung. « HCV and HIV co-infection : mechanisms and management ». In : *Nature reviews Gastroenterology & hepatology* 11.6 (2014), p. 362-371 (cf. p. 68).

[Cha11] Jasvinder Chawla. « Stepwise approach to myopathy in systemic disease ». In : *Frontiers in neurology* 2 (2011) (cf. p. 18).

[Cla+85] Frans HJ Claas, Jan J van der Poel, Ria Castelli-Visser, Jos Pool, Chen Renbiao, Xu Keyu et Jon J van Rood. « Interaction between des-Tyr 1-γ-endorphin and HLA class I molecules : Serological detection of an HLA-A2 subtype ». In : *Immunogenetics* 22.4 (1985), p. 309-314 (cf. p. 40).

[Cof+95] John M Coffin et al. « HIV population dynamics in vivo : implications for genetic variation, pathogenesis, and therapy ». In : *Science-AAAS-Weekly Paper Edition* 267.5197 (1995), p. 483-489 (cf. p. 59).

[Con+04] International Human Genome Sequencing Consortium et al. « Finishing the euchromatic sequence of the human genome ». In : *Nature* 431.7011 (2004), p. 931-945 (cf. p. 6).

[Con+15a] 1000 Genomes Project Consortium et al. « A global reference for human genetic variation ». In : *Nature* 526.7571 (2015), p. 68 (cf. p. 29).

[Con+15b] Gene Ontology Consortium et al. « Gene ontology consortium : going forward ». In : *Nucleic acids research* 43.D1 (2015), p. D1049-D1056 (cf. p. 29).

[CP89] Benjamin P Chen et Peter Parham. « Direct binding of influenza peptides to class I HLA molecules ». In : *Nature* 337.6209 (1989), p. 743-745 (cf. p. 23).

[Cro+10] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal et al. « Reactome : a database of reactions, pathways and biological processes ». In : *Nucleic acids research* 39.suppl_1 (2010), p. D691-D697 (cf. p. 97).

[CS93] Mario Clerici et Gene M Shearer. « A TH1→TH2 switch is a critical step in the etiology of HIV infection ». In : *Immunology today* 14.3 (1993), p. 107-111 (cf. p. 61).

[Dan+11] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry et al. « The variant call format and VCFtools ». In : *Bioinformatics* 27.15 (2011), p. 2156-2158 (cf. p. 30).

[Dau58] Jean Dausset. « Iso-leuco-anticorps ». In : *Acta haematologica* 20.1-4 (1958), p. 156-166 (cf. p. 21).

[Dea+96] Michael Dean, Mary Carrington, Cheryl Winkler, Gavin A Huttley, Michael W Smith, Rando Allikmets, James J Goedert, Susan P Buchbinder, Eric Vittinghoff, Edward Gomperts et al. « Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene ». In : *Science* (1996), p. 1856-1862 (cf. p. 20).

[Dil+11] Alexander T Dilthey, Loukas Moutsianas, Stephen Leslie et Gil McVean. « HLA* IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes ». In : *Bioinformatics* 27.7 (2011), p. 968-972 (cf. p. 42).

[Dil+13] Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R Nelson et Gil McVean. « Multi-population classical HLA type imputation ». In : *PLoS computational biology* 9.2 (2013), e1002877 (cf. p. 42).

[Dil+16] Alexander T Dilthey, Pierre-Antoine Gourraud, Alexander J Mentzer, Nezih Cereb, Zamin Iqbal et Gil McVean. « High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs ». In : *PLoS computational biology* 12.10 (2016), e1005151 (cf. p. 42).

[DMZ12] Olivier Delaneau, Jonathan Marchini et Jean-François Zagury. « A linear complexity phasing method for thousands of genomes ». In : *Nature methods* 9.2 (2012), p. 179-181 (cf. p. 30, 45).

[Dug+13] Priya Duggal, Chloe L Thio, Genevieve L Wojcik, James J Goedert, Alessandra Mangia, Rachel Latanich, Arthur Y Kim, Georg M Lauer, Raymond T Chung, Marion G Peters et al. « Genome-wide association study of spontaneous resolution of hepatitis C virus infection : data from multiple cohorts ». In : *Annals of internal medicine* 158.4 (2013), p. 235-245 (cf. p. 66).

[Dun12] Heather Dunckley. « HLA typing by SSO and SSP methods ». In : *Immunogenetics : Methods and Applications in Clinical Practice* (2012), p. 9-25 (cf. p. 40).

[DZM13] Olivier Delaneau, Jean-Francois Zagury et Jonathan Marchini. « Improved whole-chromosome phasing for disease and population genetic studies ». In : *Nature methods* 10.1 (2013), p. 5-6 (cf. p. 30, 43, 90).

[Erl+08] Henry Erlich, Ana Maria Valdes, Janelle Noble, Joyce A Carlson, Mike Varney, Pat Concannon, Josyf C Mychaleckyj, John A Todd, Persia Bonella, Anna Lisa Fear et al. « HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk ». In : *Diabetes* 57.4 (2008), p. 1084-1092 (cf. p. 23).

[Est+94] JtR6ME Estaquier, Thierry Idziorek, Frédéric De Bels, FRANVOISE Barré-Sinoussi, Bruno Hurtrel, Anne-Marie Aubertin, Alain Venet, Majid Mehtali, Elisabeth Muchmore et Philippe Michel. « Programmed cell death and AIDS : significance of T-cell apoptosis in pathogenic and nonpathogenic primate lentiviral infections ». In : *Proceedings of the National Academy of Sciences* 91.20 (1994), p. 9431-9435 (cf. p. 61).

[Fel+03] Ernst Feldtkeller, Muhammad Khan, Désirée Van Der Heijde, Sjef Van Der Linden et Jürgen Braun. « Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis ». In : *Rheumatology international* 23.2 (2003), p. 61-66 (cf. p. 23).

[Fel+07] Jacques Fellay, Kevin V Shianna, Dongliang Ge, Sara Colombo, Bruno Ledergerber, Mike Weale, Kunlin Zhang, Curtis Gumbs, Antonella Castagna, Andrea Cossarizza et al. « A whole-genome association study of major determinants for host control of HIV-1 ». In : *science* 317.5840 (2007), p. 944-947 (cf. p. 20, 62).

[Fie10] M Isabel Fiel. « Pathology of chronic hepatitis B and chronic hepatitis C ». In : *Clinics in liver disease* 14.4 (2010), p. 555-575 (cf. p. 66).

[Fre01] Eric O Freed. « HIV-1 replication ». In : *Somatic cell and molecular genetics* 26.1-6 (2001), p. 13-33 (cf. p. 60).

[Gib+03] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, FL Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen et al. « The international HapMap project ». In : (2003) (cf. p. 28).

[GIS98] Michael E Greenberg, A John Iafrate et Jacek Skowronski. « The SH3 domain-binding surface and an acidic motif in HIV-1 Nef regulate trafficking of class I MHC complexes ». In : *The EMBO journal* 17.10 (1998), p. 2777-2789 (cf. p. 61).

[GMM16] Sara Goodwin, John D McPherson et W Richard McCombie. « Coming of age : ten years of next-generation sequencing technologies ». In : *Nature Reviews Genetics* 17.6 (2016), p. 333-351 (cf. p. 89).

[Gri+06] Sam Griffiths-Jones, Russell J Grocock, Stijn Van Dongen, Alex Bateman et Anton J Enright. « miRBase : microRNA sequences, targets and gene nomenclature ». In : *Nucleic acids research* 34.suppl_1 (2006), p. D140-D144 (cf. p. 6).

[HBE05] John L Hopper, D Timothy Bishop et Douglas F Easton. « Population-based family studies in genetic epidemiology ». In : *The Lancet* 366.9494 (2005), p. 1397-1406 (cf. p. 17, 19).

[HDM09] Bryan N Howie, Peter Donnelly et Jonathan Marchini. « A flexible and accurate genotype imputation method for the next generation of genome-wide association studies ». In : *PLoS genetics* 5.6 (2009), e1000529 (cf. p. 30, 45).

