Imputation HLA dans des populations d'ancestralité composite grâce à des méthodes de réduction de dimension

Venceslas Douillard

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

Nantes Université

THESE DE DOCTORAT DE

NANTES UNIVERSITE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Bioinformatique*

Par
Venceslas DOUILLARD

Imputation HLA dans des populations d'ancestralité composite grâce à des méthodes de réduction de dimension

Thèse présentée et soutenue à Nantes, le 15 novembre 2022
Unité de recherche : UMR1064 Centre de Recherche Translationnelle en Transplantation et Immunologie

Rapporteurs :

Emmanuelle Génin	Directrice de recherche, Université de Brest
David Courtin	Chargé de recherche, Université de Paris

Composition du Jury :

Présidente :	Élise Launay	Professeure des universités – Praticienne hospitalière, CHU de Nantes
Examinateur :	Slim Karkar	Maître de conférences, Université de Bordeaux
Dir. de thèse :	Nicolas Vince	Chargé de recherche, Inserm, Nantes Université
Co-en. de thèse :	Élise Launay	Professeure des universités – Praticienne hospitalière, CHU de Nantes
Co-en. de thèse :	Pierre-Antoine Gourraud	Professeur des universités – Praticien hospitalier, Nantes Université

Remerciements

« And all the faces that I know
Have that same familiar glow.
I think I must've known then somewhere once before
All the faces that I know»

Je tiens à remercier toutes les personnes que j'ai pu rencontrer pendant mes cinq années au CRTI puis CR2TI, celles qui m'ont accompagnées tout au long de ma vie et m'ont soutenu pendant cette période et également un mot pour toutes les personnes que je risque d'oublier, merci à vous.

Je remercie le Dr. Nicolas Vince, Nico, pour ses conseils depuis le Master 2, pour son encadrement, pour m'avoir transmis son expertise et sa passion pour le HLA ainsi que son intérêt pour la génétique des populations. Je le remercie également pour son calme en toute situation qui m'a permis de vivre cette thèse sereinement.

Je remercie également la Pr. Élise Launay pour son engagement initial à diriger cette thèse qui m'a tout simplement permis de commencer mon doctorat, mais aussi pour sa réactivité et pour l'intérêt porté à mes travaux. Je remercie le Pr. Pierre-Antoine Gourraud pour les opportunités qu'il m'a offert tout au long de ma présence au laboratoire ainsi que de sa confiance apportée durant ces quatre années. Je remercie enfin ma directrice d'équipe, Dr. Sophie Limou, pour avoir suivi de près ma thèse et donné l'impulsion dans cette voie.

Je tiens à remercier la Dr. Emmanuelle Génin ainsi que le Dr. David Courtin pour l'intérêt porté à mes travaux et à l'engagement qu'ils ont pris en acceptant d'être rapporteurs de cette thèse. J'associe à ces remerciements le Dr. Slim Karkar pour sa participation en tant qu'examinateur au jury de cette thèse.

Je remercie aussi les membres de mon comité de suivi de thèse, Gilles Blancho et Richard Redon pour avoir accepté de suivre l'évolution de ma thèse durant ces trois années, pour leurs questions pertinentes qui ont aidé à façonner ce projet de thèse et leur bienveillance face à mes questionnements.

Je souhaite remercier l'ensemble du CR2TI et son directeur Régis Josien pour leur accueil dans l'unité. J'ai pu y côtoyer des techniciens, médecins, immunologistes, ingénieurs et gestionnaires avec des expertises multiples et étendre mes connaissances au-delà de mon champ de recherche grâce à eux.

Je souhaite remercier la région Pays de la Loire et L'Inserm pour m'avoir donné l'opportunité d'effectuer une thèse en m'attribuant une bourse. Je veux plus particulièrement remercier, Camille Sicot et Pierre Da Silva de l'Inserm pour leur aide précieuse, leur travail et leur enthousiasme autour de notre projet Transplant'Action, ainsi que la Dr. Lucie Clarysse pour avoir donné de la couleur et embelli ce projet.

Je remercie Alexandra Elbakyan pour son travail engagé et pour ces années de vie économisées pour construire ma bibliographie.

Un merci chaleureux aux premières stagiaires avec lesquelles j'ai commencé mon périple au CR2TI et qui m'ont ouvert la voix pour la thèse : Estelle, merci d'avoir respecté mes choix contestables de couleurs dans mes posters et présentations sans trop te moquer, et qu'est-ce qu'on a ri avec nos jeux de mots improbables, plus personne n'a compris mes blagues après ton départ ; Rokhaya, je me souviens des heures de debug, des chansons improvisées au milieu de l'après-midi et bien sûr des soirées musicales intenses d'apprentissage de la guitare, enfin, on se comprend.

Je remercie également toutes les personnes que j'ai vues moins souvent pendant ma thèse mais que je n'oublie pas : Anaïs pour les (nombreuses) pauses spontanées de milieu d'après-midi et les photos de Laïka pour donner du cœur à l'ouvrage ; Magalie pour les petits passages remplis de bonne humeur au début de mon arrivée au laboratoire ; Sirine pour les réunions « traduction » où on a pu bien rire ; Chadia, malgré la distance, ça a toujours été un plaisir de pouvoir discuter et boire des cafés ensemble ; à la clinique des données pour leur accueil temporaire et Stanislas pour les brefs passages dans le bureau.

J'ai une pensée pour tous les stagiaires/médecins d'année recherche qui nous ont suivi, avec qui j'ai parfois travaillé et partagé de nombreux bons moments dans cette équipe : Hélène, Elgeta, Lina, Ayan, Rémi, Marie, Chloé, Aymeric, Amandine, Julien, Claire, Raphaël, Valentin, Clémence, Morgane, Flavie, Lucas, Justine, Augustin, Ketsia, Nabilah et Antoine.

Un grand merci à vous Axelle, Irène, Léo, Martin, Olivia, Sonia, Vincent, qui m'avez accompagné dans cette fin de thèse. Je garderai un excellent souvenir de pouvoir bosser et rire avec vous et surtout de notre playlist légendaire, de nos squads Fortnite invincibles, des dessins improbables, des drifts endiablés sur Mario Kart, de tous les jeux du midi et particulièrement le GeoGuessr (sans les heures de clic dans des routes interminables), de Jérémie Decouchant, des parties de basket impromptues, des défaites cuisantes à Magic, des discussions autour des meilleurs RPG, ... Au final, peut-être que le véritable interlude, ce sont les heures de travail entre tout ça.

Nayane, eu espero que essas frases não têm muitos erros. Muito obrigado para todas as discussões, sua ajuda para aprender português e outras coisas sobre o Brasil. Boa sorte para continuar o SHLARC. Eu sei que você vai fazer uma tese excepcional.

Je remercie aussi tous les gens en dehors de ce monde de la recherche que je n'ai pas vu aussi souvent que je le pouvais mais avec qui je pouvais décompresser le temps d'une soirée, merci Enzo, Flavien, Élodie, Valentin, Johanna, Manon, Adrien, Éole, Valentin, Maxime et Élodie.

Je remercie grandement Michaël pour les calls Discord de la détente et du gaming, Abel pour les heures de theorycraft sur différents jeux et les soirées infinies à me faire découvrir des œuvres obscures, Sylvain pour les discussions musico-littero-philosophiques qui aident à prendre du recul et Kévin pour m'avoir carry sur mes stages autant que sur nos parties de CS. Merci les amis.

Je tiens à remercier les Fenouils Coopératifs pour l'ensemble de leur œuvre : les mèmes, les réactions sur les chapitres OP mais surtout pour les retrouvailles virtuelles et réelles qui ont ponctué ma thèse de rires et de bonne humeur. À Sébastien, Laura et Jérôme, les quelques 7 Wonders et autres soirées jeux ont été autant de coups de boost au moral. Merci Célia pour les accueils chaleureux de fin de journée à discuter de tout et de rien à la boutique. Merci pour tout Arnaud, des découvertes ciné, aux discussions, aux parties de Risk of Rain ou Smash infinies, et surtout merci de prendre soin de ton petit frère comme tu le fais.

Je remercie évidemment mes parents qui ont tout fait pour que je sois ici, à ce moment, qui ont découvert en même temps que moi ce qu'était une thèse, se sont inquiétés et ont continué d'apporter leur soutien indéfectible. Je ne suis pas passé assez souvent à mon goût sur ces 3 années, ça devrait bientôt s'arranger.

Enfin, Éléa, j'aurais pu remplacer chacune des pages de cette thèse par des remerciements pour toi. Soi-disant ça ne fait pas partie des recommandations de l'école doctorale. Merci de ta bienveillance et de ta force, de l'inspiration et de la motivation que tu m'apportes chaque jour. Je t'aime.

Table des matières

Liste des acronymes et des abréviations

1KGP *1,000 Genomes Project*

ACP *Analyse en Composante Principale*

ADN *Acide Désoxyribonucléique*

AFND *Allele Frequency Net Database*

ARN *Acide Ribonucléique*

ASHG *American Society of Human Genetics*

CAAPA *Consortium on Asthma among African-ancestry Populations in the Americas*

CD *Cluster of Differentiation*

CLIP *Class II-association Invariant chain Peptide*

CMH / MHC *Complexe Majeur d'Histocompatibilité / Major Histocompatibility Complex*

CMHx / xMHC *Complexe Majeur d'Histocompatibilité étendu/ Extended Major Histocompatibility Complex*

COL11A2 *Collagen Type XI Alpha 2 Chain*

COVID-19 *Coronavirus Disease 2019*

CPU *Central Processing Unit*

CTL *Cytotoxic T Lymphocyte*

CWD *Common and Well-Documented*

EBV *Epstein-Barr Birus*

EFI *European Federation for Immunogenetics*

ERAP1 *Endoplasmic Reticulum Aminopeptidase 1*

ERp57 *Endomplasmic Reticulum protein 57*

Fst *Fixation Index*

GPU *Graphical Processing Unit*

GWAS *Genome-Wide Association Study*

HBV *Hepatitis B Virus*

HCV *Hepatitis C Virus*

HIP *HLA Imputation Portal*

HIV-1 *Human Immunodeficiency Virus 1*

HLA *Human Leukocyte Antigen*

HSP *Heat-Shock Protein*

IBD *Identity By Descent*

IBS *Identity By State*

IEDB *Immune Epitope DataBase*

IGSR *The International Genome Sample Resource*

IHIW *International HLA & Immunogenetics Workshop*

IMGT/HLA *IMmunoGeneTics/HLA*

IPD *Immune Polymorphism Database*

KIR *Killer-cell Immunoglobulin-like Receptor*

LTC *Lymphocyte T Cytotoxique*

MAR *Missing At Random*

MCAR *Missing Completely At Random*

MERS *Middle-East Respiratory Syndrom*

MICA/B *MHC class I polypeptide related sequence A/B*

MIIC *MHC class II Compartment*

MNAR *Missing Not At Random*

MOG *Myelin Oligodendrocyte Glycoprotein*

NF-κB *Nuclear Factor – kappa B*

NK *Natural Killer*

NMR *Nuclear Magnetic Resonance*

PCA *Principal Component Analysis*

PCR *Polymerase Chain Reaction*

PCR-SSO *PCR with Sequence Specific Oligonucleotides*

PCR-SSP *PCR with Sequence Specific Primers*

PSSM *Position-Specific Score Matrix*

RE *Réticulum Endoplasmique*

SARS *Severe Acute Respiratory Syndrom*

SEP *Sclérose En Plaques*

SNP *Single Nucleotide Polymorphism*

TAP *Tapasin*

TCR *T-cell Receptor*

TNF *Tumor Necrosis Factor*

TUB *Tubulin*

UMAP *Unfiform Manifold Approximation Projection*

Introduction

« Human racial classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of virtually no genetic or taxonomic significance either, no justification can be offered for its continuance. »

Lewontin RC. 1972, The apportionment of human diversity

I - L'évolution du rapport entre l'humain et ses génomes

La génomique définit l'organisation des génomes sur leur support, l'ADN (ou l'ARN chez les virus), ainsi que la manière dont ils régissent l'ensemble des mécanismes moléculaires comme l'expression d'ARN, ou la production de protéines dans les organismes. Cette discipline s'est progressivement immiscée dans la vie humaine dans les dernières décennies, aussi bien dans les domaines des professionnels de santé que de manière récréative dans le grand public.

Par exemple, depuis 2019 et jusqu'à aujourd'hui en 2022, la pandémie de SARS-CoV-2 a vu croître le nombre d'informations génomiques délivrées au grand public, notamment sur l'existence de variants viraux et de leur impact sur l'infection (1). De plus, les années 2000 sont le début des tests génétiques « récréatifs », qui pour la plupart peuvent renseigner sur les ancestralités d'une personne en se basant sur les similarités avec plusieurs populations humaines. Bien qu'ils soient interdits en France (2), environ 100 000 à 200 000 y aurait recours annuellement (3). Ainsi, des entreprises comme 23andme ont même étendu leur rôle récréatif en obtenant l'autorisation de tester certaines pathologies (4).

Ces situations montrent l'augmentation de l'accès et de l'attrait de la population pour la génomique mais soulignent également la diversité des génomes étudiés. Le monde de la médecine et de la recherche connaît cette diversité. Ainsi les progrès continuels de la génomique humaine ont modifié le parcours de santé pour y incorporer des tests de pathologies génétiques (5) ou évaluer la compatibilité en transplantation par le groupe sanguin ou les gènes du *Human Leukocyte Antigen* (*HLA*) (6).

Malgré l'omniprésence de la génomique dans la recherche en santé, notamment dans l'étude des pathologies et la connaissance de l'hétérogénéité des génomes humains, l'étude de l'interaction entre les pathologies et les différences génétiques entre les populations reste minoritaire. Dans le cas de l'insuffisance rénale, la différence génétique des populations est intégrée dans certains scores de prédiction de la fonction rénale mais leur impact réel à long terme sur le traitement des patients est questionné (7,8).

Ainsi, la génomique semble prendre une importance de plus en plus grande pour comprendre les pathologies. Son inclusion de la diversité génétique humaine reste pourtant limitée par le manque de connaissance des séquences de populations non-européennes et par son utilisation dans les maladies complexes. Afin de mieux comprendre l'état actuel des connaissances en génomique, il est intéressant de se pencher brièvement sur les principes de bases de l'hérédité et de la génétique humaine pour évaluer comment les variations dans les populations humaines sont prises en compte.

I.1 - D'une vision péremptoire de la génétique mendélienne aux multiples intrications de la génétique polygénique

Les premiers modèles concernant l'hérédité précèdent d'un siècle la découverte de la structure et du rôle de l'ADN. Historiquement, Charles Darwin est à l'origine d'une des premières descriptions de la diversité des espèces au sein même d'une population et de la manière dont une sélection naturelle s'opère sur des petites variations favorables à des individus (9). En 1865, Gregor Mendel propose un modèle d'hérédité sur les bases de son observation de la transmission des caractères chez les plantes (10). Ses travaux liminaires restent sans support physique observé pendant de nombreuses années (11) jusqu'à de multiples découvertes dans les années 1950 : de l'idée de la molécule d'ADN comme support de l'information (12) aux premières démonstrations (13) jusqu'à la résolution de la structure physique de l'ADN (14–16).

Les années 1980 voient l'arrivée des technologies de séquençage (17) qui permettent d'identifier les bases nucléotidiques qui forment l'ADN à un locus particulier. Un locus correspond à un segment de taille variable sur l'ADN qui peut être associé à une fonction, un groupe de gènes ou simplement une position sur le génome. Ces technologies se concentrent sur l'identification de maladies rares monogéniques (*i.e.* reliées à la mutation d'un seul gène) découvertes dans des familles, et dans les années 2000 plus de 1000 mutations pathogéniques sont alors connues (18), notamment grâce à la cartographie du déséquilibre de liaison (19). Le déséquilibre de liaison est l'association non-aléatoire d'allèles à différents loci, en d'autres termes c'est la corrélation entre la présence d'une variation de l'ADN avec une autre, à l'échelle d'une population. Ce déséquilibre de liaison a comme origine différents phénomènes liés au mécanismes de transmission de l'ADN sur plusieurs générations. La liaison génétique, par exemple, est la tendance qu'ont deux loci proches sur l'ADN à être transmis ensemble dans la descendance et impacte ainsi le déséquilibre de liaison observé dans une population. Cette liaison est mesurée en centimorgans (cM) contrairement à une distance physique comme le nombre de paires de bases nucléotidiques (pb). Ce déséquilibre de liaison apparaît aussi par des phénomènes de mutation ou de sélection naturelle d'un ensemble de modifications de l'ADN. Elle permet notamment d'identifier la mutation causale d'une maladie monogénique en suivant la corrélation de mutations proches avec la maladie (20).

L'assemblage du premier génome humain complet en 2001 a ensuite lancé une nouvelle vague d'étude de la génétique en fournissant une première référence pour séquencer d'autres génomes entiers (21). Cela a permis de basculer de l'étude génétique à l'étude génomique qui évalue la part d'héritabilité d'une pathologie sans a priori sur la région du génome impliquée Ces dernières peuvent aussi bien relever des liens statistiques avec des variants communs de l'ADN qu'avec des variants plus rares ou

des mutation. Les chercheurs ont ainsi pu décrire des maladies complexes, des « phénotypes affectés par l'action simultanée de plusieurs facteurs génétiques, épigénétiques (*i.e* liés aux mécanismes de transcription de l'ADN) et environnementaux, ainsi que leurs interactions », comme décrit par Manfredi (22). L'essor des technologies de génotypage et de séquençage de l'ADN dans les années 2000 a ainsi pu faciliter la récolte de données génétiques : le nombre de SNP (*Single Nucleotide Polymorphism*), polymorphismes courants de l'ADN (présents à plus de 1% dans la population), d'insertions/délétions, et de variants structuraux (*i.e.* des morceaux d'ADN répétés, ou absents) n'ont alors cessé de croître et ils ont été reconnus comme influençant de nombreux traits, de la taille aux maladies auto-immunes (23).

En 2022, un nouvel assemblage du génome humain, le CHM13 a vu le jour, comblant les parties les plus ardues à séquencer et signant la fin d'un long travail sur l'organisation du génome (24). Cependant, un pan important de la génomique consiste à comprendre la diversité des polymorphismes existant dans les populations et leur répartition. L'étude de cette diversité bat encore son plein.

I.2 - L'impact de la diversité génétique et les limites actuelles de son étude

Le projet fondateur de l'analyse de la diversité en génomique humaine est le 1,000 Genomes Project (1KGP ou 1KG), une extension du projet international HapMap (25). Comme son nom l'indique, le 1KGP a eu pour ambition de réunir et séquencer 1 000 génomes humains provenant de plusieurs continents (Figure I-1, Auton *et al.* (26)).

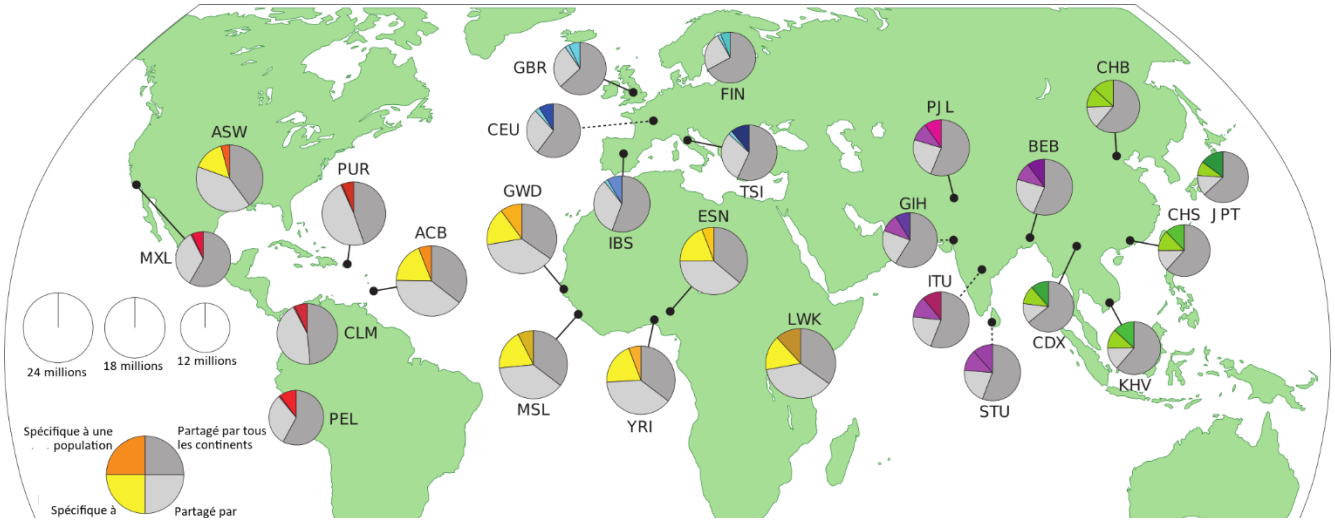


Figure I-1 Répartition des 26 populations séquencées par le 1,000 Genomes Project (1KGP). Cinq groupes, appelés super-populations, sont représentées avec des individus d'Amérique (rouge), d'Europe (bleu), d'Afrique (jaune), ainsi que de l'Asie du Sud (violet) et de l'Est (vert). La taille de chaque cercle dépend du nombre de variants identifiés. Les proportions de chaque cercle définissent la part des variants partagés par tous les continents (gris foncé), par plusieurs continents étudié (gris clair)*, exclusive au continent (couleur claire) ou à la population étudiée (couleur foncée). Traduit de Auton et al.*

Dans les faits, ce sont 2 504 individus non-apparentés, issus de 26 populations différentes, qui ont finalement été séquencés lors de la phase finale du projet (26,27). Ce projet est maintenant dirigé par l'International Genome Sample Resource (IGSR) et ses données sont encore une référence en terme de diversité (28). L'ISGR a récemment séquencé à nouveau les données 1KG pour observer les variants les plus rares (29). Ils prévoient également d'étendre le nombre d'individus dans le futur.

Ce projet a permis de remettre la diversité au centre de l'analyse génomique mais également de démontrer la répartition de ces polymorphismes entre les populations en suivant les découvertes de Lewontin (30). Dès 1972, Richard Lewontin prenait position sur les différences observées entre les populations et l'inintérêt de leur attribuer des catégories, malgré de nombreux détracteurs (31). Même si les populations mondiales présentent des fréquences de polymorphismes différentes, au regard de la diversité à l'intérieur même de ces populations, elles sont mineures. Par ailleurs, à l'occasion du 50ème anniversaire de cet article, Maróstica *et al.* ont démontré cette même répartition dans une des régions les plus complexes du génome, comportant de nombreux polymorphismes : le Complexe Majeur d'Histocompatibilité (CMH) (32).

Malgré ces constats, l'analyse génomique souffre actuellement d'un problème de représentation. En effet, plus de 90% des données de séquences concernent des individus européens (33). Un observatoire de cette diversité, le GWAS Diversity Monitor, a même été mis en place pour suivre l'évolution de la composition des données de génomique (34). Ce manque d'inclusion a un impact direct sur la fiabilité des scores de risque polygéniques par exemple, qui sont censés évaluer le fardeau génétique d'individus vis-à-vis d'une pathologie. Or, les différences de fréquence de SNP entre les populations génétiques mènent à des valeurs de risque ininterprétables lorsque les individus ne sont pas européens. Cependant, ils sont pour la plupart inadaptés aux populations non-européennes (35). Cette crise de la diversité a un impact également pour le grand public où la génétique des populations reste un fort vecteur de discriminations que l'American Society of Human Genetics a dénoncé en 2018 (36).

La diversité dans le génome humain est visible à l'œil nu mais ce n'est en réalité que la partie émergée de l'iceberg. Les personnes non-apparentées partageant un visage similaire ont ainsi des similarités génétiques (37). Tous les êtres humains diffèrent entre eux de mille et une façons qui les rend uniques. La résolution du premier génome humain, ses annotations et versions consécutives ont permis d'estimer cette diversité, de comprendre comment elle peut distinguer et décortiquer des populations grâce aux polymorphismes apportés par leurs ancêtres. Malheureusement, le manque de moyens mis en œuvre pour évaluer l'impact de la diversité sur des pathologies ou des algorithmes reste un frein à la démocratisation réelle de la génomique humaine. Cette thèse s'efforcera ainsi de mettre la lumière

sur la pluralité de la génomique humaine et son analyse. Et si l'on ne devait choisir qu'un seul locus génomique pour présenter cette incroyable diversité, le complexe majeur d'histocompatibilité (CMH) serait un choix évident. Cette région est la plus dense en gènes du génome, joue un rôle crucial dans l'immunité humaine et dispose de milliers de polymorphismes qui posent de nombreuses questions quant à leur élucidation. Les travaux de cette thèse vont donc se concentrer sur cette région, et en particulier sur la famille des gènes du *HLA* qu'elle contient.

II - La diversité du HLA et son intrication dans la biologie humaine

II.1 - Le Complexe Majeur d'Histocompatibilité (CMH), un génome de l'immunité dans un génome

Les génomes humains regorgent de diversité : des polymorphismes simples, mutations, insertions/délétions, ou changements structuraux ; à des fréquences différentes entre des populations géographiquement éloignées mais également au sein de populations proches. Cette diversité a un impact sur tous les traits pouvant affecter l'humain. L'exemple parfait de l'impact de la diversité génétique sur les humains est le Complexe Majeur d'Histocompatibilité (CMH). Le CMH est une région génomique située sur le chromosome 6 et son nom découle de sa découverte comme un facteur de compatibilité important lors de la transplantation.

II.1.1 - La découverte du HLA en transplantation

Dès 1936, l'immunologiste Peter Gorer parvient à différencier des souris consanguines à partir des anticorps contenus dans du sérum de lapin, ces différences sont appelées types antigéniques car les anticorps reconnaissent des molécules nommées alors antigènes (38,39). Il faut cependant attendre 1958 pour que Jean Dausset (40), Rose Payne (41) et Jon Van Rood (42) identifient de manière concomitante des « iso-leuco-anticorps » permettant d'agglutiner les cellules du sang. Tout comme les différents groupes sanguins, ces facteurs semblent reconnus par des anticorps qui fixent les molécules considérées comme étrangères au corps. Plusieurs années de recherche et congrès immunogénétiques ont été nécessaires pour élucider l'organisation génétique de ces antigènes, nommés par la suite Human Leukocyte Antigen (HLA).

Il a été rapidement démontré que ces antigènes étaient essentiels en transplantation et, à l'instar des groupes sanguins, que le HLA était un facteur d'histocompatibilité (i.e. compatibilités des tissus entre eux). Lors de greffes de peau, Ceppellini *et al.,* ainsi que Amos *et al.* ont démontré que le greffon était conservé plus longtemps entre des frères et sœurs qui partageaient leurs gènes HLA (43–45). En 1965, l'idée de l'importance de la correspondance HLA entre deux individus pour la greffe de rein a aussi émergé (46). Bien que la médecine actuelle, notamment l'amélioration des traitements

immunosuppresseurs, a permis d'augmenter la survie du greffon entre donneurs et receveurs de HLA différents (47), le HLA reste un marqueur essentiel dans les greffes. La greffe de cellules souches hématopoïétiques, à l'origine de toutes les cellules sanguines, est nécessaire dans de nombreux cancers et maladies auto-immunes. Elle doit strictement suivre cette compatibilité HLA (48,49). La nécessité de trouver un donneur compatible HLA dans ce cas est toujours essentielle et des outils sont développés pour trouver ces donneurs (50). La dénomination d'antigène donnée au HLA vient de sa capacité à être reconnu par le système immunitaire, chargé de morceaux de protéines, ce qui mène à la génération d'anticorps contre lui. Plus tard, on trouvera sa fonction de présentation d'antigènes.

Le HLA est encore à ce jour le système majeur de la région du CMH, que ce soit pour leur rôle d'antigène dans la transplantation, sa diversité de gènes et son interaction avec l'évolution des pathogènes ou dans les maladies auto-immunes. Malgré tout, le reste de la région du CMH en elle-même regorge de gènes cruciaux dans l'immunité humaine et même dans d'autres voies de signalisation, comme le souligne très justement Eric Thorsby : « En faisant le constat qu'il est trop tard pour le renommer actuellement, il est admis qu'au lieu de nommer le complexe HLA, ainsi que les autres complexes similaires dans d'autres espèces le complexe majeur d'histocompatibilité, ces complexes de gènes auraient dû être appelés complexe majeur de la réponse immunitaire » (45) (traduit de l'anglais).

II.1.2 - Le HLA est l'antigène qui cache la forêt

Chez l'humain, bien qu'il soit courant de voir les termes HLA et CMH utilisés de manière indifférenciée, ils désignent cependant des concepts différents. Le HLA désigne les molécules impliquées dans la présentation d'antigènes, de manière directe (grâce à *HLA-A*, *HLA-DRB1*, etc.) ou indirecte (*HLA-DOA*, etc.), ainsi que certains pseudogènes (*HLA-S*, etc.). Le CMH fait référence à la région complète dans laquelle se trouve les gènes du *HLA* (45).

La distinction entre les molécules et la région du CMH est plus rare chez les autres espèces référencées dans l'IPD-MHC (*Immune Polymorphism Database*), la base de données de référence pour les molécules non-humaines de présentation d'antigènes (51), où l'on parle souvent de CMH de l'espèce concernée. Celui-ci correspond également à une région similaire à celle des humains, riche en gènes malgré des différences d'organisation (52). Des dénominations particulières existent pour plusieurs espèces, pour le système analogue au HLA, comme BoLA chez les bovins, ou DLA chez les chiens. Cependant, les gènes nommés dans ces systèmes ne suivent pas systématiquement cette nomenclature et présentent souvent la sous-espèce dans laquelle est trouvée le gène : il existe un gène *BoLA-DQB* pour *Bos sp.*, mais aussi un gène *Bogr-DQB* pour *Bos grunniens*. Cette nomenclature, appliquée à plusieurs dizaines d'espèces actuellement est vouée à évoluer au fur et à mesure des découvertes comme le HLA le fait.

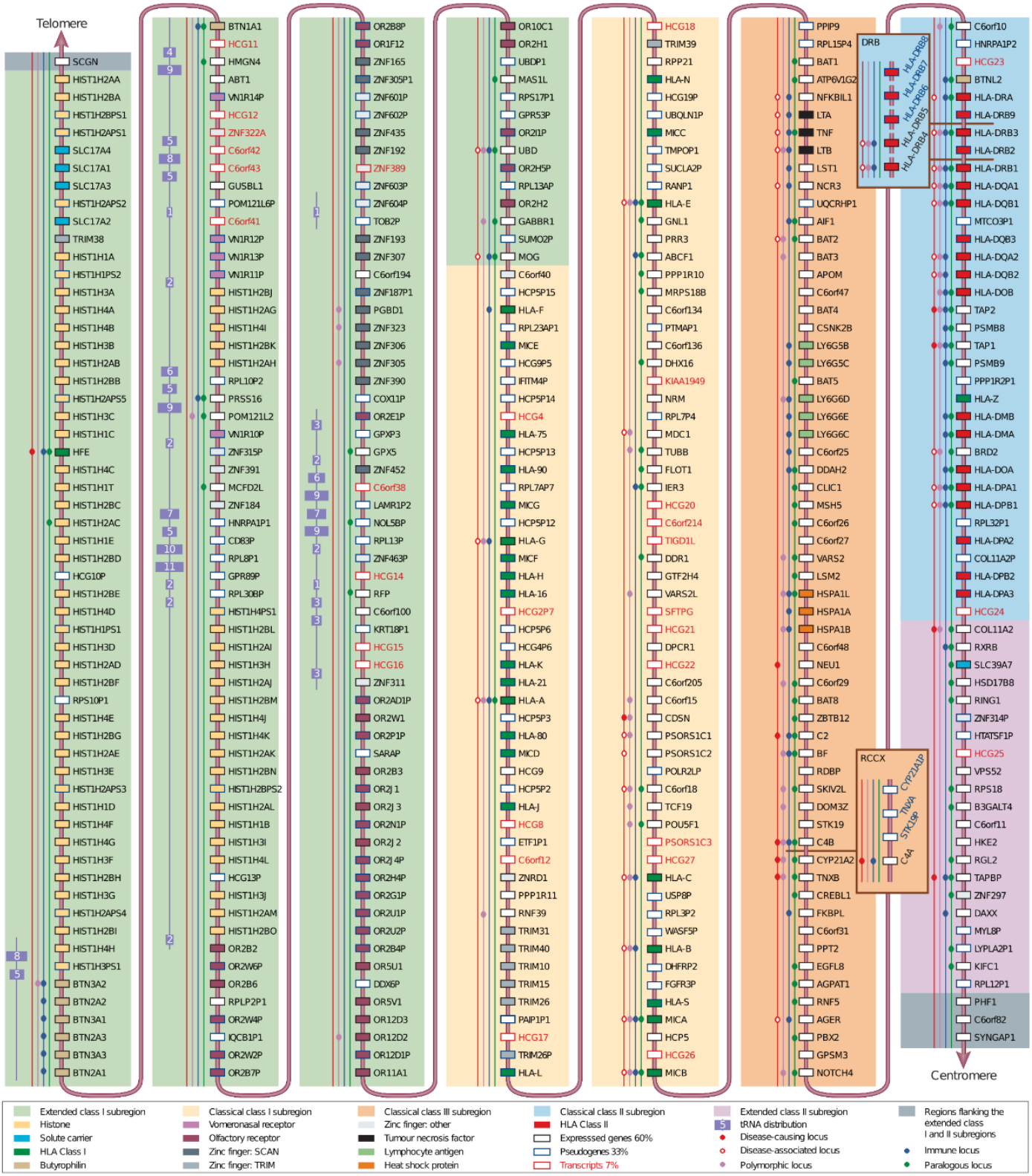


Figure II-1 Les gènes de la région étendue du CMH (CMHx). Le CMHx est divisé en trois sous-régions : classe I, III et II. Le CMH classique est plus court, alors que sa version étendue gagne 3Mb en amont de la classe I (côté télomère) et en aval de la classe II (côté centromère). Issu de Horton et al.

Comme l'atteste la confusion entre CMH et HLA, on amalgame souvent la région entière à la fonction de présentation d'antigènes du HLA. Pourtant, sur l'ensemble des gènes du CMH, seulement 39 sont du HLA, soit moins d'un quart (Figure II-1, Horton *et al.* (53)). Survoler le CMH permet de voir qu'il contient d'autres gènes, non-HLA, mais reliés au système HLA comme *TAP1*, *TAP2* et *TAPBP* (54–56) qui participent à la sélection et au transport des peptides ainsi qu'à leur chargement dans certaines molécules HLA. La région entière joue aussi un rôle central dans l'immunité humaine grâce à de nombreux gènes en dehors du système HLA. Les gènes *C2*, *C4A*, ou *CFB* font partie de différentes branches du système du complément et sont essentiels dans la formation d'un complexe d'attaque membranaire, par la reconnaissance d'anticorps, ce qui permet de lyser des cellules endommagées ou des bactéries (57). Sur un autre pan de la réponse immunitaire, le gène *TNF* code pour la cytokine TNFα, connue pour son activation de voies de signalisation liées à l'inflammation et à la mort cellulaire (58). Enfin, d'autres fonctions plus ubiquitaires sont représentées dans la région. Les gènes *HSPA1L*, *HSPA1A*, et *HSPA1B* codent notamment pour des protéines de la famille des Hsp70 qui exercent un rôle essentiel de chaperonne, dans le repliement et la dégradation des autres protéines. Le gène *TUBB* code quant à lui pour la tubuline-β, qui forme un hétérodimère avec la tubuline-α pour donner les microtubules, des cylindres protéiques structurant le cytosquelette des cellules neuronales et intervenant lors de la mitose et de la méiose (59).

Ces exemples démontrent bien que la région du CMH dispose, en dehors de la diversité exceptionnelle des molécules HLA, de nombreuses particularités : une organisation différente dans les autres espèces, et des loci avec des rôles divers chez l'humain. Il est important de garder en tête l'intérêt de la région entière d'un point de vue génétique et immunologique. Néanmoins la suite des travaux présentés dans cette thèse explorera le *HLA* et les gènes impliqués dans la fonction de présentation d'antigènes.

II.2 - Le système HLA est la clef de voûte du système immunitaire adaptatif

II.2.1 - Les généralités de l'immunité humaine

Le HLA a été identifié pour son rôle de carte d'identité de la cellule. La diversité d'allèles existants peut entraîner une reconnaissance de ces derniers par le système immunitaire, provoquant la production d'anticorps dirigés contre ces molécules lors de transplantations, chez les personnes enceintes, ou lors de transfusion fréquentes, on parle alors de rejet. Cependant, ces rejets sont une conséquence du rôle principal des molécules HLA dans la présentation d'antigènes, essentielle à l'immunité adaptative.

L'immunité humaine correspond à l'ensemble des cellules et de leurs actions qui constituent les mécanismes de défense face à tout corps étranger qui entre dans l'organisme. Les bactéries, les champignons, les parasites, les virus, les cellules cancéreuses ou les toxines, sont généralement considérés comme les principales causes biologiques d'une réaction immunitaire (60). L'immunité

apporte une réponse immédiate et non-spécifique, dite innée, à l'encontre de ces corps étrangers grâce aux barrières physico-chimiques naturelles ainsi qu'à certaines cellules comme les macrophages ou les cellules Natural Killer (NK) (61). Une immunité adaptative est également mise en place après environ 7 jours, le temps nécessaire pour que les lymphocytes B et T reconnaissent spécifiquement des pathogènes et puissent se multiplier. Une mémoire immunitaire se construit à la suite de cette première rencontre avec un pathogène, ce qui permettra une réponse plus rapide à l'avenir (62). Ces deux parties ne sont pas indépendantes et interagissent en s'activant l'une et l'autre (63).

Le HLA intervient de manière centrale dans l'immunité adaptative car ce sont ces molécules qui présentent les antigènes, les molécules potentiellement délétères, aux lymphocytes T. Chacun des lymphocytes T dispose d'un gène qui code pour un T-Cell Receptor (TCR), réarrangé différemment, qui va reconnaître spécifiquement un antigène, un couple HLA-peptide de l'individu. Cet ensemble forme un large répertoire capable de reconnaître virtuellement tous les pathogènes (64). Deux grandes catégories de molécules HLA vont ainsi stimuler le système immunitaire adaptatif par le biais de ces TCR : les HLA de classe I et ceux de classe II, nommés ainsi car ils sont situés dans deux loci différents du CMH.

II.2.2 - Les HLA de classe I veillent à l'intégrité de toutes les cellules

II.2.2.1 - La structure et la fonction générale des HLA de classe I

Les HLA de classe I sont des molécules retrouvées dans la plupart des tissus de l'organisme, à la surface des cellules nucléées (65). Cette localisation quasi-ubiquitaire lui permet de remplir un rôle assez large de surveillance immunitaire. Les HLA de classe I présentent les peptides endogènes, situées dans le cytosol. En situation physiologique, ce sont des fragments de protéines du soi, appartenant à l'individu et tolérés, qui sont présentés sans effet direct. Cependant, en situation pathologique, lors d'infections par des virus ou des bactéries, des peptides extérieurs ou peptides du non-soi vont alors être présentés aux lymphocytes T et éliciter leur activation.

D'un point de vue structurel, ce sont des molécules avec une queue cytosolique, une partie transmembranaire et des domaines semblables à ceux des immunoglobulines dotées d'un sillon peptidique pour accommoder des morceaux de protéine (Figure II-2, biorender.com). Les molécules de HLA de classe I plus particulièrement se joignent à la molécule de β2-microglobuline. La molécule de HLA en elle-même dispose de trois domaines extracellulaires : α_1, α_2 et α_3. Les domaines α_1 et α_2 forment un sillon fermé sur les extrémités dans lequel sera accommodé le peptide. Ce sillon peptidique est formé d'un feuillet de huit brins β antiparallèles (*i.e.* parallèles mais de sens contraire) sur lesquels sont superposés deux hélices antiparallèles (66).

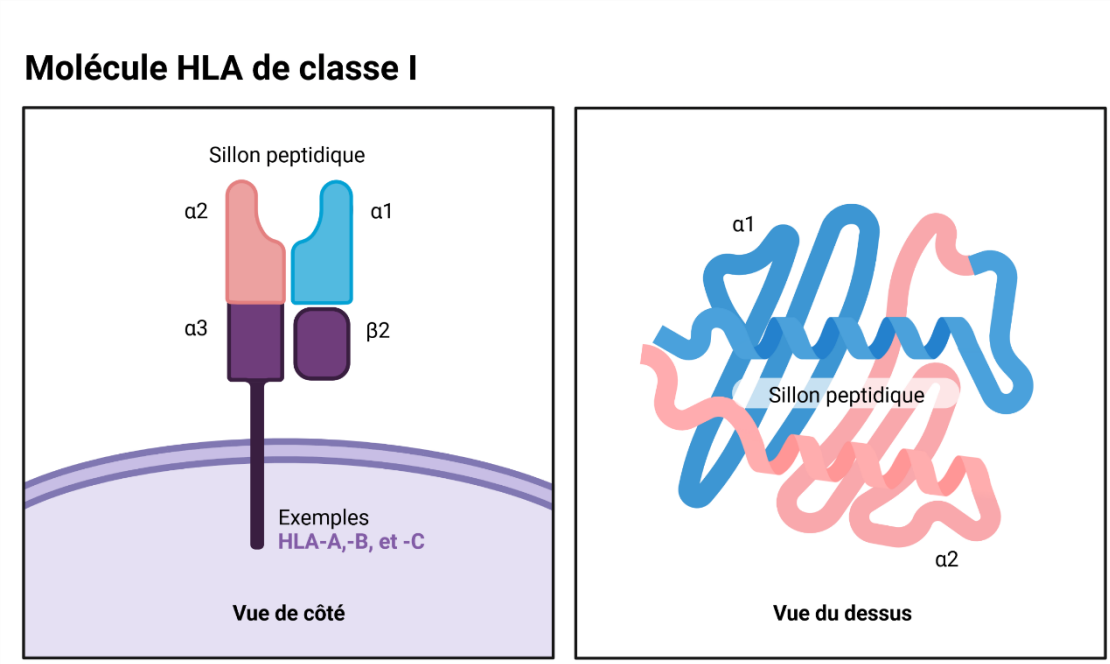


Figure II-2 Structure simplifiée d'une protéine de HLA de classe I. Les HLA de classe I se lient à une autre protéine, la β-2 microglobuline et leur sillon peptidique est formé par les domaines α_1 et α_2. Aucun peptide n'est représenté sur le schéma mais il faut noter que la stabilisation de la protéine dépend de la présence de celui-ci. Traduit d'un template biorender.com.

Dans le cas des HLA de classe I, ce sont les lymphocytes T $CD8^+$ qui vont pouvoir interagir avec leur corécepteur CD8, s'activer, se multiplier et devenir des lymphocytes T cytotoxiques (LTC). Les LTC lysent alors les cellules infectées par les peptides reconnus grâce à la perforine et au granzyme B (67). Cet effet des LTC est particulièrement efficace pour endiguer la progression des infections. Afin que cette réponse soit mise en place, une molécule de HLA chargée d'un peptide viral doit en premier lieu être amenée à la surface de la cellule. Ce processus crucial implique de nombreux acteurs protéiques.

II.2.2.2 - Le chargement des peptides

La présentation des peptides est un événement obligatoire pour la stabilisation et l'adressage à la surface de la plupart des molécules de HLA (68). Toutes les protéines présentes dans le cytosol sont soumises à l'action du protéasome (69,70), un ensemble protéique qui les découpent en peptides, ces peptides sont ensuite dégradés par les peptidases (Figure II-3, traduit et adapté d'un template biorender.com) (71). Cette dégradation par le protéasome se produit naturellement pour réguler l'expression des protéines et éliminer celles résultant d'erreur de traduction (72).

Certains peptides échappent à cette seconde dégradation et traversent la membrane du réticulum endoplasmique (RE) par la protéine TAP (partie 2) (73). Les molécules de HLA sont traduites dans le RE et TAP aide à l'ancrage des peptides dans leur sillon peptidique (74). Avec l'aide d'autres chaperonnes (ex. la tapasine, la calreticuline et ERp57), elle maintient la molécule HLA pour permettre cette

association (partie 3) (75). Les peptides qui se fixent font de 8 à 10 acides aminés et l'aminopeptidase ERAP1 hydrolyse en permanence les peptides plus longs que 9 acides aminés du RE tandis que les peptides plus courts que 8 acides aminés ne se fixent pas au HLA et sont renvoyés dans le cytosol (76). Seule une petite proportion des peptides de 8 ou 9 acides aminés, se lie dans le sillon peptidique d'une molécule de HLA (77), qui est ainsi relâchée et transportée à la membrane de la cellule (partie 4).

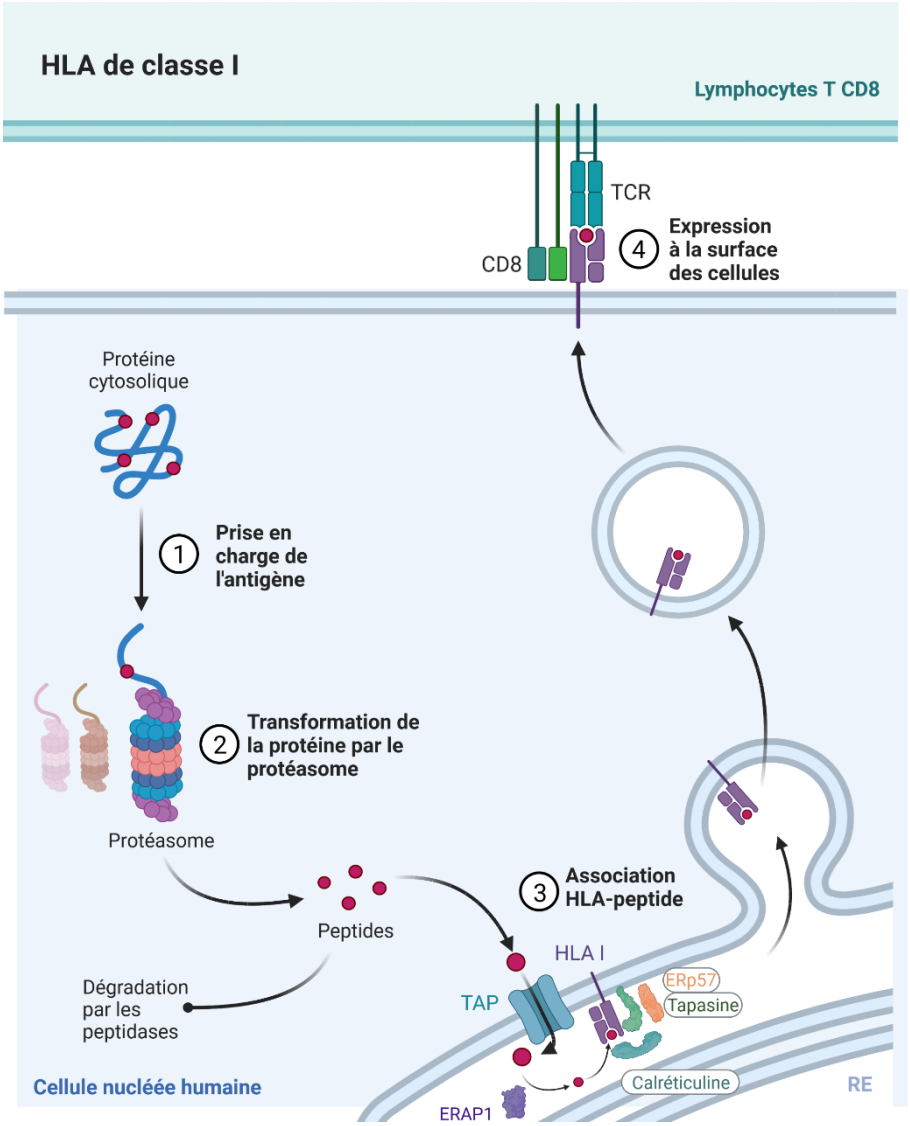


Figure II-3 Le chargement des peptides endogènes dans les HLA de classe I. Le protéasome dégrade des protéines pour réguler leur concentration, ou celles avec un mauvais repliement. Les protéines cytosoliques ont comme origine le soi, mais peuvent également être celles de virus ou de bactéries intracellulaires. Traduit et adapté d'un template de biorender.com.

II.2.3 - Les HLA de classe II alimentent la réponse immunitaire innée

II.2.3.1 - La structure et la fonction générale du HLA de classe II

Les HLA de classe II jouent un rôle similaire de présentation d'antigènes à celui des HLA de classe I. Cependant, les peptides présentés sont exogènes, c'est-à-dire d'abord phagocytés par la cellule. Les molécules HLA de classe II sont ainsi exprimées par des cellules présentatrices d'antigène : des cellules de l'immunité innée, comme les cellules dendritiques, les monocytes, et les macrophages ; des cellules de l'immunité adaptative, les lymphocytes B ; et les cellules épithéliales, mais en condition d'inflammation seulement (78).

La structure des molécules HLA de classe II est semblable à celle des molécules de classe I au niveau du sillon peptidique (66), cependant elle diffère d'un point de vue plus global (Figure II-4, traduit d'un template biorender.com). En effet, les molécules de HLA de classe II sont des hétérodimères formés de deux molécules HLA codées par des gènes différents avec une partie α et une partie β (ex. HLA-DRA1 et HLA-DRB1). Ce sont les domaines α_1 et β_1 des deux molécules qui vont créer le sillon peptidique, ouvert de part et d'autres pour accommoder des peptides de taille variable.

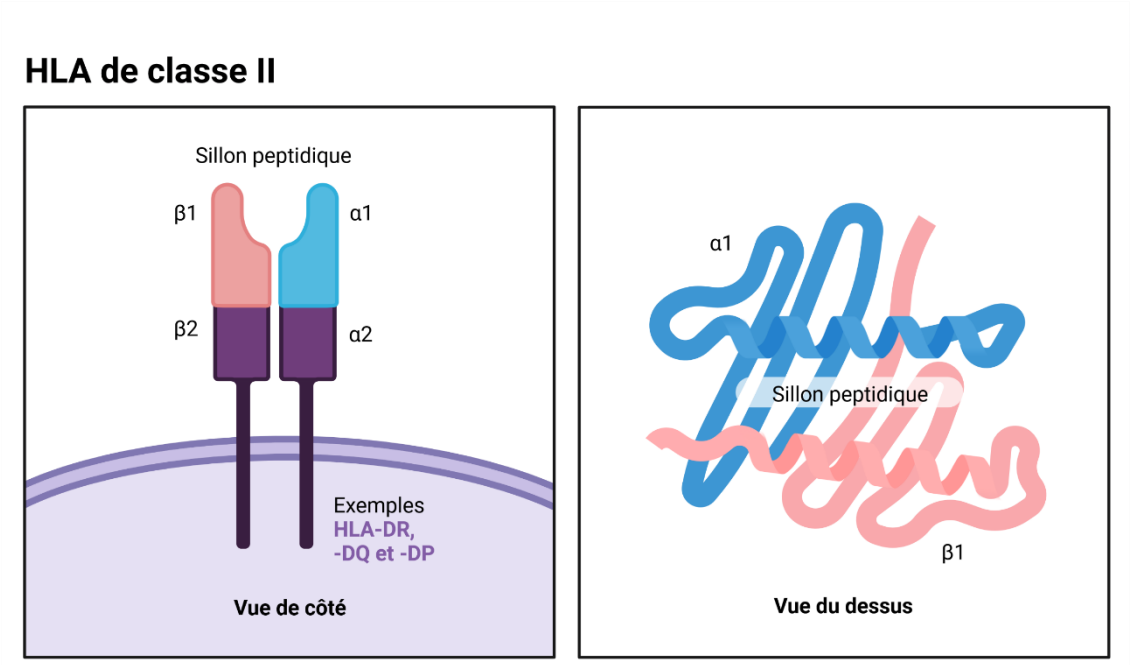


Figure II-4 Structure simplifiée d'une molécule HLA de classe II. Le HLA de classe II est constitué de deux molécules de HLA différentes α et β qui s'associent pour former le sillon peptidique. Le polymorphisme caractéristique des molécules HLA est le plus souvent restreint à la partie β de la molécule. Traduit d'un template de biorender.com.

Seuls les lymphocytes T $CD4^+$ interagissent avec les HLA de classe II par le biais de leur TCR et de leur corécepteur CD4 : l'activation entraîne une modification de leur profil d'expression pour devenir des lymphocytes T Helper (T_h). Ces lymphocytes T_h disposent d'un phénotype différent selon le cocktail de cytokines présent dans leur environnement lors de l'activation. Deux types majeurs, T_h1 et T_h2,

favorisent les réponses cellulaires et humorales respectivement (79). D'autres populations ont été découvertes comme les T_h17 (80) qui sont spécialisés dans l'inflammation des tissus, ou les T_{regs} (81) qui ont la capacité de réduire la réponse immunitaire et même d'atteindre la tolérance par homéostasie immunitaire.

II.2.3.2 - Le chargement des peptides

De la même façon que pour le HLA de classe I, les molécules HLA de classe II sont traduites dans le RE. Cependant elles en sortent rapidement en se liant à la chaîne invariante Ii (Figure II-5, traduit et adapté d'un template biorender.com). Cette protéine dispose d'un domaine CLIP (*Class II-associated invariant chain peptide*) qui remplit le rôle du peptide en se logeant dans le sillon peptidique (82). Une vésicule contenant les molécules de HLA de classe II fusionne alors avec un endolysosome, issu de la

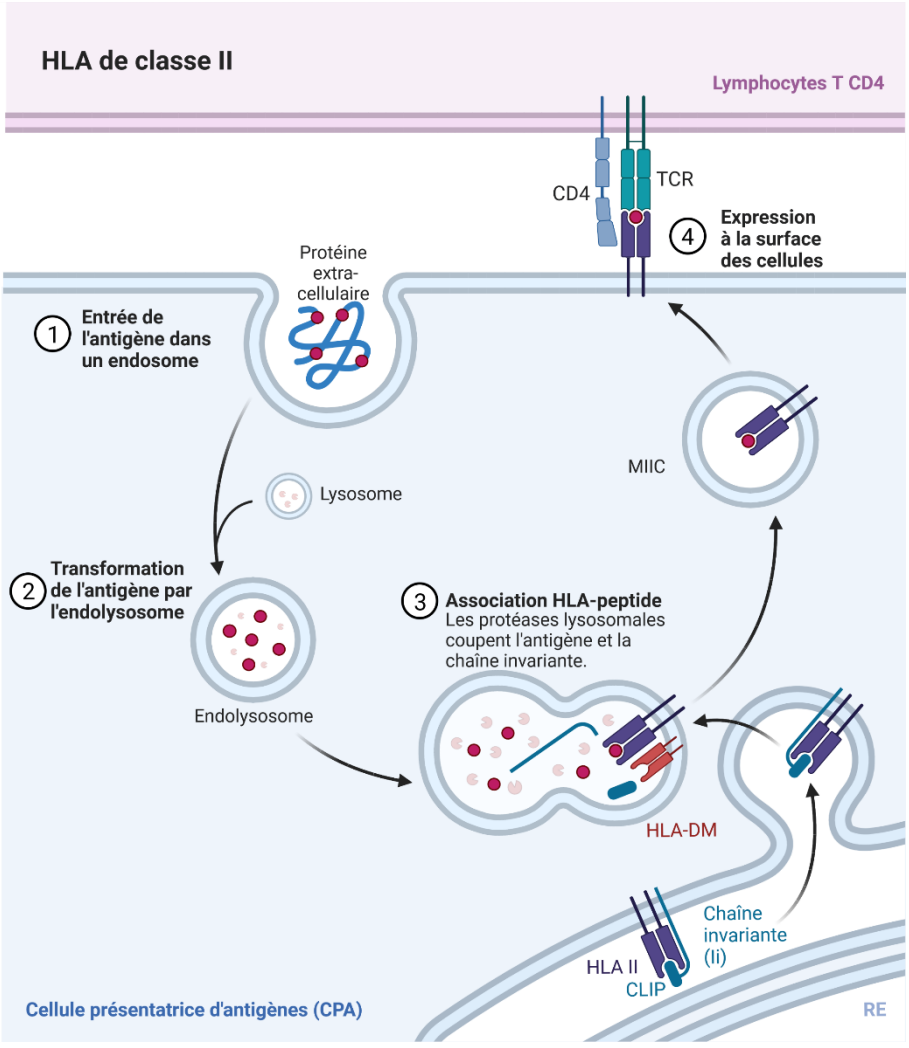


Figure II-5 Le chargement des peptides exogènes dans les HLA de classe II. Les protéines extracellulaires sont intégrées dans un endosome après phagocytose. Les protéines sont dégradées par des protéases, puis après la fusion avec une vésicule contenant des HLA de classe II, certains peptides de meilleur affinité remplacent le peptide CLIP, natif des HLA de classe II. Traduit et adapté d'un template biorender.com.

phagocytose de protéines externes par la cellule, pour devenir un endosome tardif, le compartiment HLA classe II, MIIC (83). Les protéases contenues dans l'endolysosome dégradent ainsi les protéines phagocytées mais également une partie de la chaîne invariante, laissant la partie CLIP dans le sillon peptidique. La molécule HLA-DM, à l'instar de TAP pour les HLA de classe I, va servir de chaperonne, facilitant l'échange de CLIP pour des peptides de plus forte affinité, à savoir les peptides étrangers, permettant ainsi l'adressage à la membrane d'un complexe CMH-II/peptide du non-soi (84).

Le processus diffère dans les lymphocytes B où un hétérodimère HLA-DM/HLA-DO modifie le répertoire peptidique en limitant l'accès de certains peptides (85). Ce répertoire peptidique est néanmoins très différent des HLA de classe I car le sillon peptidique de la molécule est plus large et ouvert de chaque côté. Les peptides accommodés sont généralement plus grands, de 12 à 25 acides aminés (86).

Le chargement du peptide est une étape cruciale, menée par de nombreux acteurs protéiques qui permettent à la molécule de HLA d'être exprimée à la surface des cellules. Cependant, en dehors des restrictions amenées par TAP ou HLA-DM, pour comprendre comment sont choisis les peptides présentés par les différents HLA, il faut s'intéresser à l'immense polymorphisme du sillon peptidique de ces molécules.

II.3 - La description des gènes et de la diversité du HLA : un marathon génomique et technologique

Le système HLA permet d'organiser la réponse immunitaire adaptative contre le non-soi : des virus, des bactéries ou des cellules humaines différentes issues d'une transplantation. Cette fonction immunitaire essentielle se concentre dans une faible portion du génome, qui contient un nombre important de gènes et une diversité allélique unique.

II.3.1 - La situation génomique des gènes HLA

L'ensemble du système HLA est situé au niveau du locus 6p21, sur le bras court du chromosome 6 dans le CMH. Cette région est la plus dense du génome en terme de gènes, plus de 200, soit 1% de tous les gènes humains, alors qu'elle ne compte que pour 0,1% de la longueur du génome (87). La région du CMH s'étend classiquement sur 4 millions de bases nucléotidiques (Mb) (88) du gène *MOG* au gène *COL11A2* (29 657 002–33 192 499, GRCh38.p13). En tenant compte des motifs particuliers de déséquilibre de liaison de la région, il est possible de définir un CMH étendu (CMHx / *xMHC* en anglais) selon la définition proposée par Horton *et al.* (53). Cette région peut s'étendre sur 8Mb (25 726 063–33 400 556, GRCh38.p13).

Comme présenté sur la Figure II-1, le CMH a été divisé en trois régions nommées classe I, II, et III. Aucun gène HLA n'est présent dans le CMH de classe III, ils sont ainsi divisés entre les classes I et II. La variété de structure et de fonctions HLA évoquée plus tôt correspond aussi à une localisation différente. L'ensemble des gènes du système HLA (Figure II-6, créé avec biorender.com) est situé au début et à la fin du CMH, sur les deux brins d'ADN.

Les molécules HLA de classe I sont un groupe hétérogène car elles sont traduites à partir de 6 gènes différents : *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F*, et *HLA-G*. Les trois premiers gènes (-A, -B et -C) sont dits « classiques » car ils représentent les molécules HLA exprimées dans toutes les cellules et leurs produits interagissent avec le TCR et le corécepteur CD8 des lymphocytes T $CD8^+$. Les trois derniers (-E, -F, et -G) sont dits « non-classiques » car ils ont des expressions et interactions spécifiques. Elles sont néanmoins toutes présentes sous forme d'hétérodimère en association avec une molécule de β-2-microglobuline, dont le gène est situé sur le bras long (15q21.1) du chromosome 15. Les HLA-E et -F sont ubiquitaires mais les cellules les expriment moins. De plus, ils n'interagissent pas avec des TCR (ex. le HLA-E interagit avec l'hétérodimère CD94 / NKG2A ou CD94 / NKG2C sur les cellules NK (89)). Quant à la molécule HLA-G, elle est connue pour jouer un rôle de tolérance à l'interface entre le fœtus et l'utérus, en interagissant avec le CD8 des lymphocytes T et LILRB1, LILRB2 et KIR2DL4 chez les cellules NK.

Pour les molécules de classe II, les gènes HLA classiques sont au nombre de 12, avec : *HLA-DRA*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DQB2*, *HLA-DQB3*, *HLA-DPA1*, et *HLA-DPB1*. Ce sont également des hétérodimères, mais chaque chaîne (α et β) est encodée par un gène HLA (ex. les complexes HLA-DP sont obtenus à partir des produits de *HLA-DPA1* et *HLA-DPB1*). Le cas de la chaîne β de la molécule HLA-DR est particulier puisque plusieurs gènes DRB peuvent s'associer avec DRA. DRB3/4/5 ont une variation dans leur nombre de copies (*copy number variation, CNV*), ce sont des variants structurels et peuvent être présents ou absents ; seul le gène DRB1 est présent dans tous les cas. L'organisation génomique de ces gènes peut être représentée en plusieurs haplotypes (Figure II-7, traduit de Trowsdale et al. (90)), qui peuvent être statistiquement liés à certains allèles de DRB1). Ce locus reste encore aujourd'hui un des défis de la compréhension des loci HLA.

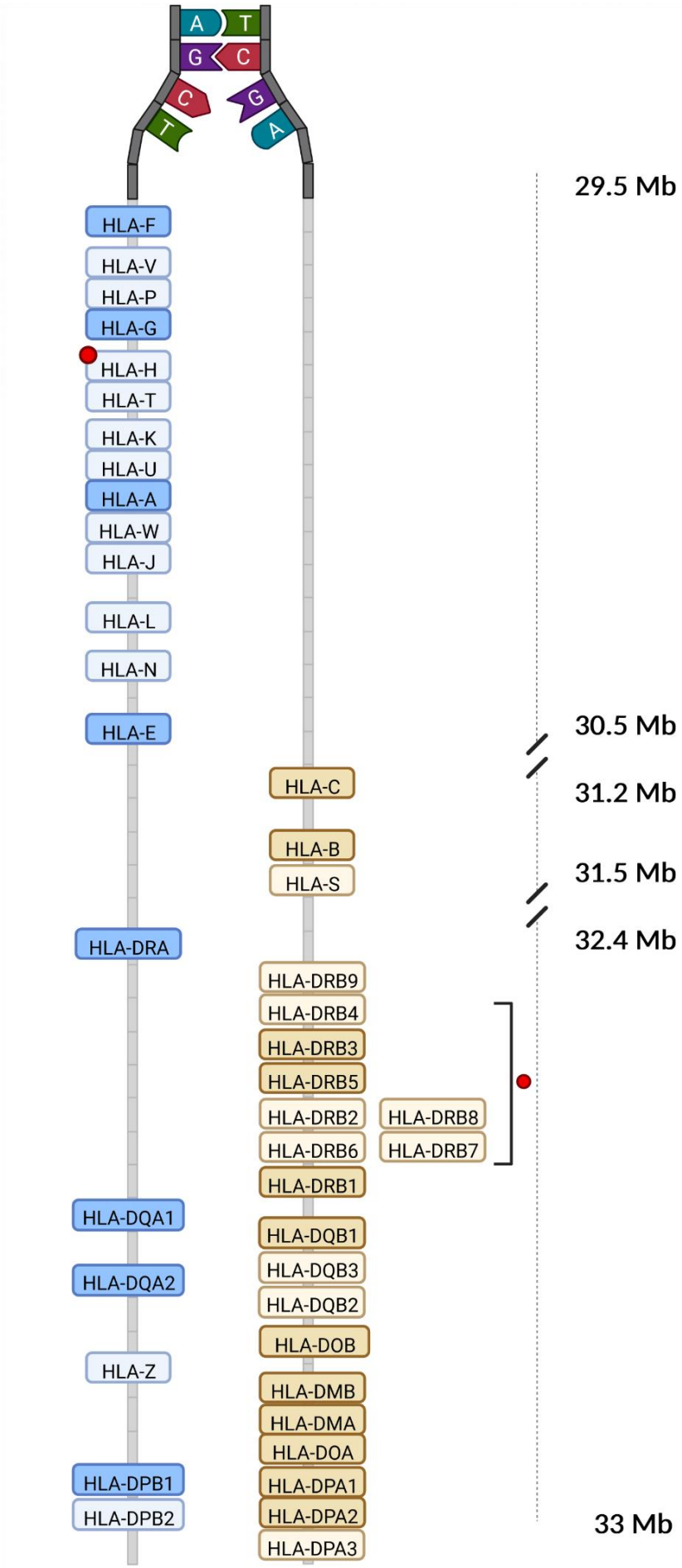


Figure II-6 Organisation génomique des gènes HLA. Le brin sens (bleu) comporte notamment le HLA-A, la plupart des pseudogènes (nuance claire) des classe I, avec les chaînes alpha de DR et DQ, ainsi que les chaînes beta de DP. Le brin anti-sens (orange) présente le reste des classe I, ainsi que le locus particulier de la chaîne beta de DR avec de nombreux variants structuraux (indiqués par un point rouge), les chaînes beta des autres classe II et les gènes HLA chaperons (DO/DM). Créé avec biorender.com.

Les 4 gènes non-classiques des molécules de classe II sont *HLA-DOA*, *HLA-DOB*, *HLA-DMA* et *HLA-DMB*. La distinction entre classique et non-classique, dans ce cas, tient dans le fait que les gènes non-classiques ne sont pas impliqués dans la présentation d'antigènes. HLA-DM et HLA-DO sont des molécules chaperonnes impliquées dans le chargement des peptides dans les molécules HLA de classe II « classiques ». HLA-DM permet l'échange du peptide invariant avec un peptide circulant (84,91), alors que HLA-DO interagit avec HLA-DM et entre en compétition avec ces autres antigènes pour le chargement dans la molécule de HLA, modifiant ainsi le répertoire de peptides pouvant être présentés (92).

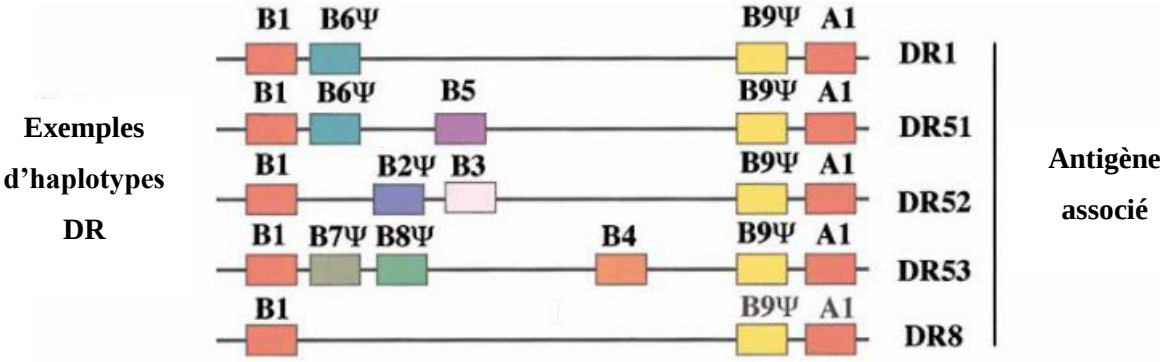


Figure II-7 Les haplotypes HLA-DRB majeurs. La région génomique de HLA-DR est connue pour sa complexité de séquençage due à ses nombreux variants structuraux. Les gènes HLA-DRB2 à HLA-DRB8 sont présents ou absents selon les haplotypes, seuls HLA-DRB3 à HLA-DRB5 sont des gènes, les autres sont des pseudogènes sans protéine associée. Traduit de Trowsdale et al.

En outre, chacune des classes dispose de pseudogènes HLA. Ces derniers sont des gènes, à l'origine codant pour une protéine, qui ont reçu une ou plusieurs mutations inactivant leur fonction principale (93). La région CMH de classe I dénombre 12 pseudogènes : *HLA-H, HLA-J, HLA-K, HLA-L, HLA-N, HLA-P, HLA-S, HLA-T, HLA-U, HLA-V, HLA-W, HLA-Y* ainsi que *HLA-X* et *HLA-Z*. Dans la région classe II, il y a 8 pseudogènes HLA : *HLA-DPA2, HLA-DPA3, HLA-DPB2, HLA-DRB9* et *HLA-DRB2/6/7/8* qui sont des variants structuraux.

Tout comme certains allèles, certains de ces pseudogènes ne sont pas décrits dans la base de données officielle IMGT-HLA (IMunoGeneTics – HLA) (94) mais sont néanmoins présents dans la dernière version annotée du génome humain ainsi que dans la littérature. Dans le cas des pseudogènes HLA, même si leur rôle direct dans la présentation d'antigène est rendu caduque, le transcrit de certains ou leur produit tronqué peuvent avoir une fonction. En cela, HLA-H a été décrit comme transcrit de manière importante, autant que HLA-G et dans les mêmes types cellulaires, avec un effet de son peptide signal sur la mobilisation à la membrane de HLA-E. Son allèle HLA-H*02:07 pourrait même être traduit en une protéine fonctionnelle (95–97).

D'un point de vue génomique, certains pseudogènes ont été décrits comme absents de certains haplotypes HLA. Ainsi, suite à une délétion de 50kb entre HLA-G et HLA-A, HLA-H peut ne pas être présent dans le génome humain (98–101). Bien que sa position exacte ne soit pas connue, HLA-Y a également été identifié dans seulement 10% des haplotypes HLA chez l'humain (102,103). Enfin, HLA-V et HLA-P ont été décrits comme le même gène, cependant aucune autre référence n'a pu être trouvée à ce sujet (104).

Ainsi, la diversité allélique et structurelle des pseudogènes est encore peu étudiée et leur impact réel sur les mécanismes moléculaires du système HLA reste à évaluer.

II.3.2 - Les allèles HLA, ou la complexité par le nombre

La présentation des antigènes repose sur le système HLA, codé par un grand nombre de gènes. Cependant, la plupart des gènes HLA jouent un rôle redondant de présentation, bien que des différences de localisation et d'intensité d'expression soient connues (87). D'autres systèmes immunitaires comme la voie de signalisation NF-κB (105), qui coordonne les réponses inflammatoires, ou encore le complément (106), qui réagit aux infections et à l'apoptose, peuvent également comporter de nombreux gènes. Le caractère unique du système HLA se situe dans le nombre d'allèles de chacun de ses gènes. D'après la base de données IMGT-HLA, plus de 30 000 allèles ont d'ores et déjà été recensés. Leur répartition (Tableau II-1, IMGT-HLA (107)) correspond à deux tiers d'allèles dans le CMH de classe I et le tiers restant dans la région classe II.

Tableau II-1 Récapitulatif du nombre d'allèles différents découverts et répertoriés dans l'IMGT-HLA, rangés par nombre décroissant de polymorphismes. Extrait de IMGT-HLA le 28.09.2022.

Gène	B	A	C	DRB1	DQB1	DPB1	DQA1	DPA1	DRB3	E
Nombre d'allèles	9 000	7 562	7 513	3 298	2 278	2 067	483	455	436	311

Ces nombres sont une sous-estimation du nombre réel car : certains allèles sont décomptés alors qu'ils ne sont pas décrits à l'échelle protéique, de nombreuses variations peuvent ainsi exister à l'échelle nucléique ; une procédure est nécessaire pour ajouter de nouveaux allèles au site de référence IMGT-HLA, ainsi certains allèles sont découverts mais ne sont pas homologués ; de nombreux allèles HLA recensés ont une fréquence très faible en population, la faible couverture de typage de la population mondiale ne permet pas de connaître certains allèles ; enfin, l'ADN n'est pas figé, de nouveaux allèles sont donc créés en continu, par recombinaison ou variations ponctuelles.

Ainsi, le nombre d'allèles recensés n'a cessé d'augmenter et connaît même une accélération continue au gré de l'évolution des technologies de séquençage de l'ADN. En 2012, nous comptions environ 8 000 allèles HLA. Ce nombre a doublé cinq ans après, en 2017, puis doublé à nouveau pour atteindre les 35 000 allèles actuels (Figure II-8, extrait de IMGT-HLA le 28.09/2022 (107)). Si le développement des technologies de séquençage récentes facilite maintenant le recensement des allèles HLA, sans le rendre trivial, cela a longtemps été freiné par la complexité de cette région génomique.

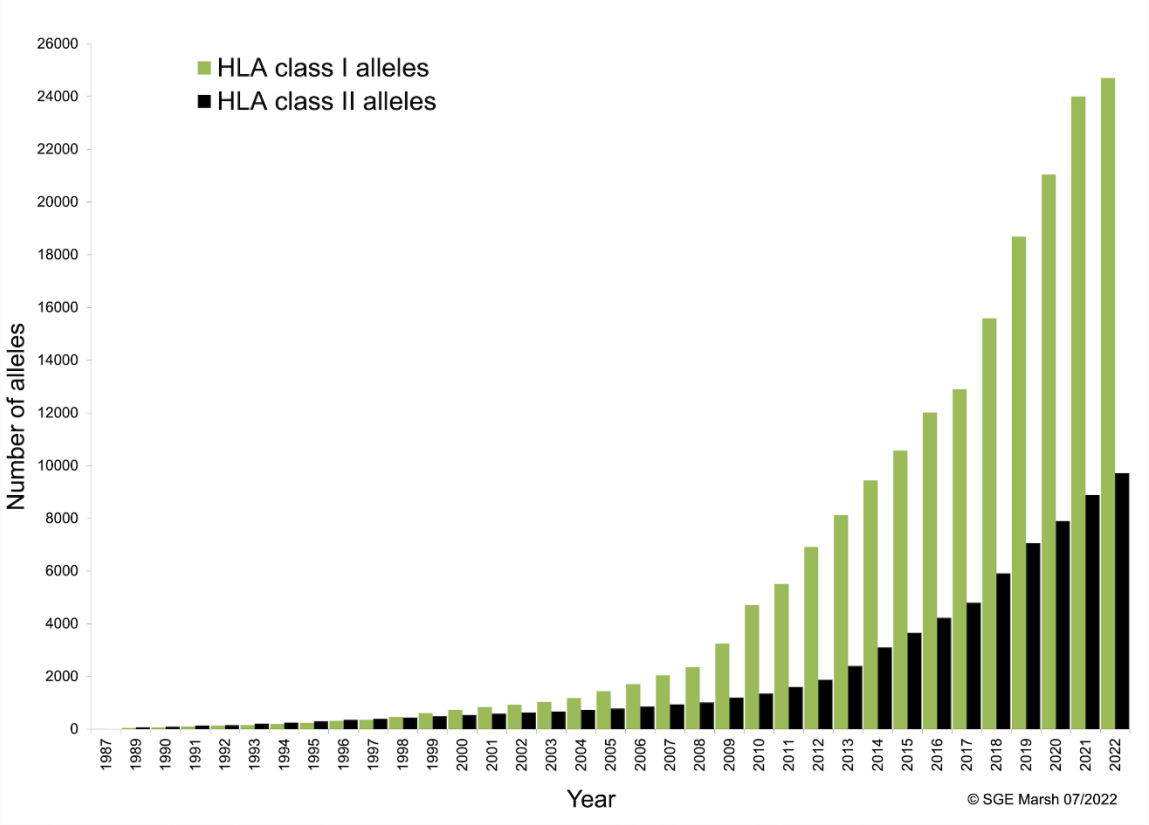


Figure II-8 Somme cumulée du nombre d'allèles HLA recensés dans la base de données IMGT-HLA chaque année, stratifiée par la région génomique d'appartenance. Dans les dix dernières années, le nombre d'allèles a quadruplé, grâce aux changements des technologies de séquençage et leur application au HLA. Ce graphique est tiré des statistiques du site IMGT-HLA, et couvre la période de 1987 à 2022. Extrait de IMGT-HLA le 28.09.2022.

II.3.3 - L'historique de l'élucidation des allèles HLA

Le typage HLA est l'identification des allèles HLA portés par un individu. Il est aussi appelé génotypage HLA, mais contrairement aux puces de génotypage qui informent sur un polymorphisme unique, celui-ci en indique plusieurs. Aujourd'hui, le séquençage HLA est le standard de typage, cependant, il a fallu plusieurs dizaines d'années avant de pouvoir connaître exactement l'organisation ainsi que la séquence exacte des gènes HLA.

II.3.3.1 - La sérologie, l'interaction anticorps-HLA

À l'origine, l'ensemble du HLA était connu comme un unique antigène découvert en 1958 par Jean Dausset (40). L'ajout du sérum d'un patient à un extrait de moelle osseuse d'un autre a mené à l'agglutination des lymphocytes présents (108), une réaction analogue à celle entre des individus de groupes sanguins différents. Cependant, les individus testés étaient de même groupe sanguin, indiquant la présence d'un antigène différent entraînant cette réaction. Celui-ci a été nommé MAC, d'après les initiales des patients possédant l'anticorps contre cet antigène (109). L'agglutination leucocytaire a été observée sur près de 60% des patients, et l'existence de cet antigène a rapidement impliqué celle d'autres antigènes similaires (40). Dans les années suivantes, plusieurs techniques basées sur cette réaction ont servi pour identifier l'ensemble des antigènes HLA, dont les tests de microlymphocytotoxicité par Terasaki *et al.* (110).

Cette technique repose sur des tests systématiques d'anti-sera ou d'anticorps monoclonaux connus sur des lymphocytes T et B de patients (pour le HLA de classe I) ou de lymphocytes B (pour le HLA de classe II). Les anticorps interagissent avec les molécules HLA correspondantes, puis un lavage élimine les anticorps qui n'interagissent pas. En ajoutant un sérum de lapin en guise de complément, celui-ci est activé par le fragment Fc des anticorps, les cellules sont ainsi lysées. En ajoutant un fluorochrome qui s'illumine seulement en présence de l'ADN libéré par la lyse puis en comparant les résultats entre les différents anticorps, il est possible d'inférer les allèles HLA de l'individu (111). Cependant, la complexité de cette méthode, additionnée aux résultats mitigés obtenus lors de son application en transplantation (112) l'ont vue remplacée par les techniques moléculaires.

II.3.3.2 - La PCR-SSO / PCR-SSP, obtenir un allèle polymorphisme par polymorphisme

Contrairement à la sérologie, les techniques moléculaires ne se basaient plus sur des reconnaissances de structure mais sur des polymorphismes connus de la séquence ADN. Elles ont permis de différencier les allèles HLA qui étaient reconnus par les mêmes anticorps mais qui disposaient de séquences différentes. La PCR-SSO et la PCR-SSP sont les premières techniques de typage moléculaire utilisées. Elles reposent sur la PCR (Polymerase Chain Reaction), un ensemble de réactions basées sur des cycles de dénaturation et hybridation des amorces (*i.e.* des séquences nucléotidiques complémentaires) situées aux abords d'un locus d'intérêt pour l'amplifier grâce à l'ADN polymérase thermorésistante *Taq*.

La PCR-SSO, pour PCR-Sequence Specific Oligonucleotide, amplifie un locus entier HLA à l'aide d'amorces biotinylées. À l'origine, ces séquences amplifiées étaient immobilisées sur plusieurs membranes, puis mises en contact de sondes oligonucléotidiques. Chaque polymorphisme était révélé

indépendamment en ajoutant de la streptavidine, se fixant à de la biotine ajoutée sur les amorces, fluorescent ou liée à une enzyme permettant une réaction colorée (113,114). Cependant, cette technique de transfert (*dot blot* en anglais) nécessite un nombre important de membranes et est compliquée à mettre en œuvre.

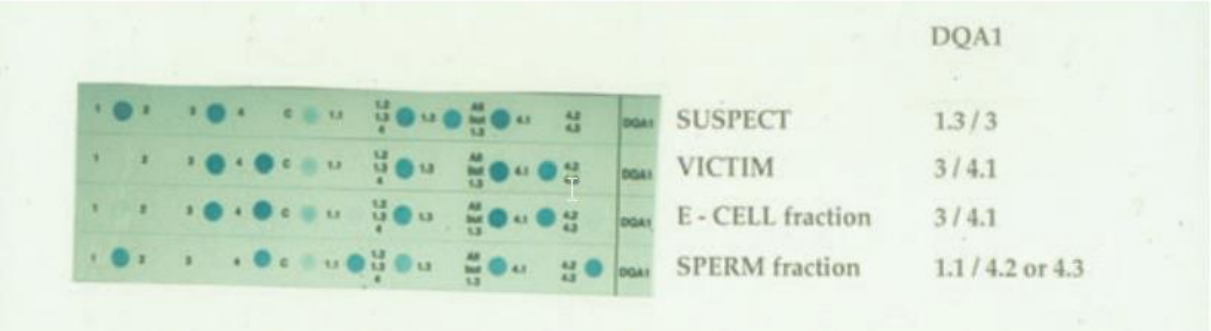


Figure II-9 Un des premiers kit commerciaux de "reverse dot blot" pour HLA-DQA1, pour les analyses médico-légales. Les sondes spécifiques des allèles de DQA1 sont immobilisées sur une membrane de nylon, puis des amplicons biotinylés venant des individus à tester sont déposés. L'ajout d'une peroxidase liée à de la streptavidine, et de son substrat permet d'identifier les polymorphismes de chaque individu à l'oeil nu. Issu de Erlich et al.

Pour faciliter le typage, le *reverse dot blot* (Figure II-9, Erlich *et al.* (115)) fixe les sondes oligonucléotidiques sur une première membrane et l'amplicon du gène HLA étudié est ensuite ajouté, réduisant le nombre d'expériences (115). Le nombre de sondes a pu augmenter au fil des années, rendant cette technologie de plus en plus efficace (116).

La PCR-SSP, pour Sequence Specific Primers, évalue les polymorphismes HLA directement par l'utilisation d'amorces spécifiques aux polymorphismes dans plusieurs PCR en parallèle (117,118). Ainsi, les paires d'amorces peuvent être complémentaires au polymorphisme ou non-complémentaires par substitution d'un nucléotide en 3', empêchant tout amplification par la *Taq*. Chaque PCR effectuée est révélée classiquement sur un gel d'agarose par bromure d'éthidium en même temps qu'un gène de ménage amplifié en parallèle (113,115). Selon les ensembles d'amorces disponibles, les produits de PCR obtenus indiquent avec plus ou moins de précision les allèles HLA possibles.

La PCR-SSO et la PCR-SSP sont toutes deux des méthodes couramment utilisées mais limitées. Le typage HLA peut être ambigu sans accès à de nombreuses sondes ou amorces. La PCR-SSP peut régler ces ambiguïtés mais n'est pas adaptée pour plusieurs individus (119). Enfin, elles nécessitent beaucoup de mise en place, elles se concentrent sur les exons les plus polymorphes, et il est impossible de typer des allèles qui n'ont pas encore été découverts.

II.3.3.3 - La PCR-SBT, les premières séquences complètes

La génération des banques de sondes et d'amorces a été possible grâce à la découverte des polymorphismes par des méthodes de séquençage, notamment la PCR-SBT, pour Sequence Based Typing ou séquençage de Sanger. Une amplification par PCR des exons des gènes HLA d'intérêt utilise des nucléotides classiques ainsi que des nucléotides terminateurs fluorescents de quatre couleurs (pour chaque nucléotide), empêchant toute élongation. Le passage des produits de PCR dans un capillaire d'électrophorèse permet ainsi de lire les copies des exons, stoppées à différentes tailles, et donc de connaître leur séquence exacte (120–122).

La PCR-SBT est restée le standard de génération de données HLA pendant plusieurs années, mais plusieurs problèmes sont liés à son utilisation. Bien qu'il soit possible de séquencer l'entièreté du gène, en pratique cela se limite aux exons les plus polymorphes (123) car il s'agit d'une technique longue et coûteuse.

II.3.3.4 - Le Next-Generation Sequencing (NGS)

Les technologies NGS, ou technologies de séquençage de deuxième génération (après Sanger), sont un ensemble de machineries et techniques de séquençage développé par plusieurs industriels, comme Roche (124), Illumina, et Life Technologies (125). Elles sont entrées dans le domaine de la recherche entre 2005 et 2007 et ont simplifié l'accès au séquençage, notamment en réduisant le nombre d'étapes nécessaires pour l'obtention des résultats et en permettant le séquençage en parallèle de millions de séquences courtes, ou reads, le long du génome pour le reconstruire entièrement (126).

En 2012, l'application des NGS pour le typage HLA a été évaluée par le 16ème International HLA and Immunogenetics Workshop (IHIW) qui a mis en lumière leur intérêt dans l'exploration de la région du CMH mais également leurs limites. Les NGS sont capables de résoudre certaines ambigüités des PCR-SBT mais manquent encore d'outils d'analyse, et la taille des reads est parfois limitée pour identifier certains loci (123). Cela explique le succès continu des PCR-SSO/SSP et SBT. Plus tard, en 2014, Gabriel *et al.* soulèvent la difficulté d'assemblage du génome au niveau des gènes HLA après un séquençage (127). Puis, en 2015, Hosoimichi *et al.* saluent l'arrivée des premiers kits NGS pour le HLA mais les problèmes de résolution de la phase, donc de détermination des haplotypes de toutes les régions, ralentissent encore la démocratisation des NGS (128). La recherche continue en bioinformatique et en immunogénétique a mené vers de nouveaux algorithmes d'alignement des reads de la région du CMH sur un génome de référence. Le 17ème IHIW a ainsi statué positivement sur l'utilisation des NGS sur le typage HLA, toujours avec des réserves sur l'analyse des séquences répétées (129) ; ce nouveau standard a été confirmé plusieurs fois, mettant en avant un taux d'erreur plus faible que la PCR-SBT

(130,131). Les technologies NGS, notamment Illumina, permettent maintenant de typer les allèles HLA en routine (132–135).

Une troisième génération de séquenceurs a récemment fait surface et elle permet notamment de lever les problèmes de phase et les ambiguïtés en travaillant avec reads plus longs, comme PacBio SMRT (136) ou même en analysant l'ADN en molécule entière avec Oxford Nanopore Technology MinION (137). Bien que des problèmes analogues à la seconde génération ont ralenti leur usage, Mosbruger *et al.* ont récemment pu les utiliser pour séquencer 11 loci HLA avec un taux d'erreur faible de 0.02% en moyenne (138).

Les nouvelles technologies de séquençage sont maintenant un maillon essentiel de la chaîne de production des données HLA, que ce soit dans la recherche fondamentale ou en clinique. Des améliorations continues sur les algorithmes permettent maintenant l'assemblage du génome dans la région du CMH. Actuellement, huit séquences de références décrivent la région du CMH, cependant une seule est utilisée lors de l'alignement (104). Dilthey *et al.* ont démontré qu'augmenter le nombre de séquences de référence améliore le typage. Ainsi, pour intégrer toutes les variations de la région, ils ont proposé un modèle de référence de génome par graphe qui prend en compte tous les allèles connus, sans faire de séquence consensus (139).

Le séquençage HLA est donc en constante évolution. L'antigène unique MAC identifié par Jean Dausset est alors successivement devenu HLA-A2, puis HLA-A*02:01, et enfin HLA-A*02:01:01:01. Ces différentes notations représentent l'évolution de la nomenclature du HLA qui a été créée au fur et à mesure des changements technologiques et de la meilleure compréhension de la région du CMH.

II.3.4 - La nomenclature HLA

L'évolution progressive des technologies a ainsi permis de caractériser de plus en plus précisément la séquence des allèles HLA. Contrairement à des gènes dont les mutations ponctuelles sont souvent notées par le nom de la substitution, comme *Y192L* qui indiquerait une substitution d'une tyrosine par une leucine en position 192 de la protéine, des milliers d'allèles HLA diffèrent par plusieurs acides aminés. L'utilisation du terme « allèle » est par ailleurs particulier, car dans le HLA un allèle peut désigner un ensemble de variations de séquence menant à la production d'une même protéine, mais aussi bien une seule de ces variations de séquence. Ces différentes manières de nommer le HLA sont appelées des résolutions et n'ont cessé d'évoluer depuis la découverte du HLA.

II.3.4.1 - Les différentes résolutions HLA

À l'origine, l'antigène MAC n'est qu'un seul facteur causant une réponse immunitaire. Cependant, le polymorphisme associé à cette région a rapidement été identifié et le WHO Nomenclature Committee

for Factors of the HLA System a été créé. Il a nommé le HLA d'après *Hu* et *LA*, les deux noms utilisés pour les antigènes jusqu'alors découverts (110), puis une nomenclature spécifique a été proposé pour tous les produits du HLA (gènes, protéines et allèles) (140).

La première nomenclature s'est donc basée sur un accord entre les différents laboratoires pour nommer chacun des facteurs identifiés par « HL-A » suivi d'un nombre. Cela a permis d'homogénéiser les notations, tout en prenant en compte les deux versions du facteur, molécule reconnue par un anticorps, porté par un individu. De fait, une personne positive pour les facteurs 1 à 4 seulement peut être notée HL-A1,2,3,4. Le typage HLA de ses parents pouvait permettre de distinguer les haplotypes des différents facteurs, sous la forme HL-A(1,3/2,4) par exemple si les couples de facteurs 1,3 et 2,4 sont identifiés comme portés par des chromosomes différents. Des facteurs ont pu être ajoutés à mesure de leur découverte. Puis, la découverte de dizaines de gènes HLA et de milliers d'allèles a poussé à étendre la nomenclature pour pouvoir nommer les allèles. Les nouveaux noms sont formés du locus HLA (*i.e.* HLA-A, HLA-DQB1, …), suivi de quatre chiffres pour définir les protéines différentes (141). La dernière modification de 2010 (142) introduit la notion de champ pour définir plus facilement les polymorphismes synonymes, ainsi que ceux situés en dehors des régions codantes.