[Hea+17] National Institutes of Health et al. « Talking glossary of genetic terms ». In : *Retrieved June* 8 (2017), p. 2017 (cf. p. 9).

[Hec99] Stephen S Hecht. « Tobacco smoke carcinogens and lung cancer ». In : *JNCI : Journal of the National Cancer Institute* 91.14 (1999), p. 1194-1210 (cf. p. 19).

[Hen+01] Houria Hendel, Cheryl Winkler, Ping An, Elisabeth Roemer-Binns, George Nelson, Philippe Haumont, Steve O'brien, Kamel Khalilli, Daniel Zagury, Jay Rappaport et al. « Validation of Genetic Case-Control Studies in AIDS and Application to the CX3CR1 Polymorphism [Epidemiology] ». In : *JAIDS Journal of Acquired Immune Deficiency Syndromes* 26.5 (2001), p. 507-511 (cf. p. 62).

[Hen+96] H Hendel, YY Cho, N Gauthier, J Rappaport, F Schächter et JF Zagury. « Contribution of cohort studies in understanding HIV pathogenesis : introduction of the GRIV cohort and preliminary results ». In : *Biomedicine & pharmacotherapy* 50.10 (1996), p. 480-487 (cf. p. 62).

[Hen+99] Houria Hendel, Sophie Caillat-Zucman, Hélene Lebuanec, Mary Carrington, Steve O'Brien, Jean-Marie Andrieu, François Schächter, Daniel Zagury, Jay Rappaport, Cheryl Winkler et al. « New class I and II HLA alleles strongly associated with opposite patterns of progression to AIDS ». In : *The Journal of Immunology* 162.11 (1999), p. 6942-6946 (cf. p. 20, 23).

[Het+02] Seth Hetherington, Arlene R Hughes, Michael Mosteller, Denise Shortino, Katherine L Baker, William Spreen, Eric Lai, Kirstie Davies, Abigail Handley, David J Dow et al. « Genetic variations in HLA-B region and hypersensitivity reactions to abacavir ». In : *The Lancet* 359.9312 (2002), p. 1121-1122 (cf. p. 20).

[Hil+10] Joan E Hilner, Letitia H Perdue, Elizabeth G Sides, June J Pierce, Ana M Wägner, Alan Aldrich, Amanda Loth, Lotte Albret, Lynne E Wagenknecht, Concepcion Nierras et al. « Designing and implementing sample and data collection for an international genetics study : the Type 1 Diabetes Genetics Consortium (T1DGC) ». In : *Clinical trials* 7.1_suppl (2010), S5-S32 (cf. p. 29).

[HIV16] UN Joint Programme on HIV/AIDS (UNAIDS). « Global AIDS Update - 2016 ». In : *Online at http://www.refworld.org/docid/574e8d394.html* (June 2016) (cf. p. 59).

[Hog+08] Clive J Hoggart, Taane G Clark, Maria De Iorio, John C Whittaker et David J Balding. « Genome-wide significance for dense SNP and resequencing data ». In : *Genetic epidemiology* 32.2 (2008), p. 179-185 (cf. p. 33).

[How+12] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini et Gonçalo R Abecasis. « Fast and accurate genotype imputation in genome-wide association studies through pre-phasing ». In : *Nature genetics* 44.8 (2012), p. 955 (cf. p. 90).

[Hüt+09] Gero Hütter, Daniel Nowak, Maximilian Mossner, Susanne Ganepola, Arne Müssig, Kristina Allers, Thomas Schneider, Jörg Hofmann, Claudia Kücherer, Olga Blau et al. « Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation ». In : *New England Journal of Medicine* 360.7 (2009), p. 692-698 (cf. p. 20).

[Ill] Illumina. *Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the $ 100 Genome*. URL : https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383 (cf. p. 28).

[Iva+98] Rayna Ivanova, Nicolas Hénon, Virginia Lepage, Dominique Charron, Eric Vicaut et François Schächter. « HLA-DR alleles display sex-dependent effects on survival and discriminate between individual and familial longevity ». In : *Human molecular genetics* 7.2 (1998), p. 187-194 (cf. p. 23).

[Jea+17] Marc Jeanmougin, Josselin Noirel, Cédric Coulonges et Jean-François Zagury. « HLA-check : evaluating HLA data from SNP information ». In : *BMC bioinformatics* 18.1 (2017), p. 334 (cf. p. 47).

[Jia+13] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J Concannon, Stephen S Rich, Soumya Raychaudhuri et Paul IW de Bakker. « Imputing amino acid polymorphisms in human leukocyte antigens ». In : *PloS one* 8.6 (2013), e64683 (cf. p. 41).

[JSG82] MD JORDON, PETER STASTNY et JAMES N GILLIAM. « Serologic and HLA associations in subacute cutaneous lupus erythematosus, a clinical subset of lupus erythematosus ». In : *Annals of Internal Medicine* 97 (1982), p. 664-671 (cf. p. 23).

[Kan+09] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe et Mika Hirakawa. « KEGG for representation and analysis of molecular networks involving diseases and drugs ». In : *Nucleic acids research* 38.suppl_1 (2009), p. D355-D360 (cf. p. 97).

[Kas+96] Richard A Kaslow, Mary Carrington, R Apple, L Park, A Munoz, AJ Saah, James J Goedert, Cheryl Winkler, Stephen J O'Brien, Charles Rinaldo et al. « Influence of combinations of human major histocompatibility complex genes on the course of HIV–1 infection ». In : *Nature medicine* 2.4 (1996), p. 405-411 (cf. p. 20).

[Kla+06] Lars Klareskog, Patrik Stolt, Karin Lundberg, Henrik Källberg, Camilla Bengtsson, Johan Grunewald, Johan Rönnelid, Helena Erlandsson Harris, Ann-Kristin Ulfgren, Solbritt Rantapää-Dahlqvist et al. « A new model for an etiology of rheumatoid arthritis : smoking may

trigger HLA–DR (shared epitope)–restricted immune reactions to autoantigens modified by citrullination ». In : *Arthritis & Rheumatology* 54.1 (2006), p. 38-46 (cf. p. 23).

[Kla+84] David Klatzmann, Eric Champagne, Sophie Chamaret, Jacqueline Gruest, Denise Guetard, Thierry Hercend, Jean-Claude Gluckman et Luc Montagnier. « T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV ». In : *Nature* 312.5996 (1984), p. 767-768 (cf. p. 59).

[Kor+94] Julie R Korenberg, XN Chen, R Schipper, Z Sun, R Gonsky, S Gerwehr, N Carpenter, C Daumer, P Dignan et C Disteche. « Down syndrome phenotypes : the consequences of chromosomal imbalance ». In : *Proceedings of the National Academy of Sciences* 91.11 (1994), p. 4997-5001 (cf. p. 16).

[Kos+05] Kenneth S Kosik, Christine P Donahue, Inbal Israely, Xin Liu et Tomoyo Ochiishi. « δ-Catenin at the synaptic–adherens junction ». In : *Trends in cell biology* 15.3 (2005), p. 172-178 (cf. p. 96).

[Kum+11] Vinod Kumar, Naoya Kato, Yuji Urabe, Atsushi Takahashi, Ryosuke Muroyama, Naoya Hosono, Motoyuki Otsuka, Ryosuke Tateishi, Masao Omata, Hidewaki Nakagawa et al. « Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma ». In : *Nature genetics* 43.5 (2011), p. 455-458 (cf. p. 66).

[Lan+01] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh et al. « Initial sequencing and analysis of the human genome ». In : (2001) (cf. p. 26).

[Lap+15] Ilkka Lappalainen, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding et al. « The European Genome-phenome Archive of human data consented for biomedical research ». In : *Nature genetics* 47.7 (2015), p. 692 (cf. p. 29).

[LDM08] Stephen Leslie, Peter Donnelly et Gil McVean. « A statistical method for predicting classical HLA alleles from SNP data ». In : *The American Journal of Human Genetics* 82.1 (2008), p. 48-56 (cf. p. 41).