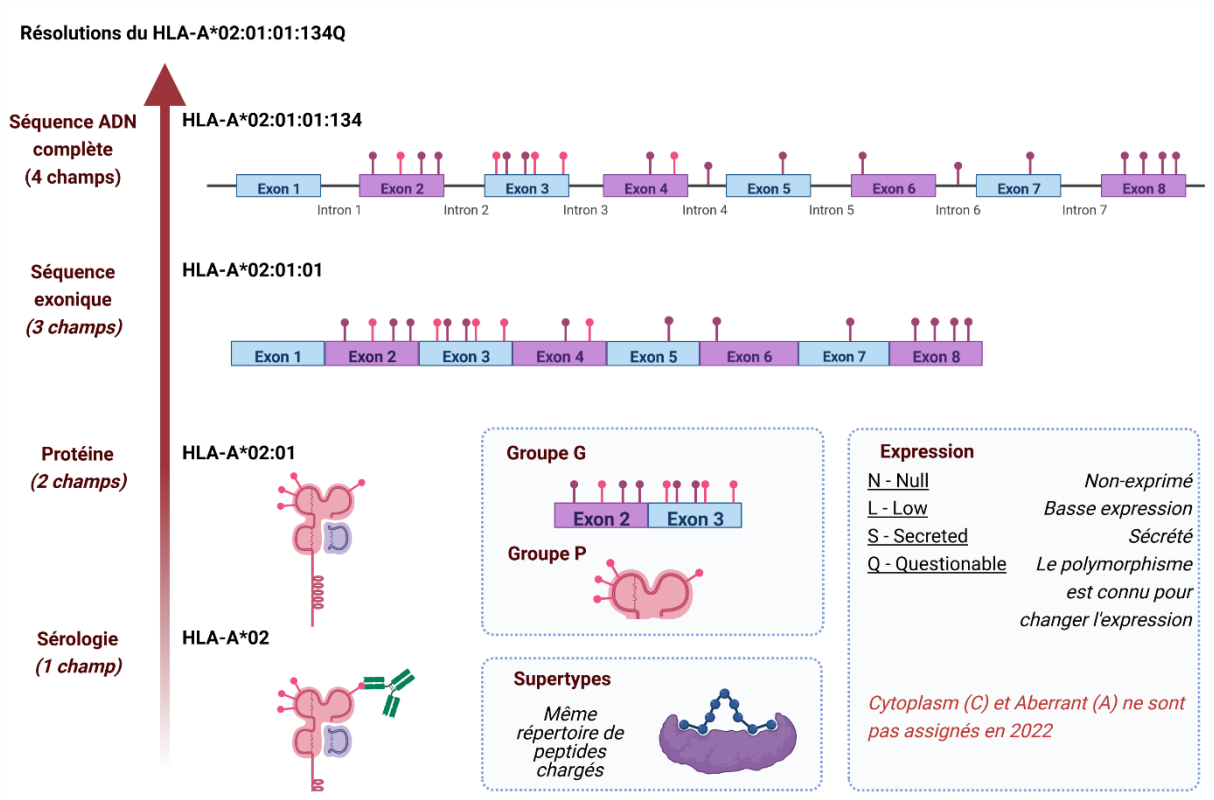


*Figure II-10 La nomenclature HLA avec l'exemple de HLA-A*02:01:01:134Q. Les résolutions correspondent à la précision de l'information concernant l'allèle HLA : de la reconnaissance par anticorps à une séquence ADN complète. L'analyse des exons 2 et 3 (ou 2 seulement pour les HLA de classe II) a longtemps été un standard, bien qu'elle ne soit pas précise car la plupart des polymorphismes HLA sont situés à ces loci. Traduit de Douillard et al.*

Le système de champs divise l'identifiant de l'allèle HLA regardé en quatre (Figure II-10, traduite de Douillard *et al.* (143)). Le premier champ conserve le système historique de sérologie, ce sont des nombres arbitraires correspondant à des reconnaissances par des anticorps spécifiques. Le second champ permet de nommer directement une molécule HLA spécifique. Cette résolution est très utilisée, il faut cependant noter que cet allèle représente plusieurs séquences HLA. Par exemple, l'allèle HLA-A*02:01 fait référence à plus de 445 versions du gène *HLA-A*. Ces versions peuvent être à nouveaux spécifiées grâce au troisième champ qui va identifier pour une protéine donnée une séquence nucléotidique unique, qui peut différer d'autres séquences par des nucléotides n'induisant pas de changement d'acide aminé. Enfin, la résolution maximale à 4 champs du typage du HLA correspond à une définition stricte d'un allèle et donne une séquence unique HLA, comprenant exons (*i.e.* la partie d'un gène qui code pour la protéine) et introns (*i.e.* la partie d'un gène qui complète la séquence du gène, souvent avec un rôle sur l'expression).

Afin de ranger les allèles selon certaines caractéristiques biophysiques, il existe également des groupes en parallèle des noms officiels. Ceux-ci ne sont pas toujours stables et ajoutent une complexité supplémentaire pour les néophytes mais peuvent trouver un intérêt dans des cas concrets.

II.3.4.2 - La terminologie du HLA en dehors de la nomenclature actuelle

Les polymorphismes des gènes HLA sont trouvés le long des gènes mais la plupart se concentrent dans le sillon peptidique, ce qui explique la diversité de peptides présentés par le HLA. Ainsi, les contraintes techniques originelles du typage HLA ont poussé à se concentrer sur ce sillon : seuls les polymorphismes des exons 2 et 3 pour les HLA de classe I, et ceux de l'exon 2 pour les HLA de classe II étaient ainsi retrouvés. Les technologies de typage ne permettaient de résoudre les ambiguïtés de séquence à d'autres niveaux, cela a donc permis de décrire des groupes d'allèles facilement. Les groupes P et G décrivent respectivement les protéines et les séquences HLA uniques au niveau des exons polymorphes.

Sidney *et al.* (144,145) ont popularisé une autre classification, celle des supertypes. Un supertype regroupe plusieurs allèles HLA avec un peptidome similaire, c'est-à-dire qu'ils présentent de nombreux peptides en commun. Une définition de Wang et Claesson (146) indique que les supertypes décrivent « des similarités structurelles, des motifs communs de fixation de peptides, et une identification de peptides permettant des réactions croisées » (146).

Le HLA a un rôle prépondérant dans la transplantation. Choisir un donneur et un receveur avec les mêmes allèles HLA est ainsi important dans la survie du greffon et du patient, notamment dans la greffe de cellules souches hématopoïétiques (147–150). Pour aller plus loin dans la correspondance, il

est possible de décrire les similarités entre les allèles HLA avec plus de précision, en utilisant les épitopes et leurs éplets (Figure II-11, traduite de Geffard *et al.* (151)).

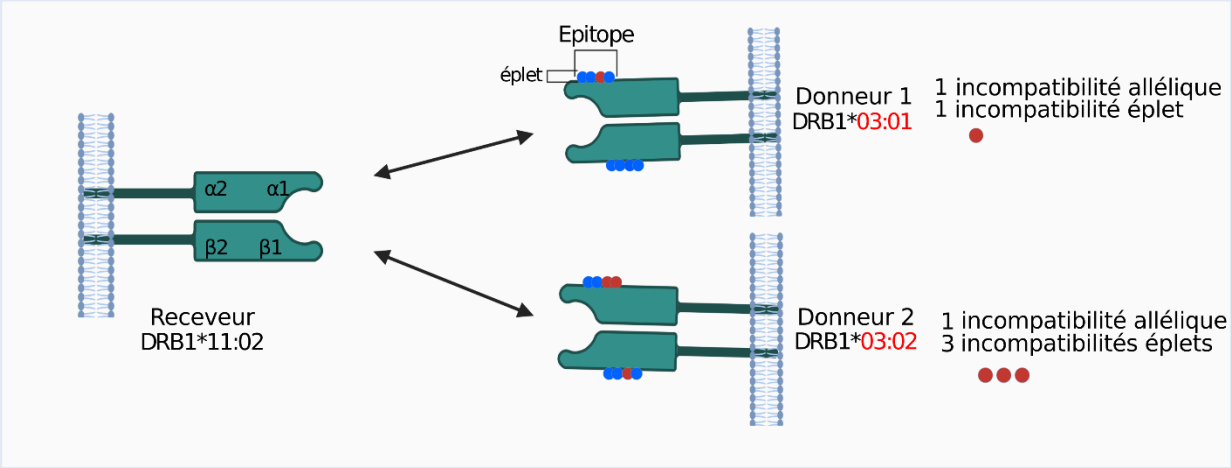


Figure II-11 Différence entre un allèle HLA, un épitope et un éplet dans le cadre d'une transplantation. Deux allèles HLA de donneurs différents peuvent être incompatibles avec un receveur mais les modifications à l'échelle locale des acides aminés de la molécule peuvent être moins nombreuses chez l'un que chez l'autre. Un épitope est un ensemble de ces modifications d'acides aminés. Traduite de Geffard et al.

Un épitope structurel est la région d'une molécule HLA reconnue par un anticorps lors d'une réaction immunitaire (152). Celui-ci contient des épitopes fonctionnels, ou éplets, qui sont des ensembles d'un à cinq acides aminés (153), consécutifs au niveau de la séquence ou proches en trois dimensions (154). Le recours à ces sous-divisions pour les allèles HLA permet ainsi de comprendre comment certains acides aminés, plutôt que des allèles entiers, jouent un rôle dans la transplantation (155).

Dans une logique semblable aux supertypes et aux éplets, Nielsen *et al.* (156,157) ont décrit les pseudo-séquences HLA comme un ensemble d'acides aminés en contact avec les peptides présentés par une molécule de HLA (ex. à une distance inférieure ou égale à 4 ångström (Å, 10^{-10} mètres). Ces pseudo-séquences n'interviennent pas dans la reconnaissance de la molécule du HLA par le TCR mais dans la capacité d'une molécule HLA à interagir avec des peptides.

Enfin, le fort déséquilibre de liaison dans la région du CMH oriente également le choix de la terminologie. Chaque allèle HLA est un haplotype de SNP, des polymorphismes tous présents sur le même chromosome. Mais le terme d'haplotype est aussi utilisé de manière plus globale. Ainsi, les allèles entiers décrivant des versions de gènes HLA différents peuvent être corrélés. Le résultat direct est que des allèles sont transmis ensemble et qu'ils sont retrouvés en population : on parle alors d'haplotypes de gènes HLA.

Depuis la découverte des gènes HLA en 1958, les immunogénéticiens ont travaillé à la caractérisation du CMH, s'adaptant au fur et à mesure aux nouvelles connaissances et aux technologies disponibles,

pour révéler la complexité intrinsèque aux milliers d'allèles HLA contenus dans cette région. Cette diversité exceptionnelle a un impact actuellement, que ce soit pour la transplantation ou pour la susceptibilité à des maladies, et son origine évolutive est une question essentielle dans la compréhension de son rôle dans le génome, toujours en débat.

II.4 - La diversité populationnelle des gènes HLA et leurs origines évolutives

II.4.1 - Les bases de données de la recherche HLA

Tout au long des progrès technologiques et des changements de paradigme autour du système HLA, les chercheurs ont également construit plusieurs bases de données afin de répertorier et disséminer les connaissances autour de sa génétique. En 1998, l'Immunogenetics Database (IMGT Database) offre la possibilité de consulter les séquences des gènes HLA, complètes ou non, et propose des séquences pour de nouveaux allèles avec son volet IMGT/HLA (158) ; elle est encore à ce jour la ressource principale qui sert de référence (159). La base de données, qui contenait déjà des données sur le polymorphisme d'autres acteurs de l'immunité comme les TCR ou les CMH d'autres vertébrés, fusionne avec l'Immune Polymorphism Database (IPD) en 2003 et acquiert entre autres des informations sur les KIR (*Killer Immunoglobulin-like Receptors*) et les antigènes de plaquettes (160,161).

Pour suivre la distribution de l'ensemble des allèles HLA découverts, mais également pour les KIR ou les TAP, Middleton *et al.* ont créé en 2003 l'AlleleFrequencies.Net Database (AFND) (162,163). Celle-ci dispose d'une dizaine de millions de fréquences alléliques pour le HLA réparties sur plus de 1 600 populations géographiques (163).

Les immunogénéticiens ont ainsi stabilisé le paysage HLA grâce à ces outils d'identification et d'étude du HLA. Après des années à dévoiler la complexité génomique du HLA, ces ressources ont permis d'évaluer la qualité, la diversité et la répartition dans les populations de ces allèles, leur apportant un regard évolutif.

II.4.2 - La diversité au-delà de la multitude des allèles

II.4.2.1 - Le sillon peptidique est le foyer des polymorphismes HLA

La complexité inhérente au système HLA ne tient pas seulement au nombre d'allèles existants mais aussi aux différentes positions des polymorphismes possibles sur l'ensemble de la structure (Figure II-12, Robinson *et al.* (159)). En effet, les molécules HLA interagissent avec les peptides qu'ils présentent et avec le TCR des lymphocytes T CD4 et CD8 (164,165) et ce par le biais des domaines α_1/β_1, et α_1/α_2 codés par l'exon 2 (pour les HLA de classe II), et les exons 2 et 3 (pour les HLA de classe I) (166). De manière analogue, les molécules KIR présentes sur les cellules NK interagissent par les

domaines α_1/α_2 (majoritairement α_1) sur les HLA de classe I (167–169). Ainsi, Parham *et al.* ont noté, dès 1995, que les exons 2 et 3 cumulent le plus de polymorphismes par rapport aux séquences HLA entières (170).

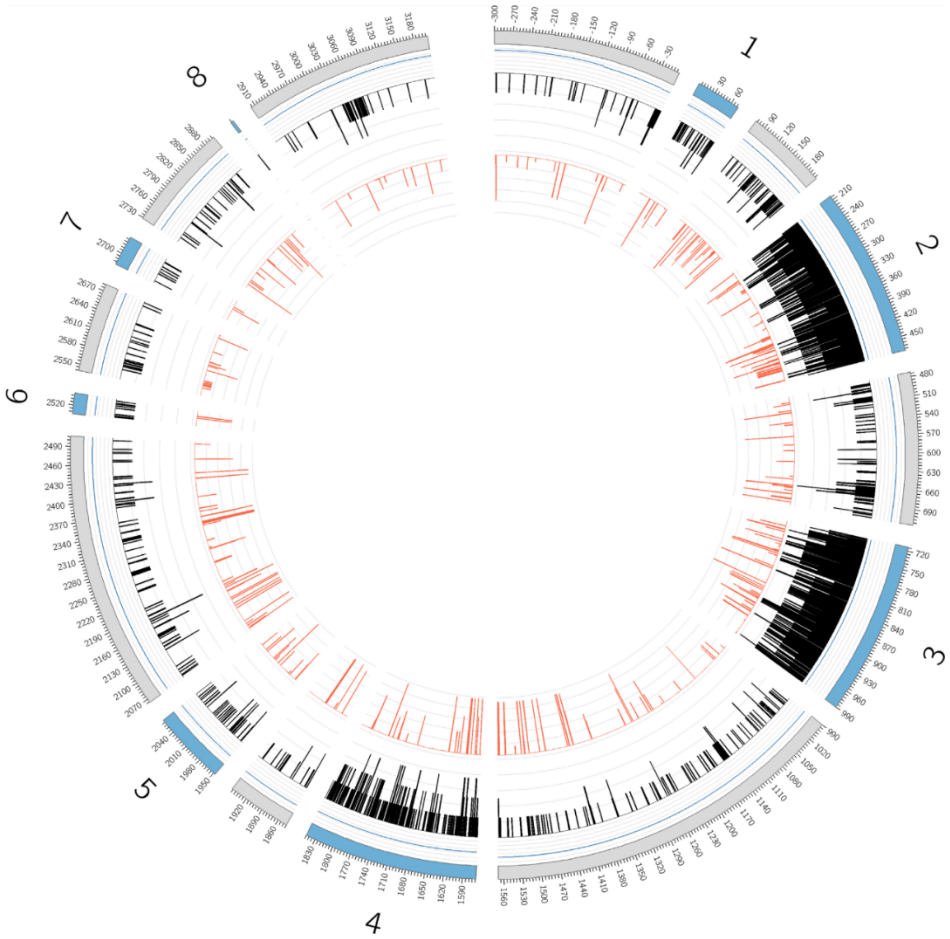


Figure II-12 La distribution des polymorphismes du HLA-A. Quatre cercles concentriques représentent, de l'extérieur vers l'intérieur : la structure, la couverture par IMGT-HLA, le nombre de polymorphismes différents et la fréquence des polymorphismes, à chaque position pour tous les allèles HLA. Chaque exon (bleu) est numéroté, et les introns (gris) sont également représentés ; la couverture de séquence de tous les allèles HLA est de 100% au niveau des exons 2 et 3, obligatoires pour être soumis à IMGT-HLA ; jusqu'à quatre nucléotides différents peuvent être observés ; la fréquence de la seconde base nucléotidique la plus fréquente est représentée en rouge (entre 0 et 50%). Issu de Robinson et al.

Ces polymorphismes sont parfois partagés entre les différents loci, comme HLA-A et HLA-B dont certains allèles sont considérés comme un seul antigène en sérologie (171). Au contraire, les exons codant pour le domaine transmembranaire et la partie cytoplasmique α_3 sont conservés entre les allèles d'un même locus mais diffèrent entre les loci (172).

Cependant, d'autres interactions se font par le biais du domaine α_3, notamment le corécepteur CD8 (64) et la molécule LILRB des cellules NK (169). En dehors des domaines α, la Figure II-12 met en évidence les polymorphismes du segment transmembranaire des molécules HLA (173,174) mais aussi de nombreux polymorphismes dans les régions non-codantes. Ces derniers sont étudiés extensivement pour le gène *HLA-G* dans lequel la région 3' UTR semble avoir un impact général sur l'expression de la molécule HLA-G (175), et particulièrement dans des cas de malaria (176) ou d'infection au BK polyomavirus (177,178). Ces polymorphismes sont également étudiés dans les gènes HLA classiques (179) : les régions 5' et 3' de *HLA-A* semblent ainsi avoir un fort impact sur l'expression des allèles (180,181). En s'éloignant un peu des gènes, le rôle des polymorphismes dans les régions promotrices qui les contrôlent a été mis en évidence, notamment pour *HLA-C* (182,183).

II.4.2.2 - Les allèles HLA ont une répartition inégale en population

Lorsque l'on regarde l'ensemble de la séquence des gènes HLA, la plupart des variants se concentrent sur les exons 2 et 3, et c'est l'ensemble de ces variants qui définit les milliers d'allèles découverts jusqu'alors. Cependant, même en se concentrant sur ces deux seuls exons, le nombre théorique d'allèles HLA de classe I différents est à des centaines d'ordres de grandeur du nombre d'individus dans la population mondiale : il atteint $4{,}3 \times 10^{381}$ d'après Robinson *et al.* (171). D'après les fréquences observées en population et celles estimées de l'apparition de nouveaux variants, il existerait actuellement environ 3 millions d'allèles différents pour chacun des gènes HLA-A, HLA-B, et HLA-C (171,184).

Tableau II-2 Allèles HLA-A du registre français des donneurs de moelle osseuse répertoriés dans allelfrequencies.net, rangés par ordre décroissant et cumulées (n=42 623).

Allèle	**02:01**	**01:01**	**03:01**	**24:02**	**29:02**	**11:01**	**32:01**	**26:01**	**23:01**	**31:01**
Fréquence	0,2901	0,1301	0,1226	0,0935	0,0566	0,0500	0,0339	0,0298	0,0270	0,0267
Cumul	0,2901	0,4202	0,5428	0,6363	0,6929	0,7429	0,7768	0,8066	0,8336	0,8603

La différence observée entre l'estimation et le nombre d'allèles actuellement connus s'explique par les fréquences déséquilibrées de certains allèles. En effet, malgré les millions d'allèles possibles pour HLA-A par exemple, l'AFND répertorie dans le jeu de données du registre français des donneurs de moelle osseuse seulement 23 allèles différents. En les rangeant par fréquence, on observe ainsi qu'un cinquième de tous ces allèles représente environ 2/3 de tous les allèles portés par ces 42 623 individus (Tableau II-2, allelefrequencies.net).

Malgré la diversité apparente des allèles HLA, les fréquences indiquent que beaucoup d'allèles sont rares et donc portés par seulement quelques individus. L'étude d'un jeu de données avec plusieurs

populations d'origines différentes permet de confirmer cette observation. Ainsi, en regardant le projet 1,000 Genomes qui contient 2 504 individus issus de 26 populations différentes (Figure II-13), 17 allèles HLA-A forment 80% de tous les allèles HLA-A typés et seulement 10 pour HLA-DQB1.

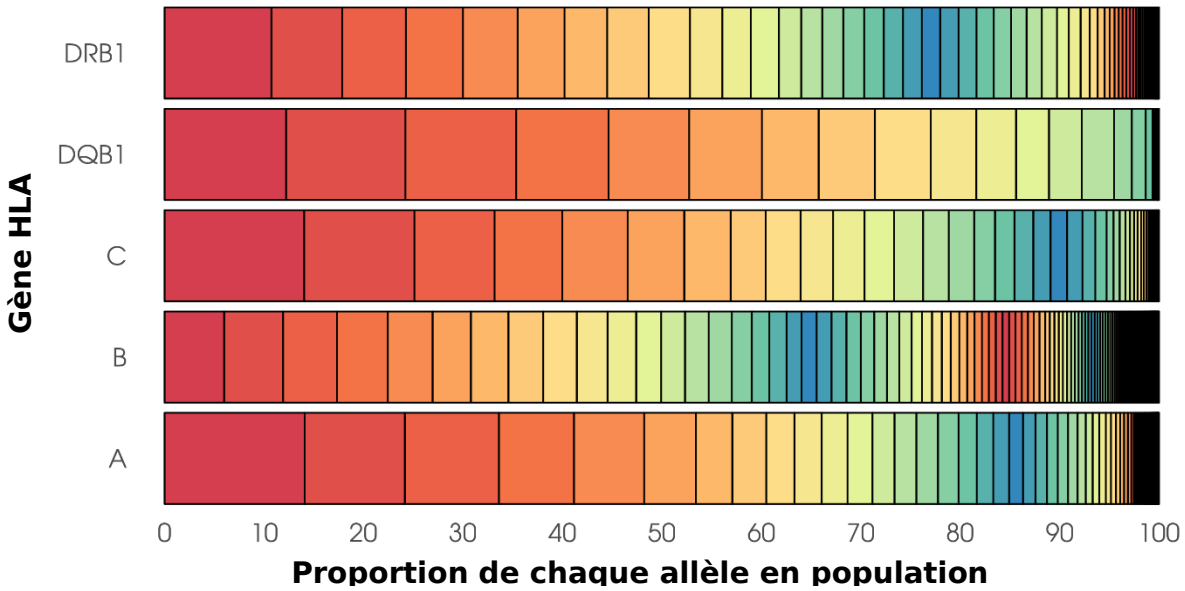


*Figure II-13 Fréquences cumulées des allèles HLA du 1KGP. Ce graphe met en avant la distribution particulière de la fréquence. Un faible nombre d'allèles HLA constituent la plus grande partie de la population (*ex. *10 allèles englobent 80% de la population pour HLA-DQB1), alors que des centaines d'allèles HLA existent à de faibles fréquences (inférieures à 1%).*

Ainsi, les allèles HLA se divisent en des allèles très rares, voire unique, et des allèles extrêmement communs. Les allèles rares et très rares posent problème pour le typage car il est difficile de détecter les erreurs de séquençage sur un allèle observé une seule fois. À partir du 15ème IHIW, des catégories « Très rare » (1 seule observation), « Rare » (entre 2 et 4 observations), et « Commun » (plus de 4 observations), ont été définies (185). En 2009, plus de 40% des allèles n'avaient été observés qu'une seule fois, et un tiers de ceux-ci étaient pourtant présents dans d'autres bases de données (186). La profondeur de séquençage est assez importante pour s'assurer de la véracité de ces allèles, mais les erreurs sont possibles. Par ailleurs, pour les allèles communs, le Common and Well-Documented allele catalog (CWD) a adapté ses catégories de fréquence avec la croissance du nombre d'allèles identifiés, amenant à une fréquence minimale de 1/10 000 pour les allèles communs, 1/100 000 pour les allèles intermédiaires, et 5 occurrences pour les allèles bien documentés (187).

Ces nombreux efforts de caractérisation se heurtent à la complexité d'obtenir des informations de fréquence pour tous les allèles. Ainsi, en mai 2020, seulement 20% des allèles HLA-A avaient une fréquence connue (188).

La fréquence des haplotypes HLA est une donnée tout aussi importante que celle des allèles uniques, qui sert notamment dans les greffes de cellules souches hématopoïétiques (189,190). Les haplotypes

sont des ensembles de loci HLA où plusieurs allèles sont en déséquilibre de liaison. Leur diversité est par définition bien supérieure aux allèles simples. Ainsi Maiers *et al.* ont montré que 54% des haplotypes des Américains d'origine européenne faisaient partie des 100 haplotypes les plus fréquents, et ce chiffre atteint 30% chez les Américains d'origine africaine (191). Un haplotype est ainsi considéré fréquent pour une fréquence supérieure à 1% (192).

Ces informations de fréquence expliquent la découverte continue de nouveaux allèles HLA et mettent en lumière la diversité HLA à l'intérieur d'une population, ainsi qu'entre des populations d'ancestralité différente.

II.4.2.3 - Des fréquences qui varient différemment selon les ancestralités

Comme rappelé par Robinson *et al.* (171), le polymorphisme des HLA est tel qu'il n'existe pas d'allèle de référence, ce qui complique la compréhension de la génétique de cette région. En effet, de nombreuses populations doivent être étudiées pour découvrir de nouveaux allèles et comprendre leur origine. Ces études de populations sont limitées en taille par leur coût et les allèles étudiés sont souvent les allèles déjà connus.

Cependant, l'augmentation des typages HLA a permis d'observer des changements de fréquence des allèles HLA entre les populations. Si cela s'explique facilement pour les allèles rares relevés seulement quelques fois, l'AFND (188) a montré récemment que cela s'applique également pour des variants communs. En effet, il existe une grande variabilité de fréquence des allèles à l'intérieur des populations et entre les différentes populations. Des différences d'hétérozygotie aux loci HLA démontrent aussi une diversité variable selon les régions du globe.

De plus, en prenant les allèles HLA-A les plus communs de deux régions du globe, deuxième gène le plus polymorphe, pour le Nord-Est et le Sud de l'Asie (Figure II-14, http://allelefrequencies.net/top10freqsb.asp), les données de l'AFND mettent en évidence une disparité. L'allèle le plus fréquent au Nord-Est (HLA-A*24:02) a une fréquence moyenne deux fois plus élevée qu'au sud de l'Asie, et les deux populations ont seulement 4 allèles sur 10 en commun.

À la multitude d'allèles HLA vient donc s'ajouter une diversité d'emplacement, de fréquence et de répartition des polymorphismes dans les populations. La distribution des polymorphismes sur l'ensemble de la molécule impacte la liaison avec de nombreux autres acteurs de l'immunité, quant à la divergence des fréquences entre certains allèles et entre les populations, elle implique des réponses différentes face aux pathogènes ou aux maladies en général, qui peuvent trouver une origine dans l'histoire évolutive des gènes HLA.

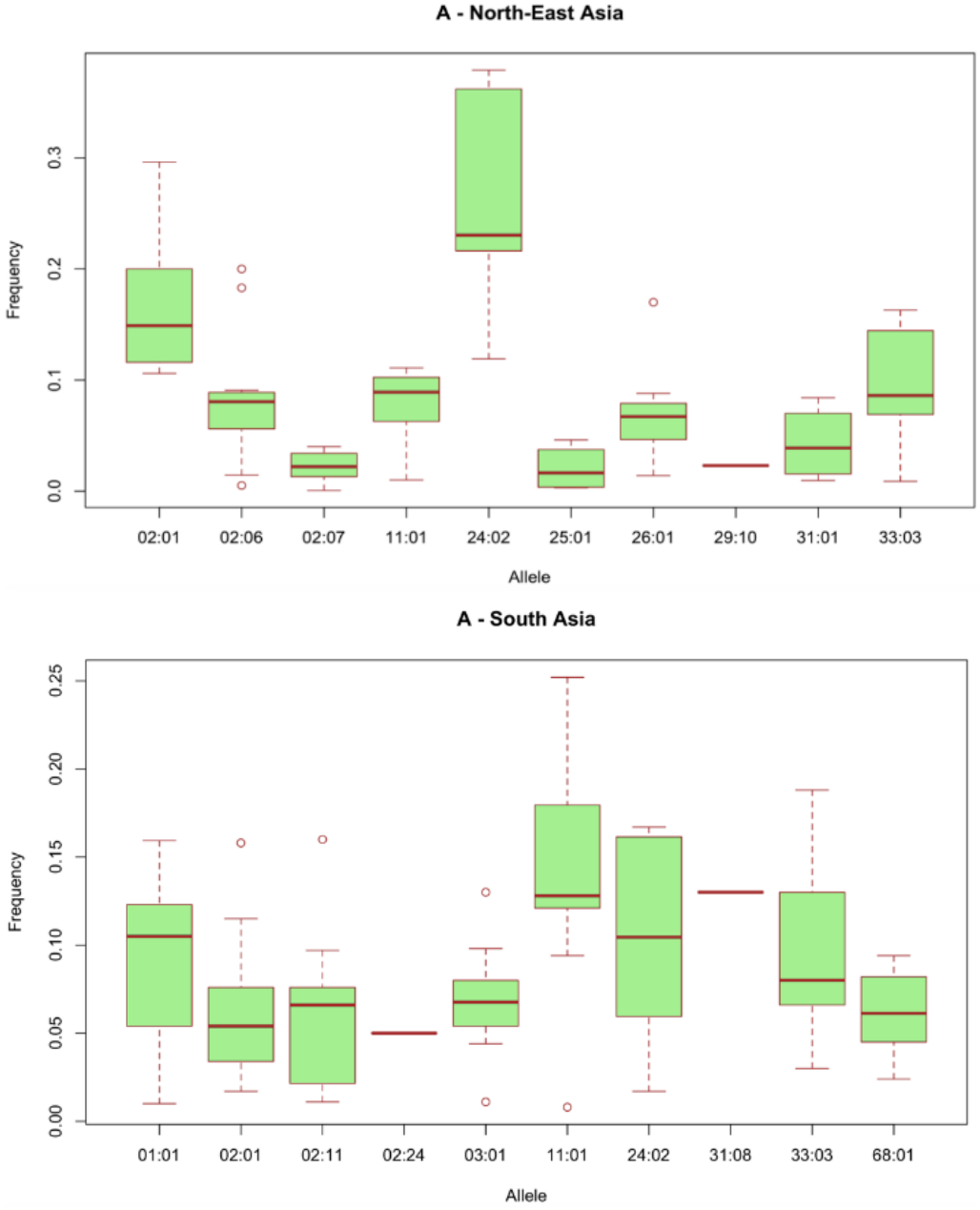


Figure II-14 Distribution des fréquences des 10 allèles les plus fréquents dans deux populations asiatiques, au nord-est et au sud. Des allèles sont communs entre les deux populations (ex. *HLA-A*02:01, HLA-A*11:01, HLA-A*24:02, HLA-A*33:03) mais peuvent fortement différer au niveau des fréquences. Au contraire, certains allèles sont totalement absents d'une population à l'autre. Issu de allelfrequencies.net.*

II.4.3 - L'origine évolutive du CMH et de la diversité des gènes HLA

II.4.3.1 - Une région ancestrale partagée par le règne animal

La région du CMH fait partie d'un ensemble de gènes de l'immunité adaptative, auquel les TCR et les immunoglobulines appartiennent. Dans l'évolution, ces gènes ont été mis en évidence par la génomique comparative dès l'apparition des poissons cartilagineux. Bien qu'il existe des régions paralogues dans l'embranchement des Cordés, ce sont seulement les animaux avec une mâchoire qui disposent de ces gènes (193). L'apparition des gènes de l'immunité adaptative daterait donc d'il y a 520 millions d'années, et elle serait arrivée sur une courte échelle de temps (194) par d'importantes duplications génomiques (195).

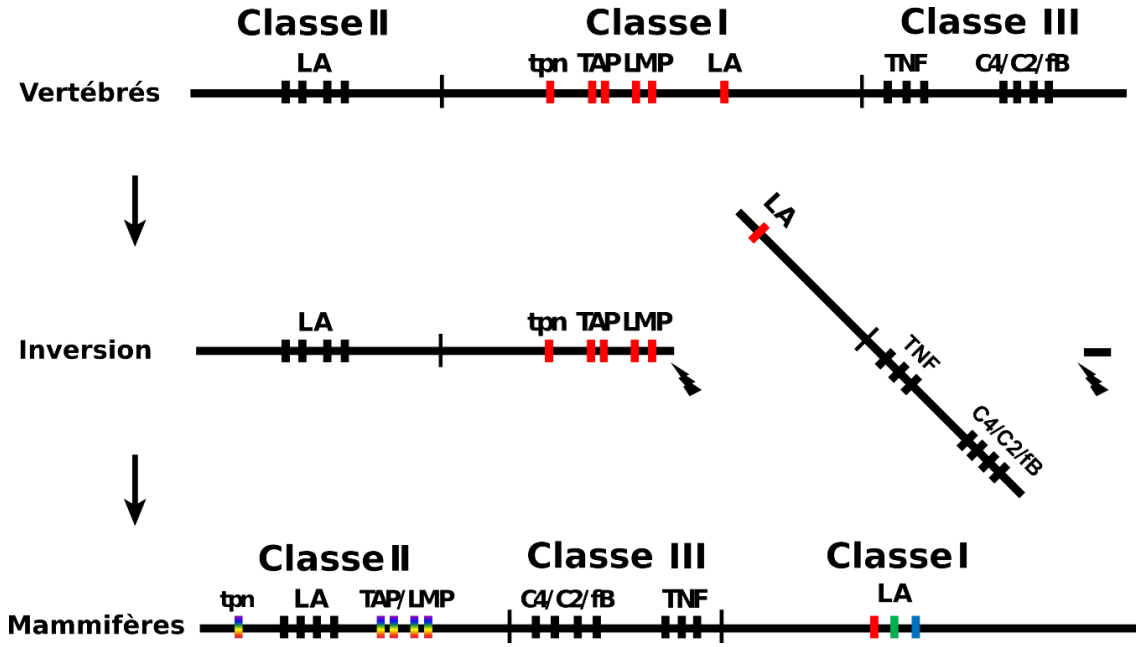


Figure II-15 L'hypothèse de l'inversion dans la région du CMH depuis son existence dans le groupe des vertébrés jusqu'à sa forme chez les mammifères. La proximité du système de sélection et de présentation des peptides favorise un système avec un allèle majoritaire pour la présentation et le polymorphisme sur la sélection du peptide. Une inversion pourrait expliquer l'éloignement des deux systèmes menant à une sélection de peptides adaptées pour tous les allèles HLA, et une présentation soutenue par de nombreux gènes et allèles avec leur propre peptidome. Traduite de Kaufman et al.

Kaufman *et al.* ont ainsi décrit le CMH du poulet en 1995, qui ne dispose que de 19 gènes ; ce CMH minimal permet ainsi de mieux comprendre l'évolution de l'organisation de ces gènes (196). Ainsi, le gène codant pour la protéine présentant les peptides est peu polymorphique mais plus proche, par rapport à l'humain, des protéines TAP qui présentent un plus fort polymorphisme. Ainsi, la région du CMH ancestrale aurait pu être inversée dans l'évolution (Figure II-15, traduite de Kaufman *et al.* (197)), isolant les gènes classes I des TAP polymorphiques et poussant sélectivement vers l'apparition de gènes dupliqués couvrant un plus grand répertoire de peptides (197).

II.4.3.2 - Une diversité d'hypothèses pour comprendre le polymorphisme HLA

Les informations de l'évolution génomique permettent de comprendre l'organisation humaine de la région du CMH, mais pas le nombre important d'allèles. L'apparition des allèles HLA et leur conservation sur de longues périodes peut s'expliquer par une pression de sélection équilibrée (*balancing selection*). La sélection naturelle peut favoriser l'existence d'une variation qui avantagerait une population dans son environnement, ou à l'inverse baisser la diversité pour les variations délétères. Le CMH est lui sous une pression de sélection équilibrée dont la raison principale semble être la réponse aux pathogènes de l'environnement, ce qui favorise la présence de plusieurs allèles d'un même gène dans la population (198). Cette pression de sélection équilibrée existe par plusieurs mécanismes : l'avantage hétérozygote, la sélection négative par fréquence et la sélection par variation spatio-temporelle (199). Les gènes HLA sont codominants, ainsi chaque allèle est transcrit puis traduit en protéine. L'hétérozygotie à un locus donné peut augmenter la diversité en population dans le cas où posséder deux versions différentes d'un même gène confère un avantage de survie ou de reproduction. Dans le cas des HLA, l'efficacité de la présentation de peptides dépend des polymorphismes du sillon peptidique : avoir différents allèles permettrait de couvrir une plus grande variété de pathogènes (200).

La sélection négative par fréquence considère que ce sont les pathogènes qui subissent une forte sélection pour contourner le système immunitaire des allèles HLA les plus communs. Cela donnerait un avantage aux allèles rares. Ceux-ci seraient tour à tour favorisés ou non, selon les adaptations du pathogène, leur permettant de rester en population (201).

La sélection par variation spatio-temporelle, quant à elle, statue que des événements indépendants aux pathogènes façonnent leur fonctionnement différemment selon le temps et la zone géographique, ce qui dirige une sélection de certains allèles HLA au fil du temps. L'influence des pathogènes sur la région du CMH serait donc unilatérale (202).

Enfin, Spurgin *et al.* indiquent que toutes ces hypothèses peuvent exister ensemble et interagir dans la population (199). De la même manière, d'autres solutions peuvent être apportées pour comprendre la diversité du HLA. Kaufman introduit ainsi dans son modèle le répertoire peptidique des allèles HLA et leur expression à la surface pour expliquer deux types de diversité (197) : les allèles qui confèrent une protection contre de nombreux pathogènes courant mais peu exprimés (*i.e.* les généralistes), et ceux qui confèrent une protection contre un pathogène en particulier et exprimés plus fortement (*i.e*. les spécialistes). Les pathogènes bactériens ou viraux ont exercé une pression de sélection et ainsi créé la diversité des gènes et des allèles du HLA. Cette évolution a encore maintenant un impact fort sur les

pathologies humaines, infectieuses, mais également auto-immunes et parfois sur des traits non-immuns, qui peuvent être étudié par association génétique.

III - Les associations génétiques avec les pathologies : le revers de la médaille du HLA

La présentation d'antigènes est une fonction qui existe depuis des millions d'années et a subi de nombreuses transformations. Actuellement, elle permet aux populations humaines de se protéger contre une grande diversité de pathogènes. Cependant, la diversité allélique populationnelle des gènes HLA peut avoir des effets secondaires individuels sur d'autres processus biologiques. La transplantation d'organes entre des individus portant des allèles différents conduit ainsi à un rejet du greffon. Les antigènes qui peuvent être accomodés dans le sillon peptidique du HLA sont différents selon les polymorphismes de la molécule. Ainsi, sans même d'intervention humaine, une meilleure présentation de certains antigènes viraux par exemple peut améliorer la protection de certains individus. A l'inverse, cela peut mener à une présentation d'antigènes du soi de façon immunogène et une moins forte présentation de ces antigènes peut augmenter le risque contre certaines infections. Les études d'association génétique peuvent estimer le risque ou la protection apportés par un variant génétique, et *a fortiori* par un allèle HLA, dans une pathologie particulière.

III.1 - L'association génétique dans la région du CMH

Une représentation classique de l'impact de la génétique sur le phénotype d'un individu est celle dite mendélienne. Une mutation (*i.e.* une variation généralement rare) dans la séquence codante d'un gène vient perturber la fonction principale de la protéine traduite, en changeant un acide aminé essentiel ou en la tronquant : une pathologie peut alors en découler ; c'est le cas de la mucoviscidose par exemple (203).

Pourtant les variants rares ne sont pas les seuls à avoir un impact sur les phénotypes humains observés. La plupart des traits complexes sont ainsi significativement associés à des variants communs, dont la fréquence allélique est supérieure à 0,5-1% (204). L'association en génétique peut aussi avoir un effet faible ou modéré, protecteur ou délétère, conféré par un variant commun à propos d'un trait étudié (Figure III-1, inspirée de Tam *et al.* (205) et créée avec biorender.com). L'ensemble des SNP individuels définit un fardeau génétique. Ce terme comprend l'ensemble des changements courants sur l'ADN qui sont associés statistiquement à une pathologie et permettent aussi d'en élucider les mécanismes physiopathologiques (206). Dans les études d'associations pangénomiques comme les GWAS (*Genome-Wide Association Study*), les SNP sont des polymorphismes bi-alléliques communs, présents à 1% dans la population étudiée. Un variant est associé à un trait catégoriel lorsque sa fréquence est

statistiquement différente entre deux populations (ex. diabète, sclérose en plaques), ou avec un trait continu quand sa fréquence varie linéairement avec son intensité. Il est possible de parler de trait ou de phénotype pour désigner n'importe quel paramètre biologique ou clinique. Les associations

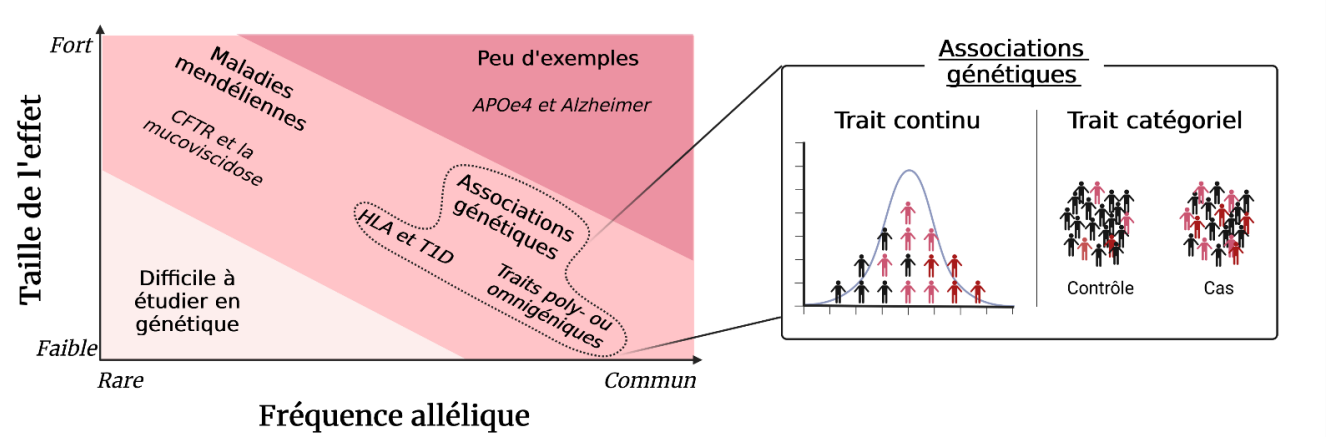


Figure III-1 Les variants génétiques peuvent avoir une fréquence différente en population et leur effet change selon leur nature et leur position. Les mutations sont rares et celles identifiées modifient généralement une protéine, jusqu'à la rendre non-fonctionnelle, les mutations sans effet sont rarement recherchées et leur fréquence réduit la possibilité de les trouver par hasard. Les polymorphismes génétiques sont couramment définis par une fréquence supérieure à 1%, ils sont retrouvés dans tout le génome et tous les traits, cependant leur effet est limité. Les polymorphismes à effet fort sont souvent liés à une pression sélective équilibrée. Les individus peuvent avoir aucuns polymorphismes ou mutations (en noir) par rapport à une partie de la population qui peut en avoir plus ou moins (rose à rouge, selon le nombre). Inspirée de Tam et al. et créé avec biorender.com.

polygéniques représentent souvent des changements de fréquence subtils et des tailles d'effet, donc un impact sur le phénotype par la génétique, faibles. Il est alors essentiel d'étudier des centaines voire plusieurs milliers d'individus pour s'assurer de la véracité du lien statistique.

Le GWAS Catalog est une base de données (https://www.ebi.ac.uk/gwas/) qui recense des résultats d'associations génétiques pangénomiques validés (207,208). La base de données est remplie de *summary statistics*, des résumés d'études d'associations qui décrivent pour chaque publication, le ou les traits étudiés, les SNP associés et leur position génomique, entre autres. Ces résumés peuvent être utilisés librement par d'autres chercheurs dans le cas des méta-analyses, pour combiner les études et augmenter leur puissance statistique.

Le GWAS Catalog met en lumière des milliers de corrélations entre des traits et des variants. La région du CMH est associée à de nombreuse pathologies (209), parfois changeant la prise en charge clinique notamment dans l'hypersensibilité aux médicaments, où des tests génétiques HLA sont effectués pour éviter des effets secondaires délétères comme le syndrome de Stevens Johnson (210,211). Parmi tous ces signaux, la région étendue du CMH a une importance disproportionnée par rapport à sa taille (Figure III-2, issue de Douillard *et al.* (212)).

En effet, 2,5% de toutes les associations significatives du GWAS Catalog se retrouvent dans cette région de 5Mb, soit un tiers des associations dans le chromosome 6. De plus, près de 20% des traits recensés disposent d'au moins une association dans la région du CMHx. Les associations retrouvées sont pour la plupart reliées à l'immunité, que ce soit des pathologies infectieuses (213) ou auto-immunes (214).

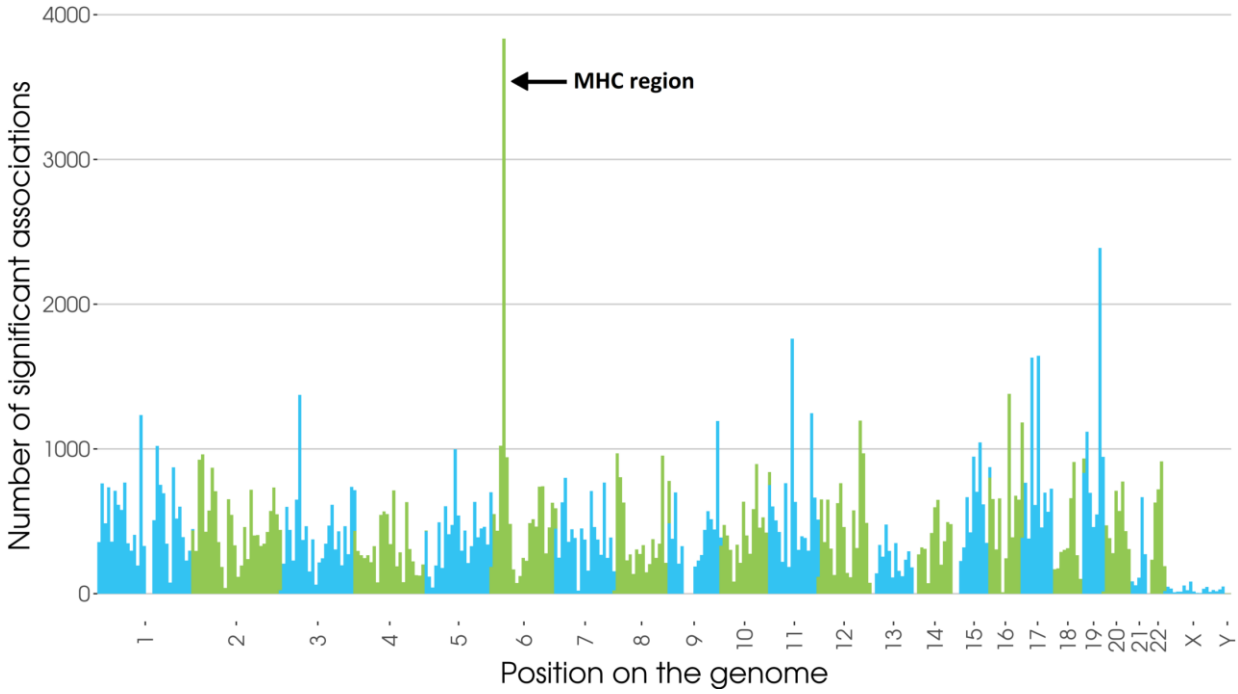


Figure III-2 Le CMH est la région du génome avec le plus d'associations génétiques statistiquement significatives (à p >= 5e10^{-8}). Elle correspond à 2% de toutes les associations répertoriées. Issue de Douillard et al.

III.2 - Une protection efficace des populations contre les infections, mais faillible à l'échelle individuelle

La diversité des allèles HLA dans leur fonction de présentation d'antigènes explique une partie de ces associations. Puisque la diversité se situe dans le sillon peptidique des molécules HLA, les variants observés dans la région du CMH changent le répertoire de peptides présentés. Ils sont logiquement associés à une protection ou un risque accru d'infection par des pathogènes.

III.2.1 - Les associations majeures du HLA avec des pathologies infectieuses

L'association découverte entre le HLA et le virus de l'immunodéficience humaine (Human Immunodeficiency Virus HIV-1) est la plus connue des associations à une maladie infectieuse. Des SNP en déséquilibre de liaison presque complet avec l'allèle *HLA-B*57:01* ont été mis en évidence pour la première fois par Fellay *et al.* (215) pour leur association positive avec le contrôle de ce virus.

Les associations des allèles *HLA-B*57:01* et *HLA-B*57:03* et la protection qu'ils confèrent contre le HIV-1 ont été répliqués en 2010 par Pereyra *et al.* (Figure III-3, issue de Pereyra *et al.* (216)), chez des individus d'ancestralité européenne et des individus d'ancestralité africaine, respectivement (216,217). Le locus du HLA-C a également été retrouvé associé et son rôle protecteur semble lié à une expression accrue de HLA-C plutôt qu'à un allèle particulier (215,218,219). Mais ces résultats sont encore discutés et pourraient également être en lien avec l'expression du miRNA-148a (220).

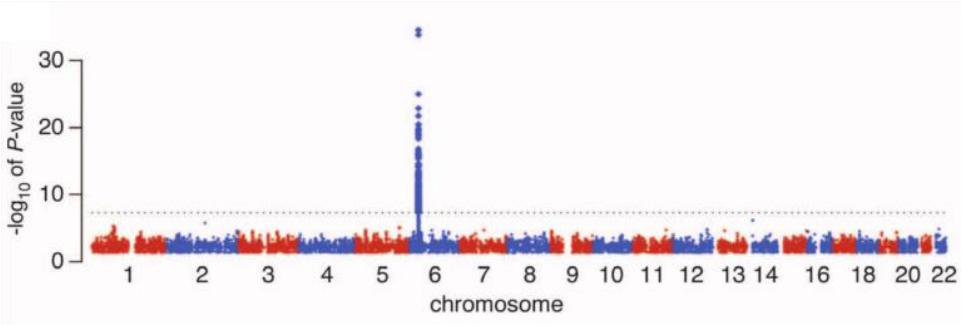


Figure III-3 Manhattan plot des résultats des tests d'association de 1,3 millions de SNP autosomaux (i.e. les chromosomes non-sexuels) avec le contrôle du HIV-1. L'abscisse représente les positions génomiques des SNP et l'ordonnée la significativité du test statistique en –log10 ; une valeur haute indique une p-value très faible. Le CMH est la seule région qui dépasse le seuil de Bonferroni de 5.10^{-8}(indiqué en pointillé), soit environ 7,3 en –log10. Issue de Pereyra et al.

À l'inverse du HIV-1, le virus de l'hépatite C a été associé au HLA de classe II. S'il n'est pas éliminé au fil du temps, l'infection par ce virus peut devenir chronique et déclencher une cirrhose ou un cancer hépatique (221). Cependant, 30% des individus infectés éliminent entièrement le virus. Duggal et *al.* ainsi que Vergara *et al*. ont pu associer cette capacité à éliminer le virus à des signaux génétiques situés près de *HLA-DQA2* et *HLA-DQB1* (222,223). Une étude d'association au niveau des allèles et des acides aminés a récemment confirmé l'association spécifique avec *HLA-DQB1*03:01*, et plus particulièrement, avec les acides aminés leucine (Leu26) et proline (Pro55), en positions 26 et 55 respectivement sur la molécule du HLA (224).

Ces deux exemples ne sont qu'une petite partie des associations retrouvées dans le HLA. D'autres virus comme le HPV (225), la dengue (226) ou l'HBV (227,228), et des infections bactériennes comme la tuberculose ou l'infection par *Mycobacterium* (229) ont une association avec le HLA, que ce soit pour l'infection ou la réponse immunitaire face à l'infection. Récemment, des études d'association ont recherché des liens génétiques entre la région du CMH et le virus SARS-CoV-2.

III.2.2 - Les associations élusives de la pandémie de SARS-CoV-2

La pandémie de SARS-CoV-2 a commencé à s'étendre vers la fin de l'année 2019 et continue d'évoluer plus de deux ans après. L'infection par SARS-CoV-2 provoque la maladie du COVID-19 ; elle est en grande partie asymptomatique. Cependant, elle peut causer un syndrome de détresse respiratoire aigu (*acute respiratory distress syndrome*, ARDS) létal, notamment chez les patients âgés ou présentant des

comorbidités (230,231). On dénombre plus de 6,5 millions de décès liés à cette maladie depuis 2020 (à la date du 22.09.2022) (232). Les immunogénéticiens ont donc cherché à comprendre les facteurs de risque génétiques. La région du CMH n'a pas dérogé à une investigation car, comme il a été montré pour diverses infections, le HLA peut avoir un impact dans le développement de la pathologie.

Le typage HLA est une tâche compliquée et coûteuse. Ainsi les premières études se sont basées sur la corrélation entre un phénotype et la distribution des allèles HLA en population. Puis d'autres études se sont concentrées sur l'affinité des allèles HLA pour les peptides du SARS-CoV-2. Nguyen *et al.* (233) ont ainsi identifié, par l'analyse de présentation du peptidome de SARS-CoV-2, le facteur de risque HLA-B*46:01 (présentation diminuée) et le facteur de protection HLA-B*15:03 (présentation accrue), tandis que Romero-López a identifié d'autres allèles HLA de classe I et classe II (234). Les résultats de corrélations ont tous trouvé des allèles HLA significatifs, mais de manière discordante entre les études (235–238).

Les premières études cas-contrôle basées sur les associations HLA ont commencé en 2020 et se poursuivent encore aujourd'hui. Elles ont souffert dans un premier temps d'un nombre d'individus limité et d'une grande hétérogénéité des phénotypes observés. Ces études s'intéressent tour à tour à des patients séropositifs pour le SARS-CoV-2, hospitalisés ou à des degrés de sévérité de la maladie différents. Dans tous les cas, les associations HLA révèlent des associations hétérogènes qui ne sont pas retrouvées systématiquement : certaines identifient l'allèle *HLA-A*11:01* (231,239) ou *HLA-DRB1*15:01* (240), entre autres.

Si de nombreux signaux prometteurs ont été découverts dans les études préliminaires avec un faible pouvoir statistique, en se concentrant sur les études les plus robustes, nous pouvons constater que la plupart des associations disparaissent (Tableau III-1, basé sur Douillard *et al.,* Deb *et al.,* Dobrijević *et al.* (212,241,242)). Une méta-analyse permet de combiner les *p-value* obtenues dans des études différentes pour évaluer globalement les associations. Seule la méta-analyse de Dobrijević *et al.* (242) semble conserver des signaux d'association, mais celle-ci se base sur des études très différentes en nombre d'individus et en méthodologie. De plus, la GWAS du Severe COVID-19 GWAS Group (243) et la méta-analyse de Degenhardt *et al.* (244) n'identifient aucun signal, et ceux précédemment identifiés n'atteignent même pas le seuil nominal de significativité et ont des effets contraires.

Le 18ème IHIW en 2022 a permis de confronter les différents/divers points de vue de chercheurs. Il existe un scepticisme sur l'intérêt de continuer les études d'association entre le HLA et le COVID-19 après le manque de résultats significatifs révélé par l'analyse des jeux de données les plus importants. Pourtant, l'absence du HLA paraît difficile à expliquer et il existe une demande d'efforts pour produire des associations d'allèles spécifiques au HLA plutôt qu'une association SNP qui reflète l'association

statistique de toute la région du CMH. Le rôle du HLA ne peut pas être affirmé dans le cas de l'infection par SARS-CoV-2 ou de la sévérité du COVID-19.

Tableau III-1 Récapitulatif des associations avec des allèles HLA. Ces études sont des études cas-contrôle avec plus de 250 individus positifs pour le COVID-19, des données HLA pour HLA-A, HLA-B, HLA-C, HLA-DQB1 et HLA-DRB1. Les études marquées d'une astérisque (*) ont testé plusieurs phénotypes, avec des effectifs différents. Basé sur les études recoupées par les revues de Douillard et al., Deb et al., et Dobrijević et al.

Étude	Type	Allèle HLA	Cas de COVID-19
Astbury *et al.* (245)	Association	DRB1*13:02, DRB1*15:01, DRB1*15:02,	272
Wang *et al.* (246)	Association	A*11:01, B*51:01, C*14:01	332
Weiner *et al.* (247)	Association	C*04:01	435
Gutiérrez-Bautista *et al.* (248)	Association	/	450
Marco *et al.* (249)	Association	/	720
Augusto *et al.* (250)	Association	B*15:01, DRB1*04:01	640-788*
Ellinghaus *et al.* (243)	Association	/	775-835*
Nguyen *et al.* (251)	Association	/	3 235
Shachar *et al.* (252)	Association	/	6 413
Dobrijević *et al.* (242)	Méta-analyse	A*01, A*03, A*11, A*23, A*31, A*68, A*68:02, B*07:02, B*14, B*15, B*40:02, B*51:01, B*53, B*54, B*54:01, C*04, C*04:01, C*06, C*07:02, DRB1*11, DRB1*15, DQB1*03, DQB1*06	10 551
Degenhardt *et al.* (244)	Méta-analyse	/	3 255 – 14 467*

III.3 - Les maladies auto-immunes : un excès de zèle immunitaire ? L'exemple de la sclérose en plaques (SEP)

Si l'inégalité des individus devant des maladies infectieuses est la conséquence évidente de la diversité allélique du HLA, les maladies auto-immunes sont, elles aussi, très bien représentées dans les associations HLA (23,253,254).

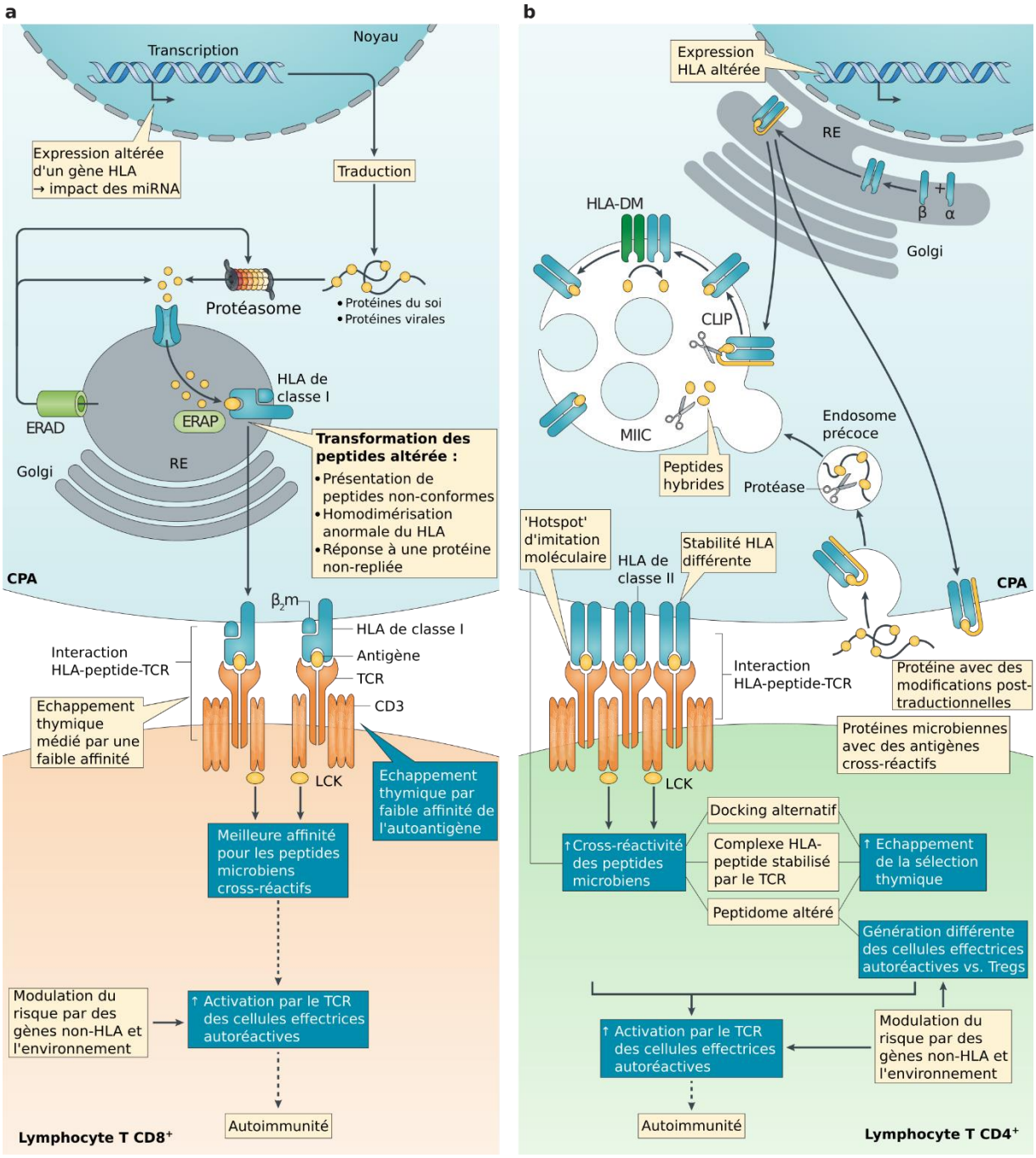


Figure III-4 Les allèles HLA peuvent provoquer des réactions auto-immunitaires par le biais de plusieurs mécanismes moléculaires. Traduite de Dendrou et al.

Parmi ces maladies, le diabète de type 1 a une association avec des haplotypes DRB1-DQA1-DQB1 (255), notamment la position 57 de HLA-DQB1, connue depuis plusieurs dizaines d'années (256) ; *HLA-*

*C*06:02* et *HLA-DQB1*06:02* ont quant à eux été associés au psoriasis (257,258) et à la narcolepsie (259,260), respectivement. Malheureusement, comme souligné par Trowsdale en 2013, peu de ces associations sont en réalité reliées à des phénomènes biologiques (254). Plusieurs hypothèses ont néanmoins fait surface et expliquent comment une molécule HLA peut, dans son interaction avec le TCR ou non, mener au développement de lymphocytes T auto-réactifs, comme le mimétisme moléculaire local (*hotspot molecular mimicry*) et la fixation alternative (*i.e*. alternative *docking*) (Figure III-4, traduite de Dendrou *et al.* (214)).

Dans le cas de la sclérose en plaques (SEP), *HLA-DRB1*15:01* est l'acteur principal de l'association (261,262). Cependant, l'association se retrouve également dans des haplotypes entiers classe I et classe II (261,263). De plus, l'infection par le virus d'Epstein-Barr (EBV) serait un facteur de risque supplémentaire (261,264). Le mécanisme derrière ces associations pourrait être le *hotspot molecular mimicry* (265). Ce terme définit la capacité de certains TCR à reconnaître seulement une partie spécifique d'un peptide, ce qui augmente la probabilité qu'un TCR auto-réactif reconnaisse par erreur un peptide du soi ressemblant à un peptide issu d'un pathogène. Ainsi, les lymphocytes T auto-réactifs pourraient être activés par un peptide de l'EBV porté par HLA-DRB5*01:01, en déséquilibre de liaison avec DRB1*15:01 (266), car il ressemble localement au peptide de la myéline (protéine ciblée dans la SEP ou auto-antigène cible de la SEP). De plus, HLA-DRB1*15:01 peut présenter un peptide dérivé de la myéline et certains TCR interagissent de biais avec le couple HLA-peptide : c'est le docking alternatif, qui entraîne l'activation des lymphocytes T auto-réactifs (267,268).

III.4 - Des exemples de la diversité des associations dans le CMH

Des traits ou pathologies qui ne sont pas reliés habituellement à l'immunité semblent s'ajouter à la longue liste des associations HLA (23). Des mécanismes similaires à ceux des maladies auto-immunes pourraient être impliqués dans des maladies neuro-dégénératives, comme la maladie de Parkinson. En effet, il existe une association positive avec *HLA-DRB1*15:01* qui semble éliciter une réponse inflammatoire des lymphocytes T par présentation des peptides de l'α-synucléine, une protéine qui s'agrège dans le cerveau des patients (269).

Les associations dans la région du CMH sont complexes et il est parfois difficile de statuer de la véracité de celles-ci, comme avec le SARS-CoV-2. De plus, dans certains cas, les associations du CMH sont attribuées à tort au HLA. Des GWAS ont alors mis en évidence l'impact de la région du CMH dans la schizophrénie, et plus particulièrement de certains allèles HLA (254,270,271). Sekar *et al*. ont écarté cette hypothèse en partie, en montrant qu'une grande partie de cette association reposait en fait sur un polymorphisme de C4, un gène du complément présent dans la région de CMH classe III (272).

En dehors de ces pathologies, une littérature extensive sur l'impact de la région du CMH sur le choix de partenaire chez les humains, et les vertébrés de manière plus large, semble pencher pour l'existence d'une reconnaissance des molécules du HLA entre individus. Ce phénomène permettrait de conserver une diversité de molécules HLA dans la descendance et favoriser l'hétérozygotie et donc étendre le répertoire des peptides présentés (273,274).

La diversité inhérente aux gènes HLA est donc non-seulement centrale dans l'immunité humaine mais elle régit aussi des centaines d'interactions connexes, devenant des facteurs de risque ou de protection au-delà de l'immunité. Il est parfois difficile de vérifier ou de comprendre les associations dans la région du CMH car ce ne sont que des variations de fréquences de SNP entre des groupes d'individus. Ces associations SNP sont parfois confirmées par une association avec des allèles HLA directement. Heureusement, le champ d'étude du HLA s'est largement développé et il est maintenant possible d'utiliser d'autres informations telles que l'expression ou la structure moléculaire pour investiguer le HLA. L'attribution du rôle biologique de la région du CMH est ainsi souvent donnée au HLA, mais parfois à tort. Des méthodes d'inférence statistiques des allèles HLA sont à l'œuvre depuis une dizaine d'années et sont une solution pour étudier la région génomique du CMH.

IV - Une grande diversité implique de grandes dimensions : quand la statistique et l'informatique se mettent au service de la génétique

L'émulation autour de la région du CMH tient à sa diversité allélique qui est expliquée par son évolution, entraîne sa complexité génomique et l'implique dans des pathologies extrêmement diverses. En ce sens, le CMH et les gènes du HLA sont le fer de lance de nombreuses améliorations en génomique, en génétique des populations et en épidémiologie génétique. Ainsi, l'exploration de la diversité HLA passe par des études différentes, dont l'association HLA. Ces études sont complétées par des outils permettant d'analyser l'ancestralité génétique d'individus ou pour prédire les génotypes HLA pour générer de nouvelles données, par exemple.

IV.1 - Les différentes facettes de l'analyse bioinformatique du HLA

Comme la pandémie de SARS-CoV-2 a pu le montrer en temps accéléré, la mobilisation de la région du CMH dans la recherche se fait à différents niveaux : des réponses rapides *in silico*, des études d'association, puis l'extension des recherches à des paramètres immunogénétiques plus larges.

IV.1.1 - Les réponses *in silico* sont en première ligne pour étudier le HLA

IV.1.1.1 - Les corrélations HLA-trait et leurs multiples failles

Bien qu'incomplète, la connaissance de la distribution des allèles HLA dans les différentes populations ne cesse de grandir au fil des années. D'un point de vue mondial, l'Allele Frequency Net Database (AFND) (163) représente une base de données essentielle pour suivre cette évolution. Mais il existe aussi au niveau national de nombreuses biobanques et autres registres de donneurs qui permettent de mobiliser très rapidement les fréquences d'allèles ou d'haplotypes HLA (191,252,275,276).

Dans le cadre de l'étude de pathologies, ces données en population peuvent être utilisées pour évaluer la corrélation statistique entre un trait quantitatif (*i.e.* le nombre de cas de COVID-19 par région) et des fréquences d'allèles HLA dans différentes populations (*i.e.* les fréquences d'allèles HLA dans plusieurs pays). Cette cooccurrence est souvent calculée avec le coefficient de Pearson qui représente une relation linéaire entre deux variables. Dans le cas de l'incidence du COVID-19 dans le monde, cette corrélation peut être calculée à partir de la covariance de l'incidence (Inc_{pays}) et de la fréquence ($Freq_{pays}$) d'un allèle en particulier dans chaque pays, comparée à celle observée mondialement ($Inc_{mondiale}$ et $Freq_{mondiale}$) divisée par le produit des écarts-types (σ) (Équation IV-1).

Équation IV-1 Corrélation en population entre l'incidence d'une pathologie et la fréquence d'un allèle HLA

$$\rho_{Inc,Freq} = \frac{\mathbb{E}\left[(Inc_{pays} - Inc_{mondiale}) \times (Freq_{pays} - Freq_{mondiale})\right]}{\sigma_{Inc} \times \sigma_{Freq}}$$

Bien que ces études puissent être obtenues plus rapidement et facilement que de nouveaux génotypes HLA, elles ne rendent pas compte de réelles associations statistiques car elles décrivent deux populations distinctes. Cela accroit le risque d'effectuer une corrélation sur la position géographique ou sur tous les facteurs de confusion génétiques et environnementaux pouvant être liés à celle-ci (236). Les nombreux tests statistiques en association HLA (que ce soit selon le nombre de SNP ou le nombre d'allèles HLA) nécessitent également une correction liée à l'utilisation de tests statistiques multiples. L'absence de cette correction dans des publications résulte en la présentation des associations fortuites. Ces tests doivent évaluer si les associations obtenues sont dues au hasard ou véritables et l'absence de correction tend à surestimer les associations. Enfin, cette méthode est fondamentalement biaisée car la recherche d'un déséquilibre entre la présence d'un allèle chez des individus porteurs d'une pathologie ou non, ne peut pas se baser sur des fréquences d'allèles HLA de populations générale. D'autant plus que la faible fréquence des allèles HLA, ou pire des haplotypes HLA, pousse à observer seulement les plus communs, introduisant de nouveaux biais.

IV.1.1.2 - Sur la piste des allèles HLA à risque et protecteurs grâce à l'exploration du peptidome

L'une des hypothèses les plus simples pour envisager le rôle du HLA dans une pathologie est à travers son peptidome, soit l'ensemble des peptides que l'allèle est susceptible de présenter. Dans le cas d'une maladie infectieuse, il serait attendu qu'une présentation accrue permettrait aux lymphocytes d'interagir plus rapidement avec un peptide du pathogène, et réciproquement, un allèle pourrait être à risque s'il présente moins d'allèles d'un pathogène donné. L'inverse est attendu pour des maladies auto-immunes, où la présentation d'un peptide du soi antigénique est délétère. Dans le cas des cancers, la présentation de néo-antigènes, des peptides issus de protéines du soi mutées, par le système HLA est également d'un intérêt particulier (277–280).

Depuis la fin des années 90, les scientifiques tentent de rassembler les interactions HLA-peptide dans des bases de données (281) et ont coordonné leurs efforts pour créer l'Immune Epitope Database (IEDB), la plus grande bases de données peptidiques. Les premiers résultats sont obtenus en calculant l'affinité des peptides grâce à leur cinétique de liaison. Les peptides étudiés sont mis en compétition avec un peptide radioactif de haute affinité pour la molécule de HLA. La quantité du peptide identifié nécessaire pour déloger le peptide radioactif sert ainsi de mesure (282). La spectrométrie de masse est depuis devenue standard pour l'identification des peptides présentés et facilite la génération de ces données (283).

La nécessité de générer des modèles capables de décrire l'interaction HLA-peptide est née de la diversité limitée des bases de données : la plupart des allèles décrits étaient communs dans les populations européennes, mais rares dans le reste du monde (284). Des outils de prédiction du peptidome ont donc été développés pour pallier ce manque, tout d'abord pour les HLA de classe I. Les outils pour les HLA de classe II ont nécessité plus de temps car l'interaction HLA-peptide est plus complexe : les peptides ont des tailles variables et deux molécules différentes forment le sillon peptidique (285). La prédiction se base sur l'information déjà présente dans les bases de données HLA-peptide pour inférer le peptidome d'allèles HLA non-décrits. Mei *et al.* (286) ont décrit plusieurs méthodes différentes avec ce même but : les scores, l'apprentissage statistique (*machine learning*) et le consensus (Figure IV-1, adaptée de Mei *et al.* (286), cristallographie par Ling *et al.* (287), créée avec biorender.com).

Un exemple d'une méthode de score est l'utilisation d'une matrice de score par position (*Position-Specific Score Matrix, PSSM*). Cette méthode assigne, à partir des données HLA-peptide, des scores individuels aux acides aminés dans une matrice *PSSM*, et ce pour chacun des résidus du peptide.

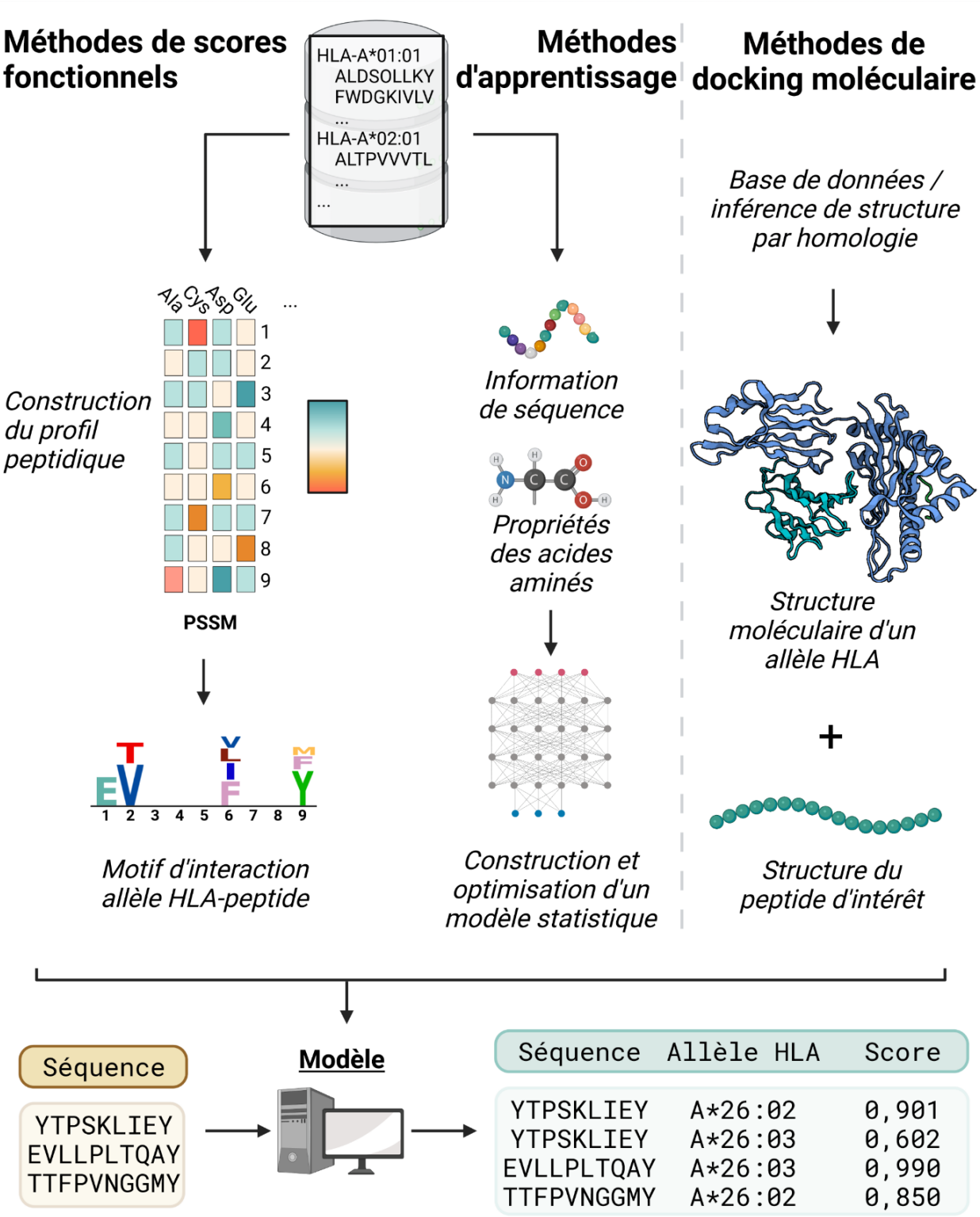


Figure IV-1 Prédiction des interactions peptide-HLA pour tous les allèles. Les méthodes de scores fonctionnels et d'apprentissage se reposent sur des données de séquence, et d'élution ou de spectrométrie de masse connues pour des allèles HLA. Les méthodes de docking simulent l'interaction moléculaire entre les peptides et une structure 3D connue (structure 4HWZ de Ling et al.(287) de la Protein Data Bank, pour HLA-A68), ou prédite, d'un allèle HLA. Toutes ces méthodes décrivent un modèle qui peut inférer l'affinité théorique entre un peptide donné et une molécule HLA. Adapté de Mei et al. et créée avec biorender.com

Elle est parfois complémentée d'une matrice de distance (*i.e.* BLOSUM 62) qui contient des informations de similarité entre les acides aminés selon leurs propriétés physico-chimiques.

Parmi les méthodes d'apprentissage statistique pour la prédiction de peptidome, les réseaux de neurones sont les plus utilisés. Brièvement, ces réseaux sont un ensemble de couches de neurones que traversent les variables d'entrée d'un jeu de données d'entraînement (*i.e.* séquence HLA, propriétés des acides aminés, etc.), qui sont modifiées et passées à la couche suivante. A chaque variable est assignée un poids selon son importance et ce poids est modifié au fur et à mesure des itérations pour correspondre à la sortie voulue (*i.e.* l'affinité que l'on cherche à prédire). Comme son nom l'indique, le consensus optimise plusieurs types de modèles à la fois (score et *machine learning*) pour maximiser les résultats de prédiction. Un quatrième type d'algorithme fonctionne avec les structures moléculaires obtenues par cristallographie aux rayons X ou rayonnement NMR (*Nuclear Magnetic Resonance*). Cependant, pour obtenir de tels résultats, il faut au préalable obtenir une structure en trois dimensions de l'allèle HLA étudié, soit par expérimentation, soit par prédiction. Cette dernière possibilité requiert des structures homologues et nécessite donc des structures adaptées pour tous les allèles HLA.

Bien que ce champ de recherche soit en évolution constante, mixMHCpred (288,289) et netMHCpan 4.1 (156,290) sont les algorithmes les plus performants actuellement pour les HLA de classe I, en méthode de score et d'apprentissage statistique, respectivement.

IV.1.2 - L'association statistique dans la région du CMH

Les analyses *in silico* donnent une première idée de la relation possible entre des allèles HLA et une pathologie étudiée. L'accès à des génotypes HLA permet de faire un pas en avant et d'effectuer des études pangénomiques d'association de SNP (*genome-wide association studies,* GWAS). Ces GWAS ont permis d'identifier la région du CMH comme extrêmement associée aux pathologies humaines, mais on distingue deux types d'associations : l'association SNP et l'association HLA. La première se base sur des ensembles de SNP présents dans la région du CMH, sans désigner forcément un gène en particulier, la seconde est plus précise et se base directement sur des allèles HLA.

IV.1.2.1 - Le CMH est le berceau de milliers d'associations génétiques emmêlées dans des motifs de déséquilibre de liaison

Les GWAS sont des méthodes d'identification de polymorphismes communs de l'ADN comme facteurs de risque ou de protection pour des phénotypes divers, allant de la taille au diabète de type 1. Les premières GWAS ont été introduites entre 2005 et 2006 (291,292) grâce à l'essor des puces de génotypage qui permettent de connaître des milliers de polymorphismes préalablement sélectionnés, par individu. Ces puces sont équipées de simples brins d'ADN complémentaire correspondant à des

polymorphismes connus du génome, placés de manière ordonnée. L'ADN simple brin est coupé et une réaction de fluorescence se produit lorsque l'échantillon de l'individu a pu s'apparier à celui présent sur la puce, indiquant le polymorphisme (293). Les génotypes obtenus sont composés de 0, 1 ou 2, selon le nombre de copies du polymorphisme retrouvé. Les nouvelles technologies de séquençage sont maintenant utilisées pour obtenir les génotypes mais les puces sont toujours populaires car généralement moins coûteuses et plus facile à mettre en place.

Équation IV-2 Les régressions linéaires (a) et logistiques (b) permettent d'évaluer l'association statistique de SNP et d'autres covariables à un phénotype continu et catégoriel, respectivement. La transformation ln(y/1-y) permet de considérer un trait binaire comme continu.

$$(a)\ y = G\beta_G + X\beta_X + \varepsilon$$
$$(b)\ \ln\left(\frac{y}{1-y}\right) = G\beta_G + X\beta_X + \varepsilon$$

L'association génétique est obtenue avec ces génotypes par régression linéaire ou logistique, selon le trait observé (Équation IV-2). La régression attribue à chaque génotype SNP (G) observé un coefficient (β_G) et ajoute un résidu ε tenant compte de la différence entre la valeur du phénotype réel (y) et celle modélisée. Il est également possible d'ajouter des covariables (X), avec leur coefficient (β_X), qui peuvent expliquer une partie du phénotype (Figure IV-2, créée avec biorender.com). Sans ces covariables, les coefficients des génotypes pourraient également refléter des facteurs confondants. L'âge, le sexe et des variables supplémentaires prenant en compte l'ancestralité génétique des individus sont ainsi régulièrement ajoutées. L'ancestralité correspond à une représentation de la similarité génétique entre les individus. Chaque individu est alors défini comme un génome-mosaïque issu de populations génétiques ancestrales. Ces populations sont elles-mêmes définies par des différences génétiques acquises au cours de l'évolution.

Lors d'une GWAS, la région du CMH est étudiée comme le reste du génome, cependant, elle a souvent été exclue des analyses d'association à cause de sa densité de gènes, de ses long motifs de déséquilibre de liaison, de la fréquence de ses variants ou simplement pour mieux visualiser les associations sur le reste du génome (23,294). Les GWAS relient généralement les SNP significativement associés à des gènes proches mais peuvent également rechercher les voies de signalisation communes ou chercher un impact sur des données de transcriptomique (295).

Dans des cas très précis, certains SNP (les *tag SNP*) indiquent non seulement un gène HLA probablement impliqué mais également un allèle HLA spécifique. Ceci est possible lorsqu'un SNP et un allèle HLA sont en déséquilibre de liaison parfait. Le cas du HIV et du HLA-B*57:01/rs2395029 (296) est l'exemple parfait d'un *tag SNP*.

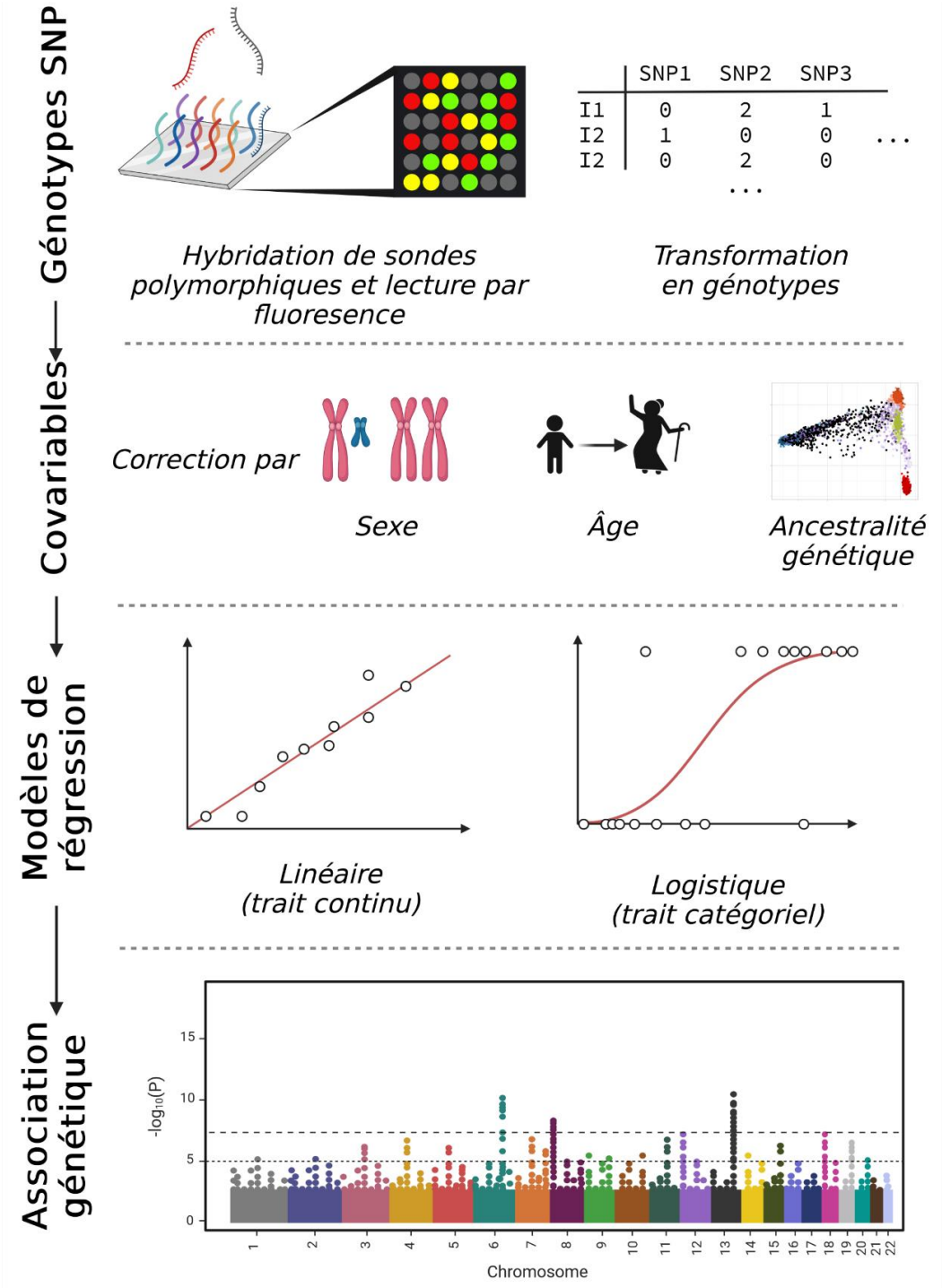


Figure IV-2 Du génotypage à l'analyse d'association. Les puces de génotypage (ou le séquençage) indiquent par fluorescence la présence des polymorphismes recherchés, qui sont transformés en 0, 1 ou 2 pour chaque SNP pour chaque individu. Les modèles de régression intègrent les SNP, et toutes les covariables susceptibles d'influencer le phénotype observé. La régression donne une p-value à chaque variable du modèle, ainsi qu'une taille d'effet, non-représentée sur le Manhattan plot. Créée avec biorender.com.

Néanmoins, cette propriété est rare et n'est pas forcément conservée selon les populations étudiée (297). Le déséquilibre de liaison et les nombreux gènes de la région du CMH causent ainsi des difficultés d'interprétation des signaux d'association ou d'attribution à des allèles HLA spécifiques.

IV.1.2.2 - Les associations HLA lient directement les traits et les allèles

Le séquençage HLA permet de s'affranchir en partie de l'incertitude d'association qui entoure les SNP de la région du CMH. En effet, si l'on dispose d'allèles HLA typés pour des individus, il est possible d'effectuer une étude d'association classique.Certaines biobanques et registres de donneurs peuvent donc mobiliser des données HLA individuelles et, en les couplant à des données cliniques, effectuer des analyses HLA avec de grands effectifs (252).

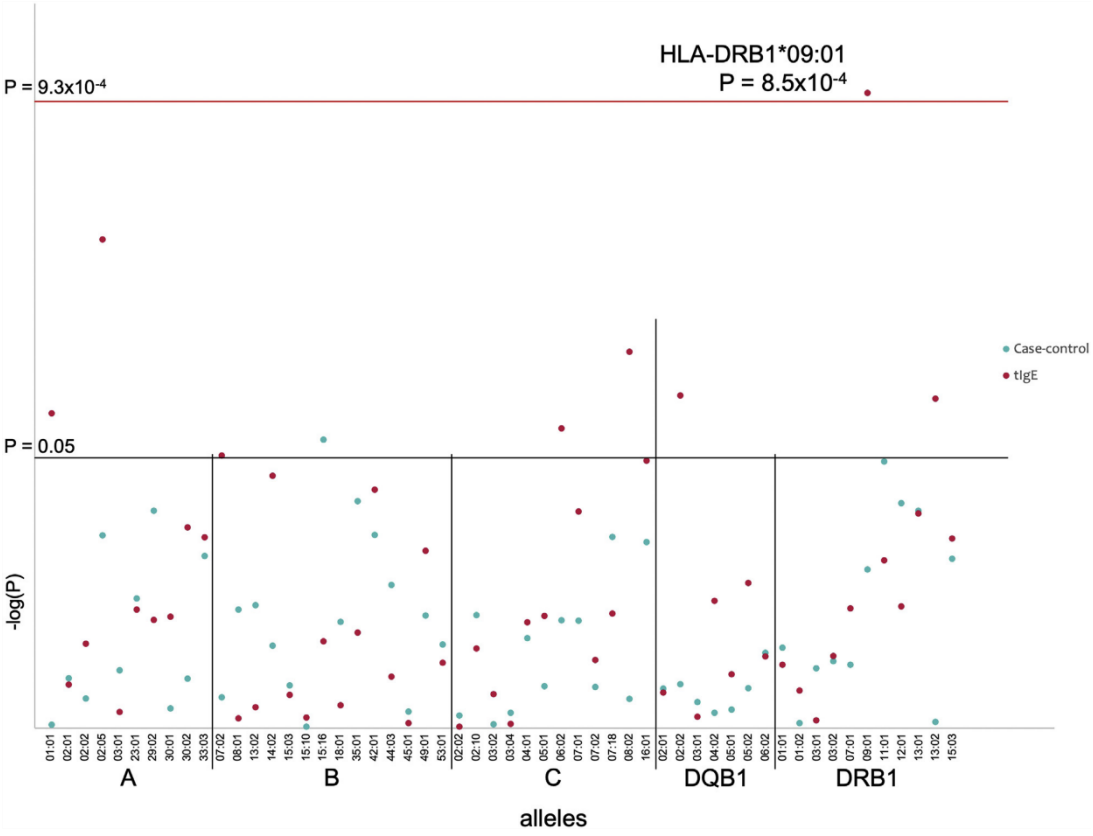


*Figure IV-3 Exemple d'association HLA dans une population afro-américaine. HLA-DRB1*09:01 est positivement associé à la présence d'immunoglobulines E dans le sérum total des individus asthmatiques. Issue de Vince et al.*

Certains ajustements néanmoins doivent être suivis. Chaque locus HLA dispose de milliers d'allèles. Il est donc indispensable de transformer ces allèles en variables de génotype : pour chaque individu, on dispose donc d'un vecteur de milliers d'allèles, annotés 0, 1 ou 2 selon leur nombre d'observations. De plus, les allèles HLA suivent un mode d'expression codominant (298), où les deux allèles s'expriment dans la cellule, il convient alors d'étudier l'association génétique du point de vue dominant ou allélique. Le premier considère la présence d'un ou deux allèles identique comme équivalente alors que la seconde est proportionnelle au nombre de copies de l'allèle. Le résultat est ensuite présenté,

comme pour les GWAS, avec chaque valeur de significativité d'association en ordonnée, l'abscisse ne reflète plus la position génomique mais simplement l'allèle HLA étudié (Figure IV-3, issue Vince *et al.* (299)). Ces études se concentrent généralement sur les allèles fréquents pour garder des effectifs statistiquement viables.

Techniquement, l'étude d'association HLA réussit là où l'étude GWAS échoue, en mettant en avant un allèle HLA responsable du lien statistique. Cependant, il faut être vigilant car même un allèle HLA peut être en déséquilibre de liaison avec un polymorphisme causal proche (272). De plus, le polymorphisme important des allèles HLA nécessite un effectif conséquent pour les études (212). Le séquençage limite pourtant le nombre d'individus inclus dans les études à cause de son coût global, ce qui entraîne la conduite de petites études. Pour contourner ce problème, l'inférence statistique d'allèles HLA est une solution de *machine learning* qui devient maintenant de plus en plus commune.

IV.1.3 - L'extension des études immunogénétiques à partir du HLA

L'association avec des allèles HLA est une facette importante de l'immunogénétique, mais ce n'est pas la seule. De nombreux paramètres immunogénétiques connexes peuvent être dérivés ou étudiés en parallèle pour enrichir les analyses HLA (143).

Concernant le HLA, les études d'association peuvent dépasser la résolution allélique et se concentrer jusqu'aux acides aminés. Domenighetti *et al.* ont ainsi pu redéfinir l'interaction entre le gène HLA-DRB1, la maladie de Parkinson et le statut tabagique par l'intermédiaire de la valine 11 (300). Certains allèles HLA ont seulement quelques acides aminés différents, en partager un dans le sillon peptidique peut donc les rapprocher d'un point de vue mécanistique. Ces mêmes acides aminés peuvent aussi être d'intérêt dans la transplantation, notamment en recherche d'épitopes, pour analyser l'immunogénicité potentielle d'allèles HLA discordants (151).