[Le +09] Sigrid Le Clerc, Sophie Limou, Cédric Coulonges, Wassila Carpentier, Christian Dina, Lieng Taing, Olivier Delaneau, Taoufik Labib, Rob Sladek, ANRS Genomic Group et al. « Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03) ». In : *The Journal of infectious diseases* 200.8 (2009), p. 1194-1201 (cf. p. 63, 96).

[Le +11] 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 et al. « Screening low frequency SNPS from genome wide association study reveals a new risk allele for progression to AIDS ». In : *Journal of acquired immune deficiency syndromes (1999)* 56.3 (2011), p. 279 (cf. p. 63).

[LE97] Todd M Lowe et Sean R Eddy. « tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence ». In : *Nucleic acids research* 25.5 (1997), p. 955-964 (cf. p. 6).

[Lew+08] Charlotte Lewden, Thierry May, Eric Rosenthal, Christine Burty, Fabrice Bonnet, Dominique Costagliola, Eric Jougla, Caroline Semaille, Philippe Morlat, Dominique Salmon et al. « Changes in causes of death among adults infected by HIV between 2000 and 2005 : the "Mortalite 2000 and 2005" surveys (ANRS EN19 and Mortavic) ». In : *JAIDS Journal of Acquired Immune Deficiency Syndromes* 48.5 (2008), p. 590-598 (cf. p. 67).

[Lim+09] Sophie Limou, Sigrid Le Clerc, Cédric Coulonges, Wassila Carpentier, Christian Dina, Olivier Delameau, Taoufik Labib, Lieng Taing, Rob Sladek, ANRS Genomic Group et al. « Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02) ». In : *The Journal of infectious diseases* 199.3 (2009), p. 419-426 (cf. p. 20, 23, 63).

[Lim+10] Sophie Limou, Cédric Coulonges, Joshua T Herbeck, Daniëlle Van Manen, Ping An, Sigrid Le Clerc, Olivier Delaneau, Gora Diop, Lieng Taing, Matthieu Montes et al. « Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS ». In : *The Journal of infectious diseases* 202.6 (2010), p. 908-915 (cf. p. 63).

[Lok+10] Marc-Arthur Loko, Dominique Salmon, Patrizia Carrieri, Maria Winnock, Marion Mora, Laurence Merchadou, Stéphanie Gillet, Elodie Pambrun, Jean Delaune, Marc-Antoine Valantin et al. « The French national prospective cohort of patients co-infected with HIV and HCV (ANRS CO13 HEPAVIH) : early findings, 2006-2010 ». In : *BMC infectious diseases* 10.1 (2010), p. 303 (cf. p. 29).

[LR05] Brett D Lindenbach et Charles M Rice. « Unravelling hepatitis C virus replication from genome to function ». In : *Nature* 436.7053 (2005), p. 933 (cf. p. 65).

[LTG59a] Jerome Lejeune, Raymond Turpin et Marthe Gautier. « Le mongolisme, maladie chromosomique (trisomie) ». In : *Bull Acad Natl Med* 143.11-12 (1959), p. 256-265 (cf. p. 15).

[LTG59b] JTRGM Lejeune, Raymond Turpin et M Gautier. « Le mongolisme, premier exemple d'aberration autosomique humaine ». In : *Ann Genet* 1.4 (1959), p. 1-49 (cf. p. 8).

[Lu10] Qun Lu. « δ-Catenin dysregulation in cancer : interactions with E-cadherin and beyond ». In : *The Journal of pathology* 222.2 (2010), p. 119-123 (cf. p. 96).

[LV86] Richard M Lawn et Gordon A Vehar. « The molecular genetics of hemophilia ». In : *Scientific American* 254.3 (1986), p. 48-57 (cf. p. 17).

[Mac+17] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales et al. « The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog) ». In : *Nucleic acids research* 45.D1 (2017), p. D896-D901 (cf. p. 16).

[Man+09] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti et al. « Finding the missing heritability of complex diseases ». In : *Nature* 461.7265 (2009), p. 747-753 (cf. p. 21, 95).

[Man09] Teri A Manolio. « Cohort studies and the genetics of complex disease ». In : *Nature genetics* 41.1 (2009), p. 5-6 (cf. p. 32).

[MD11] Shuangge Ma et Ying Dai. « Principal component analysis based methods in bioinformatics studies ». In : *Briefings in bioinformatics* 12.6 (2011), p. 714-722 (cf. p. 34).

[Mes+15] Jane P Messina, Isla Humphreys, Abraham Flaxman, Anthony Brown, Graham S Cooke, Oliver G Pybus et Eleanor Barnes. « Global distribution and prevalence of hepatitis C virus genotypes ». In : *Hepatology* 61.1 (2015), p. 77-87 (cf. p. 63).

[Mig+00] Stephen A Migueles, M Shirin Sabbaghian, W Lesley Shupert, Maria P Bettinotti, Francesco M Marincola, Lisa Martino, Clair W Hallahan, Sara M Selig, David Schwartz, John Sullivan et al. « HLA B* 5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors ». In : *Proceedings of the National Academy of Sciences* 97.6 (2000), p. 2709-2714 (cf. p. 20).

[Mik+11] Daiki Miki, Hidenori Ochi, C Nelson Hayes, Hiromi Abe, Tadahiko Yoshima, Hiroshi Aikata, Kenji Ikeda, Hiromitsu Kumada, Joji Toyota, Takashi Morizono et al. « Variation in the DEPDC5 locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers ». In : *Nature genetics* 43.8 (2011), p. 797-800 (cf. p. 66).

[Mik+13] Daiki Miki, Hidenori Ochi, Atsushi Takahashi, C Nelson Hayes, Yuji Urabe, Hiromi Abe, Tomokazu Kawaoka, Masataka Tsuge, Nobuhiko Hiraga, Michio Imamura et al. « HLA-DQB1* 03 confers susceptibility to chronic hepatitis C in Japanese : a genome-wide association study ». In : *PloS one* 8.12 (2013), e84226 (cf. p. 66).

[Mil+05] Melissa Farmer Miller, Clinton Haley, Margaret James Koziel et Christopher F Rowley. « Impact of hepatitis C virus on immune restoration in HIV-infected patients who start highly active antiretroviral therapy : a meta-analysis ». In : *Clinical Infectious Diseases* 41.5 (2005), p. 713-720 (cf. p. 67).

[MMR97] AJ Mighell, AF Markham et PA Robinson. « Alu sequences ». In : *FEBS letters* 417.1 (1997), p. 1-5 (cf. p. 17).

[Moh+13] Khayriyyah Mohd Hanafiah, Justina Groeger, Abraham D Flaxman et Steven T Wiersma. « Global epidemiology of hepatitis C virus infection : New estimates of age-specific antibody to HCV seroprevalence ». In : *Hepatology* 57.4 (2013), p. 1333-1342 (cf. p. 63).

[Mor+02] Yasuo Morishima, Takehiko Sasazuki, Hidetoshi Inoko, Takeo Juji, Tatsuya Akaza, Ken Yamamoto, Yoshihide Ishikawa, Shunichi Kato, Hiroshi Sao, Hisashi Sakamaki et al. « The clinical significance of human leukocyte antigen (HLA) allele compatibility in patients receiving a marrow transplant from serologically HLA-A, HLA-B, and HLA-DR matched unrelated donors ». In : *Blood* 99.11 (2002), p. 4200-4206 (cf. p. 38).

[MT16] Kentaro Matsuura et Yasuhito Tanaka. « Host genetic variants influencing the clinical course of hepatitis C virus infection ». In : *Journal of medical virology* 88.2 (2016), p. 185-195 (cf. p. 66).

[MTS13] Giulia Marchetti, Camilla Tincati et Guido Silvestri. « Microbial translocation in the pathogenesis of HIV infection and AIDS ». In : *Clinical microbiology reviews* 26.1 (2013), p. 2-18 (cf. p. 61).

[Nej+07] Sergey Nejentsev, Joanna MM Howson, Neil M Walker, Jeffrey Szeszko, Sarah F Field, Helen E Stevens, Pamela Reynolds, Matthew Hardy, Erna King, Jennifer Masters et al. « Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A ». In : *Nature* 450.7168 (2007) (cf. p. 23).

[Ng+08] Pauline C Ng, Samuel Levy, Jiaqi Huang, Timothy B Stockwell, Brian P Walenz, Kelvin Li, Nelson Axelrod, Dana A Busam, Robert L Strausberg et J Craig Venter. « Genetic variation in an individual human exome ». In : *PLoS genetics* 4.8 (2008), e1000160 (cf. p. 33).

[NIH16] NIH. *The Cost of Sequencing a Human Genome*. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. Updated : 2016-07-06. 2016 (cf. p. 28).

[Nov+08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson et al. « Genes mirror geography within Europe ». In : *Nature* 456.7218 (2008), p. 98 (cf. p. 13, 14).

[OH13] Stephen J O'Brien et Sher L Hendrickson. « Host genomic influences on HIV/AIDS ». In : *Genome biology* 14.1 (2013), p. 201 (cf. p. 62).

[Pat+12] Etienne Patin, Zoltán Kutalik, Julien Guergnon, Stéphanie Bibert, Bertrand Nalpas, Emmanuelle Jouanguy, Mona Munteanu, Laurence Bousquet, Laurent Argiro, Philippe Halfon et al. « Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection ». In : *Gastroenterology* 143.5 (2012), p. 1244-1252 (cf. p. 66).

[Pat+13] Nikolaos A Patsopoulos, Lisa F Barcellos, Rogier Q Hintzen, Catherine Schaefer, Cornelia M Van Duijn, Janelle A Noble, Towfique Raj, Pierre-Antoine Gourraud, Barbara E Stranger, Jorge Oksenberg et al. « Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis : HLA and non-HLA effects ». In : *PLoS genetics* 9.11 (2013), e1003926 (cf. p. 23).

[PIP11] Orestis A Panagiotou, John PA Ioannidis et Genome-Wide Significance Project. « What should the genome-wide significance threshold be ? Empirical replication of borderline genetic associations ». In : *International journal of epidemiology* 41.1 (2011), p. 273-286 (cf. p. 33).

[Pri+06] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick et David Reich. « Principal components analysis corrects for stratification in genome-wide association studies ». In : *Nature genetics* 38.8 (2006), p. 904 (cf. p. 34).

[PTM06] Kim D Pruitt, Tatiana Tatusova et Donna R Maglott. « NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins ». In : *Nucleic acids research* 35.suppl_1 (2006), p. D61-D65 (cf. p. 28).

[Pur+07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly et al. « PLINK : a tool set for whole-genome association and population-based linkage analyses ». In : *The American Journal of Human Genetics* 81.3 (2007), p. 559-575 (cf. p. 30).

[Qui+98] Caroline Quillent, Estelle Oberlin, Joséphine Braun, Dominique Rousset, Gustavo Gonzalez-Canali, Patricia Métais, Luc Montagnier, Jean-Louis Virelizier, Fernando Arenzana-Seisdedos et Alberto Beretta. « HIV-1-resistance phenotype conferred by combination of two separate inherited mutations of CCR5 gene ». In : *The Lancet* 351.9095 (1998), p. 14-18 (cf. p. 20).

[Ram+12] Vijay K Ramanan, Li Shen, Jason H Moore et Andrew J Saykin. « Pathway analysis of genomic data : concepts, methods, and prospects for future development ». In : *TRENDS in Genetics* 28.7 (2012), p. 323-332 (cf. p. 97, 98).

[Rap+97] Jay Rappaport, Yi-Yun Cho, Houria Hendel, Elissa J Schwartz, François Schachter et Jean-François Zagury. « 32 bp CCR-5 gene deletion and resistance to fast progression in HIV-1 infected heterozygotes ». In : *The Lancet* 349.9056 (1997), p. 922-923 (cf. p. 20).

[Rat+12] Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero et Segolene Ayme. « Representation of rare diseases in health information systems : the Orphanet approach to serve a wide range of end users ». In : *Human mutation* 33.5 (2012), p. 803-808 (cf. p. 18).

[Rau+10] Andri Rauch, Zoltán Kutalik, Patrick Descombes, Tao Cai, Julia Di Iulio, Tobias Mueller, Murielle Bochud, Manuel Battegay, Enos Bernasconi, Jan Borovicka et al. « Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure : a genome-wide association study ». In : *Gastroenterology* 138.4 (2010), p. 1338-1345 (cf. p. 66).

[Ric+09] SS Rich et al. « Special Issue : Fine mapping of the MHC Region for Type 1 diabetes genes. » In : *Diabetes, Obesity and Metabolism* 11.s1 (2009) (cf. p. 29).

[Ric12] Brian D. Tait (eds.) Richard J. N. Allcock (auth.) Frank T. Christiansen. *Immunogenetics : Methods and Applications in Clinical Practice*. 1re éd. Methods in Molecular Biology 882. Humana Press, 2012. isbn : 978-1-61779-841-2,978-1-61779-842-9 (cf. p. 40).

[Ris+99] Neil Risch, Donna Spiker, Linda Lotspeich, Nassim Nouri, David Hinds, Joachim Hallmayer, Luba Kalaydjieva, Patty McCague, Sue Dimiceli, Tawna Pitts et al. « A genomic screen of autism : evidence for a multilocus etiology ». In : *The American Journal of Human Genetics* 65.2 (1999), p. 493-507 (cf. p. 20).

[RKD08] Guy-Franck Richard, Alix Kerrest et Bernard Dujon. « Comparative genomics and molecular dynamics of DNA repeats in eukaryotes ». In : *Microbiology and Molecular Biology Reviews* 72.4 (2008), p. 686-727 (cf. p. 10).

[Rob+09] Richard J Roberts, Tamas Vincze, Janos Posfai et Dana Macelis. « REBASE—a database for DNA restriction and modification : enzymes, genes and genomes ». In : *Nucleic acids research* 38.suppl_1 (2009), p. D234-D236 (cf. p. 18).

[Rob+14] James Robinson, Jason A Halliwell, James D Hayhurst, Paul Flicek, Peter Parham et Steven GE Marsh. « The IPD and IMGT/HLA database : allele variant databases ». In : *Nucleic Acids Research* 43.D1 (2014), p. D423-D431 (cf. p. 39).

[RRK58] William L Russell, Liane Brauch Russell et Elizabeth M Kelly. « Radiation dose rate and mutation frequency ». In : *Science* 128.3338 (1958), p. 1546-1550 (cf. p. 7).

[Sam+96] Michel Samson, Frédérick Libert, Benjamin J Doranz, Joseph Rucker, Corinne Liesnard, Claire-Michèle Farber, Sentob Saragosti, Claudine Lapouméroulie, Jacqueline Cognaux, Christine Forceille et al. « Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene ». In : *nature* 382.6593 (1996), p. 722-725 (cf. p. 20).

[Say+01] D Sayer, R Whidborne, B Brestovac, F Trimboli, C Witt et F Christiansen. « HLA-DRB1 DNA sequencing based typing : an approach suitable for high throughput typing including unrelated bone marrow registry donors ». In : *HLA* 57.1 (2001), p. 46-54 (cf. p. 40).

[SBR98] Richard A Sturm, Neil F Box et Michele Ramsay. « Human pigmentation genetics : the difference is only skin deep ». In : *Bioessays* 20.9 (1998), p. 712-721 (cf. p. 17).

[She+01] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski et Karl Sirotkin. « dbSNP : the NCBI database of genetic variation ». In : *Nucleic acids research* 29.1 (2001), p. 308-311 (cf. p. 9, 29).

[Smi+07] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall et al. « The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration ». In : *Nature biotechnology* 25.11 (2007), p. 1251 (cf. p. 28).

[Sne12] George Snell. *Histocompatibility*. Elsevier, 2012 (cf. p. 21).

[Sob+01] Satoshi Sobue, Tomoyuki Nomura, Tetsuya Ishikawa, Satomi Ito, Katsuhisa Saso, Hirotaka Ohara, Takashi Joh, Makoto Itoh et Shinichi Kakumu. « Th1/Th2 cytokine profiles and their relationship to clinical features in patients with chronic hepatitis C virus infection ». In : *Journal of gastroenterology* 36.8 (2001), p. 544-551 (cf. p. 66).