En prenant du recul sur la région du CMH, il est également possible d'envisager l'étude de tous les allèles HLA dans un haplotype complet. Si l'association directe avec un haplotype peut être ralentie par la faible fréquence de ceux-ci, la prédiction d'haplotypes est essentielle dans la transplantation de cellules souches hématopoïétiques, afin d'identifier de potentiels donneurs haploidentiques (113). Ainsi, l'outil HLA-2-Haplo de la suite Easy-HLA infère les haplotypes HLA à partir de génotypes existants et permet d'étendre les connaissances sur la région (50).

Par ailleurs, si les allèles HLA classiques, et plus particulièrement *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* et *HLA-DRB1*, sont les plus polymorphes et aussi les plus étudiés, le reste du système HLA entier ainsi que ses partenaires protéiques ne sont pas en reste. Les polymorphismes de HLA-G peuvent ainsi modifier la réponse à la malaria (176), et ils jouent un rôle important dans la tolérance materno-fœtale (301).

Cette dernière propriété fonctionne de concert avec les gènes KIR du chromosome 19, connus pour leur polymorphisme à l'échelle du CMH, et qui ont pour ligands des HLA de classe I (169).

IV.2 - Les données manquantes en HLA : contourner l'absence d'information par le contexte génétique

Connaître les allèles HLA d'individus est la porte d'entrée à de multiples analyses immunogénétiques et à la meilleure compréhension de centaines de pathologies. Malheureusement, si les GWAS ont pu profiter des puces de génotypage pour devenir omniprésentes avec des effectifs toujours plus grands, l'association HLA se heurte au coût du séquençage, ce qui conduit vers des études plus petites ou absentes. Les données manquantes sont un sujet de recherche en statistique et leur place dans le domaine médical et génétique ont pavé la voie pour résoudre de tels problèmes.

IV.2.1 - Les concepts généraux de l'imputation de données

En 1976, Rubin théorise les raisons derrière les données manquantes d'un jeu de données et les divise en trois catégories (302) : *Missing Completely At Random* (MCAR), *Missing At Random* (MAR), et *Missing Not At Random* (MNAR). Les données MCAR sont des données dont l'absence est entièrement indépendante du reste du jeu de données, comme des erreurs sporadiques d'expérimentation ou une maintenance technique imprévue d'un instrument de mesure. Les données MAR, quant à elles, sont partiellement dues au hasard car elles sont liées de manière conditionnelle à une autre variable du jeu de données, ainsi une mesure clinique de pression sanguine est souvent absente chez des personnes plus jeunes. Enfin, pour les données MNAR la probabilité d'observer la donnée manquante est associée à la donnée en elle-même. La distribution de la donnée est donc différente entre les individus qui l'ont et ceux qui ne l'ont pas (303). En conséquence, seuls les scenarios MCAR et MAR permettent l'inférence de la valeur manquante à partir des données restantes.

Il est possible de s'affranchir de ces données manquantes en excluant de l'analyse les individus (ou les variables) aux données manquantes ou de compléter d'une manière ou d'une autre. La moyenne, la médiane, les régressions linéaires ou des algorithmes plus complexes, comme les forêts aléatoires, sont des solutions possibles pour inférer statistiquement la donnée manquante à partir du reste du jeu de données. En pratique, l'imputation multiple connaît une grande popularité (304). Elle consiste à observer la variabilité de plusieurs imputations pour déterminer le résultat final, mais la méthodologie appliquée entre les études est souvent variable (305).

Les mathématiques et la statistique font partie intégrante de l'histoire de la génétique, de l'hérédité et de la compréhension de la génétique en population (306). Ainsi, ces méthodes ont très rapidement trouvé leur utilité en médecine, puis en génomique, pour compléter des jeux de données.

IV.2.2 - L'imputation SNP, une application statistique de grande envergure en génomique

Les GWAS ont été portées par la technologie des puces de génotypage, mais ces puces ne contiennent qu'une fraction choisie des polymorphismes humains, réduisant le pouvoir statistique de telles études pour découvrir de nouvelles associations.

Pour répondre à ce problème, des outils statistiques ont été construits à partir de bases de données d'haplotypes en s'appuyant sur le déséquilibre de liaison pouvant exister entre les SNP. Deux polymorphismes proches sur l'ADN ont statistiquement moins de risque d'être séparés au fur et à mesure des générations, ils sont transmis en bloc pour former des haplotypes : statistiquement, la présence d'un SNP de l'haplotype aiguille sur la présence (ou non) de SNP alentours (Figure IV-4, traduite de Marchini *et al.* (307)).

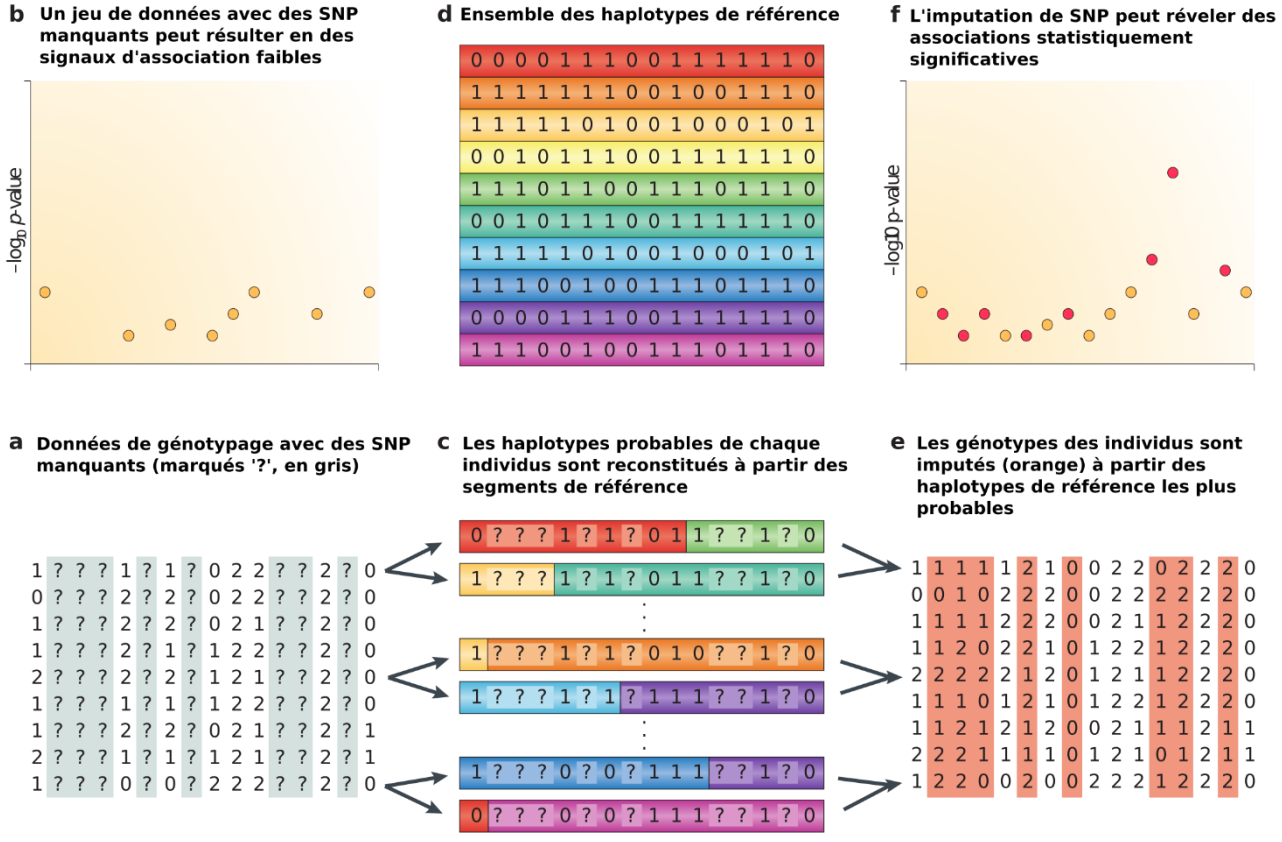


Figure IV-4 Principe de l'imputation de SNP à partir d'haplotypes SNP d'une population de référence. Chaque individu a un génotype SNP (a) qui peut être représenté par deux haplotypes qui sont des mosaïques de ceux de référence (c,d). Le génotype peut ainsi être complété selon les SNP présents sur la référence (e), pouvant révéler de nouvelles associations génétiques (b,f). Traduite de Marchini & Howie.

De multiples outils d'imputation ont implémenté des algorithmes d'imputation de SNP au fil des années comme : Impute (308,309), MaCH (310), fastPHASE (311) ou Beagle (312,313), qui reposent

notamment sur des modèles de Markov cachés pour décrire les relations probabilistes entres les SNP d'un haplotype.

Le succès de l'imputation SNP vient de sa grande précision mais également de son accessibilité. Les serveurs d'imputation de l'université du Michigan (314), le Wellcome Trust Sanger Institute (315) et maintenant TopMed (316) ont contribué grandement à la démocratisation de l'imputation de SNP dans les études d'association en permettant aux chercheurs de téléverser leurs données de génotypage à distance pour obtenir les génotypes imputés.

Les méthodes d'inférence statistique de SNP à partir de SNP ont permis l'essor des GWAS et la découverte de milliers d'associations, dont des signaux dans la région du CMH. Seulement, ces signaux identifient difficilement des gènes à cause de leur densité et du fort déséquilibre de liaison à travers toute la région.

IV.2.3 - L'inférence statistique HLA, vers l'immunogénétique pour tous

Très rapidement les méthodes d'imputation SNP-SNP ont été adaptées dans un but plus précis : celui de l'imputation HLA. En effet, cela a permis d'examiner en détail la région du CMH qui se retrouvait déjà associée dans la plupart des GWAS (317). Alternativement, l'imputation HLA peut décrire l'obtention de génotypes HLA à partir d'un séquençage de génome, ou exome, entier.

IV.2.3.1 - La déséquilibre de liaison de la région du CMH, un faisceau d'indice pour retrouver l'identité d'un allèle HLA

Les allèles HLA sont des collections de SNP. Ainsi, l'imputation HLA à partir de SNP fonctionne sur la même base que l'imputation de SNP : des jeux de données de référence SNP et HLA nourrissent des algorithmes qui permettent à leur tour de prédire de nouveaux allèles HLA.

L'imputation HLA à l'aide de SNP permet de capitaliser sur les données GWAS déjà générées en grand nombre pour obtenir une information de génotype HLA (315). Ceci a un intérêt énorme pour les études d'association qui peuvent alors investiguer plus en détail l'association du CMH dans des pathologies (318).

Les premières méthodes d'imputation HLA sont en réalité concomitantes avec l'imputation SNP mais elles n'ont cessé d'évoluer en termes de données acceptées, de techniques utilisées, et de précision (Tableau IV-1, e-HLA issu de Pappas *et al.* (319)). En 2006, de Bakker *et al.* (296) font l'une des premières mentions de la déduction d'allèles HLA et décrit assez simplement des polymorphismes corrélés aux allèles HLA, les *tag SNP*, qui permettaient de les prédire. Cette méthode reste encore assez limitée dans son utilisation.

La première génération d'outils d'inférence statistique s'est inspirée des méthodes d'imputation SNP et nécessite souvent des bases de données de référence, contenant des individus avec des données SNP phasées en haplotypes, ainsi que des génotypes HLA, comme la méthode LDMhc (317). Ce sont des méthodes qui vont tenter de représenter les haplotypes de SNP d'individus comme une mosaïque d'haplotypes de référence et leur assigner l'allèle HLA de référence associé.

L'imputation HLA connaît de nombreuses améliorations dans les années suivantes. Cette deuxième génération est majoritairement représentée par HIBAG (320), SNP2HLA (321) et HLA*IMP:02 (322). Brièvement, même si leurs algorithmes diffèrent de manière assez fondamentale, ils reposent sur un principe global d'imputation (Figure IV-5, traduite de Douillard *et al.* (143)). Des individus avec des génotypes SNP non-phasés, ainsi que des génotypes HLA, composent des panels de référence, qui sont utilisés pour déduire un modèle statistique liant SNP et HLA.

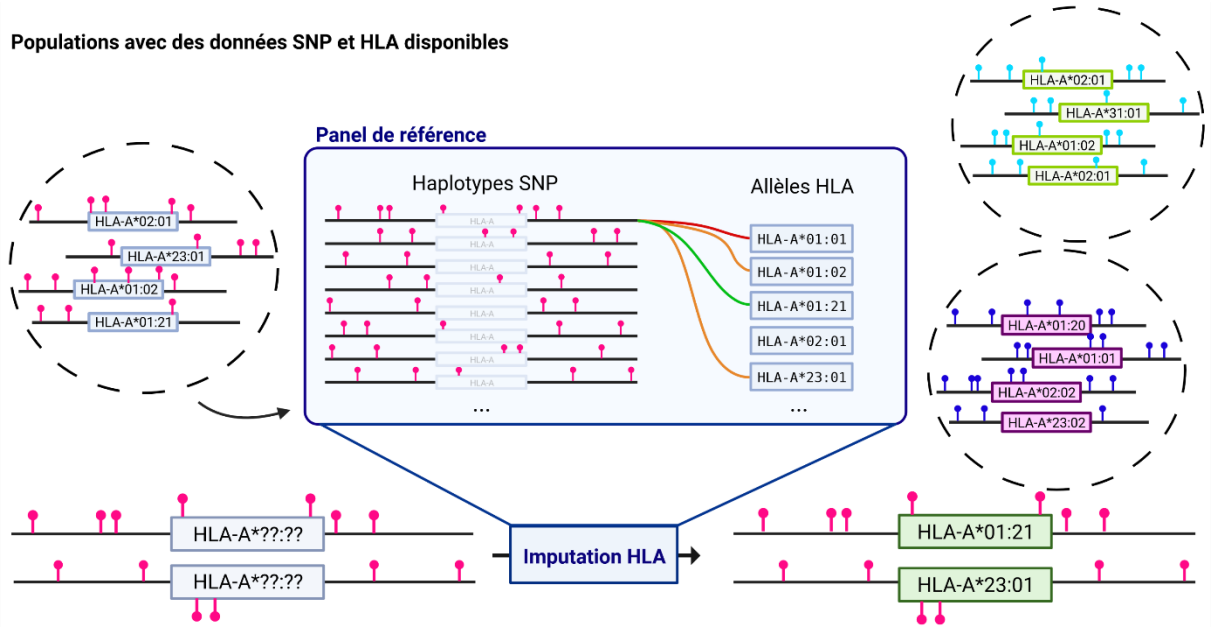


Figure IV-5 Principe de l'imputation HLA. Les données de génotypes SNP et HLA servent de panel de référence, ils sont phasés en plusieurs haplotypes avec des probabilités d'observation d'un allèle HLA pour un haplotype SNP spécifique. A partir d'un nouveau génotype SNP, il est possible de le représenté par plusieurs couples d'haplotypes SNP dans le panel de référence. Des allèles HLA sont associés à ces haplotypes, ce qui permet de sélectionner le génotype HLA le plus probable. Les résultats d'une imputation HLA sont dépendants du panel de référence, qui peut changer les allèles et les haplotypes SNP liés aux allèles selon les populations. Traduite de Douillard et al.

La vraisemblance des haplotypes de SNP-HLA est estimée à partir des données de référence : grâce au déséquilibre de liaison, un même allèle HLA a souvent un contexte polymorphique semblable chez plusieurs individus. Cependant, ce contexte peut changer : la vraisemblance quantifie alors cette probabilité conditionnelle de chaque couple haplotype SNP-HLA d'après les données de référence. Cet ensemble constitue un panel de référence. Le génotype SNP non-phasé d'un nouvel individu est

confronté à ce panel de référence. Alors, tous les couples d'haplotypes SNP pouvant satisfaire le génotype SNP sont étudiés et la vraisemblance est utilisée pour retrouver le génotype HLA le plus probable.

HLA*IMP:02 et SNP2HLA fonctionnent avec un graphe d'haplotypes et des chaînes cachées de Markov, inspirées des évolutions du logiciel Beagle, pour l'imputation SNP. HIBAG repose sur un algorithme *d'attribute bagging*, une méthode utilisant plusieurs sous-ensembles d'un jeu de données. Son originalité principale vient de sa fonctionnalité de création de modèles d'imputations qu'il est alors le seul à proposer.

*Tableau IV-1 L'historique de l'imputation HLA, de l'utilisation détournée de l'imputation des SNP au deep learning. EM = Expectation-Maximization. HMM = Hidden Markov Model. *e-HLA a été comparé à d'autres logiciels par Pappas et al., cependant sa référence est un article non-publié.*

Nom	Méthode	Date	Auteur
-	Tag SNP	2006	De Bakker (296)
LDMhc	IBD & HMM	2008	Leslie (317)
e-HLA*	-	2010	Biesiade
WSG-HI	Graphe de similarité pondéré	2010	Minzhu (323)
-	Parcours de graphe d'IBD	2010	Setty (324)
HLA*IMP	LDMhc & sélection de SNP modifiée	2011	Dilthey (325)
MAGPrediction	EM	2011	Li (326)
-	EM & Fréquence d'haplotypes HLA	2012	Paunić (327)
SNP2HLA	Beagle	2013	Jia (321)
HLA*IMP:02	Graphe d'haplotype du CMH	2013	Dilthey (322)
HIBAG	Attribute bagging & EM	2014	Zheng (320)
Amb-EM	EM & fréquence d'haplotypes HLA	2014	Paunić (328)
HLA*IMP:03	Forêt aléatoire	2016	Motyer (329,330)
MHC*IMP	Forêt aléatoire	2020	Squire (331)
CookHLA	SNP2HLA & distance génétique	2021	Cook (332)
DeepHLA	Deep learning	2021	Naito (333)

SNP2HLA et HIBAG sont les algorithmes les plus populaires dans l'imputation HLA. Une comparaison indépendante de ces logiciels par Karnes *et al.* (334) montre de meilleurs performances pour SNP2HLA, mais d'autres études ultérieures donnent HIBAG en avance (319,335,336). Amb-EM (328) quant à lui n'est pas retrouvé dans la littérature mais a l'unique particularité d'imputer directement à partir d'allèles HLA ambigus. Ces allèles ambigus correspondent à une liste de plusieurs allèles désignés comme probable, et AmbEM utilise les fréquences d'haplotypes HLA pour sélectionner le résultat final.

La popularité de cette deuxième génération de logiciels d'imputation HLA tient à leur facilité d'accès : il suffit de génotypes SNP non-phasés, ceux d'une puce de génotypage, pour imputer. Ils sont plus accessibles aux chercheurs, faciles à installer et à utiliser. L'augmentation de la puissance de calculs des ordinateurs les rend utilisables par tous.

Néanmoins, au fil des améliorations méthodologiques de l'imputation HLA, le typage HLA a aussi connu un essor important : le nombre d'allèles connus. Les allèles HLA peuvent être présents à de très faibles fréquences. D'autant plus que ces fréquences diffèrent entre les populations, ce qui a posé la question de la représentation des populations non-européennes dans les panels de référence. Par ailleurs, l'hétérogénéité des SNP présents sur les puces de génotypage ont ralenti leur utilisation.

Récemment, une troisième génération d'imputation HLA a apporté des améliorations majeures de précision. CookHLA dérive son algorithme de SNP2HLA mais prend également en compte la distance génétique entre les polymorphismes, selon un ensemble de populations de référence, pour mieux imputer les différentes ancestralités. L'algorithme de DeepHLA repose sur la technologie de *deep learning*, ou réseaux de neurones profonds, qui permet de capter de subtils motifs de corrélation entre SNP et HLA. Il maximise ainsi la précision d'imputation grâce aux relations entre les acides aminés des allèles HLA, et aussi grâce au déséquilibre de liaison entre les différents loci des gènes HLA.

Malgré ces efforts, l'imputation SNP-HLA est dépendante de ses panels de référence dans la diversité des allèles représentés. Comme il est également impossible de prédire des allèles *de novo*, le séquençage est ainsi le standard pour typer les allèles HLA.

IV.2.3.2 - Le séquençage HLA : une solution infaillible ?

Le typage précis des allèles HLA est possible par NGS en se concentrant sur les loci HLA. En parallèle, de nombreux séquençages de génome (*whole-genome sequencing*, WGS) ou d'exome (*whole-exome sequencing*, WES), sont faits pour des études plus générales d'association dans le génome entier sans hypothèse préalable sur la région génomique d'intérêt.

Ces derniers ne sont pas adaptés pour le typage HLA car l'alignement est fait avec un seul génome de référence, alors qu'un individu peut avoir de nombreuses différences génétiques dans la région du

CMH : les fréquences alléliques et les génotypage obtenus ne sont alors pas fiables (337). Les logiciels d'alignement, utilisés tels quels, ne sont pas suffisant pour les dizaines de polymorphismes parfois partagés entre les allèles HLA (338). De plus, les gènes HLA ont une forte similarité à cause de longs segments conservés entre eux, les algorithmes ont alors tendance à confondre les gènes séquencés, causant des erreurs (339).

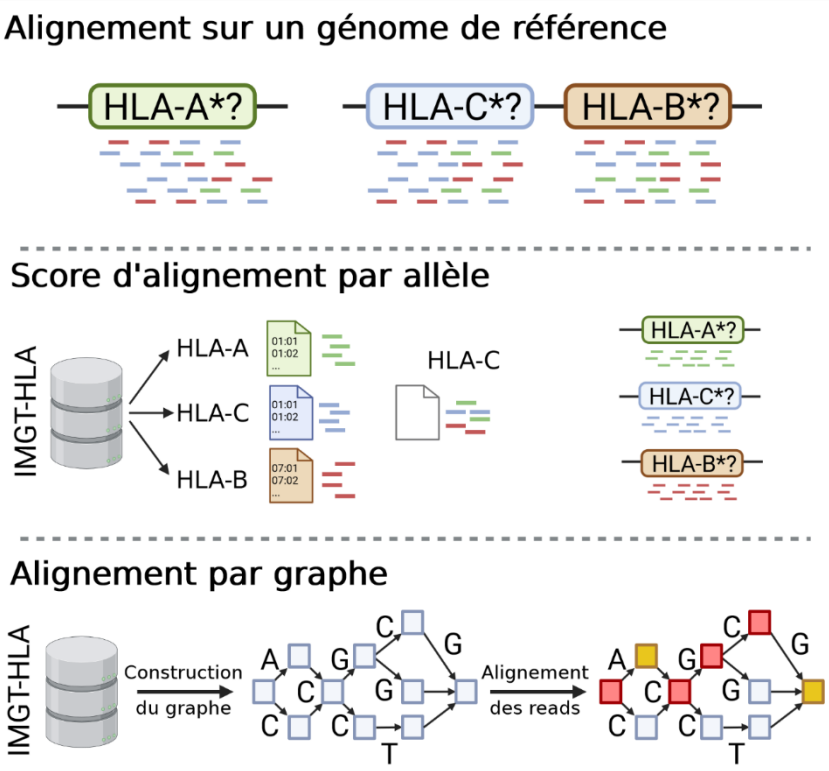


Figure IV-6 L'alignement de reads lors du séquençage de la région du CMH. La forte homologie entre les gènes HLA permet à des reads de s'aligner sur n'importe quel gène, de manière incomplète et il est impossible de typer les allèles avec certitude. Des algorithmes basés sur l'alignement avec une base de données de séquences des allèles HLA, ou sur un graphe de référence améliorent l'alignement. Créée avec biorender.com.

Le typage HLA à partir de ces données n'est pas impossible mais nécessite des algorithmes spécifiques d'alignement et un travail statistique pour inférer les génotypes HLA avec assurance (128). Klasberg *et al.* (340) divisent ces algorithmes entre ceux qui effectuent un alignement classique mais avec un score et des étapes de qualité supplémentaires (ex. Hla-mapper (339)) et ceux basés sur des graphes de populations de référence (ex. HLA*PRG:LA (341)), qui évaluent des probabilités entre plusieurs polymorphismes pour déterminer la probabilité d'emplacement des reads (Figure IV-6, créée avec biorender.com).

En 2016, les premières revues indépendantes d'inférence statistique d'allèles HLA notaient un manque important de précision à une résolution de 2-champs (entre 30 et 81%) à partir de données WGS ou WES (338). Actuellement, ce sont des alternatives solides pour le typage HLA. Chen *et al.* conseillent ainsi HLA-HD (342) dont la précision globale a été estimée à 97%.

Des méthodes qui n'étaient de prime abord pas ajustées pour la complexité du HLA ont été améliorées au fil des années pour atteindre des précisions d'inférence du HLA élevées. Néanmoins, toutes ces méthodes, qu'elles tiennent du génotypage SNP ou du séquençage, n'atteignent pas la précision d'un typage direct.

IV.2.3.3 - Les limites de l'inférence statistique

Le typage HLA permet d'identifier avec une grande précision les allèles HLA d'un individu, dans le but d'observer des incompatibilités en transplantation ou d'effectuer des études d'association. Cependant, pour ces dernières, des milliers d'individus sont nécessaires et le typage peut ne pas être réalisable. Les méthodes d'inférence statistique à partir de SNP, ou en adaptant des séquençages peu profonds, peuvent combler ces lacunes mais des obstacles restent à surmonter.

Pour l'imputation à partir de SNP, un premier problème a été l'hétérogénéité des puces de génotypage en termes de SNP : des puces différentes génotypent des SNP différents, ce qui peut grandement réduire la précision d'imputation. Ceci a pu être réglé par l'imputation SNP (343) ou la sélection préalable de SNP communs (344).

L'obstacle fondamental à l'obtention d'allèles HLA fiables est la composition du panel de référence. Ce panel contient un nombre limité d'allèles, ceux des individus de référence, et est dans l'incapacité d'imputer correctement ceux qu'il n'a pas. Ces problèmes n'étaient pas identifiés avec la première génération d'outils d'inférence grâce au plus faible nombre d'allèles connus et à l'imputation d'individus européens à partir de panels de référence d'européens. L'imputation de n'importe quelle autre population avec une ancestralité éloignée résultait en des erreurs systématiques, à cause des allèles spécifiques et des fréquences différentes (345). En réponse, des chercheurs ont constitué des panels multi-ethniques (346) ou spécifiques d'une population (336,347). Ces derniers sont utiles car des allèles HLA identiques peuvent être retrouvés dans des contextes génomiques différents selon les populations (296).

L'inférence statistique à partir de données de séquençage permet d'outrepasser cette limite d'allèles HLA dans la référence mais il faut alors composer avec les erreurs ou les allèles aberrants dans les bases de données (342). De plus, contrairement aux typages, aucune de ces techniques ne permet la découverte d'allèles *in silico*.

Malgré l'incertitude inhérente à de telles méthodes, l'imputation HLA offre aux chercheurs la capacité de tirer profit de données existantes pour augmenter la puissance statistique des études HLA (Figure IV-7, traduite et adaptée de Douillard *et al.* (143)).

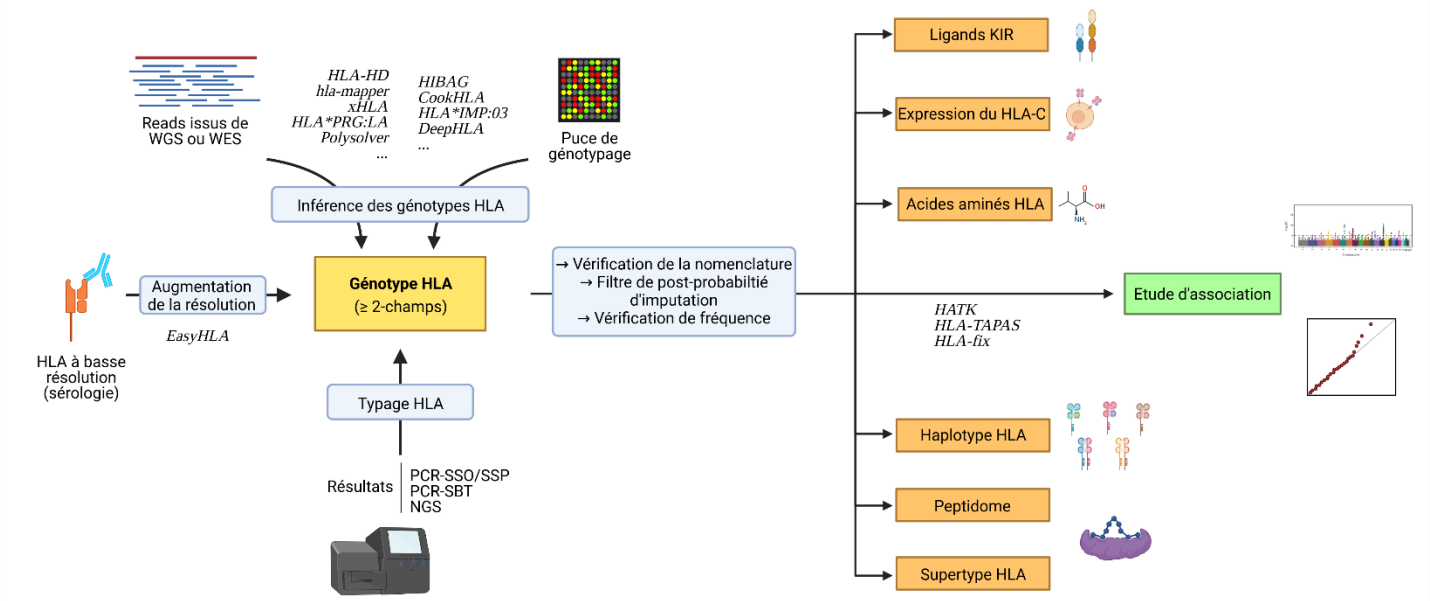


Figure IV-7 Vue d'ensemble de l'analyse HLA. Les données de génotype HLA ont des sources diverses, par séquençage direct, ou inférence statistique à partir de SNP ou d'autres allèles HLA. Il est ensuite possible de valoriser ces données HLA par une étude d'association sur les allèles HLA, mais aussi sur les haplotypes ou les acides aminés. Traduite et adaptée de Douillard et al.

Afin de comprendre l'importance de l'ancestralité dans les erreurs de l'imputation HLA par les SNP, il est intéressant de comprendre comment la structure génétique des populations peut être étudiée et comment ces informations peuvent venir supporter l'analyse génétique et HLA d'individus.

IV.3 - L'exploration de la structure génétique des populations humaines et la répartition de la diversité

La génétique des populations humaines est un domaine scientifique au croisement de l'histoire, de la génétique, de la géographie et de la biologie. Son étude permet aussi bien de suivre l'évolution de multiples espèces (348), de comprendre les mécanismes évolutifs (349), de suivre la migration de populations d'une espèce (350) ou de comprendre leurs différences en médecine (32).

IV.3.1 - Les distances génétiques, de la parenté à l'ancestralité

Afin de faciliter l'interprétation d'événements complexes comme les milliers d'années d'évolution et de migration des populations humaines, il est parfois nécessaire de réduire leur complexité en la résumant dans des scores ou représentations.

IV.3.1.1 - La quantification de la proximité génétique entre des individus ou des populations

La proximité génétique de deux individus est souvent décrite par les principes d'*identity by state* (IBS) et d'*identity by descent* (IBD). Lorsque deux segments d'ADN identiques sont retrouvés entre des individus distincts, ils sont appelés *identical by state*. Cette similitude peut survenir par convergence aléatoire de leurs séquences ADN respectives, ou elle peut être issue d'un ancêtre commun, on parle d'*identity by descent* (351). En effet, les humains reçoivent la moitié d'un génome par chacun de leurs parents et, lors de la méiose, chacun de ces génomes peut se mélanger localement par recombinaison homologue (Figure IV-9, issue de Wikimedia Commons par Gklambauer (352)). Au fil des générations, ces recombinaisons réduisent la longueur des fragments IBD entre deux individus apparentés.

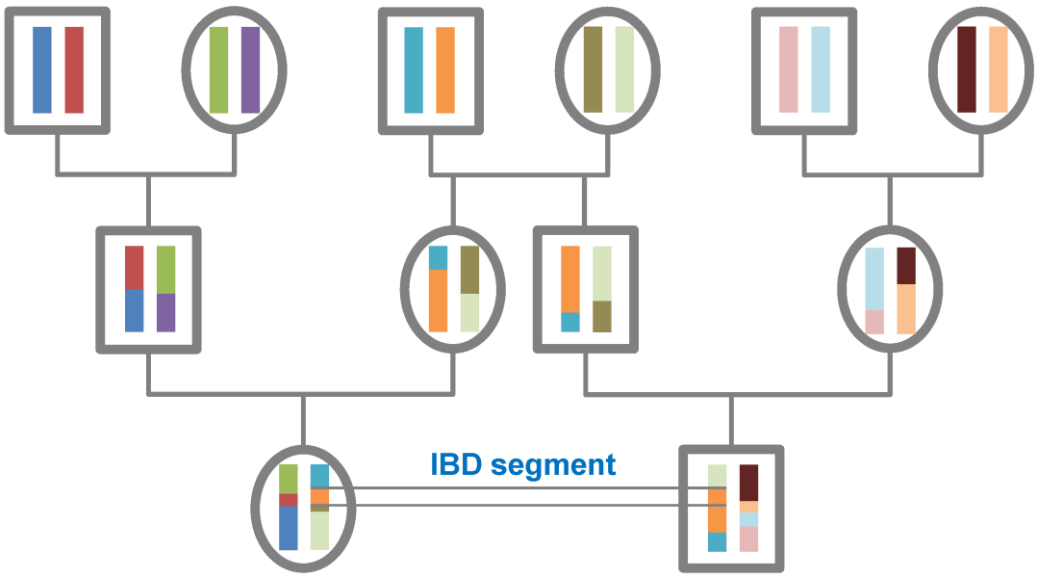


Figure IV-8 Arbre phylogénétique simplifié pour visualiser la transmission de l'ADN sur plusieurs générations. Chaque nouvelle génération hérite d'une version de génome de chaque parent. Cette version peut en réalité contenir l'information de deux chromosomes différents si une recombinaison homologue à lieu. Au fil des générations, des fragments d'ADN en commun sont retrouvés parmi tous les descendants. Issue de Wikimedia Commons par Gklambauer.

Pour prendre du recul et comprendre les différences génétiques en population, l'index de fixation (F_{st}) est utilisé (Équation IV-3 (a)) et en pratique l'estimateur de Hudson (353) est utilisé pour connaître sa valeur (Équation IV-3 (b)). Wright a défini le F_{st} par la proportion de variance génétique due à des différences de fréquences alléliques entre les populations (354,355). Cette valeur est comprise entre 0 et 1, et Cavalli-Sforza l'ont évaluée à un minimum de 0,0021 entre deux populations européennes (danois et anglais) et à un maximum de 0,4573 entre deux populations de la République Démocratique

du Congo et de la Nouvelle Guinée (356), une valeur basse indique alors un partage élevé des allèles entre les populations.

Équation IV-3 Le F_{st} est le ratio de la moyenne des variances d'un allèle dans des populations (σ_S^2) comparé à sa variance dans la population entière (σ_T^2). Celui-ci peut être calculé à partir de l'estimateur de Hudson. Cet estimateur est la proportion de différences de bases nucléotidiques entre deux individus issus de populations différentes (Π_{ext}) par rapport à deux individus issus de la même population (Π_{int}).

$$(a)\ F_{st} = \frac{\sigma_S^2}{\sigma_P^2}$$

$$(b)\ \widehat{F_{st}} = \frac{\Pi_{ext} - \Pi_{int}}{\Pi_{ext}}$$

Le F_{st} donne une estimation moyenne de la proportion de diversité expliquée par les différences entre populations, mais il existe aussi un F_{st} spécifique d'une population : la moyenne de plusieurs F_{st} spécifiques d'une population revient au F_{st} standard (357). Maróstica *et al.* ont ainsi pu démontrer que pour les allèles HLA, 90% de la diversité est détenue à l'intérieur des populations, plutôt qu'entre elles (32).

Ce résultat, partagé dans le reste du génome (30), ne remet pas en cause les différences visibles d'allèles entre des populations mais indique que cette diversité n'est pas la principale. Il est d'ailleurs possible d'identifier chez un individu des correspondances avec plusieurs ADN ancestraux.

IV.3.1.2 - La diversité génétique individuelle à travers le prisme des générations

Le concept d'ancestralité en génétique est un concept qui décrit l'origine des segments IBD pour un individu. Selon les populations ancestrales contribuant à l'ADN d'une personne. Il est ainsi possible de la rapprocher génétiquement, sans prendre en compte sa géographie ou son identité socio-culturelle.

Ainsi, des logiciels tels que STRUCTURE (358,359), ADMIXTURE (360) ou RFMix (361) arrivent à estimer l'origine ancestrale des différentes parts du génome d'un individu. Un nombre de populations ancestrales est choisi et un modèle mathématique estime comment chaque génome individuel peut suivre ce découpage (Figure IV-8, issue de Naslavsky *et al.* (362)).

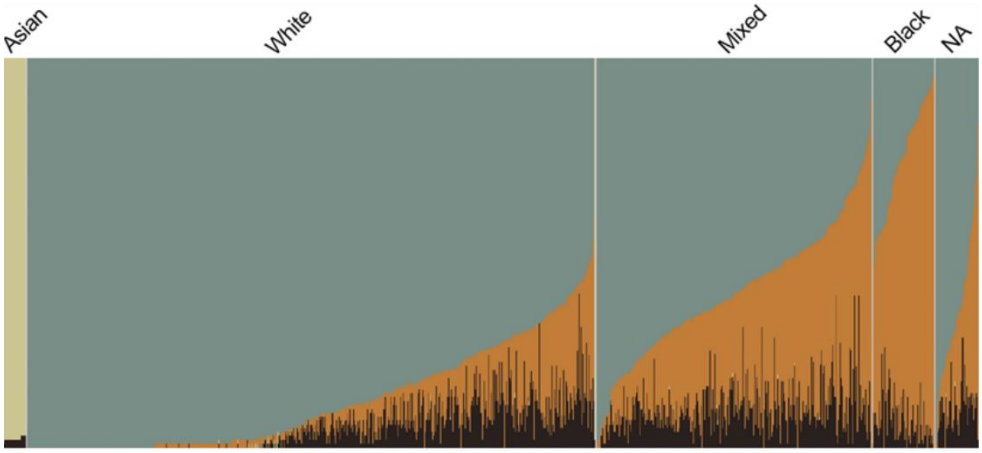


Figure IV-9 Découpage des génomes de SABE (n=1 171) en plusieurs ancestralités génétiques par Admixture. Chaque ligne verticale est un individu et la proportion de chaque ancestralité est indiquée par une couleur : européenne (gris), africaine (orange), native d'Amérique (noir) et est-asiatique (vert). Le jeu de données représenté est composé d'individus de São Paulo au Brésil répartis en groupes ethniques auto-identifiés. Ces individus ont une ancestralité composite variables allant d'une proportion égale de génome européen, africains et autochtone des Amériques, à une homogénéité européenne ou asiatique. Cette dernière correspond à une communauté japonaise connue de São Paulo. Issue de Naslavsky et al.

Ces proportions définissent une structure de population qui est notamment essentielle dans les études d'association, car cela peut entraîner des biais, notamment lorsque les individus cas et contrôle ont des ancestralités différentes. Cependant, le temps de calcul de ces méthodes les rend peu pratique et les analyses de réduction de dimensions leur sont souvent préférées (363).

IV.3.2 - Les méthodes de réduction de dimensions et leurs utilisations en génomique

Lors de l'analyse de jeux de données de grandes dimensions (*i.e.* avec un grand nombre de variables), il est difficile de représenter l'ensemble de l'information car nous ne pouvons visualiser et représenter qu'un nombre limité de variables : par les dimensions spatiales, la couleur, la forme, etc. La réduction de dimensions est une approche mathématique qui permet de résumer la variabilité de l'entièreté d'un jeu de données, généralement dans le but de la visualiser ou de tirer des conclusions descriptives sur une structure sous-jacente.

Dans le cadre de la génomique, où il existe des millions de polymorphismes possibles entre deux individus humains, la réduction de dimensions offre la possibilité de modéliser de nombreux effets qui peuvent rapprocher ou éloigner les individus d'un point de vue génétique. Deux méthodes principales sont utilisées en génétique des populations : l'analyse en composante principale (ACP, ou *principal component analysis, PCA*) et la projection et approximation de variétés uniformes (*Uniform Manifold Approximation and Projection*, *UMAP*).

IV.3.2.1 - La construction d'un nouvel espace avec l'ACP

L'ACP est utilisée depuis plusieurs décennies pour modéliser la structure génétique d'une population mais ce n'est pas sa seule propriété. La réduction de dimensions par ACP reflète également la parenté, les motifs de déséquilibre de liaison ou même les effets de lot (363).

Brièvement, l'ACP repose sur une matrice de corrélation. À partir de cette matrice, l'ACP construit de nouveaux axes (les composantes principales), qui sont simplement des combinaisons linéaires de génotypes SNP pondérés qui maximisent la variance. Chaque nouvel axe est orthogonal aux autres et contient donc une information différente, basée sur une nouvelle combinaison de variables. Ainsi, en représentant les individus dans cette dimension particulière, les individus avec des SNP corrélés sont regroupés alors que si les SNP diffèrent ils se retrouvent éloignés.

En répétant ce processus avec un nombre restreint de dimensions il est alors possible de voir des groupes émerger (Figure IV-10, données issues de Byrska-Bishop *et al.* (29) et CAAPA (364)), notamment selon l'ancestralité (365).

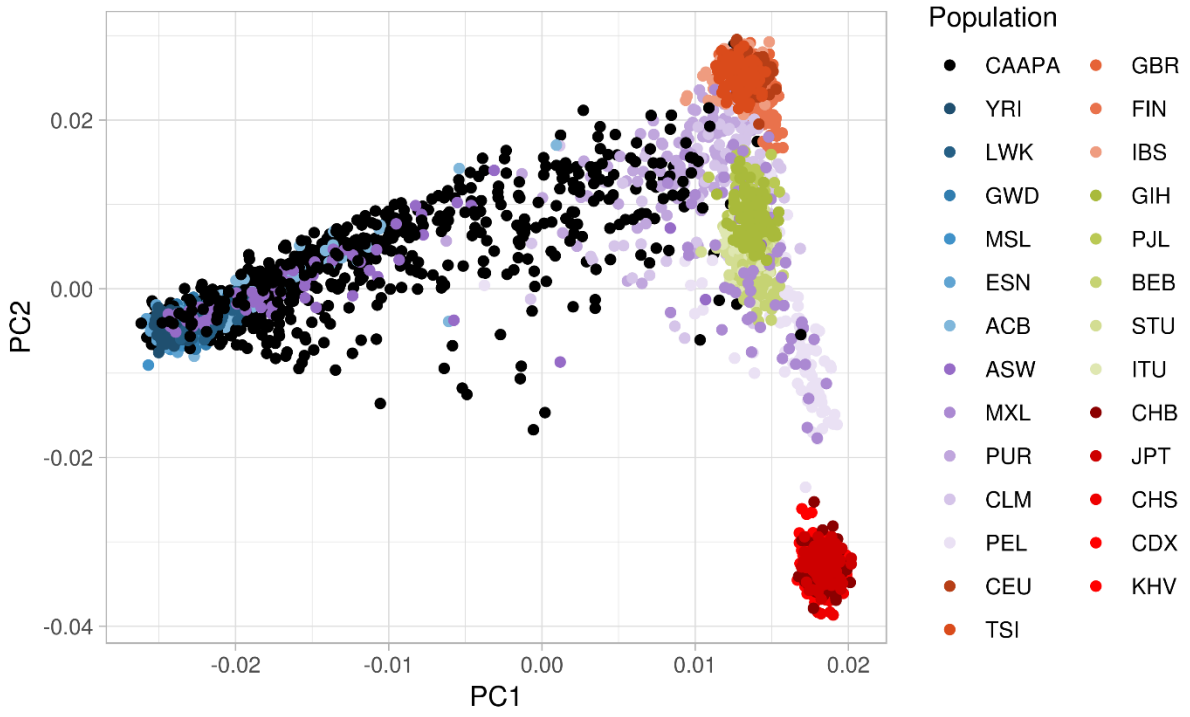


Figure IV-10 ACP de génotypes SNP des individus du 1KGP (n=2 504) et de CAAPA (n=880), une cohorte d'individus d'ancestralités européenne et africaine. Les individus de CAAPA sont représentés en noir. Pour 1KGP, les européens (orange), les américains (violet), les africains (bleu), les sud-asiatiques (vert) et les est-asiatique (rouge) sont en différentes nuances selon la sous-population. Les populations avec le plus de différence sur les SNP des deux premières composantes sont répartis sur la PCA dans des groupes distincts, les individus d'ancestralité composite sont retrouvés à différentes positions entre ces groupes (CAAPA et américains). Données issues de Byrska-Bishop et al. & CAAPA.

L'utilisation principale de l'ACP en génomique est la correction de biais dans les GWAS. En effet, l'association étudiée entre un génotype et un trait peut être différente selon les groupes génétiques, soit : si le risque génétique ou la fréquence de certains variants, sont liés à l'ancestralité, ou si la répartition de ces sous-groupes entre cas et contrôle diffère (366). En ajoutant les composantes principales au modèle de régression, ces facteurs sont ainsi pris en compte.

Cependant, les axes sont limités par leur représentation linéaire des données. De plus, l'utilisation l'ACP pour distinguer les différents sous-groupes génétiques peut cacher des structures génétiques plus fines (367).

IV.3.2.2 - L'UMAP et la topologie des données comme point de repère

L'UMAP est une technique de réduction de dimensions qui a connu un essor important avec la technologie des analyses *single-cell* (368). Une des applications des données *single-cell* est d'étudier le transcriptome de milliers de cellules d'une personne, afin d'identifier des populations cellulaires. Cette tâche nécessite alors d'incorporer l'expression de milliers de gènes à la fois. L'UMAP est par ailleurs une méthode concurrente, plus performante, de la t-SNE (*t-distributed Stochastic Neighbor Embedding*) bien que leurs performances respectives semblent débattues (369).

Récemment, l'UMAP est aussi devenue une nouvelle alternative à l'ACP dans les études de génétique des populations (370,371). Sakaue *et al.* (370) ont ainsi pu mettre en évidence les différences de structure génétique entre les populations japonaises vivant sur les archipels par rapport à celles de l'île principale (Figure IV-11, issue de Sakaue *et al.* (370)).