[Spa+15] Jean-Louis Spadoni, Pierre Rucart, Sigrid Le Clerc, Daniëlle van Manen, Cédric Coulonges, Damien Ulveling, Vincent Laville, Taoufik Labib, Lieng Taing, Olivier Delaneau et al. « Identification of Genes Whose Expression Profile Is Associated with Non-Progression towards AIDS Using eQTLs ». In : *PloS one* 10.9 (2015), e0136989 (cf. p. 63).

[SS02] Bernhard Schölkopf et Alexander J Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT press, 2002 (cf. p. 46).

[Sub+05] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander et al. « Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles ». In : *Proceedings of the National Academy of Sciences* 102.43 (2005), p. 15545-15550 (cf. p. 97).

[Ted+03] Ellen M Tedaldi, Katherine Huppler Hullsiek, Carlos D Malvestutto, Roberto C Arduino, Evelyn J Fisher, Paul J Gaglio, Elizabeth R Jenny-Avital, Joseph P McGowan, George Perez et Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA). « Prevalence and characteristics of hepatitis C virus coinfection in a human immunodeficiency virus clinical trials group : the Terry Beirn Community Programs for Clinical Research on AIDS ». In : *Clinical Infectious Diseases* 36.10 (2003), p. 1313-1317 (cf. p. 67).

[Ulv+16] Damien Ulveling, Sigrid Le Clerc, Aurélie Cobat, Taoufik Labib, Josselin Noirel, Vincent Laville, Cédric Coulonges, Wassila Carpentier, Bertrand Nalpas, Markus H Heim et al. « A new 3p25 locus is associated with liver fibrosis progression in human immunodeficiency virus/hepatitis C virus-coinfected patients ». In : *Hepatology* 64.5 (2016), p. 1462-1472 (cf. p. 68, 95, 101).

[UNA09] UNAIDS. « AIDSinfo ». In : *Online at http://aidsinfo.unaids.org/* (2009) (cf. p. 59).

[Ura+13] Yuji Urabe, Hidenori Ochi, Naoya Kato, Vinod Kumar, Atsushi Takahashi, Ryosuke Muroyama, Naoya Hosono, Motoyuki Otsuka, Ryosuke Tateishi, Paulisally Hau Yi Lo et al. « A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region ». In : *Journal of hepatology* 58.5 (2013), p. 875-882 (cf. p. 66, 98).

[Van+90] JAR Van den Hoek, HJA Van Haastrecht, J Goudsmit, F De Wolf et Roel A Coutinho. « Prevalence, incidence, and risk factors of hepatitis C virus infection among drug users in Amsterdam ». In : *Journal of Infectious Diseases* 162.4 (1990), p. 823-826 (cf. p. 64).

[Ven+01] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt et al. « The sequence of the human genome ». In : *science* 291.5507 (2001), p. 1304-1351 (cf. p. 26).

[Vin12] Jon C. Aster Vinay Kumar Abul K. Abbas. *Robbins Basic Pathology*. 9e éd. Robbins Pathology. Saunders, 2012. isbn : 1437717810,9781437717815 (cf. p. 38).

[Vis+17] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown et Jian Yang. « 10 Years of GWAS discovery : biology, function, and translation ». In : *The American Journal of Human Genetics* 101.1 (2017), p. 5-22 (cf. p. 94).

[Wal+87] Bruce D Walker, Sekhar Chakrabarti, Bernard Moss, Timothy J Paradis, Theresa Flynn, Amy G Durno, Richard S Blumberg, Joan C Kaplan, Martin S Hirsch et Robert T Schooley. « HIV-specific cytotoxic T lymphocytes in seropositive individuals ». In : *Nature* 328.6128 (1987), p. 345-348 (cf. p. 61).

[Wan+16] Feiran Wang, Yong Qiang, Lirong Zhu, Yasu Jiang, Yinda Wang, Xian Shao, Lei Yin, Jiahui Chen et Zhong Chen. « MicroRNA-7 downregulates the oncogene VDAC1 to influence hepatocellular carcinoma proliferation and metastasis ». In : *Tumor Biology* 37.8 (2016), p. 10235-10246 (cf. p. 97).

[WC+53] James D Watson, Francis HC Crick et al. « Molecular structure of nucleic acids ». In : *Nature* 171.4356 (1953), p. 737-738 (cf. p. 4).

[Wel+13] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff et al. « The NHGRI GWAS Catalog, a curated resource of SNP-trait associations ». In : *Nucleic acids research* 42.D1 (2013), p. D1001-D1006 (cf. p. 29, 94).

[Whe+07] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen et al. « Database resources of the national center for biotechnology information ». In : *Nucleic acids research* 36.suppl_1 (2007), p. D13-D21 (cf. p. 29).

[Whi+90] Thomas J White, Thomas Bruns, SJWT Lee, JL Taylor et al. « Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics ». In : *PCR protocols : a guide to methods and applications* 18.1 (1990), p. 315-322 (cf. p. 6).

[Wik17] Wikipedia. *Caryotype — Wikipedia, The Free Encyclopedia*. 2017. url : https://fr.wikipedia.org/w/index.php?title=Caryotype (cf. p. 5).

[Yat+15] Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil et al. « Ensembl 2016 ». In : *Nucleic acids research* 44.D1 (2015), p. D710-D716 (cf. p. 28).

[Zag+86] D Zagury, J Bernard, R Leonard, R Cheynier, M Feldman, PS Sarin et Robert C Gallo. « Long-term cultures of HTLV-III-infected T cells : a model of cytopathology of T-cell depletion in AIDS ». In : *Science* 231 (1986), p. 850-854 (cf. p. 59).

[Zha+15] Lina Zhang, Jinzhao Zhao, Guanglin Cui, Hong Wang et Dao Wen Wang. « Genotyping on ALDH2 : Comparison of four different technologies ». In : *PloS one* 10.3 (2015), e0122745 (cf. p. 27).

[Zhe+14] Xiuwen Zheng, Judong Shen, Charles Cox, Jonathan C Wakefield, Margaret G Ehm, Matthew R Nelson et Bruce S Weir. « HIBAG—HLA genotype imputation with attribute bagging ». In : *The pharmacogenomics journal* 14.2 (2014), p. 192 (cf. p. 42, 46, 91).

Annexe A

HIV/HCV article : Supplementary material

Characteristic	All patients (n = 280)	F0F1F2 (n = 205)	F4 (n = 75)	p
Sex (Male/Female)	193/87	138/67	55/20	0.41
Age (years) Median ± SD	45.4±5.87	45.3±5.89	46.3±5.8	0.04
Infection time (years) Median ± SD	24±7.78	23±8.4	25±5.5	0.001
Alcohol status (never/past/current)	48/76/156	37/52/116	11/24/40	0.51
HCV genotype 1 vs other	160/120	116/89	44/31	0.86
METAVIR score, n (%) 0 or 1 2 3 4	149 (53%) 56 (20%) 0 (0%) 75 (27%)	149 56 0 0	0 0 0 75	N/A
SVR after treatment (Yes/No) n (%)	26 (9%)	18 (9%)	8 (10%)	0.80
CD4+ (cell/mm3) Median ± SD	456.5±304	468±304	406±298	0.08
HIV viral load * Median ±sd	3.69±1.88	3.69±1.7	3.69±2.3	0.31
ARVT ** (years) Median ± SD	10.1±4.5	9.8±4.5	10.6±4.6	0.19

TABLE A.1 – Demographic statistics

based on the date of patients inclusions in the cohort. Statistical differences are determined between F0F1F2 and F3F4 groups using the t-test (quantitatives variables) or the χ^2 test (qualitatives variables), included in the R software version 3.0.2 (https ://cran.r-project.org/).

1p region	5p region	19p region
rs72727113	rs5865934	rs11669331
rs5779589	rs2158099	rs11669473
	rs35268064	rs16992567
	rs11367210	rs16992570
	rs1423493	rs34105873
	rs2074260	rs114019858
	rs17802039	rs115300990
	rs2032879	rs8112813
	rs715126	rs8103240
		rs143620761
		rs556921174
		rs561332451
		rs74554920
		rs137946930

TABLE A.2 – table of SNPs
in high linkage disequilibrium ($r^2 > 0.8$) with significant SNPs ($p - value < 5 \times 10^{-8}$)
in the 1p, 5p and 19p regions.

SNP	Gene	Genetic context
rs11367210	*CTNND2*	intronic
rs1423493	*CTNND2*	intronic
rs17802039	*CTNND2*	intronic
rs2032879	*CTNND2*	intronic
rs2074260	*CTNND2*	intronic
rs2158099	*CTNND2*	intronic
rs35268064	*CTNND2*	intronic
rs5865934	*CTNND2*	intronic
rs715126	*CTNND2*	intronic
rs114019858	*MIR7-3HG*	3' downstream
rs115300990	*MIR7-3HG*	3'downstream
rs11669331	*MIR7-3HG*	Non-coding, intronic
rs11669473	*MIR7-3HG*	Non-coding, intronic
rs16992567	*MIR7-3HG*	3' downstream
rs16992570	*MIR7-3HG*	3' downstream
rs34105873	*MIR7-3HG*	3' downstream

TABLE A.3 – RefSeq annotations of SNPs
in high linkage disequilibrium ($r^2 > 0.8$) with significant SNPs ($p - value < 5 \times 10^{-8}$)
in the 1p, 5p and 19p regions.

SNP	Chr	Position	P-val HEPAVIH	Q-val HEPAVIH	P-val GENOSCAN
rs72727113	1	191555232	1.63×10^{-8}	0.06	0.74
rs141853043	1	191548566	1.25×10^{-7}	0.06	0.99
rs32123	5	11345898	3.70×10^{-7}	0.11	0.77
rs11989480	8	74361979	3.25×10^{-7}	0.10	0.41
rs11790112	9	19469751	2.71×10^{-7}	0.09	0.35
rs11790131	9	19469846	2.71×10^{-7}	0.09	0.43
rs3894272	10	127141448	6.35×10^{-8}	0.06	0.15
rs3929230	10	127140233	8.11×10^{-8}	0.06	0.15
rs74554920	19	4775195	4.79×10^{-8}	0.06	NA
rs556921174	19	4775193	4.79×10^{-8}	0.06	NA
rs11669331	19	4770167	1.24×10^{-7}	0.06	0.49
rs11669473	19	4770564	1.29×10^{-7}	0.06	0.48
rs16992567	19	4772921	1.29×10^{-7}	0.06	0.47
rs16992570	19	4773049	1.29×10^{-7}	0.06	0.47
rs34105873	19	4773715	1.29×10^{-7}	0.06	0.47
rs114019858	19	4774107	1.33×10^{-7}	0.06	0.40
rs8103240	19	4774984	1.51×10^{-7}	0.06	0.42
rs143620761	19	4775060	1.53×10^{-7}	0.06	0.42
rs561332451	19	4775194	1.55×10^{-7}	0.06	NA
rs115300990	19	4774534	1.68×10^{-7}	0.06	0.46
rs137946930	19	4775503	1.70×10^{-7}	0.06	0.40
rs8112813	19	4774734	4.51×10^{-7}	0.12	0.43
rs9982976	21	16435223	4.42×10^{-7}	0.12	0.25

TABLE A.4 – SNPs with a FDR inferior to 15%
in the additive analysis. SNPs are sorted by chromosomal position, allowing one to identify six independent signals. The p-value obtained for the GENOSCAN replication is displayed.

SNP	Chr	Position	P-value	Q-value	P-value GENOSCAN
rs72727113	1	191555232	7.76×10^{-9}	0.02	0.53
rs141853043	1	191548566	1.25×10^{-7}	0.06	0.89
rs74791205	1	191768372	7.40×10^{-7}	0.12	0.85
rs80310414	1	191839957	7.73×10^{-7}	0.12	1
rs144164086	1	191802066	7.66×10^{-7}	0.12	0.87
rs72703464	1	166734183	4.74×10^{-7}	0.11	0.66
rs72703477	1	166745107	4.74×10^{-7}	0.11	0.66
rs1925032	1	158031160	6.22×10^{-7}	0.11	0.26
rs150868533	4	38303793	3.61×10^{-7}	0.09	0.19
rs11367210	5	11362853	9.88×10^{-9}	0.02	0.83
rs1423493	5	11363645	9.95×10^{-9}	0.02	0.82
rs35268064	5	11362322	1.45×10^{-8}	0.02	0.88
rs2074260	5	11365166	3.52×10^{-8}	0.04	1
rs32124	5	11345678	1.11×10^{-7}	0.06	0.78
rs1978462	5	11361864	3.79×10^{-7}	0.09	0.80
rs39930	5	11340507	5.19×10^{-7}	0.11	0.90
rs32139	5	11333739	6.29×10^{-7}	0.11	0.79
rs17802039	5	11366830	1.05×10^{-6}	0.14	NA
rs138336017	5	157538117	6.04×10^{-7}	0.11	0.03
rs11989480	8	74361979	3.25×10^{-7}	0.09	0.25
rs114755724	8	74369365	9.96×10^{-7}	0.14	0.19
rs3894272	10	127141448	5.23×10^{-8}	0.04	0.22
rs3929230	10	127140233	7.62×10^{-8}	0.05	0.21
rs556921174	19	4775193	5.55×10^{-8}	0.04	NA
rs74554920	19	4775195	5.56×10^{-8}	0.04	NA
rs11669331	19	4770167	1.50×10^{-7}	0.06	0.57
rs34105873	19	4773715	1.56×10^{-7}	0.06	0.56
rs11669473	19	4770564	1.57×10^{-7}	0.06	0.57
rs16992567	19	4772921	1.57×10^{-7}	0.06	0.56
rs16992570	19	4773049	1.57×10^{-7}	0.06	0.56
rs114019858	19	4774107	1.61×10^{-7}	0.06	0.47
rs561332451	19	4775194	1.88×10^{-7}	0.06	NA
rs8103240	19	4774984	1.83×10^{-7}	0.06	0.50
rs143620761	19	4775060	1.86×10^{-7}	0.06	0.49
rs115300990	19	4774534	2.06×10^{-7}	0.06	0.54
rs137946930	19	4775503	2.06×10^{-7}	0.06	0.47
rs8112813	19	4774734	5.48×10^{-7}	0.11	0.51
rs143180446	19	49482903	4.96×10^{-7}	0.11	NA
rs4801785	19	49482676	6.61×10^{-7}	0.12	0.68
rs11882285	19	49483068	7.27×10^{-7}	0.12	0.64
rs11882289	19	49483086	7.29×10^{-7}	0.12	0.64
rs12980441	19	49482522	7.58×10^{-7}	0.12	0.67
rs1862492	19	49478434	8.62×10^{-7}	0.13	0.73

TABLE A.5 – SNPs with a FDR inferior to 15% in the dominant analysis. SNPs are sorted by chromosome, allowing one to identify ten independent signals. The p-value obtained for the GENOSCAN replication is displayed.

Gene nearby	SNP	Chr	Position	MAF F0F1F2	MAF F4	OR	P-val	Q-val
Additive model								
MIR7-3HG	rs74554920	19	4775195	4,00 %	16,00 %	8.4	4.79×10^{-8}	0.06
LOC105375988	rs11790131	9	19469846	16,00 %	34,00 %	3.4	2.71×10^{-7}	0.09
STAU2	rs11989480	8	74361979	1,00 %	6,00 %	36	3.25×10^{-7}	0.10
CTNND2	rs32123	5	11345898	55,00 %	31,00 %	0.17	3.70×10^{-7}	0.11
NRIP-1	rs9982976	21	16435223	23,00 %	43,00 %	3.2	4.42×10^{-7}	0.12
Dominant model								
KIRREL	rs1925032	1	158031160	22,00 %	34,00 %	0.04	6.22×10^{-7}	0.11
FMO11P	rs72703464	1	166734183	5,00 %	16,00 %	4.4	4.74×10^{-7}	0.11
CTNND2	rs11367210	5	11362853	55,00 %	31,00 %	0.17	9.88×10^{-9}	0.02
STAU2	rs114755724	8	74369365	1,00 %	6,00 %	36	9.96×10^{-7}	0.10
MIR7-3HG	rs74554920	19	4775195	4,00 %	16,00 %	8.4	5.56×10^{-8}	0.04
GYS1	rs12980441	19	49482522	46,00 %	26,00 %	1/4.5	7.58×10^{-7}	0.12

TABLE A.6 – SNPs and genes with FDR <15%.