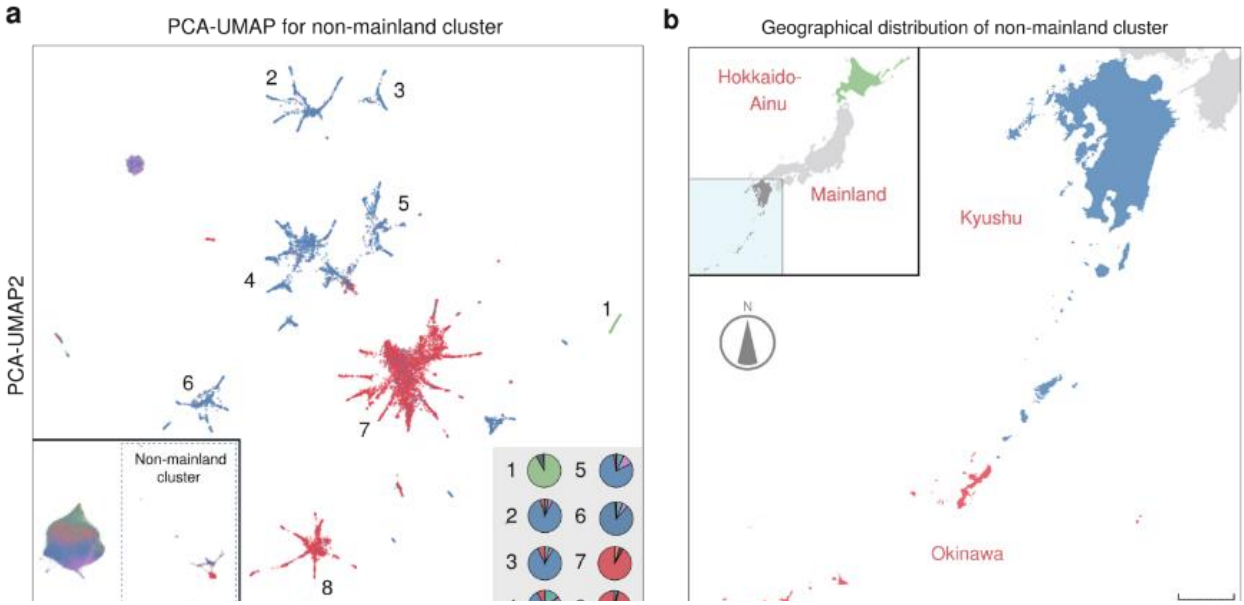


Figure IV-11 La PCA-UMAP de plusieurs populations du Japon (n=169 719), sur l'île principale (Honshu) et ses autres îles (Hokkaido, Kyushu et Okinawa). Cette représentation distingue clairement les individus de Honshu des autres îles, mais conserve également les structures génétiques internes aux différentes populations. Issue de Sakaue et al.

Contrairement à l'ACP, l'UMAP est capable de représenter tout aussi bien des structures génétiques fines que plus importantes. Brièvement, l'UMAP repose sur la notion de topologie et de voisinage des points (372). Appliqué à des données de génotypage, chaque individu est représenté dans l'espace de plus grande dimension (*i.e.* selon chacun de ses génotypes SNP), et un nombre de ses voisins est identifié. Les algorithmes de l'UMAP cherchent ensuite à définir un nouvel espace de dimension moindre tout en conservant les voisins de chaque individu.

L'immense variabilité des données génomiques est une grande source de développements algorithmiques dont les buts sont multiples : améliorer les interprétations biologiques, compléter des données manquantes pour améliorer la puissance statistique des études ou encore visualiser et investiguer les relations génétiques entre des individus. Le HLA est ainsi un exemple extrême de cette variabilité qui nécessite une attention toute particulière. Malgré les évolutions technologiques et méthodologiques qui l'entourent, le HLA se trouve être encore à l'heure actuelle un obstacle important qui nécessite une maîtrise transversale d'outils venant de domaines très divers.

V - Problématique et objectifs de la thèse

L'étude de la génomique s'est transformée lors des deux dernières décennies. Elle œuvre désormais tout aussi bien à l'identification de mutations dans les maladies monogéniques, qu'à celle de milliers de polymorphismes dans des traits complexes. Ces polymorphismes peuvent être associés à un risque plus faible ou plus élevé dans la mécanique physiopathologique de maladies complexes ou simplement influencer des traits phénotypiques divers. Dans son chemin, la génomique a rencontré l'obstacle du CMH avec plus de 200 gènes compactés dans une petite portion du génome, des milliers d'associations génomiques ainsi qu'une diversité exceptionnelle. Cette diversité tient majoritairement dans ses gènes du HLA avec plus de 30 000 allèles découverts à ce jour. Le système HLA tient un rôle central dans la mise en place de l'immunité adaptative par sa capacité de présentation des antigènes du soi et du non-soi.

Si cette importance est visible dans les GWAS, l'identification des variants potentiellement causaux est un problème majeur. En effet, le déséquilibre de liaison dans la région du CMH empêche l'attribution d'une association à un gène en particulier. Pour contourner cela, des études d'association HLA sont menées à bien. Bien qu'elles permettent d'identifier plus précisément l'implication d'allèles HLA dans l'association, leur pouvoir statistique est souvent limité par la taille de l'échantillon et la diversité des populations observées.

Les techniques d'inférence statistique d'allèles HLA à partir de données SNP permettent de pallier ces limites, mais elles sont soumises à plusieurs contraintes. L'accès aux outils bioinformatiques d'imputation HLA n'est pas forcément possible pour toutes les infrastructures informatiques, et une expertise est nécessaire pour les chercheurs. L'imputation HLA est également limitée par des problèmes de diversité dans les données. Tout comme les données génétiques SNP, les données HLA sont disponibles majoritairement pour des populations européennes. Des panels de référence multi-ethniques, ou alors spécifiques d'une population, existent et améliorent la précision d'imputation HLA. Cependant, de nombreuses populations, notamment celles où les individus ont plusieurs ancestralités génétiques, comme les afro-américains ou les brésiliens, restent sous-représentées. Ce problème de diversité est à la fois organisationnel, car elles nécessitent des efforts de séquençage, et technique, car l'imputation de plusieurs populations à la fois reste un défi.

Mon projet de thèse s'inscrit ainsi dans le SNP-HLA Reference Consortium (SHLARC), un projet mondial autour de l'imputation HLA avec trois objectifs principaux : l'acquisition de nouvelles données, l'amélioration de l'imputation HLA, ainsi que le partage des infrastructures et expertise autour de l'imputation et de l'analyse HLA en général.

Le volet d'amélioration de l'imputation HLA est l'objectif majeur de ma thèse. Si à long terme, la récolte de nouvelles données est essentielle pour augmenter la précision d'imputation, je me suis concentré sur des changements de méthodologies sur les données actuelles pour gagner en précision. J'ai cherché à tirer profit au maximum de la diversité déjà contenue dans les données disponibles. En se basant sur des algorithmes de réduction de dimension, j'ai utilisé la structure génétique des populations pour créer des panels de référence spécifiques des populations génétiques, plutôt que sur des panels choisis par la proximité géographique. Je me suis concentré sur les populations d'ancestralité mixtes et j'ai éprouvé nos panels de référence spécifiques face à des panels multi-ethniques ou venant d'une population géographique semblable. De plus, les allèles rares HLA sont souvent moins bien imputés par les logiciels, j'ai donc adapté la façon d'évaluer la précision de l'imputation HLA afin de se concentrer sur ces allèles. D'un point de vue purement technique, cette thèse repose sur les avancées technologiques en termes de calcul pour organiser l'imputation HLA autour d'une infrastructure de calcul solide pour faciliter et augmenter la vitesse de création de panels de référence.

Mes travaux de thèse ont permis de nourrir le SHLARC, en mettant en évidence l'importance de l'imputation HLA et en communiquant dans les congrès d'immunogénétique sur son impact dans les études HLA. Cela nous a permis de nouer des partenariats pour créer des jeux de données SNP-HLA avec de nouvelles populations comme la biobanque FinnGen ou le consortium Severe COVID-19. Ceux-ci permettront *in fine* de créer des panels de référence plus divers et contenant de nouveaux allèles HLA, pour augmenter la précision de l'imputation HLA.

Dans la même lancée, l'expertise que j'ai développée en imputation HLA m'a permis de participer aux études d'association HLA menées par des génomiciens et des épidémiologistes génétiques, à aider à l'interprétation des résultats HLA et aussi à me munir d'autres outils d'immunogénétique pour explorer en profondeur la relation entre la région du CMH et des pathologies.

Mes travaux de thèse sont dans l'intérêt de la communauté scientifique, et le SHLARC dans lequel ils s'inscrivent veut également partager ces avancées de l'imputation HLA à travers un site Web. Celui-ci permettra aussi bien de mettre en commun des données SNP et HLA, dans le but de créer des panels de référence adaptés à toutes les ancestralités génétiques, que de globalement faciliter l'accès à l'imputation HLA.

Résultats

“Life is like topography, Hobbes. There are summits of happiness and success, flat stretches of boring routine and valleys of frustration and failure.”

Calvin, Calvin & Hobbes, Bill Waterson

VI - Naviguer les eaux troubles de l'imputation HLA avec le SNP-HLA Reference Consortium (SHLARC)

VI.1 - Un projet international pour démocratiser les études d'association HLA

Le SHLARC est un consortium international d'immunogénétique qui réunit plus d'une vingtaine de laboratoires répartis sur quatre continents et dont l'objectif est de capturer l'élan des études d'association en génome entier existantes pour les étendre aux études d'association HLA. Ces dernières permettront de comprendre en détail le lien réel entre la région du CMH et les pathologies associées. Afin d'atteindre ce but, le consortium s'appuie sur des modèles d'imputation HLA à partir de données SNP pour tirer profit des données GWAS déjà disponibles. Ces études sont connues pour être majoritairement européennes, or la fréquence des allèles HLA peut varier selon les populations étudiées : il y a un intérêt tout particulier à favoriser la récolte de données diverses à travers le monde, notamment pour améliorer la précision de l'imputation HLA.

VI.1.1 - La création d'un environnement propice à l'imputation HLA

Les trois objectifs du SHLARC sont la récolte de données SNP-HLA diverses, l'amélioration de l'imputation HLA et la démocratisation des études d'association HLA. Mes travaux ont pu, dans un premier temps, démontrer l'intérêt de créer des modèles d'imputation avec des individus proches génétiquement de la population à prédire. Puis, ils ont pu mettre en avant l'importance des données de grande dimension pour la création de modèles, et enfin permettre le développement d'un environnement de calcul pour l'imputation HLA.

VI.1.1.1 - CAAPA, un consortium pour la diversité génétique africaine

Le consortium sur l'asthme dans les populations d'ancestralité africaine en Amérique (*Consortium on Asthma among African-ancestry Populations in the Americas, CAAPA*) a été créé en 2012 afin de cataloguer la diversité génétique des populations d'origine africaine et en particulier la diaspora africaine en Amérique, issue de l'esclavage. Son second rôle a été la découverte de gènes de prédisposition à l'asthme chez les personnes d'ancestralité africaine (373).

CAAPA est composé de seize populations d'origine géographique similaire, l'Afrique de l'ouest, mais dont les proportions d'ancestralité varient (Figure VI-1, issue de Mathias *et al.* (374)). Les individus de CAAPA disposent ainsi d'une ancestralité génétique composite : à la fois africaine, européenne et autochtone des Amériques ; à des proportions différentes. Ainsi, le jeu de données est constitué de 880 individus non-apparentés qui sont couramment identifiés comme afro-américains (299,374).

L'accès à ces données a été essentiel pour évaluer comment des modèles d'imputation HLA déjà disponibles pouvaient être inférieurs en termes de performance, malgré une identification ethnique similaire. Nous avons pu utiliser CAAPA pour créer des modèles d'imputation grâce à une collaboration préalable de notre équipe pour l'association du HLA avec la sévérité de l'asthme (299). Ces groupes ethniques sont une réduction beaucoup trop importante de la diversité génétique réelle des populations. Ainsi, la prédiction des génotypes HLA d'une partie de CAAPA par un modèle d'imputation

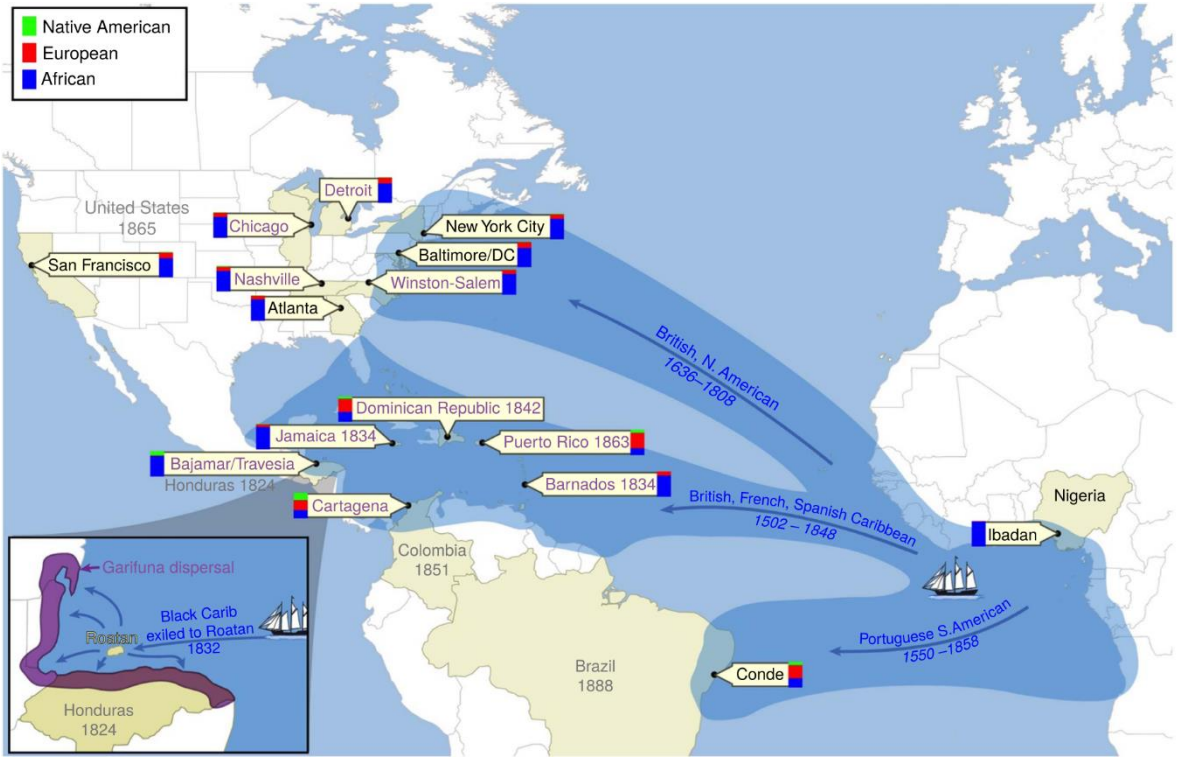


Figure VI-1 Seize populations composent le consortium CAAPA et chacune a une estimation de son ancestralité selon la proportion de génome africain, européen et autochtone des Amériques qu'elle contient. Leur ancestralité composite provient de la traite d'esclaves transatlantique par les empires coloniaux européens qui prend origine en Afrique de l'ouest au XVI^{ème} siècle et se termine à différentes dates indiquées en regard des pays, au XIX^{ème} siècle. Issue de Mathias et al.

composé d'un de ses sous-échantillons a donné de meilleurs résultats qu'un modèle de référence de HIBAG, composé d'autres personnes afro-descendantes.

VI.1.1.2 - La création de modèles d'imputation HLA avec HIBAG

L'imputation HLA dispose de nombreux algorithmes différents, pour la plupart travaillant avec des génotypes SNP et HLA non-phasés. Néanmoins, le logiciel HIBAG de Zheng *et al.* a été le premier à autoriser les chercheurs à créer leurs propres modèles d'imputation HLA (320). Cela a été un avantage essentiel dans la compréhension de l'interaction entre ancestralité génétique et précision de l'imputation HLA.

Afin de comprendre comment la diversité génétique impacte l'imputation, il est intéressant de comprendre la méthode d'imputation HLA. HIBAG est un algorithme d'apprentissage ensembliste par *attribute bagging*. Brièvement, il repose sur la création de plusieurs modèles statistiques différents (*i.e.* avec des sous-échantillons des individus et des variables) qui décrivent un lien entre un génotype

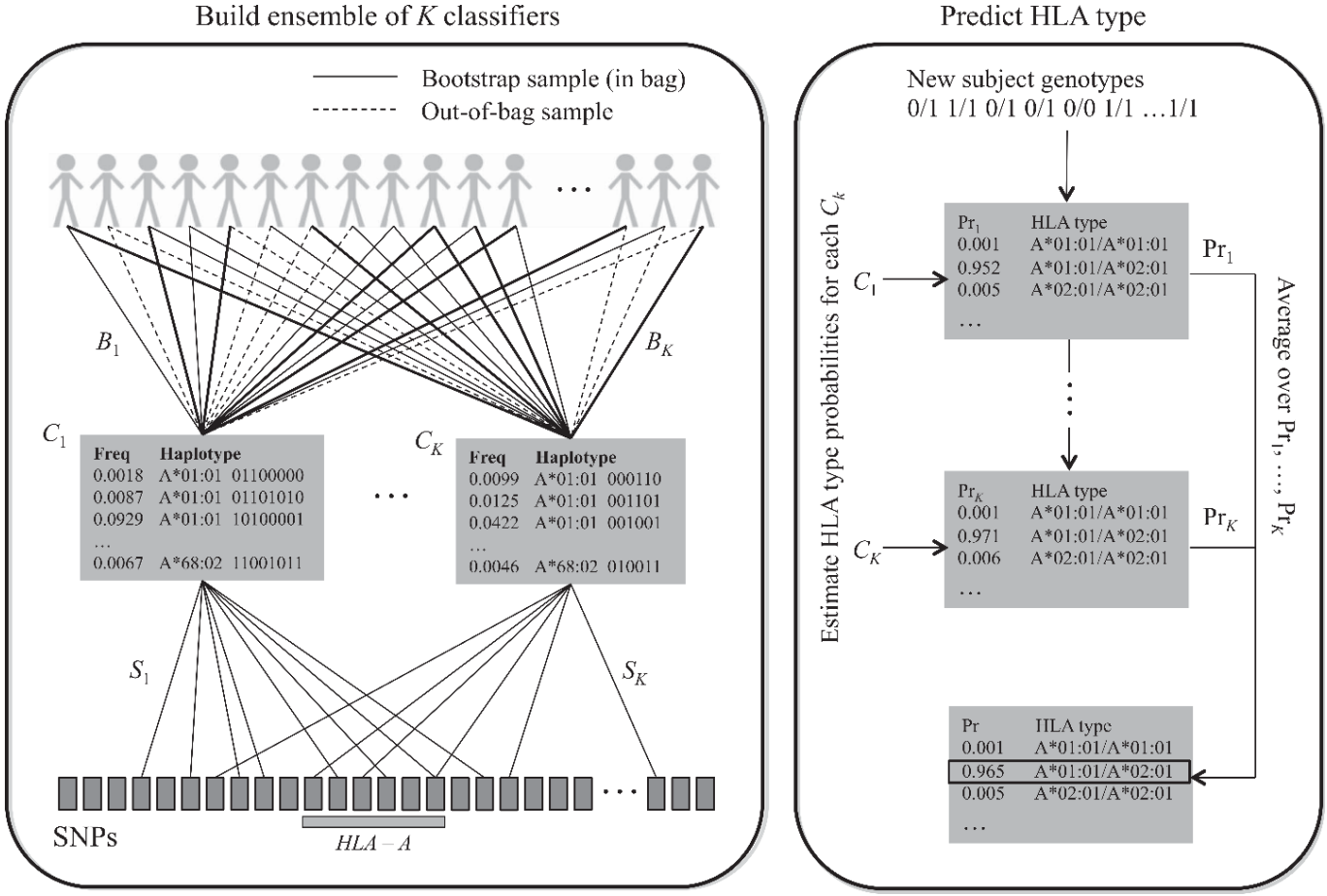


Figure VI-2 Algorithme de création d'un modèle d'imputation HLA par attribute bagging et prédiction à partir de celui-ci. Issue de Zheng et al.

HLA et plusieurs des génotypes SNP. Tous ces modèles peuvent, à partir d'un génotype SNP, déduire un génotype HLA probable et leurs résultats sont fusionnés pour obtenir une réponse consensus (Figure VI-2, Zheng *et al.* (320)).

Dans le cas d'HIBAG, la construction d'un modèle statistique commence par le choix avec remise d'autant d'individus qu'il y a dans le jeu d'entraînement. En moyenne, 37% des individus sont alors écartés à cette étape et sont appelés *out-of-bag* (375). Un sous-ensemble aléatoire de SNP est d'abord sélectionné, puis le pouvoir prédictif de chacun est évalué un par un. Pour cela, un algorithme Expectation-Maximization (EM) calcule la probabilité de l'haplotype SNP d'un individu avec un allèle HLA et ces informations sont utilisées pour prédire le HLA des individus *out-of-bag*. Le SNP le plus

prédictif pour les génotypes HLA est retenu, puis le processus d'ajout de SNP continue tant que la précision d'imputation augmente.

Ce processus est effectué pour chaque modèle différent. Pour la prédiction du génotype HLA d'un nouvel individu, le modèle nécessite un génotype SNP. Chaque modèle contient des haplotypes SNP qui sont combinés pour pouvoir ré-obtenir le génotype SNP de l'individu. Chaque haplotype SNP est associé à un allèle HLA avec une probabilité, en multipliant les probabilités des deux haplotypes on obtient celle du génotype HLA (Équation VI-1).

Équation VI-1 Calcul du génotype HLA $\langle y^{(1)}, y^{(2)} \rangle$ le plus probable pour un individu d'après les K modèles d'imputation créés, sachant le génotype SNP $g^{new}{}_j$ d'un individu, pour tous les SNP présents dans le modèle S_k

$$\underset{\langle y^{(1)}, y^{(2)} \rangle}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} Pr_k\left(\langle y^{(1)}, y^{(2)} \rangle \,\middle|\, g_j^{new}, \ldots \forall j \in S_k\right)$$

Cette création de modèles d'imputation HLA est extrêmement dépendante du nombre d'individus et de SNP présents dans le jeu de données d'entraînement, ainsi que de la diversité génétique.

VI.1.1.3 - Le façonnement d'un nouvel environnement pour faire prospérer l'imputation HLA

D'un point de vue technique, le problème majeur de l'imputation HLA n'est pas la prédiction. Celle-ci est généralement facile à réaliser sur un ordinateur personnel, tant pour la mise en place que le temps de calcul requis. En revanche, la création des modèles statistiques nécessite beaucoup de calculs informatiques, il est donc compliqué de créer des modèles sans infrastructure de calcul dédiée. Nous avons pu profiter des serveurs nantais de BiRD (Institut de Recherche en Santé), du Liger (École Centrale de Nantes) et enfin du CCIPL (Nantes Université).

L'algorithme HIBAG peut tirer parti de ces serveurs de calculs en utilisant les processeurs graphiques (*Graphics Processing Units, GPU*) plutôt que des processeurs classiques (*Central Processing Units, CPU*). L'architecture de ces processeurs les rend plus efficients pour les nombreux calculs parallèles effectués dans les algorithmes d'apprentissage statistique.

Ces travaux sont un effort international pour fournir une solution aux analyses d'association HLA, tout en mettant en lumière l'importance de la diversité génétique pour combler les lacunes qu'elles peuvent avoir actuellement. Dans ce but, nous avons étudié l'imputation HLA de CAAPA avec différents panels de référence construits sur les supercalculateurs du CCIPL. Nous avons ainsi pu montrer dans cet article qu'une infrastructure de calculs adaptée permet de créer rapidement des panels de références. La précision d'imputation de ces modèles est très dépendante du nombre d'individus et

de SNP en commun. L'augmentation de 10 à 100 individus et de 500 à 20 000 SNP dans le modèle a amélioré la précision d'imputation par 2 et 1,5, respectivement. Enfin, nous avons démontré que les panels de références basés sur des groupes ethniques larges (ici afro-américains) sont moins précis que ceux avec une plus grande proximité génétique.

VI.1.2 - Article - SNP-HLA Reference Consortium (SHLARC) : le partage de données HLA et SNP dans le but de promouvoir les analyses génomiques centrées sur la région du CMH

Erratum

- À la page 3 de l'article, la phrase « […] we created 40 different reference panels with either increasing numbers of individuals (10/20/500/1,000) […]» doit être corrigée par « we created 40 different reference panels with either increasing numbers of individuals (10/20/**50/100**) »

- La légende de l'abscisse de la figure 1.(a) « Number of individuals (log10 scale) » doit être remplacée par la légende « Number of individuals ».

Received: 6 January 2020 | Revised: 19 June 2020 | Accepted: 3 July 2020

DOI: 10.1002/gepi.22334

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics

Nicolas Vince[1] | **Venceslas Douillard**[1] | **Estelle Geffard**[1] | **Diogo Meyer**[2] | **Erick C. Castelli**[3] | **Steven J. Mack**[4] | **Sophie Limou**[1,5] | **Pierre-Antoine Gourraud**[1]

[1]Centre de Recherche en Transplantation et Immunologie, ITUN, UMR 1064, Université de Nantes, CHU Nantes, Inserm, Nantes, France

[2]University of São Paulo, São Paulo, Brazil

[3]UNESP—Universidade Estadual Paulista, Botucatu, São Paulo, Brazil

[4]Department of Pediatrics, University of California, San Francisco, UCSF Benioff Children's Hospital Oakland, Oakland, California

[5]Ecole Centrale de Nantes, Nantes, France

Correspondence
Nicolas Vince, CRTI UMR1064—ITUN, CHU Nantes Hôtel Dieu, 30 bld Jean Monnet, 44093 Nantes Cedex 01, France.
Email: nicolas.vince@univ-nantes.fr

Funding information
H2020 Marie Sklodowska-Curie Actions, Grant/Award Number: 846520

Abstract

Genome-wide associations studies have repeatedly identified the major histocompatibility complex genomic region (6p21.3) as key in immune pathologies. Researchers have also aimed to extend the biological interpretation of associations by focusing directly on human leukocyte antigen (*HLA*) polymorphisms and their combination as haplotypes. To circumvent the effort and high costs of HLA typing, statistical solutions have been developed to infer *HLA* alleles from single-nucleotide polymorphism (SNP) genotyping data. Though *HLA* imputation methods have been developed, no unified effort has yet been undertaken to share large and diverse imputation models, or to improve methods. By training the HIBAG software on SNP + HLA data generated by the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) to create reference panels, we highlighted the importance of (a) the number of individuals in reference panels, with a twofold increase in accuracy (from 10 to 100 individuals) and (b) the number of SNPs, with a 1.5-fold increase in accuracy (from 500 to 24,504 SNPs). Results showed improved accuracy with CAAPA compared to the African American models available in HIBAG, highlighting the need for precise population-matching. The SNP-HLA Reference Consortium is an international endeavor to gather data, enhance *HLA* imputation and broaden access to highly accurate imputation models for the immunogenomics community.

KEYWORDS

consortium, HLA, imputation, SNP

Nicolas Vince and Venceslas Douillard contributed equally to this work.

1 | INTRODUCTION

Beginning with the discovery of the HLA system in the 1950s, the characterization of *HLA* polymorphism and *HLA* disease associations have been performed in parallel (Dausset, 1999; Trowsdale & Knight, 2013). In the genome-wide association study (GWAS) era, the focus was shifted on single-nucleotide polymorphisms (SNP) with little to no biological relevance. Even when located in the major histocompatibility complex (MHC) region (6p21.3), these SNP associations have largely supplanted the traditional study of *HLA* allele associations. GWASs have however confirmed the crucial role of the *HLA* loci for the genetic epidemiology of nearly a quarter of all diseases and traits (MacArthur et al., 2017; Trowsdale & Knight, 2013), but SNP associations do not convey the immune-biological relevance that specific *HLA* alleles have. For example, GWASs of HIV disease identified the rs2395029 SNP near the *HCP5* gene on chromosome 6 as being the strongest associated with viral control (Fellay et al., 2007; Limou & Zagury, 2013). This SNP, which is located 100 kb from *HLA-B,* is in nearly complete linkage disequilibrium with the *HLA-B*57:01*, which can present HIV peptides crucial for HIV detection by the immune system (Chen et al., 2012; Limou & Zagury, 2013). Using novel bioinformatic approaches, we now have the ability to statistically infer *HLA* alleles from genotypic SNP data (imputation), returning HLA molecular functions to the forefront of disease-associated research (Meyer & Nunes, 2017; Pappas et al., 2018). Imputations are statistical methods that infer or predict missing information based on haplotypes. Haplotypes are a combination of genetic variants on one chromosome, they can be SNP haplotype (e.g., 011010, referring as the presence or absence of SNPs), gene haplotype (e.g., *HLA-A*01:01~HLA-B*08:01~HLA-C*07:01~HLA-DRB1*03:01~HLA-DQB1*02:01*) or a combination of different genetic variants (SNP, indels, substitution) haplotype (e.g., *HLA* alleles). In genomics, SNP imputation can infer the identity of missing SNPs that were not genotyped on GWAS arrays (Delaneau, Zagury, & Marchini, 2013; McCarthy et al., 2016) by comparing whole-genome SNP genotypes to a large reference panel of SNP haplotypes (Delaneau et al., 2013). Filling the genotyping gaps, SNP imputation performance and accuracy increased significantly when new large reference haplotype panels became available (McCarthy et al., 2016), which has contributed to a large number of discoveries over the past decade (Visscher et al., 2017).

In parallel to SNP, imputation also applies to *HLA* polymorphisms themselves, alone or in combination. It has revealed key associations in numerous diseases (Fellay et al., 2007; Limou & Zagury, 2013; MacArthur et al., 2017; Trowsdale & Knight, 2013; Vince et al., 2020) and can, as such, lead to the development of new drugs or patient-care guidance. Efforts to impute *HLA* alleles from these GWAS should be pursued to empower the community to go beyond simple SNP associations and to discover new disease associations (Khor et al., 2015; Meyer & Nunes, 2017; Shen et al., 2018); as an example, *HLA* alleles can bring new functional immunogenomics data such as prediction of amino acid, haplotypes (five genes: *A~B~C~DRB1~DQB1*) or imputed HLA-C expression easily implemented with Easy-HLA (Geffard et al., 2019; Vince et al., 2016). *HLA* allele imputation appears as a time and cost-effective alternative to the laborious *HLA* typing of all GWAS subjects. However, to rely on *HLA* imputation we must consider its accuracy, which depends on the reference panel quality (e.g., matching ancestry background, matching SNPs composition; Khor et al., 2015) and size (e.g., number of individuals with both SNP as well as *HLA* typing data, referred as SNP + HLA data; Pappas et al., 2018; Zheng et al., 2014). Successful *HLA* imputation, therefore, depends on the availability of large and diverse reference panels, which warrants a major collective effort in organizing community resources. Here, we advocate for the development of the SNP-HLA Reference Consortium (SHLARC), a new international network focused on collecting a large collection of high-quality HLA and SNP data, especially from an ethnically diverse population, with the goal to develop and share large reference panels and help worldwide researchers exploring *HLA* allelic information from their cohorts.

2 | RESULTS

We had access to the CAAPA (Consortium on Asthma among African-ancestry Populations in the Americas) data set (Daya et al., 2019; Vince et al., 2020) that consists of 880 whole-genome sequenced African American subjects with associated SNP GWAS data and typed *HLA* alleles at a two-field resolution (corresponding to the protein level). We chose the *HLA* Genotype Imputation with Attribute Bagging (HIBAG) R package (Zheng et al., 2014) to test the impact of the number of subjects and SNPs on *HLA* imputation accuracy. HIBAG demonstrates improved imputation accuracy over other available methods (Pappas et al., 2018) and allows the creation of custom reference panels, using the machine-learning technique of attribute bagging. Building reference panels requires heavy computing power which is related to the number of subjects and number of SNPs in an almost linear correlation (Zheng et al., 2014). The development of machine-learning algorithms heavily

relies on the evolution of computational power. We used graphics processing units (GPUs) as they are architecturally better suited to handle the computationally intensive tasks. For this project, we took advantage of the upgraded HIBAG version (HIBAG v1.15.3, HIBAG.gpu v0.9.1; Zheng, 2018) and used GPUs to build and compare multiple reference panels with a fivefold reduction in computation time relative to central processing units).

Starting with the complete data set ($n = 880$ individuals), we simulated scenarios of reference panel building by creating a collection of training and test sets. Each of the condition was replicated 10 times to assess the variability in the frequency of SNPs and HLA types and display confidence intervals for each prediction: (a) from a set of 100 samples ($n_{\text{training}} = 100$), we created 40 different reference panels with either increasing numbers of individuals (10/20/500/1,000) or increasing numbers of SNPs (500/1,000/5,000/10,000/24,504; see Supporting Information Methods) and (b) a test set ($n_{\text{test}} = 780$) used to assess the accuracy of *HLA* imputation from the 40 different reference panels (5 *HLA* genes × [4 different number of individuals + 4 different number of SNPs]; Figure 1). Accuracy is defined by the percentage of correct HLA allele prediction.

We observed that increasing the number of individuals in the reference panel increased *HLA* imputation accuracy (two-field resolution) for all loci (Figure 1a). As an example, accuracy rose from 60% with 10 individuals to 93% with 100 individuals for *HLA-DQB1*, and from 27% with 10 individuals to 71% with 100 individuals for *HLA-B* on average. We then compared the *HLA* imputation accuracies obtained from our CAAPA-based test set with pre-existing reference panels available on the HIBAG website (http://www.biostat.washington.edu/~bsweir/HIBAG/). These precomputed reference panels were all created with more than 100 individuals of African American ancestry (from 137 for *HLA-DQB1* to 171 for *HLA-B*) from the HLARES data and the HapMap Yoruba population. The accuracies using the precomputed HIBAG reference panels (represented as horizontal lines in Figure 1a) ranged from 70% (*HLA-DRB1*) to 87% (*HLA-A*) and were lower than those obtained using the CAAPA-based reference panels using a smaller number of individuals. This illustrates the importance of close matching of ancestry between the reference panel and the genotyped subjects, even within a single ancestry group (here African ancestry).

In addition, we reduced the number of SNPs in the training data set (500, 1,000, 5,000 and 10,000 out of the 24,504 available chromosome-6 SNPs) and observed that increasing the number of SNPs in the reference panel increased the *HLA* imputation accuracy for all genes (Figure 1b). For example, accuracy rose from 86% with 500 SNPs to 91% with the full set of 24,504 SNPs for *HLA-A*, and from 65% with 500 SNPs to 77% accuracy with the full set of SNPs for *HLA-B*. The number of SNPs in the training data set differs from the number of SNPs in the statistical model (or bag) as HIBAG does not use all SNPs provided in the input to create the reference panels (see Tables S1.1 and 1.2 for exact numbers). Indeed, HIBAG only includes SNPs within a 500-kb window around the gene of interest, and only keeps those improving the model after random selection (see Supporting Information Methods). For in-depth analysis of *HLA* imputation, we have also plotted the sensitivity and frequency of each allele to predict in the validation data set, to identify alleles decreasing the overall accuracy (see Figures S1–S5 and Table S2).

3 | DISCUSSION

Our results illustrate the importance of matching large reference panels with high SNP coverage to the input data set for efficient and accurate *HLA* allele imputation (Dilthey et al., 2016; Jia et al., 2013; Khor et al., 2015; Pappas et al., 2018). The goal of the SHLARC is to combine international expertise with data and computational resources. It will bring data to a level of interpretation that is key to solving questions on immune-related pathologies through innovative algorithms and powerful computation tool development. To achieve this goal, we determined three main objectives (Figure 2):

1. *Data*. By bringing together scientists from around the world, we will collectively increase the amount of SNP + HLA data available, both in terms of quantity and genetic diversity. Building new reference panels from these data will improve the performance of *HLA* allele imputation from SNPs as large, diverse, well-defined genomic data are the *prima materia* of successful collaborations and machine-learning applications for dissecting the genetic determinants of disease association.
2. *Applied mathematical and computer sciences*. We will further optimize SNP-HLA imputation methods using the HIBAG tool, and particularly for genetically diverse and admixed populations as (a) the higher complexity of their *MHC* region is a challenge for imputation and (b) these populations are still underrepresented in genomic studies (Sirugo, Williams, & Tishkoff, 2019). In addition, we will explore new machine-learning approaches such as deep learning to develop new, more efficient methods of *HLA* imputation.
3. *Accessibility and service to the scientific community*. Following the Haplotype Reference Consortium

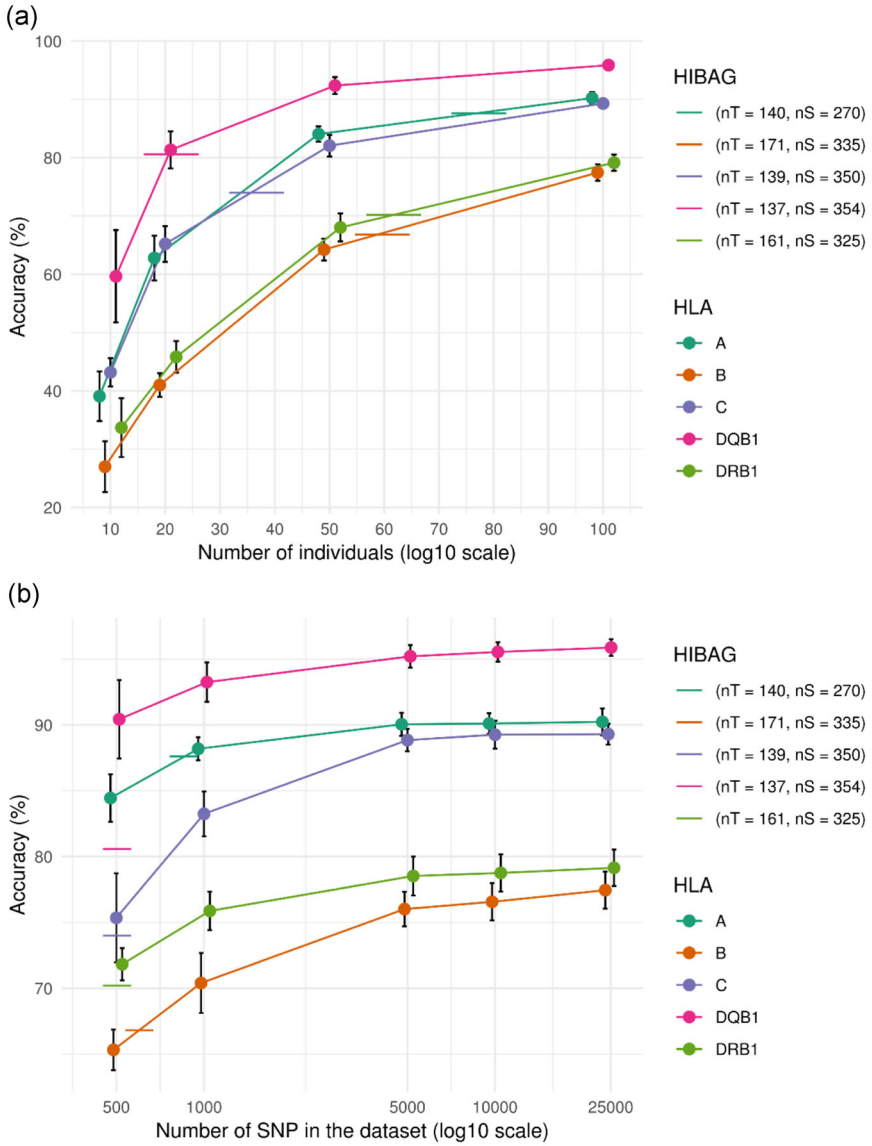


FIGURE 1 Influence of the number of individuals (a) and SNPs (b) in the HIBAG reference panel building on the accuracy of *HLA* alleles prediction. From the CAAPA data set ($N = 880$ and SNPs = 24,504), we produced a set of 10 training subsets ($n_{\text{training}} = 100$) and test ($n_{\text{test}} = 780$) sets to assess *HLA* imputation accuracy in different scenarios. Each model was validated by comparing the typed *HLA* alleles to the model-predicted *HLA* alleles across all individuals to provide an accuracy percentage (postprobability call threshold = 0). (a) By randomly selecting individuals in the training data set, we created sub-datasets containing 10, 20, and 50 individuals. Custom HIBAG models were computed for these subsets as well as for the whole 100 training individuals, using every available SNP. (b) Subsets of the training data set with 500, 1,000, 5,000, 10,000 randomly selected SNPs (out of the 24,504 available SNPs) were created and the corresponding models computed. The number of SNPs on the *x*-axis is indicative of the number of SNPs in the data set. The number of SNPs kept to create the model, which varies depending on the gene studied and the subset, is five times lower on average (see Tables S1.1 and S1.2). Note that the horizontal marks on each *HLA* gene curve indicate the accuracies obtained with the default African American HIBAG models. HIBAG, *HLA* Genotype Imputation with Attribute Bagging; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism; nS, number of SNPs in the model; nT, number of individuals in the model

initiative (McCarthy et al., 2016), our network envisions building a free, user-friendly webserver where researchers can access our improved imputation protocols by simply uploading their data and obtaining the most accurate possible *HLA* imputation for their data set. This service will offer several solutions (a) ready-to-use anonymized reference panels for researchers wishing to impute the *HLA* themselves, (b) allow the on-demand creation and sharing of tailored (customized) reference panels based on data available

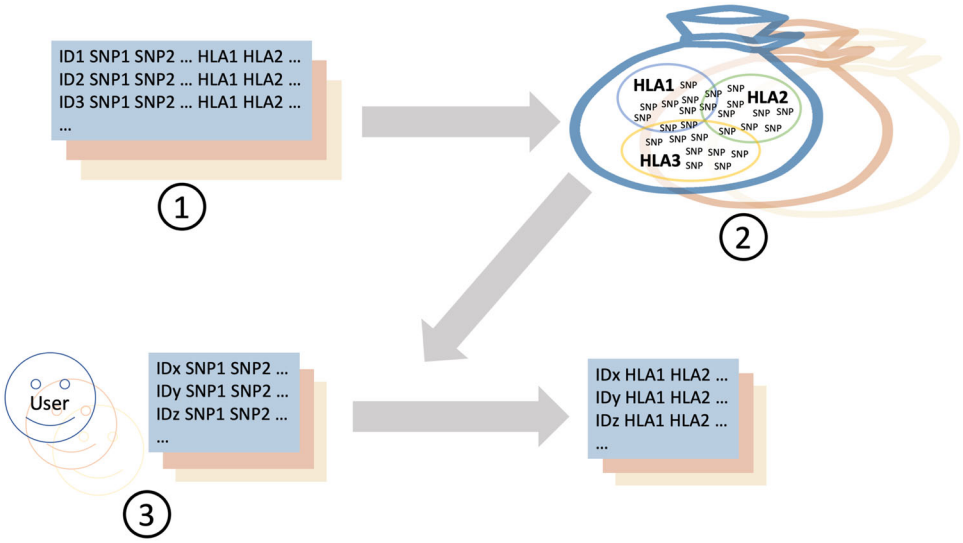


FIGURE 2 The SNP-HLA Reference Consortium (SHLARC) design. Aim 1: Increase the amount of SNP + HLA data available both in terms of quantity and diversity. Aim 2: Optimize SNP-HLA imputation methods. Aim 3: The SHLARC website will allow users from the scientific community to benefit from the data and knowledge accumulated by the consortium on SNP-to-HLA allele imputation. From a list of SNPs and a selected ethnicity of interest, or alternatively from uploading SNP genotype data sets, the best custom reference panel for *HLA* allele imputation will be built in our servers. HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism

in our database, or (c) provide a full SNP-to-HLA imputation service from uploaded raw SNP genotypes. We will also explore how to create the reference panel with the best fit for ancestry and genotyping platforms, given the queried samples, without the need for the centralization of individual data. Indeed, distributed calculation techniques may allow to create reference panels from data hosted on different servers without collecting all the information in a single place.

Our objectives require access to the extensive computation power that is readily available through several GPU servers within the Université de Nantes. For each submission, we aim to design custom reference panels, for which SNPs, *HLA*, and reference panel data will be securely stored on University's servers. Importantly, reference panels represent statistical models that do not allow individual re-identification. The current SHLARC partners share complementary expertise including but not limited to bioinformatics, population genetics, and immunogenetics. Importantly, our network is designed around data sharing to facilitate open research as we believe research can be accelerated by freely sharing knowledge and data. With this in mind, we have added this consortium as a component of the 18th International HLA and Immunogenetics Workshop (https://www.ihiw18.org/).

HLA imputation is primarily intended for research applications, as clinical applications such as hematopoietic stem cell transplantation (HSCT) cannot tolerate statistical uncertainty, even though it might be used to accelerate pre-selection of HSCT patients as well (Meyer & Nunes, 2017; Pappas et al., 2018). The 1000 Genomes project (1000 Genomes Project Consortium et al., 2015) generated a large collection of polymorphisms from 2,504 individuals of diverse ancestry (SNPs, indels, and copy number variants), along with *HLA* allele typings (Gourraud et al., 2014), providing an informative overview of genetic diversity among human populations. However, a recent study by Abi-Rached et al. (2018) highlighted the absence of several common *HLA* alleles (>1% allele frequency) from the 1000 Genomes project which shows how *HLA* imputation results could be biased by an insufficient reference panel. With the proper sampling and a shared effort in gathering diverse data, *HLA* imputation could bridge the gap between *HLA* allele diversity and the understanding of its impact on phenotypes by harnessing the latent information stored in GWAS data sets to upgrade genetic epidemiological knowledge of immune-related diseases. As shown previously (Okada et al., 2015), predicting *HLA* alleles from population-matching reference panels not only increases the confidence in the predicted *HLA* but above all, allows prediction of specific *HLA* alleles that could not be imputed otherwise. Therefore, the informed choice of the applied model would strengthen the relation between *HLA*, ancestry, and disease risk factor. By applying this customization at a general level, we would assess ancestry with SNP relatedness, a

consistent marker of population, rather than using self-reported ancestry which can be often misleading (Sanchez-Mazas et al., 2012).