First author	journal	pmid	SNPs	P $\leq$ 10-5	region	Reported genes	P-value HEPAVIH
Patin E	Gastroenterology	22841784	rs16851720	9×10^{-9}	3q23	*RNF7*	0.26
			rs4374383	1×10^{-9}	2q13	*MERTK*	0.83
			rs9380516	5×10^{-7}	6p21	*TULP1*	0.018
			rs2629751	1×10^{-7}	12q23	*GLT8D2*	0.77
			rs883924	2×10^{-6}	9q22	*LOC340515*	0.47
			rs7800244	3×10^{-6}	7p12	*PKD1L1*	0.68
Urabe Y	J Hepatol	23321320	rs910049	9×10^{-11}	6p21	*C6orf10*	0.79
			rs3135363	1×10^{-10}	6p21	intergenic	0.28
			rs3129860	1×10^{-9}	6p21	intergenic	0.18
			rs3817963	1×10^{-8}	6p21	*BTNL2*	0.038
			rs9405098	1×10^{-10}	6p21	intergenic	N/A

TABLE A.7 – GWAS associations dealing with HCV liver fibrosis.

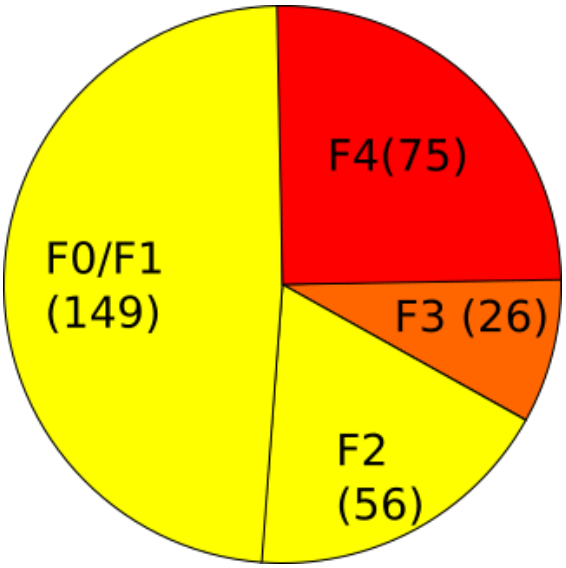

FIGURE A.1 – Distribution of METAVIR scores in the HEPAVIH sample used in analysis.

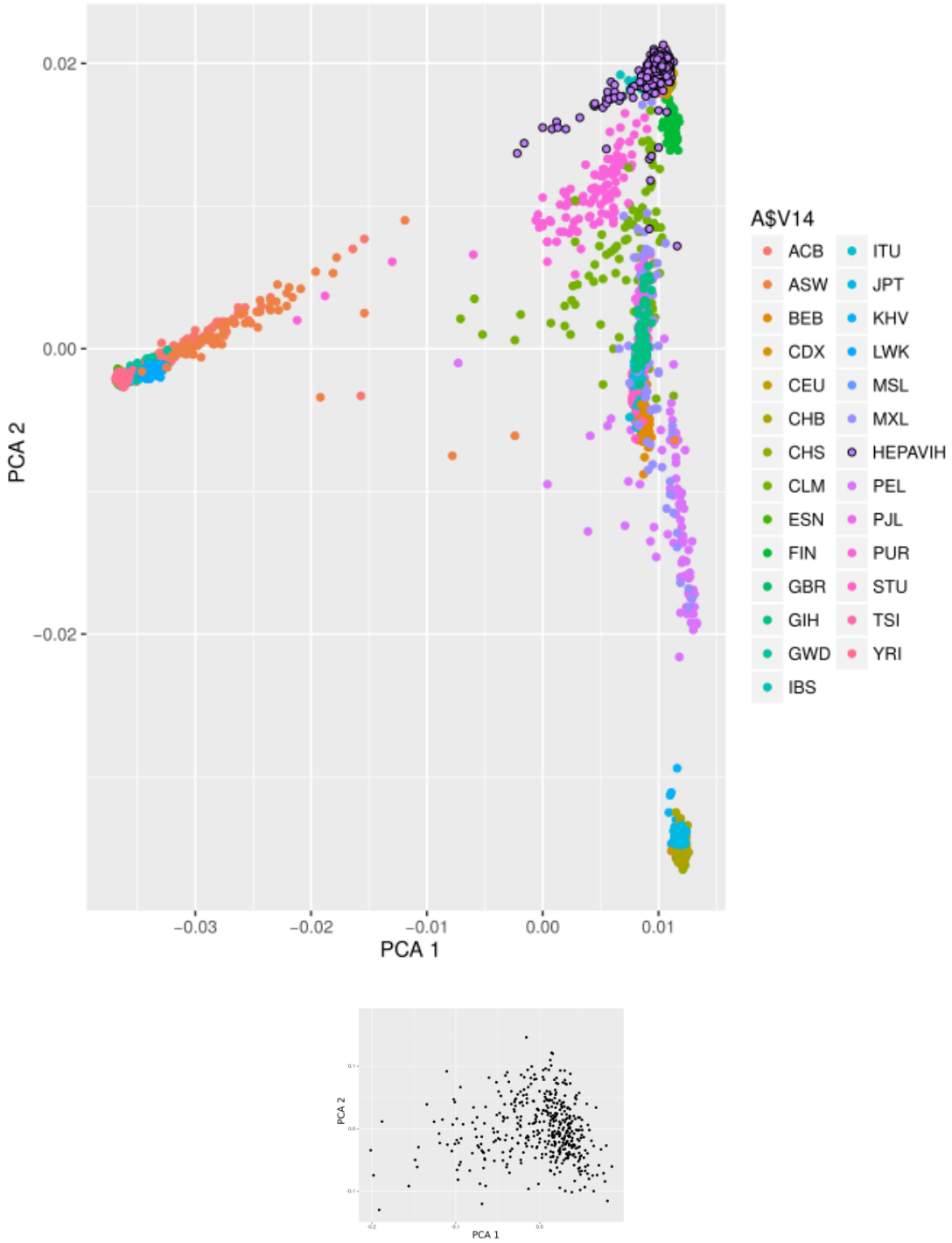

FIGURE A.2 – Principal component analysis (PCA) plots.
(A) PCA of HEPAVIH and 1000Genomes individuals (each color represents a different 1000genomes population). Each point represents a patient. HEPAVIH individuals are in violet. (B) PCA of HEPAVIH individuals after removing of non-European samples. Stratification of the individuals include in the analysis.

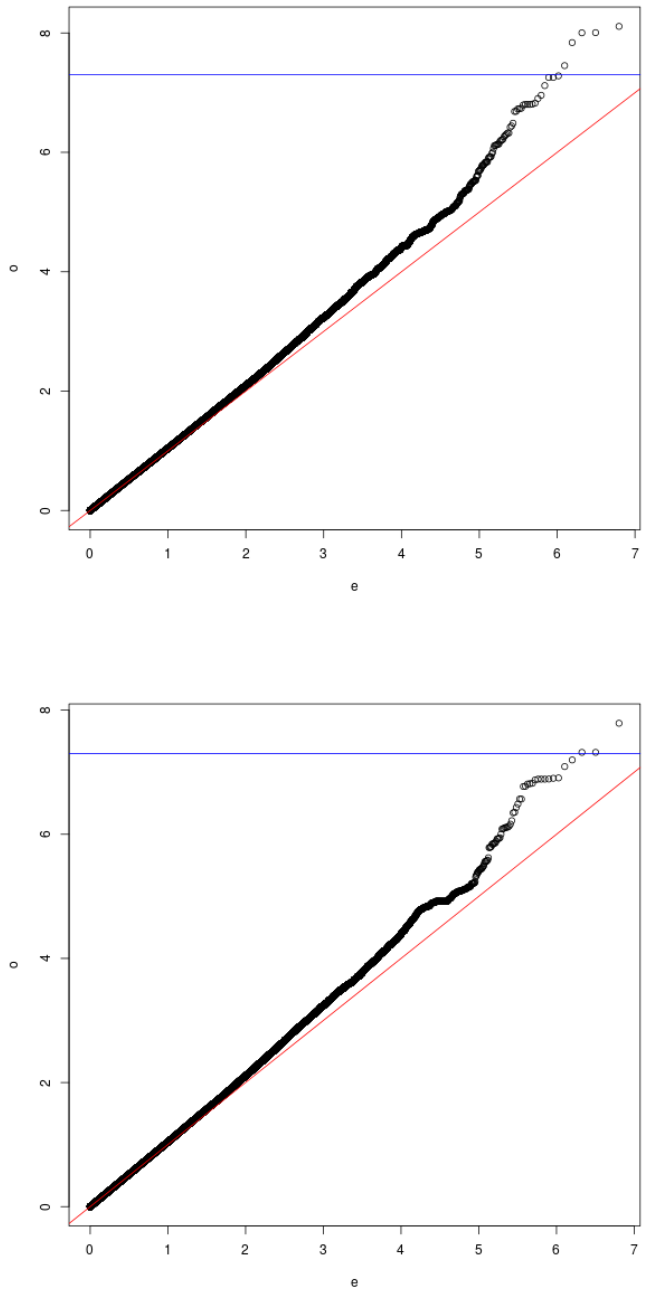

FIGURE A.3 – Quantile–quantile (QQ) plot.
The plot displays the observed (y-axis) versus expected (x-axis) $-log10(p-value)$. (a) Dominant analysis (b) Additive analysis

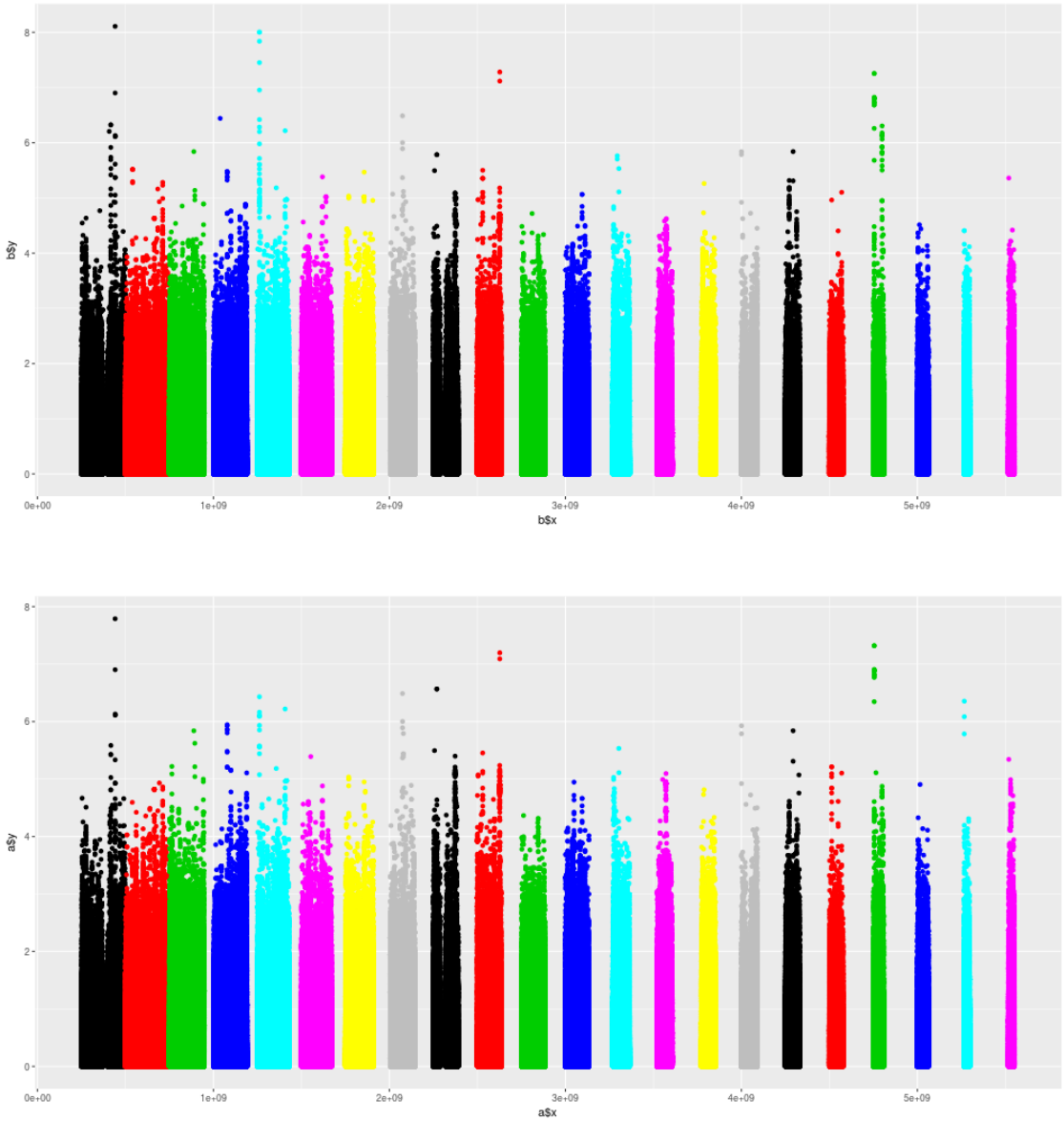

FIGURE A.4 – Manhattan plot of the genome-wide association results by comparing F0F1F2 and F3F4 METAVIR score groups.
After genotype imputation, 8,500,000 common variants were tested for association by comparing 205 subjects with low METAVIR score (< F3) with 75 subjects with high METAVIR score (F4), in logistic regression analysis with an additive (a) and dominant (b) model. Each dot corresponds to a SNP, with it chromosomal position (x axis) and $-log10(p-value)$ (y axis) as the coordinates.

le cnam

Marc JEANMOUGIN

Imputation HLA et Analyse génomique de coinfection VIH/VHC

le cnam

Résumé :
La génomique d'association cherche à déterminer des liens entre le génome et des phénotypes ou maladies. Cependant, certaines régions du génome, notamment la région HLA, sont difficiles à génotyper avec les méthodes les moins chères et les plus utilisées, et leur inférence statistique est incertaine.

J'ai donc développé un outil permettant d'évaluer, à partir de données génomiques, la plausibilité d'un typage HLA, et démontré l'efficacité de cette technique.

J'ai également effectué une étude d'association génome entier sur le déclenchement de la cirrhose chez une cohorte française de patients co-infectés par le VIH et le VHC. Cette étude a permis d'identifier trois signaux, dont deux dans des gènes (*CTNND2* et *MIR7-3HG*), qui permettront de mieux comprendre les mécanismes de déclenchement de la cirrhose chez ces patients.
Mots-clés :
GWAS, SNP, VIH/VHC, cirrhose, HLA, CMH, Imputation

Abstract :
Association genomics try to establish possible links between the genotype of individuals and some traits or illnesses. However, some regions of the genome, such as the MHC, are difficult to genotype with the cheapest and most used methods, and their statistical imputation is imprecise.

I have created a new tool, HLA-Check, to assess from genomic data the plausibility of a given HLA typing, and have demontrated the efficiency of that tool.

I also performed a genome-wide association study on cirrhosis outbrak in a french cohort of HIV/HCV coinfected individuals. This study allowed us to identify three signals, two of which are in genes (*CTNND2* and *MIR7-3HG*) and may help better understand the mechanisms of cirrhosis outbreak for those individuals.
Keywords :
GWAS, SNP, HIV/HCV, cirrhosis, HLA, MHC, Imputation