To develop this ambitious project, we encourage willing participants with available two-fields *HLA* alleles + SNPs data sets to join the SNP-HLA reference consortium (https://www.ihiw18.org/component-bioinformatics/snp-hla-reference/) to contribute empowering the immunogenetic community to move into the era of immunogenomic association.

ACKNOWLEDGEMENTS

The authors thank Labex IGO (ANR-11-LABX-0016-01) and IHU CESTI for their support. Nicolas Vince has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 846520. This study is supported by the ATIP-Avenir Inserm program, the Region Pays de Loire ConnectTalent, the ANR PIA-Investment (NExT), and the 18th International HLA and Immunogenetics Workshop. SNP-HLA Reference Consortium (SHLARC) Partners: Pierre-Antoine Gourraud, Nicolas Vince, Sophie Limou, Estelle Geffard, and Venceslas Douillard, Nantes Université, Centrale Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France; Mario Südholt, Damien Eveillard, and Fatima-Zahra Boujdad, LS2N, UMR6004 CNRS, Université de Nantes, Centrale Nantes, IMTA, Nantes, France; Luisa Rocha Da Silva, Hugues Digonnet, and Domenico Borzacchiello, Ecole Centrale de Nantes, Nantes, France; Diogo Meyer, Vitor Aguiar, Kelly Nunes, University of São Paulo, São Paulo, Brazil; Erick C. Castelli, Unesp—Universidade Estadual Paulista, Botucatu-SP, Brazil; Surakameth Mahasirimongkol, Nuanjun Wichukchinda, Nusara Satproedprai, Sukanya Wattanapokayakit, Sacarin Bunbanjerdsuk, Punna Kunhapan, Thanyapat Wanitchanon, Penpitcha Thawong, and Pundharika Piboonsiri, Medical Genetics Center, Medical Life Sciences Insitute, Department of Medical Sciences, Ministry of Public Health; Soranun Chantarangsu, Chulalongkorn University, Department of Oral Pathology, Bangkok, Thailand; Sasithorn Chotewutmontri, Faculty of Medicine and Public Health, HRH Princess Chulabhorn College of Medical Science, Bangkok, Thailand; Supichaya Boonvisut, Environmental Toxicology, Chulabhorn Graduate Institute, Chulabhorn Royal Academy, Bangkok, Thailand; Derek Middleton, University of Liverpool, Liverpool, UK; Faviel Gonzalez, University of Liverpool, Liverpool, UK and Autonomous University of Coahuila, Mexico; James Traherne and Vitalina Kirgizova, University of Cambridge, Cambridge, UK; Andre Franke, Frauke Degenhardt, David Ellinghaus, and Mareike Wendorff, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; Mehmet Dorak, Kingston University London, London, UK; Xiuwen Zheng, Department of Biostatistics, University of Washington, Seattle, WA, USA; Benedicte A. Lie, Marte Kathrine Viken, and Riad Hajdarevic, Department of Medical Genetics University of Oslo and Oslo University Hospital, Oslo, Norway; Department of Immunology, Rikshospitalet, University of Oslo and Oslo University Hospital, Oslo, Norway; Veron Ramsuran, University of KwaZulu-Natal, Durban, South Africa; Dara Torgerson and Ryan Hernandez, McGill University, Montreal, Canada; Zachary Szpiech, Auburn University, Auburn, AB, USA; Jill Hollenbach and Melissa Spear, University of California, San Francisco, CA, USA; Steven J. Mack, Department of Pediatrics, University of California, San Francisco and UCSF Benioff Children's Hospital Oakland, Oakland, CA, USA; Martin Maiers, Bioinformatics Research, Center for International Blood and Marrow Transplant Research, Minneapolis, MN, USA; Satu Koskela, Finnish Red Cross Blood Service, Helsinki, Finland; Anders Albrechtsen and Torben Hansen, The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark; Zorana Grubic, Katarina Stingl Jankovic, and Marija Maskalan, University Hospital Center Zagreb, Zagreb, Croatia; Martin Petrek and Katerina Sikorova, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czechia; Fatma Oguz, Instanbul University, Instanbul, Turkey; Jeremie Decouchant, Marcus Volp, Maria Fernandes, University of Luxembourg, Luxembourg, Luxembourg; Piotr Kusnierczyk, Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wroclaw, Poland; Blanka Vidan-Jeras and Sendi Montanic, Blood Transfusion Center of Slovenia, Ljubljana, Slovenia.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at dbGAP (CAAPA, dbGaP Study Accession: phs001123.v1.p1) and from the 1000 Genomes Project website, using the latest SNP (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/) and *HLA* data at the time (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/).

ORCID

Nicolas Vince http://orcid.org/0000-0002-3767-6210
Venceslas Douillard http://orcid.org/0000-0002-6762-4083

REFERENCES

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., ... 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abi-Rached, L., Gouret, P., Yeh, J.-H., Di Cristofaro, J., Pontarotti, P., Picard, C., & Paganini, J. (2018). Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS One*, *13*(10), e0206512. https://doi.org/10.1371/journal.pone.0206512

Chen, H., Ndhlovu, Z. M., Liu, D., Porter, L. C., Fang, J. W., Darko, S., ... Walker, B. D. (2012). TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. *Nature Immunology*, *13*(7), 691–700. https://doi.org/10.1038/ni.2342

Dausset, J. (1999). The HLA adventure. *Transplantation Proceedings*, *31*(1–2), 22–24.

Daya, M., Rafaels, N., Brunetti, T. M., Chavan, S., Levin, A. M., Shetty, A., ... CAAPA. (2019). Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. *Nature Communications*, *10*(1), 880. https://doi.org/10.1038/s41467-019-08469-7

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6.

Dilthey, A. T., Gourraud, P.-A., Mentzer, A. J., Cereb, N., Iqbal, Z., & McVean, G. (2016). High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. *PLOS Computational Biology*, *12*(10), e1005151. https://doi.org/10.1371/journal.pcbi.1005151

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., ... Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, *317*(5840), 944–947.

Geffard, E., Limou, S., Walencik, A., Daya, M., Watson, H., Torgerson, D., ... Vince, N. (2019). Easy-HLA, a validated web application suite to reveal the full details of HLA typing. *Bioinformatics*, *36*(7), https://doi.org/10.1093/bioinformatics/btz875

Gourraud, P.-A., Khankhanian, P., Cereb, N., Yang, S. Y., Feolo, M., Maiers, M., ... Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. *PLOS One*, *9*(7), e97282.

Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P. J., Rich, S. S., ... de Bakker, P. I. W. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. *PLOS One*, *8*(6), e64683. https://doi.org/10.1371/journal.pone.0064683

Khor, S.-S., Yang, W., Kawashima, M., Kamitsuji, S., Zheng, X., Nishida, N., ... Tokunaga, K. (2015). High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. *The Pharmacogenomics Journal*, *15*(6), 530–537. https://doi.org/10.1038/tpj.2015.4

Limou, S., & Zagury, J.-F. (2013). Immunogenetics: Genome-wide association of non-progressive HIV and viral load control: HLA genes and beyond. *Frontiers in Immunology*, *4*, 118. https://doi.org/10.3389/fimmu.2013.00118

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., ... Parkinson, H. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Meyer, D., & Nunes, K. (2017). HLA imputation, what is it good for? *Human Immunology*, *78*(3), 239–241. https://doi.org/10.1016/j.humimm.2017.02.007

Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., ... Kubo, M. (2015). Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nature Genetics*, *47*(7), 798–802. https://doi.org/10.1038/ng.3310

Pappas, D. J., Lizee, A., Paunic, V., Beutner, K. R., Motyer, A., Vukcevic, D., ... Maiers, M. (2018). Significant variation between SNP-based HLA imputations in diverse populations: The last mile is the hardest. *The Pharmacogenomics Journal*, *18*(3), 367–376. https://doi.org/10.1038/tpj.2017.7

Sanchez-Mazas, A., Vidan-Jeras, B., Nunes, J. M., Fischer, G., Little, A.-M., Bekmane, U., ... Tiercy, J.-M. (2012). Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. *International Journal of Immunogenetics*, *39*(6), 459–472. https://doi.org/10.1111/j.1744-313X.2012.01113.x. quiz 473–476.

Shen, J. J., Yang, C., Wang, Y.-F., Wang, T.-Y., Guo, M., Lau, Y. L., ... Sheng, Y. (2018). HLA-IMPUTER: An easy to use web application for HLA imputation and association analysis using population-specific reference panels. *Bioinformatics*, *37*(7), https://doi.org/10.1093/bioinformatics/bty730

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Trowsdale, J., & Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, *14*, 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

Vince, N., Li, H., Ramsuran, V., Naranbhai, V., Duh, F.-M., Fairfax, B. P., ... Carrington, M. (2016). HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. *American Journal of Human Genetics*, *99*(6), 1353–1358. https://doi.org/10.1016/j.ajhg.2016.09.023

Vince, N., Limou, S., Daya, M., Morii, W., Rafaels, N., Geffard, E., ... CAAPA. (2020). Association of HLA-DRB1*09:01 with tIgE levels among African ancestry individuals with asthma. *The Journal of Allergy and Clinical Immunology*, https://doi.org/10.1016/j.jaci.2020.01.011

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Zheng, X. (2018). Imputation-based HLA typing with SNPs in GWAS studies. *Methods in Molecular Biology (Clifton, NJ)*, *1802*, 163–176. https://doi.org/10.1007/978-1-4939-8546-3_11

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG—HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, *14*(2), 192–200. https://doi.org/10.1038/tpj.2013.18

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Vince N, Douillard V, Geffard E, et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. *Genetic Epidemiology*. 2020;44:733–740. https://doi.org/10.1002/gepi.22334

SNP-HLA reference consortium (SHLARC) partners.

Pierre-Antoine Gourraud, Nicolas Vince, Sophie Limou, Estelle Geffard, Venceslas Douillard. Nantes Université, Centrale Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France
Mario Südholt, Damien Eveillard, Fatima-Zahra Boujdad. LS2N, UMR6004 CNRS, Université de Nantes, Centrale Nantes, IMTA, Nantes, France.
Luisa Rocha Da Silva, Hugues Digonnet, Domenico Borzacchiello. Ecole Centrale de Nantes, Nantes, France.
Diogo Meyer, Vitor Aguiar, Kelly Nunes. University of São Paulo, São Paulo, Brazil.
Erick C. Castelli. Unesp - Universidade Estadual Paulista, Botucatu-SP, Brazil.
Surakameth Mahasirimongkol, Nuanjun Wichukchinda, Nusara Satproedprai, Sukanya Wattanapokayakit, Sacarin Bunbanjerdsuk, Punna Kunhapan, Thanyapat Wanitchanon, Penpitcha Thawong, Pundharika Piboonsiri. Medical Genetics Center, Medical Life Sciences Insitute, Department of Medical Sciences, Ministry of Public Health
Soranun Chantarangsu. Chulalongkorn University, Department of Oral Pathology, Bangkok, Thailand
Sasithorn Chotewutmontri. Faculty of Medicine and Public Health, HRH Princess Chulabhorn College of Medical Science, Bangkok, Thailand
Supichaya Boonvisut. Environmental Toxicology, Chulabhorn Graduate Institute, Chulabhorn Royal Academy, Bangkok, Thailand
Derek Middleton. University of Liverpool, Liverpool, UK.
Faviel Gonzalez, University of Liverpool, Liverpool, UK and Autonomous University of Coahuila, Mexico.
James Traherne, Vitalina Kirgizova. University of Cambridge, Cambridge, UK.
Andre Franke, Frauke Degenhardt, David Ellinghaus, Mareike Wendorff. Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany.
Mehmet Dorak. Kingston University London, London, UK.
Xiuwen Zheng. Department of Biostatistics, University of Washington, Seattle, WA, USA.
Benedicte A. Lie, Marte Kathrine Viken, Riad Hajdarevic. Department of Medical Genetics University of Oslo and Oslo University Hospital, Oslo, Norway. Department of Immunology, Rikshospitalet, University of Oslo and Oslo University Hospital, Oslo, Norway.
Veron Ramsuran. University of KwaZulu-Natal, Durban, South Africa.
Dara Torgerson, Ryan Hernandez. McGill University, Montreal, Canada.
Zachary Szpiech. Auburn University, Auburn, AB, USA.
Jill Hollenbach, Melissa Spear. University of California, San Francisco, CA, USA.
Steven J. Mack. Department of Pediatrics, University of California, San Francisco, CA and Children's Hospital Oakland Research Institute, Oakland, CA, USA.
Martin Maiers. Bioinformatics Research, Center for International Blood and Marrow Transplant Research, Minneapolis, MN, USA.
Satu Koskela. Finnish Red Cross Blood Service, Helsinki, Finland.
Anders Albrechtsen, Torben Hansen. The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark.
Zorana Grubic, Katarina Stingl Jankovic, Marija Maskalan. University Hospital Center Zagreb, Zagreb, Croatia.
Martin Petrek, Katerina Sikorova. Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czechia.

Fatma Oguz. Instanbul University, Instanbul, Turkey.
Jeremie Decouchant, Marcus Volp, Maria Fernandes. University of Luxembourg, Luxembourg, Luxembourg.
Piotr Kusnierczyk. Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wroclaw, Poland.
Blanka Vidan-Jeras, Sendi Montanic. Blood Transfusion Center of Slovenia, Ljubljana, Slovenia.

SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics.

Nicolas Vince[1*], Venceslas Douillard[1*], Estelle Geffard[1], Diogo Meyer[2], Erick C. Castelli[3], Steven J. Mack[4], SHLARC investigators[5], Sophie Limou[1,6], Pierre-Antoine Gourraud[1]

1. Université de Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France
2. University of São Paulo, São Paulo, Brazil
3. Unesp - Universidade Estadual Paulista, Botucatu-SP, Brazil
4. Department of Pediatrics, University of California, San Francisco, CA and Children's Hospital Oakland Research Institute, Oakland, CA, USA
5. Full list of investigators listed in SHLARC_investigators.docx
6. Ecole Centrale de Nantes, Nantes, France
* These authors contributed equally to this work.

Methods

Datasets building

CAAPA (The Consortium on Asthma among African-ancestry Populations in the Americas) aims to discover genetic risk factors of asthma and catalog diversity in African-Americans [1]. SNP and HLA data from CAAPA come from whole-genome sequencing (WGS). SNP genotypes are in PLINK files format and contain 24,504 SNPs (MHC region only: chromosome 6, 29Mb to 34Mb) from 917 individuals. After quality control, we excluded A/T or G/C SNPs, filtered by MAF 0.001 and by genotype missing call rate >2%. SNPs positions are based on the GRCh37/hg19 genome assembly. HLA genotypes were called from the same 917 individuals WGS data using Omixon software, they are available as two fields (4 digits) resolution in a CSV file. We excluded 37 individuals flagged as related to another in dbGAP CAAPA metadata.

We split the CAAPA dataset into a training and a test dataset ($n_{training}$=100 and n_{test}=780, respectively) and randomly repeated the operation 10 times to assess variability. To assess the effect of both the number of individuals and the number of SNPs in *HLA* imputation, we created 8 datasets from the initial training data. We applied the same method for different number of individuals or different SNP subsets: we randomly selected with PLINK 1.9 a list of 10, 20 or 50 individuals (or 500, 1,000, 5,000 or 10,000 SNPs) from the initial training dataset (n_T=100 and SNP=24,504). Besides, HIBAG selects SNPs inside a window of 500kb around each gene to build its reference panels, hence the real number of SNPs inside reference panels is different from the initial number. Finally, due to HIBAG removing monomorphic SNPs during attribute bagging, subsetting individuals may delete a polymorphism in the training set that is present in the full set (see table below).

SNP in file	500	1,000	5,000	10,000	24,504
SNP for HLA-A	90	169	837	1,677	4,108
SNP for HLA-B	82	167	829	1,682	4,123
SNP for HLA-C	87	176	866	1,755	4,3155
SNP for HLA-DQB1	109	220	1,086	2,164	5,340
SNP for HLA-DRB1	105	213	1,051	2,095	5,166

Supplementary Table 1.1. Mean effective number of SNPs in the custom HIBAG reference panels compared to number of SNPs in the full genotype set (changing the number of SNPs as depicted in Figure 1A).

Individuals	10	20	50	100
SNP in file	24,504	24,504	24,504	24,504
SNP for HLA-A	2,380	2,892	3,613	4,108
SNP for HLA-B	2,277	2,798	3,588	4,123
SNP for HLA-C	2,385	2,940	3,753	4,315
SNP for HLA-DQB1	3,553	4,184	4,866	5,340
SNP for HLA-DR	3,406	4,030	4,697	5,166

Supplementary Table 1.2. Mean effective number of SNPs in the custom HIBAG reference panels compared to number of SNPs in the full genotype set (changing the number of individuals as depicted in Figure 1B).

Supplementary Table 2 contains data used to make supplementary Figures 1 to 5. It includes the mean sensitivity or frequency for each *HLA* allele based on the 10 different subsets for the changing number of SNP and individuals, respectively.

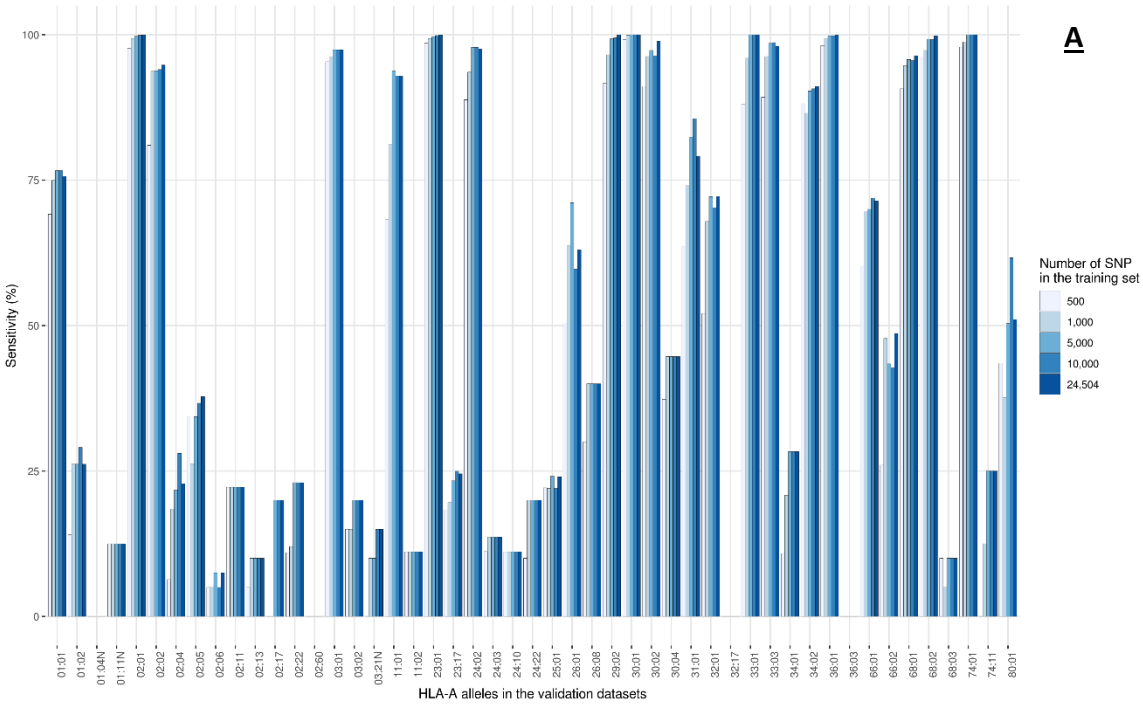
Reference panels with the HIBAG package

We built 40 reference panels of *HLA* imputation with the 8 datasets for each *HLA* genes: 3 class I genes (*HLA-A*, *B* and *C*) and 2 class II genes (*HLA-DQB1* and *DRB1*). To build a statistical model of *HLA* imputation, we need SNP and HLA genotypes information. The *HLA* imputation principle is to infer *HLA* alleles from SNP genotypes with our built reference panel models.

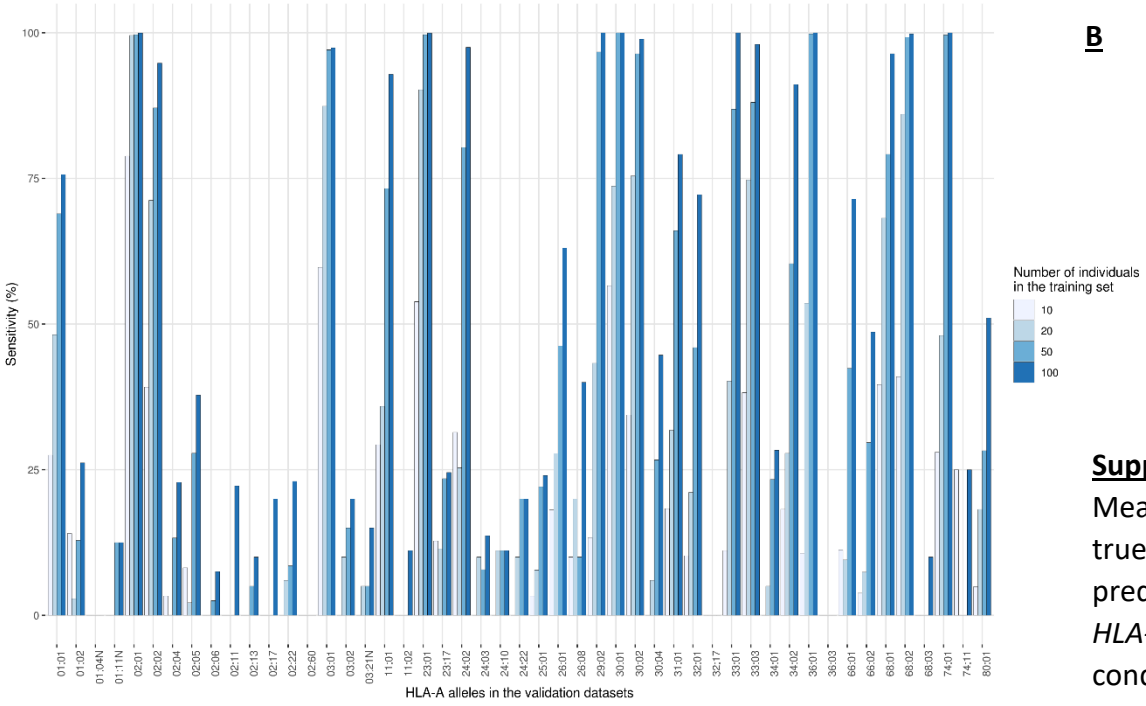
We selected the HIBAG R package as our reference *HLA* imputation software as it showed better performance compared to other software (see ref. Pappas et al., 2018 in the main manuscript). Moreover, HIBAG allows easy creation of custom reference panels. Statistical models were computed with R 3.5.1, using the HIBAG (v1.4) package and its extension HIBAG.gpu (v0.9.1). Calculations were computed on GPU nodes (Nvidia Tesla K80 cards) installed in the Liger supercomputer cluster at Ecole Centrale de Nantes. The m number of SNP retained in the process of the model building were selected within a 500kb window of the studied gene, the proportion of SNPs integrated into each classifier was the default value of HIBAG, $\sqrt{m}$.

Accuracies statistics were provided directly inside HIBAG output. We imputed *HLA* alleles for the 780 validation individuals with the 40 models of prediction. We did not set a call threshold; therefore, accuracy is calculated as the number of correctly predicted HLA alleles out of every predicted one. Results were aggregated and plotted on R 3.5.1 with the ggplot2 package.

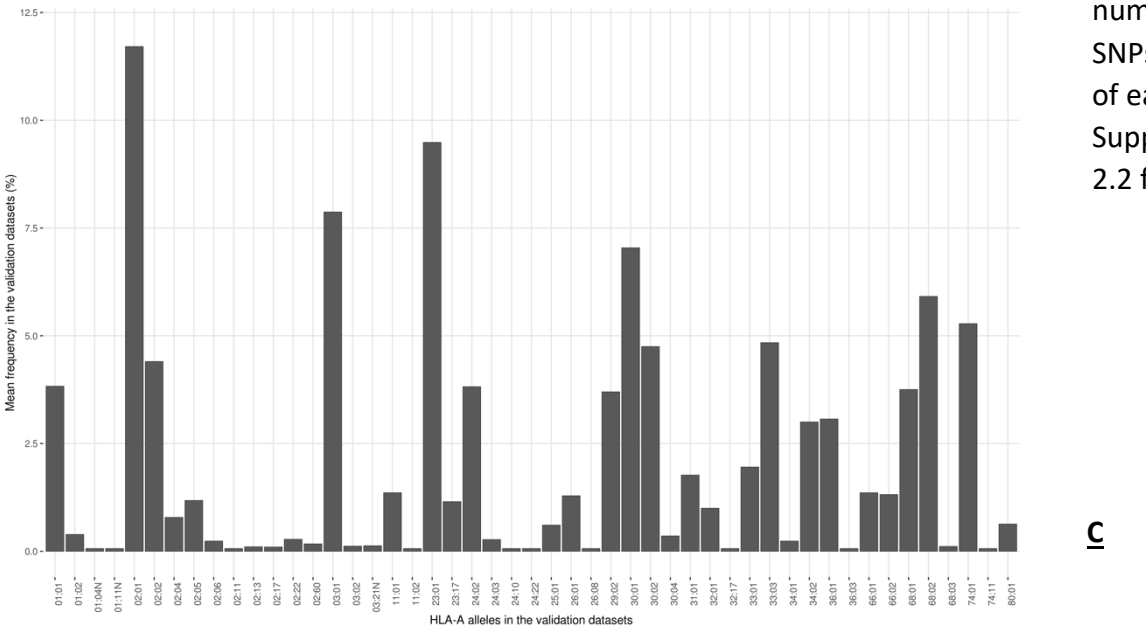
Results



A

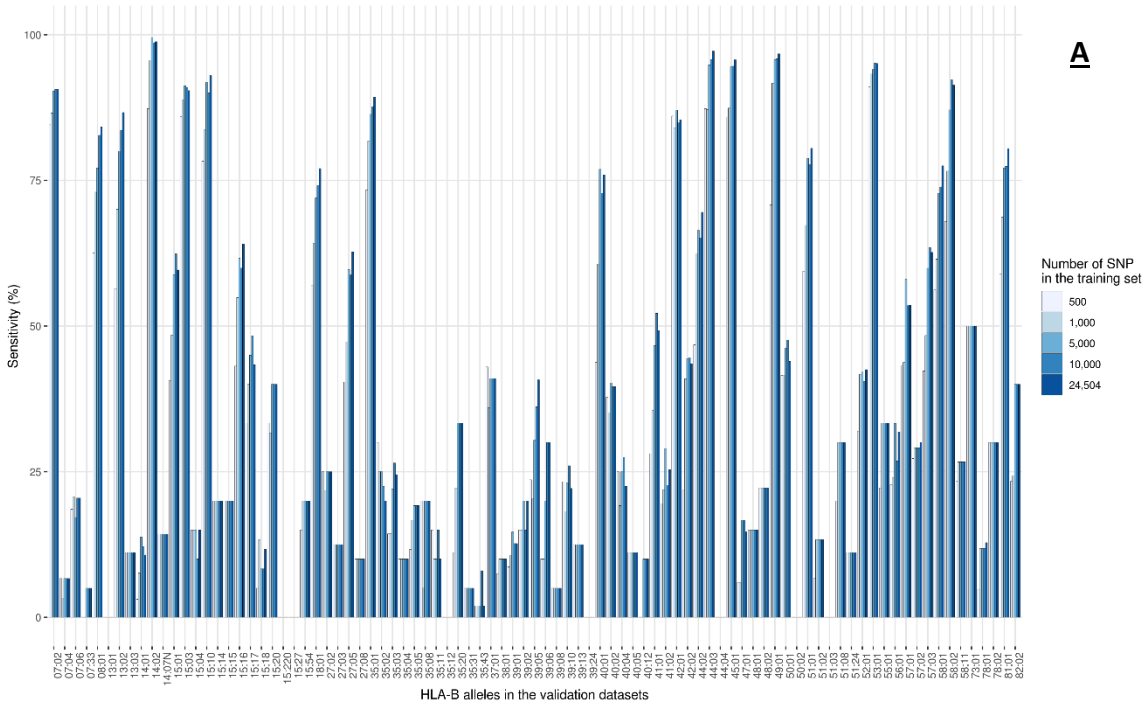


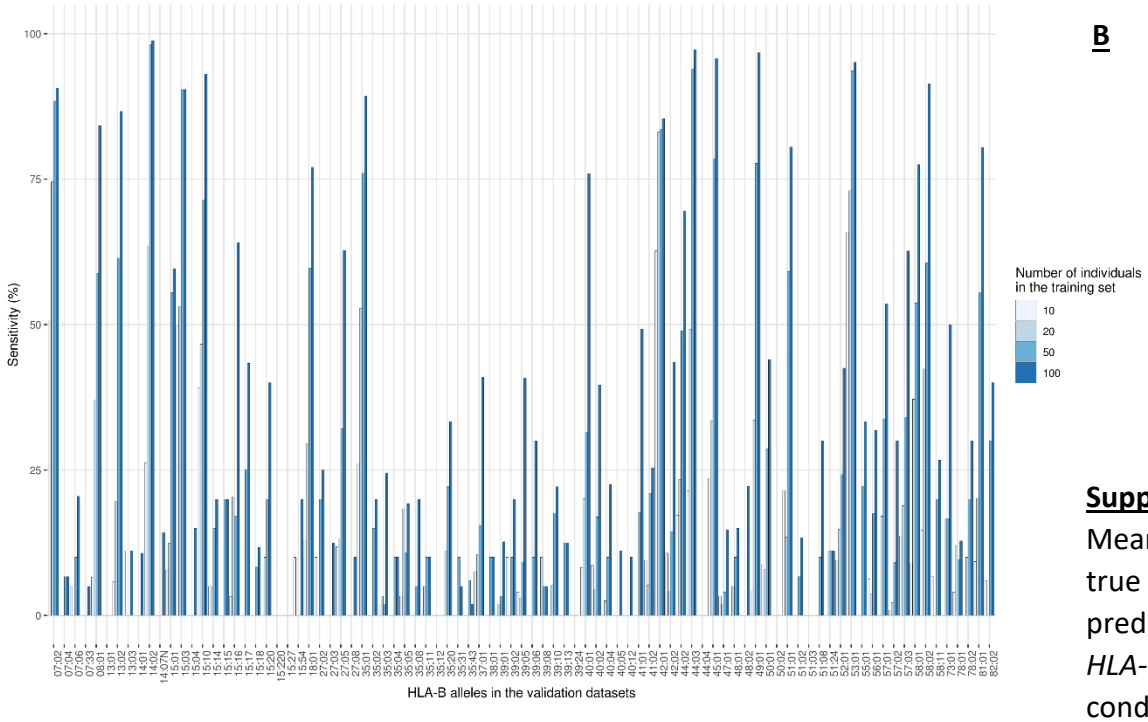
B

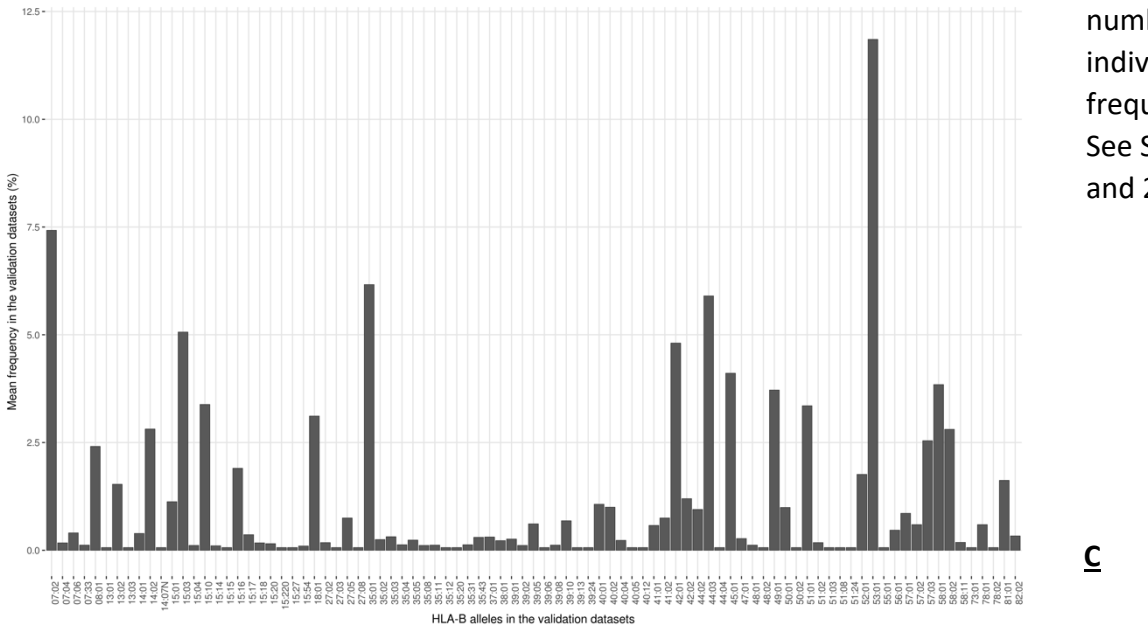


C

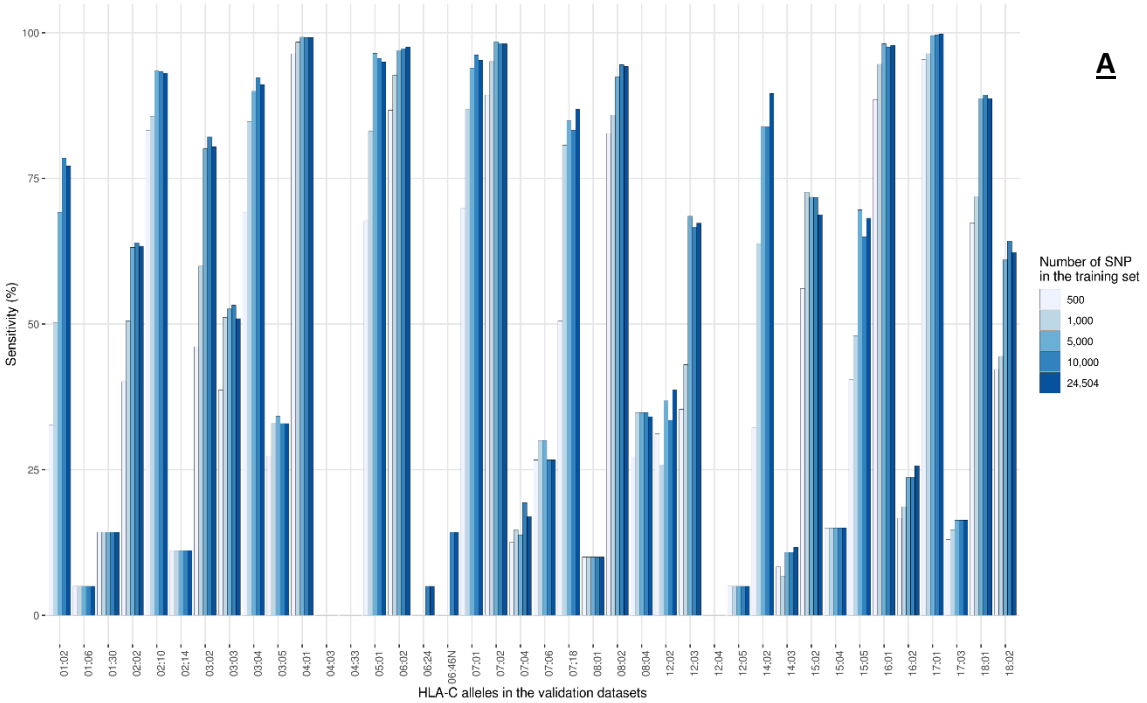
Supplementary Figure 1.
Mean sensitivity (number of true alleles correctly predicted) for each allele of *HLA-A*, according to each condition changing the number of individuals (A), SNPs (B) and mean frequency of each allele (C). See Supplementary Table 2.1 and 2.2 for data.



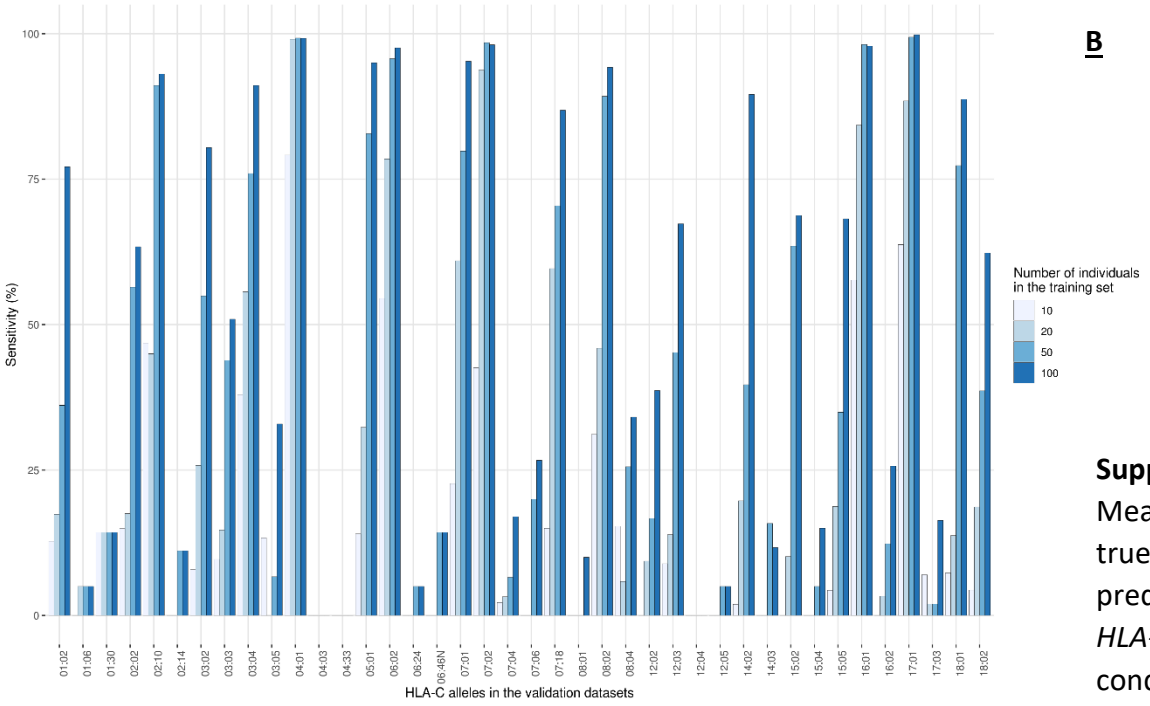




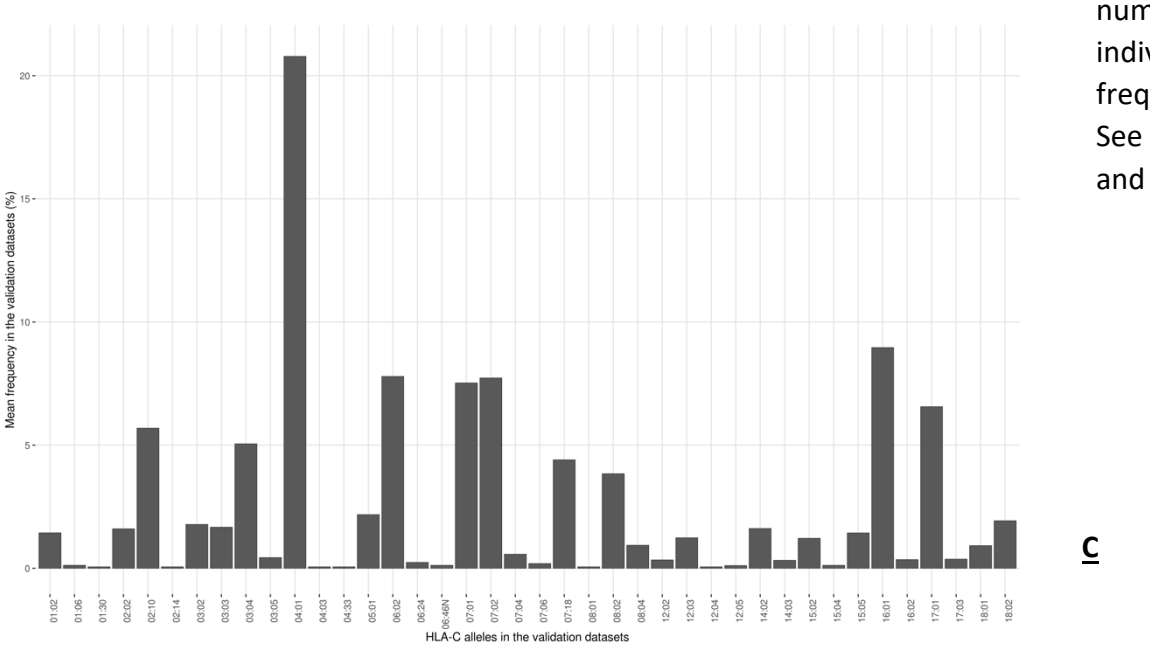
Supplementary Figure 2.
Mean sensitivity (number of true alleles correctly predicted) for each allele of *HLA-B*, according to each condition changing the number of SNPs (A), individuals (B) and mean frequency of each allele (C). See Supplementary Table 2.1 and 2.2 for data.



A

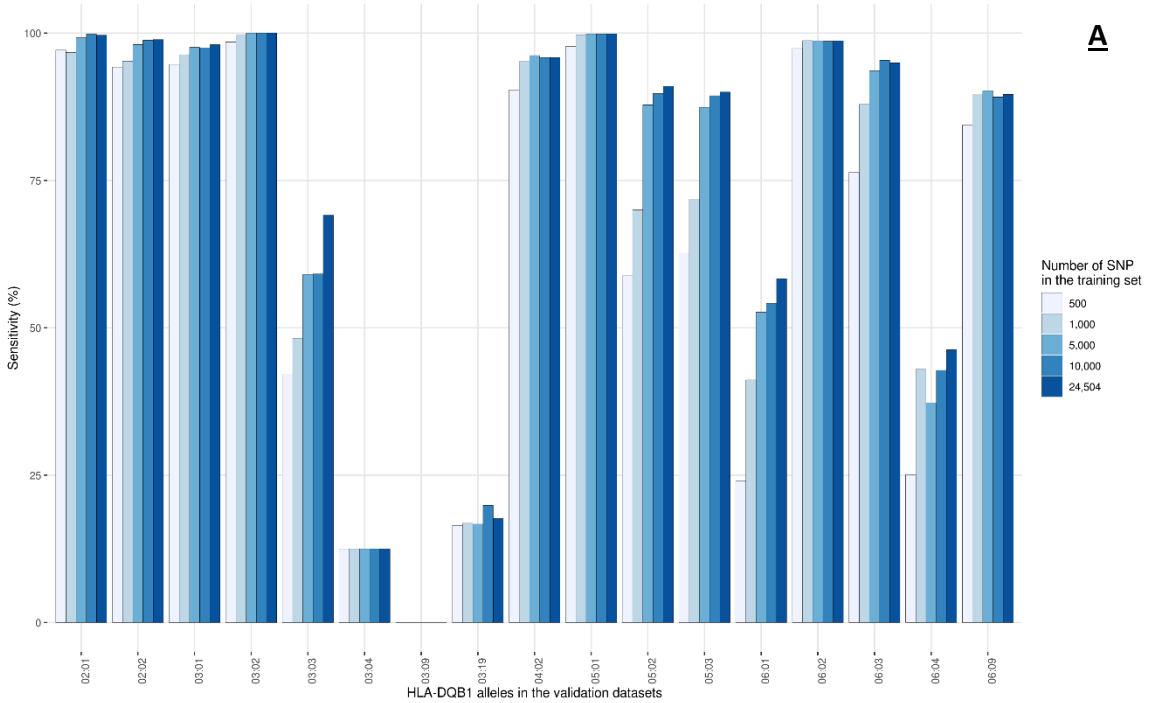


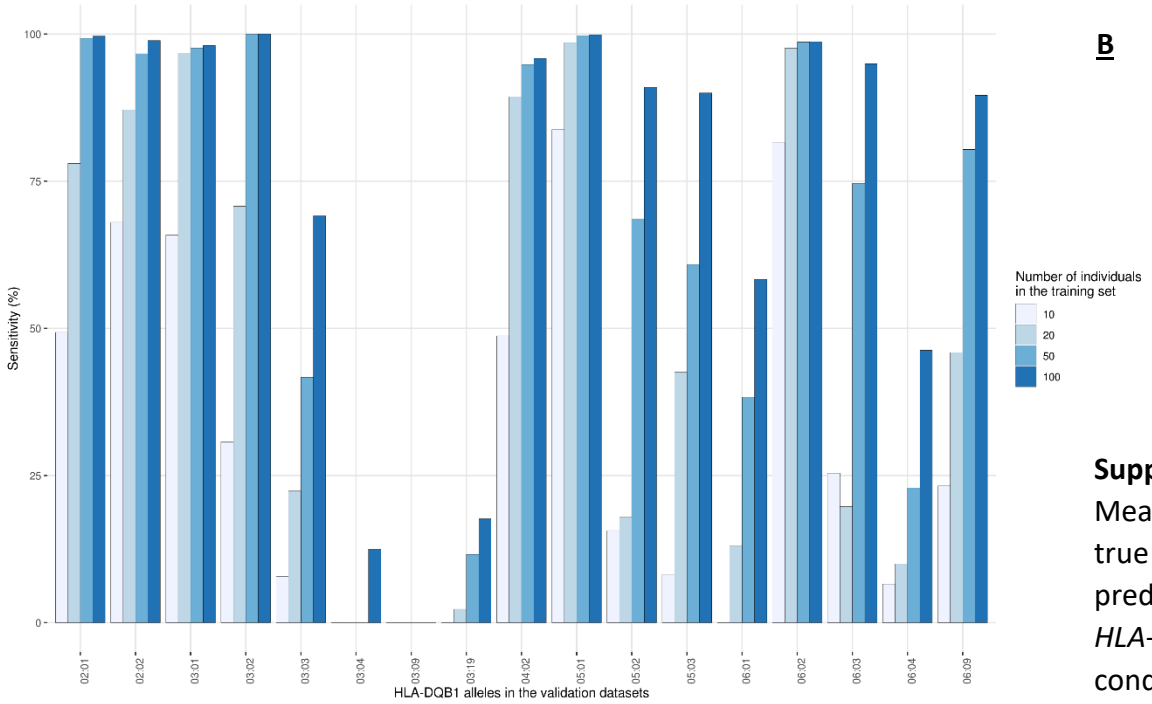
B

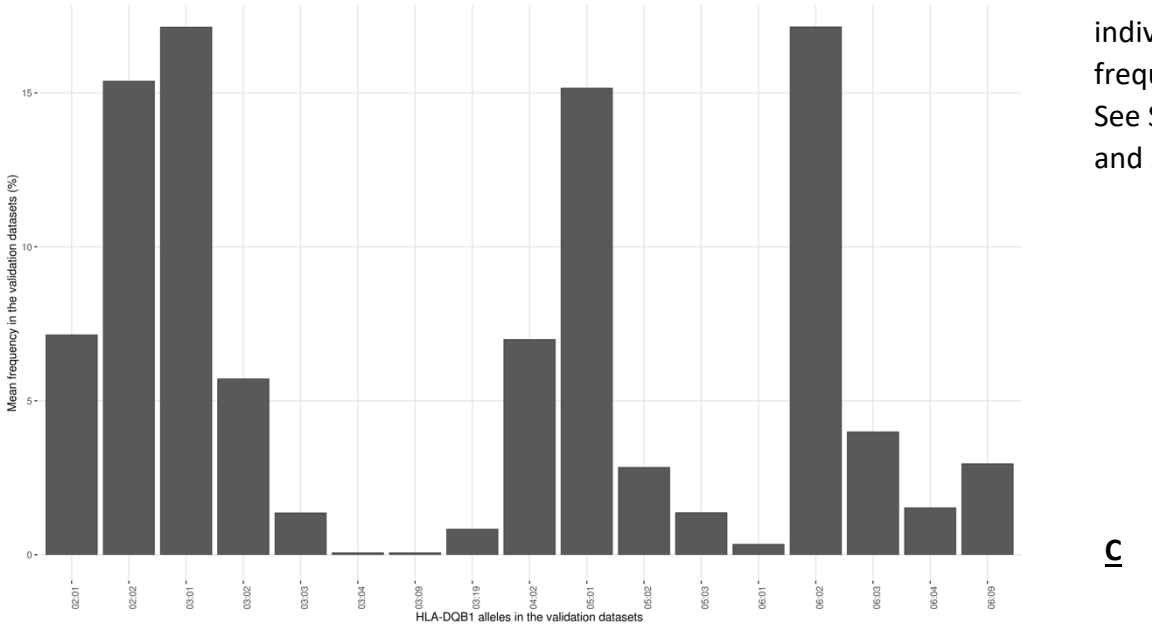


C

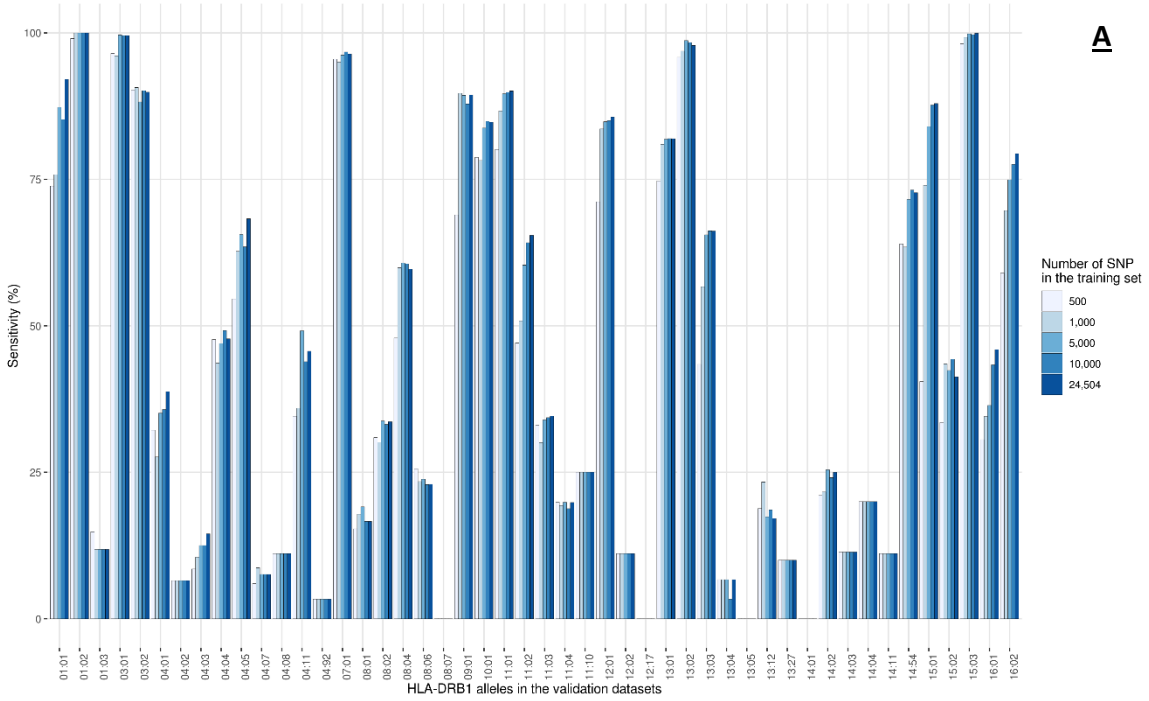
Supplementary Figure 3. Mean sensitivity (number of true alleles correctly predicted) for each allele of *HLA-C*, according to each condition changing the number of SNPs (A), individuals (B) and mean frequency of each allele (C). See Supplementary Table 2.1 and 2.2 for data.

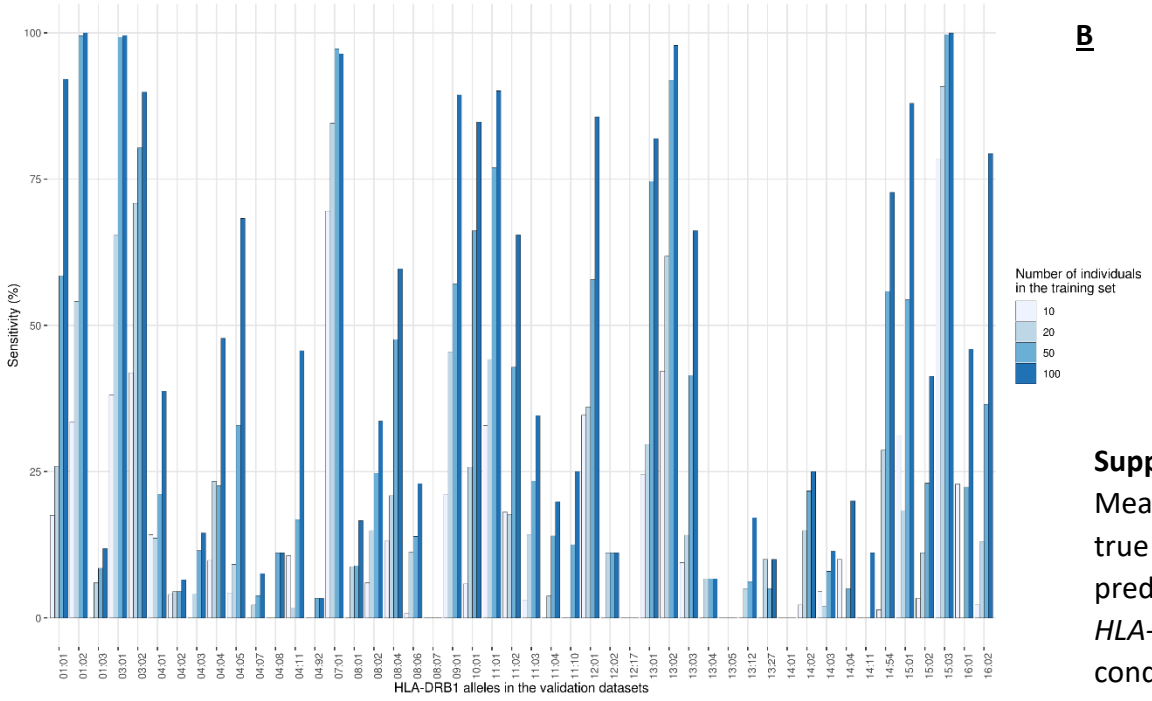


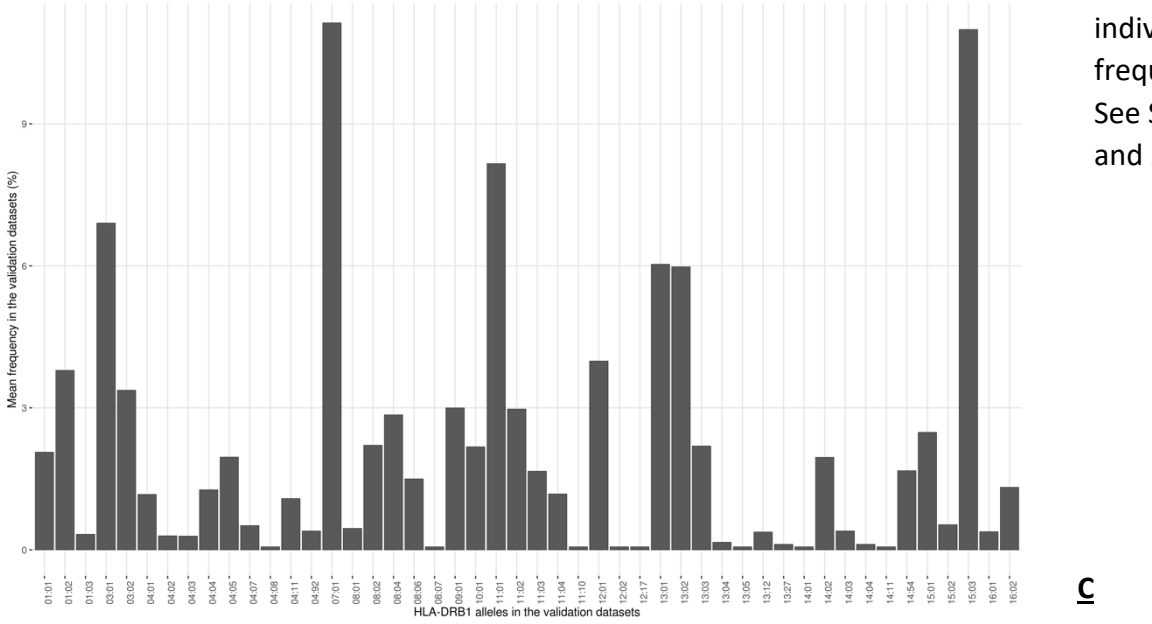




Supplementary Figure 4. Mean sensitivity (number of true alleles correctly predicted) for each allele of *HLA-DQB1*, according to each condition changing the number of SNPs (A), individuals (B) and mean frequency of each allele (C). See Supplementary Table 2.1 and 2.2 for data.







Supplementary Figure 5. Mean sensitivity (number of true alleles correctly predicted) for each allele of *HLA-DRB1*, according to each condition changing the number of SNPs (A), individuals (B) and mean frequency of each allele (C). See Supplementary Table 2.1 and 2.2 for data.

VI.2 - L'amélioration de l'imputation HLA dans des populations sous-représentées

VI.2.1 - Le contournement du manque de diversité par une nouvelle méthodologie

VI.2.1.1 - Les variations de fréquences HLA dans les populations et leur impact sur l'imputation

Pour améliorer l'imputation HLA, Ritari *et al.* proposent de créer un panel de référence avec une population spécifique, cela permet ainsi de mieux prédire les individus de cette population (336). Néanmoins, augmenter la taille et la diversité d'un jeu de données requiert du typage HLA. Or, ces expériences supplémentaires représentent un coût important dans la recherche. Afin de contourner cette limite, nous avons cherché à comprendre comment améliorer la méthodologie de l'imputation HLA à partir des données disponibles.

Pour évaluer l'impact des allèles HLA spécifiques d'une population sur la précision des modèles d'imputation HLA, nous avons utilisé les populations du projet 1,000 Genomes comme jeu de données d'entraînement pour HIBAG (Figure VI-3).

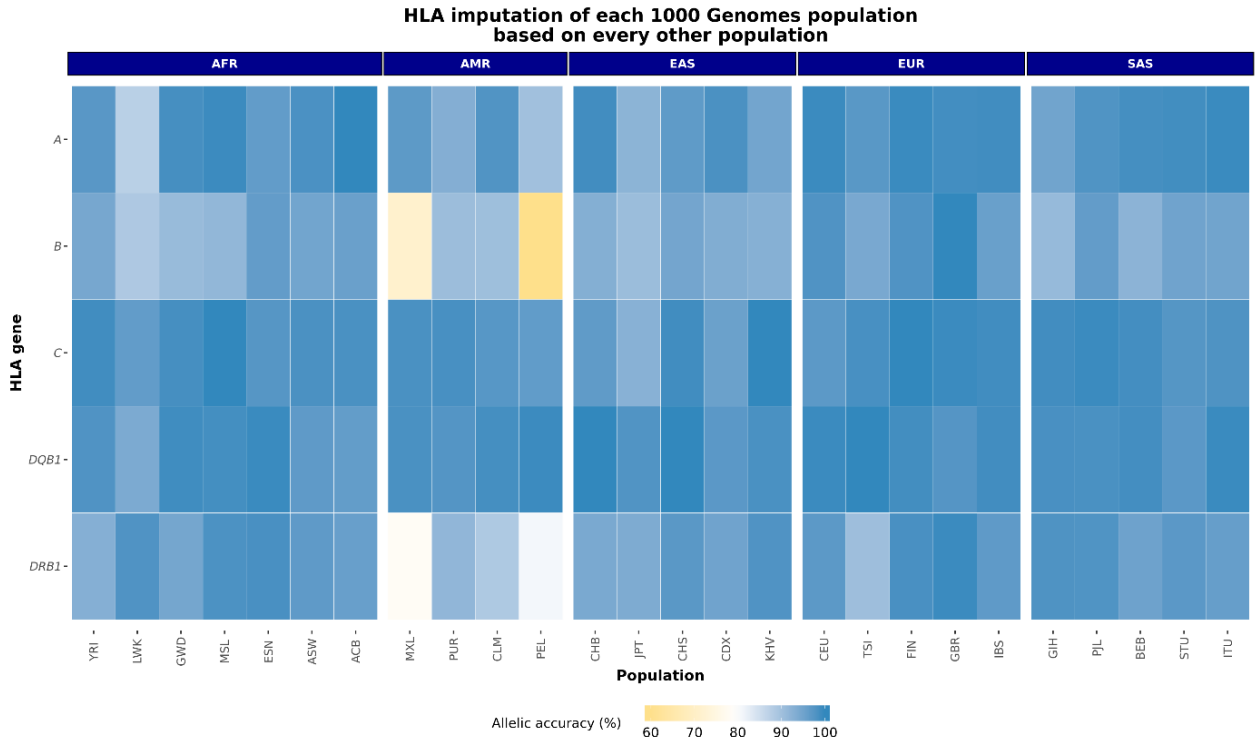


Figure VI-3 Précision d'imputation des gènes HLA-A, HLA-B, HLA-C, HLA-DQB1 et HLA-DRB1 pour 26 populations du 1KGP. Le HLA de chaque individu d'une population a été imputé à partir de tous les autres individus de 1KGP. La précision représente le nombre d'allèles correctement prédits.

En créant un modèle d'imputation avec toutes les populations de 1KGP sauf une, il est possible de mettre en avant les allèles spécifiques d'une population. Ainsi, les populations mexicaines (MXL) et péruviennes (PEL) présentent une précision d'imputation bien inférieure aux autres populations à

cause de leurs allèles HLA-B*15:15 et HLA-DRB1*09:06, respectivement, fréquents chez elles mais absents des autres populations.

VI.2.1.2 - Le changement de métrique d'imputation HLA

Un deuxième défi lié à celui des allèles spécifiques est l'imputation d'allèles rares. La métrique d'évaluation classique de l'imputation est la précision : elle décompte pour chaque allèle le nombre de bonnes prédictions (ex. lorsque l'allèle a été prédit correctement ou non-prédit à raison). Cependant, pour les allèles rares, la précision tend à être surestimée car même si l'allèle n'est jamais prédit correctement, il y a de fortes chances qu'il soit correctement « non-prédit » de multiples fois. Pour donner une vision plus équitable de l'imputation pour ces allèles, il convient de regarder d'autres métriques, comme le F_1-score (Tableau VI-1).

Tableau VI-1 Les métriques d'évaluation de la prédiction. Chacune des métriques est complémentaire aux autres et communique des informations sur certaines propriétés de la prédiction. Pour l'imputation HLA, la sensibilité donne la proportion de prédictions correctes d'un allèle spécifique.

		Allèle prédit			
	Population totale	**Positif** *PP*	**Négatif** *PN*		
Allèle réel	**Positif** *P*	**Vrai Positif** *VP*	**Faux Négatif** *FN*	**Sensibilité** $\frac{VP}{P}$	**Taux de faux négatifs** $\frac{FN}{P}$
	Négatif *N*	**Faux Positif** *FP*	**Vrai Négatif** *VN*	**Taux de faux positifs** $\frac{FP}{N}$	**Spécificité** $\frac{VN}{N}$
	Prévalence $\frac{P}{(P+N)}$	**Valeur Prédictive Positive (VPP)** $\frac{VP}{PP}$	**Taux de fausses omissions** $\frac{FN}{PN}$		
	Précision $\frac{(VP+VN)}{(P+N)}$	**Taux de fausses découvertes** $\frac{FP}{PP}$	**Valeur Prédictive Négative** $\frac{VN}{PN}$		
	Précision balancée $\frac{Sensibilité + Spécificité}{2}$	**F1-score** $2 \times \frac{VPP \times Sensibilité}{VPP + Sensibilité}$	**Index de Fowlkes-Mallows** $\sqrt{VPP + Sensibilité}$		

Le F_1-score vérifie à la fois la proportion des occurrences de l'allèle voulu qui ont été prédites (Sensibilité) et la proportion des prédictions de présence de l'allèle qui sont vérifiées (Valeur Prédictive

Positive) : cette métrique ne prend plus en compte les non-prédictions et est donc appropriée pour les allèles rares.

Mon second article a été construit sur les connaissances précédemment établies sur l'importance de la proximité génétique entre la population du panel de référence et celle que l'on cherche à imputer, ainsi que sur l'utilisation d'une métrique appropriée pour évaluer l'imputation HLA. Nous nous sommes concentrés sur l'imputation de la population CAAPA, d'ancestralité européenne-africaine-autochtone des Amériques, à partir des données 1KG qui sont multi-ethniques mais avec un ensemble limité d'individus afro-américains.

L'hypothèse de travail est que des individus avec des génotypes SNP similaires dans la région du CMH ont des génotypes HLA similaires. Nous avons donc créé des panels de référence en sélectionnant des individus de 1KG proches génétiquement de ceux de CAAPA, en se basant sur des projections par ACP et UMAP. Ces projections ont été effectuées avec des SNP du CMH ou du chromosome 6, sans enlever ceux en déséquilibre de liaison. Cela peut modifier certaines valeurs de contributions pour les axes de la réduction de dimension mais conserve l'information globale d'ancestralité. Par ailleurs, le chromosome 6 est représentatif du reste du génome dans une représentation par réduction de dimension. Dans certains cas, nous avons séparé CAAPA en plusieurs sous-échantillons de validations, lorsque la réduction de dimension indiquait la présence de différents groupes. Nous avons comparé ces modèles aux descriptifs ethniques classiques (Africain, Européen), ainsi qu'à des modèles multi-ethniques. Nos résultats ont montré l'intérêt de notre approche à nombre d'individus proches dans les panels, et notamment pour certains allèles spécifiques. D'un point de vue global, l'utilisation d'un panel de référence plus grand et multi-ethnique s'est avérée toujours supérieure en termes de précision d'imputation.

VI.2.2 - Article – Explorer l'imputation HLA des populations d'ancestralité composite avec la réduction de dimension

Exploring HLA imputation of admixed population with dimension reduction

Venceslas Douillard[1], Nayane dos Santos Brito Silva[1,2], Sophie Limou[1], Pierre-Antoine Gourraud[1], Élise Launay[3], Erick C. Castelli[2], Nicolas Vince[1], on behalf of the SNP-HLA Reference Consortium (SHLARC)

1. Nantes Université, INSERM, Ecole Centrale Nantes, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 Nantes, France.
2. São Paulo State University, Molecular Genetics and Bioinformatics Laboratory, School of Medicine, Botucatu, State of São Paulo, Brazil.
3. Department of Pediatrics and Pediatric Emergency, Hôpital Femme Enfant Adolescent, CHU de Nantes, Nantes, France.

Corresponding author
nicolas.vince@univ-nantes.fr
Nicolas Vince
CR2TI UMR1064 – ITUN, CHU Nantes Hôtel Dieu, 30 bld Jean Monnet, 44093 Nantes Cedex 01, France, +33 2 40 08 74 24

Abstract

Human genomics quickly evolved in the last decade thanks to technological advances in genotyping and sequencing, continuously fueling the growth of genome-wide association studies. Those studies repetitively brought light on the Major Histocompatibility Complex (MHC), and especially the HLA molecule which is key in immunity, for its association in infectious and autoimmune pathologies. However, the wide genetic diversity of the MHC technically hinders its investigation, therefore, statistical inference methods of the HLA were developed. The SNP-HLA Reference Consortium (SHLARC) aims to improve HLA imputation methods which are currently optimized for European populations. In this work, we trained reference panels on the 1,000 Genomes (1KG) dataset using HIBAG and evaluated the accuracy of these models on the 880 individuals of admixed European and African ancestries of the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). We investigated multiethnic, superpopulation based and specific reference panels potency to impute HLA genotypes of admixed individuals. Specific reference panels were computed with 1KG individuals considered to be the closest to the target population according to a dimension reduction method, PCA or UMAP. The biggest multi-ethnic (n=2,504) reference panel always outperformed the rest of the models, with 0.66 of F1-score in HLA-B. Nevertheless, in HLA-B again and considering models with a similar sample size (n=200-485), specific models consistently outperformed multiethnic or superpopulation reference panels with F1-scores up to 0.53, against F1-scores up to 0.42 for superpopulation references. Additionally, two specific models using UMAP and PCA showed promising results and were able to better predict specific HLA alleles, even when compared to the full 1KG reference panel (e.g. *HLA-A*02:04*). In this article, we demonstrated the real interest of using restrained but genetically specific reference models in the imputation of admixed population which are currently underrepresented. The SHLARC opens the door to HLA imputation for every genetic population. Along with additional HLA studies, it helps identifying the biological mechanisms linking HLA and pathologies.

Keywords: HLA imputation, multi-ethnic, reference panels, HIBAG, UMAP, PCA, dimension reduction

Introduction

Genome-wide association studies (GWAS) have now become a strong ally in the understanding of biology, with historic associations such as rs2395029 in HIV (1,2), or the 233 variants linked to multiple sclerosis (3); but GWASs have also been performed as first lines of research at the beginning of the SARS-CoV-2 pandemic, to evaluate the genetic part of the infection and associated phenotypes (4–6). Starting from the first GWASs of hundreds of individuals in the 2000s (7,8), multiple initiatives emerged in the last decade seeking to systematically gather clinical and genetic information, such as the UK Biobank (9), Japanese BioBank (10,11), or TOPMed (12), which gather hundreds of thousands of samples. These studies greatly improved the comprehension of the genetic impact in phenotype variation (13–15). Along with the collective organization effort, continuous advances in the domain of Single Nucleotide Polymorphism (SNP) imputation (16), and the availability of computing power from imputation servers globally helped the genomics community (17).

A bystander effect of these studies is the confirmation of the central role of the Major Histocompatibility Complex (MHC), and especially *HLA*, in immune-related diseases. The MHC was discovered in the 1950s by Jean Dausset (18) and it was identified as crucial for transplantation procedures (19). Association studies expanded our understanding of the role of MHC: the GWAS catalog counts 2.5% of all significant associations in the MHC and approximately 20% of all traits associated with at least one SNP within the MHC (6); the associations go from auto-immune diseases such as type 1 diabetes (20) or multiple sclerosis (3), neurological disorders such as Parkinson (21), to infectious diseases such as HIV (1,22), Hepatitis B (23,24) and C (25).

In this genetic association context, a parallel effort focused on direct association with *HLA* genes in order to understand the mechanisms in which HLA molecules impact human organisms. Those HLA association studies allowed to characterize protective and risk alleles, such as *HLA-DRB1*15:01* in multiple sclerosis (26), *HLA-DRB1*09:01* with tIgE levels in asthma (27), specific HLA-DQB1 amino acids in hepatitis C virus infection (28) or the HLA-DRB1 valine 11 in Parkinson's disease(29). The five most polymorphic HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1*) are exceptionally diverse, with almost 30,000 alleles combined (30), however the majority of them seem to have frequencies <1% and lower (31). Therefore, HLA typing of thousands of individuals is needed to reach a sufficient statistical power for association. The cost-efficiency of directly typing *HLA* for such cohorts is limited, thus, following the steps of SNP association, the HLA community organized multiple typing initiatives and developed imputation tools (32,33). Several *HLA* imputation tools allow to create reference panel for imputation as well as imputing HLA genotypes from SNP data: HIBAG (34) and SNP2HLA (35) are the most common choices and Pappas et al. evaluated HIBAG to be the most accurate (36). A new generation of software followed, with improvements to existing algorithms such as HLA*IMP:03 (37) and CookHLA (38), or using deep learning with DEEP-HLA (39), all of which will probably gain traction over time.

However, every *HLA* imputation results are dependent on the reference panel used to predict the target genotypes; if training and target data are not of the same ancestry, it will provide inaccurate results due to different HLA alleles and linkage disequilibrium pattern between SNP and HLA. To solve this issue, researchers advocated for both: specific reference panels such as in Japan (40), Finland (41), or SweHLA (42), and large multi-ethnic reference panels (43,44). To pursue the different efforts, we created the SNP-HLA Reference Consortium, or SHLARC (45). Our goal is to coordinate an international effort to gather HLA data and reference panels, make them available for the scientific community and improve methodology of *HLA* association studies. In general, HLA imputation works well for European origin populations as a large amount of data are available. However, the challenge is higher when focusing on admixed or underrepresented populations; indeed, fewer data are available and a clear

HLA imputation strategy is required to improve accuracy. In this study, we focused on the results of *HLA* imputation on admixed populations and investigated dimension reduction as a method to mitigate HLA imputation errors on rare alleles, by relying on the existing reference of the 1,000 Genomes Project (1KG) (46,47).

Results

HLA imputation strategy

The accuracy of HLA imputation for a target population heavily depends on the data diversity used as reference. The aim of our study is to find the preferred HLA imputation methodology when dealing with a target population whose ancestry is absent or underrepresented in the available training data. The 1KG dataset is one of the most diverse public dataset with 2,504 individuals from 26 populations (46,48), these populations are grouped in 5 super-populations, as described in Table 1: African (AFR), American (AMR), European (EUR), East Asian (EAS), South Asian (SAS). In the past years, the HLA genotyping for the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1* genes was published using an HLA calling from whole genome sequencing data (49), moreover, the SNP data has been updated with a new whole genome sequencing of 30X coverage from the New York Genome Center (47). We selected these data as training dataset to create 395 reference panels to be tested (Figure 1): 1) the full dataset (full1KG, N=2,504), 2) 10 replications from 6 conditions (1KG, AFR, AMR, EUR, EAS, SAS; N=200 for each), 3) 19 conditions for the custom reference panels (further described in the next chapter; 200<N<485); each condition replicated 5 times for each *HLA* gene (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1*).

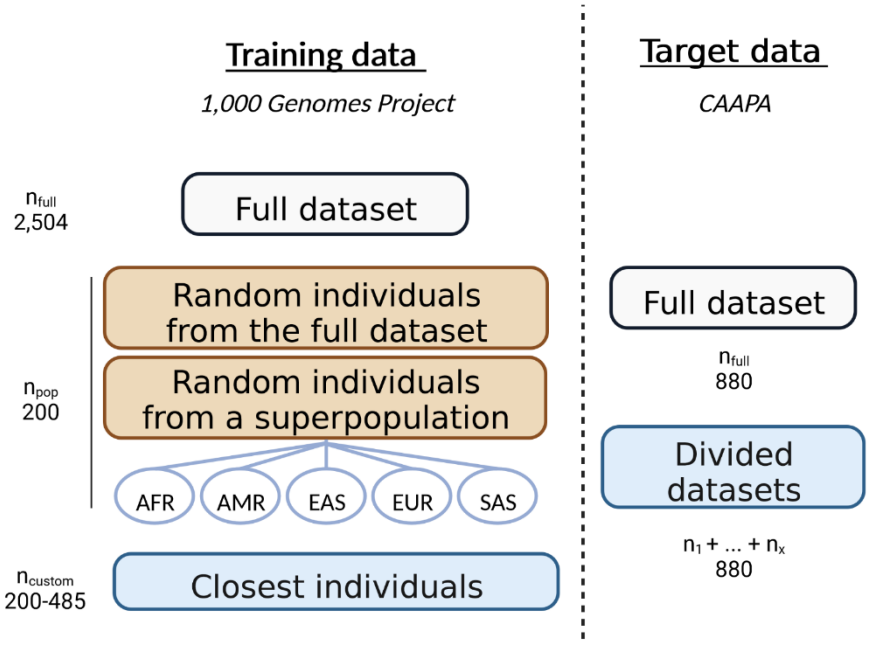


Figure 1 Selection strategy: description of the dataset selection for training and testing. Different subsets of the 1KG dataset are used as reference, selected by super-population or from genetic proximity with CAAPA dataset. The HLA genotypes from CAAPA are either imputed from a single model, or by multiple models specific to subsets of CAAPA.

CAAPA (Consortium on Asthma among African-ancestry Populations in the Americas) was created to study asthma in African-ancestry populations. From this study, we have access to 880 individuals with whole genome sequencing data of the *MHC* region as well as HLA genotypes. The CAAPA data is composed of individuals from admixed European and African ancestry, with various proportion (27), therefore we used these data to evaluate the pertinence of our models. Only a small part of these populations ancestries are represented in the 1KG dataset, and we also wanted to evaluate the impact of admixture in the imputation process. Thus, the CAAPA population was alternatively considered as a

unique dataset of 880 individuals, or as multiple subsets of it, depending on the representation with dimension reduction methods.

Table 1 Repartition of 1KG individuals in the 5 super-populations

Super-population	Population
AFR - African (n_{AFR}=661)	**ACB** - African Carribean in Barbados (n_{ACB}=96)
	ASW - African ancestry in SouthWest US (n_{ASW}=61)
	ESN - Esan in Nigeria (n_{ESN}=99)
	GWD - Gambian in Western Division, The Gambia (n_{GWD}=113)
	LWK - Luhya in Webuye, Kenya (n_{LWK}=99)
	MSL - Mende in Sierra Leone (n_{MSL}=85)
	YRI - Yoruba in Ibadan, Nigeria (n_{YRI}=108)
AMR - American (n_{AMR}=347)	**CLM** - Colombian in Medellin, Colombia (n_{CLM}=94)
	MXL - Mexican Ancestry in Los Angeles, California (n_{MXL}=64)
	PEL - Peruvian in Lima, Peru (n_{PEL}=85)
	PUR - Puerto Rican in Puerto Rico (n_{PUR}=104)
EAS - East Asian (n_{EAS}=504)	**CDX** - Chinese Dai in Xishuangbanna, China (n_{CDX}=93)
	CHB - Han Chinese in Beijing, China (n_{CHB}=103)
	CHS - Southern Han Chinese, China (n_{CHS}=105)
	JPT - Japanese in Tokyo, Japan (n_{JPT}104)
	KHV - Kinh in Ho Chi Minh City, Vietnam (n_{KHV}=99)
EUR - European (n_{EUR}=503)	**CEU** - Utah residents with Norther and Western European ancestry (n_{CEU}=99)
	FIN - Finnish in Finland (n_{FIN}=99)
	GBR - British in England and Scotland (n_{GBR}=91)
	IBS - Iberian populations in Spain (n_{IBS}=107)
	TSI - Toscani in Italy (n_{TSI}=107)
SAS - South Asian (n_{SAS}=489)	**BEB** - Bengali in Bangladesh (n_{BEB}=86)
	GIH - Gujarati Indian in Houston, Texas (n_{GIH}=103)
	ITU - Indian Telugu in the UK (n_{ITU}=102)
	PJL - Punjabi in Lahore, Pakistan (n_{PJL}=96)
	STU - Sri Lankan Tamil in the UK (n_{STU}=102)

Data selection for customized HLA imputation

We created models with individuals from 1KG genetically close to the CAAPA target data: the custom models. We decided to rely on dimension reduction, which is common in population genetics, to assess individuals' ancestry. The goal is to select 200 individuals from 1KG closest to the target data, regardless of their self-assessed ancestry. Classically, ancestry is assessed with whole genome SNPs by Principal Component Analysis (PCA), however since we focused our study on HLA, we decided to represent the populations using only SNPs within the MHC region (29-34Mb). This strategy of representation separated the African population along with a portion of the American population in one part, and the rest of 1KG on the other part (Figure 2A). The usual granularity of PCA on whole genome genotypes (Figure S1A) is not obtained and does not allow to finely group ancestries.

However, with a two-dimension UMAP (Uniform Manifold Approximation and Projection) of the MHC region (Figure 2B), we could identify well separated groups.

To fully investigate the effect of dimension reduction on HLA imputation, we tested 3 parameters for representation: the algorithm (PCA or UMAP), the number of dimensions used (2 or 10), and the

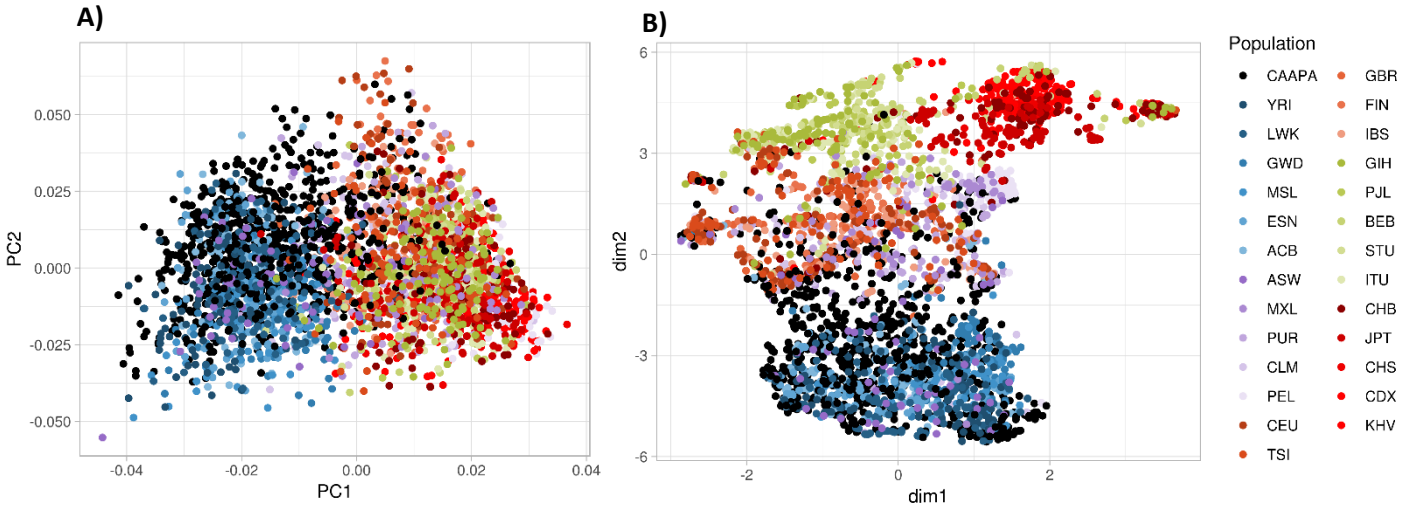


Figure 2 PCA (A) and UMAP (B) representation of 1KG and CAAPA dataset with merged genotypes of the MHC region. CAAPA is represented in black, super-populations are colored in five main colors divided in different shades for each population (Table 1), including 5 super-populations: African (AFR) in blue, American (AMR) in purple, European (EUR) in orange, South Asian (SAS) in green, East Asian (EAS) in red. PCA does not separate the population well when restrained to the MHC region, whereas UMAP creates different groups of ancestries.

genomic region covered by the genotypes dataset (the whole chromosome 6 or the MHC region). The different conditions are named after the combination of these parameters: for instance, a selection of the training data based on a UMAP, using the distance computed in 10 dimensions on every SNP available on the chromosome 6 is named UMAPnotMHC_10D.

We performed a silhouette score analysis (see Methods) to the resulting projection of the CAAPA dataset and identified that in every UMAP condition and with the ten-dimensions PCA in the MHC region, we could cluster CAAPA in more than one group. In those cases, we decided to create one model per group. We computed the average coordinates of the CAAPA individuals, then selected the 200 individuals from 1KG closest to this point (Figure 3). To avoid redundant models, we checked the overlap of selected individuals between the conditions, surprisingly, they all yielded a unique list of 1KG individuals, with low overlap between conditions. (Figure S2). For the conditions where the CAAPA dataset was separated into different subsets, we imputed the individuals separately, thus relying on multiple models, but merged the results into one table. As an example, with the two-dimension UMAP representation (genotypes from MHC region), we computed three different models of 200 1KG individuals (Figure 3). We then imputed three CAAPA groups independently (357, 344 and 179 individuals for a total of 880) and combined results into a unique table of imputation for the full dataset.

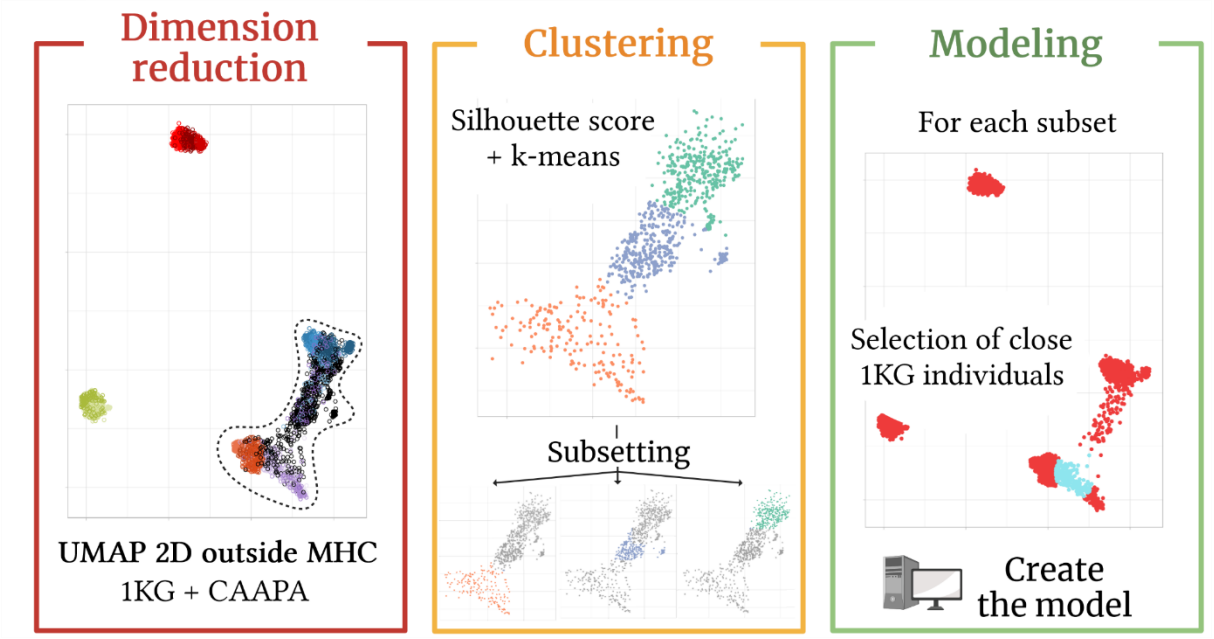


Figure 3 Creation of custom 1KG models for CAAPA imputation. 1) Dimension reduction allows to separate individuals according to ancestry, here using UMAP as an example. It is also possible to apply dimension reduction to one dataset and project another onto it. 2) The target dataset, here CAAPA, can be represented as homogeneous or not, the silhouette score allows to evaluate the preferred number of clusters, then k-means allows for subsetting. 3) The barycenter of each subset is computed, then 1KG individuals closest to the point are selected, allowing to create a custom model. Created with biorender.com

CAAPA HLA imputation comparison between classical and custom reference panels from 1KG data

We compared the different conditions by averaging the F1-score of each allele. As explained by Cook et al. (38), the F1-score has an advantage over other accuracy metrics on representing the rare alleles as it is the mean of two metrics, taking into account both the potential under- and over-prediction of an allele (see Methods). As expected, the full1KG model (N=2,504) displayed the highest F1-score in every HLA gene imputation, with scores ranging from 0.64 in *HLA-DRB1* to 0.87 in *HLA-C* (Figure 4). In *HLA-B*, full1KG has a 0.66 F1-score. However, when considering the smaller models, we found that the 1KG models (F1-score 0.42) as well as the super-populations with close ancestry to CAAPA (AFR: F1-score 0.37, AMR: F1-score 0.42) had nominally lower F1-score than some custom models (PCAnotMHC_10D: F1-score 0.52, UMAPnotMHC_10D: F1-score 0.53). This trend was also found in *HLA-A* and *HLA-DRB1*, the two other most polymorphic HLA genes. *HLA-C* and *HLA-DQB1* show a higher mean F1-score for 1KG. Low F1-scores are expected for models computed with a reduced number of individuals, however we also observe an F1-score close to 0.6 in the full dataset. F1-scores are not to be interpreted as regular accuracies, as a matter of fact, when the same methodology is applied and average accuracies of each allele are computed, these accuracies obtain more than 98% (represented as error rates in Figure S3). Additionally, individual and haplotype accuracies, respectively corresponding to the proportion of correct genotypes (individuals can be counted as 0 or 1), and the proportion of correct allele (individuals can be counted as 0, 0.5, or 1) also show values above 80% (Figure S4).

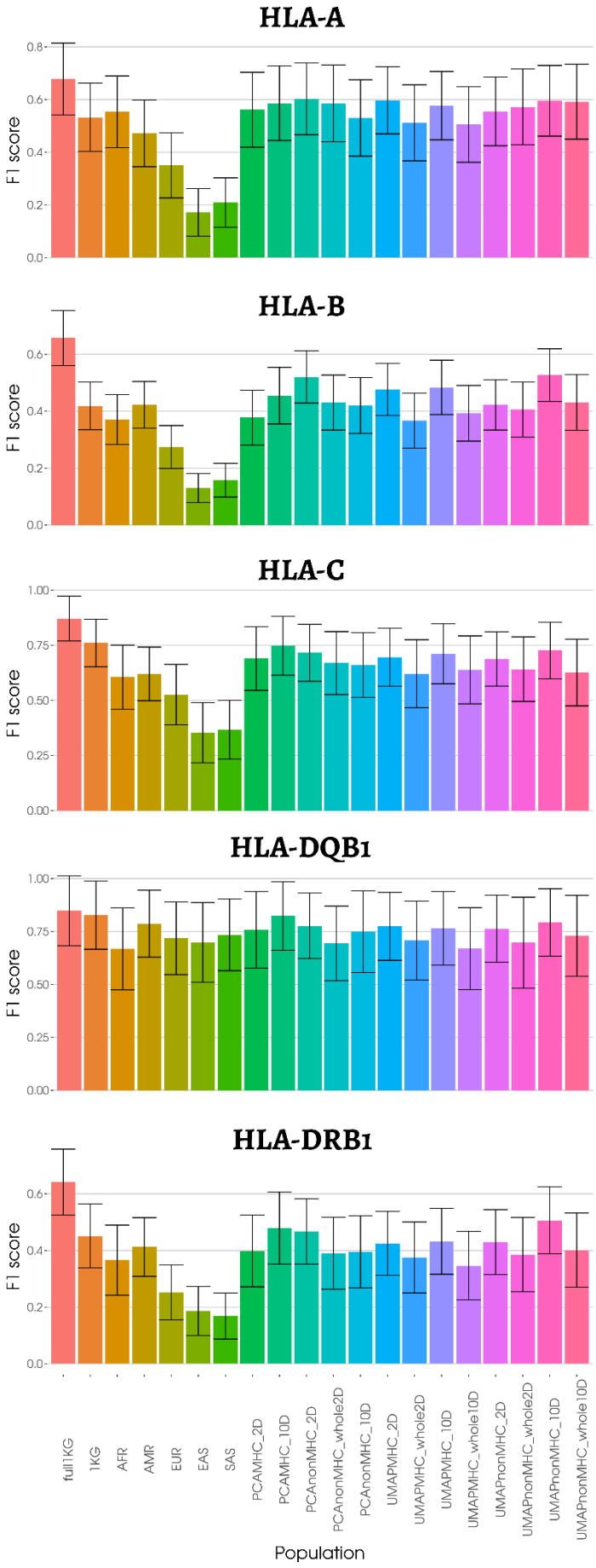


Figure 4 Average F1-score of HLA allele predictions for HLA-A, HLA-B, HLA-C, HLA-DQB1, and HLA-DRB1 based on imputation of CAAPA dataset, with training models ranging from the full 1KG dataset, to the defined super-populations, to selections of individuals based on the proximity to CAAPA (using different parameters). Alleles absent from the training datasets were removed to obtain these values.

To investigate the impact of those custom models on imputation and why they seemed to perform better for polymorphic genes, we stratified the mean F1-score metric by HLA allele frequency (Figure 5). The model full1KG (N=2,504) yields consistently higher F1-score through all allele frequencies. Custom models performed equally or marginally better for the rarest alleles (frequency <= 0.1%) and the most common alleles (frequency >10%) but had higher score for every other category compared to super-population models. For *HLA-B*, UMAPnotMHC_10D (N=485) has a F1-score of 0.30, 0.70, 0.85 and 0.91 for the categories from 0.1 to 10% frequency, whereas the multi-ethnic model (1KG, N=200) shows scores of 0.18, 0.45, 0.78 and 0.91. Notably, the reference panel based on the African

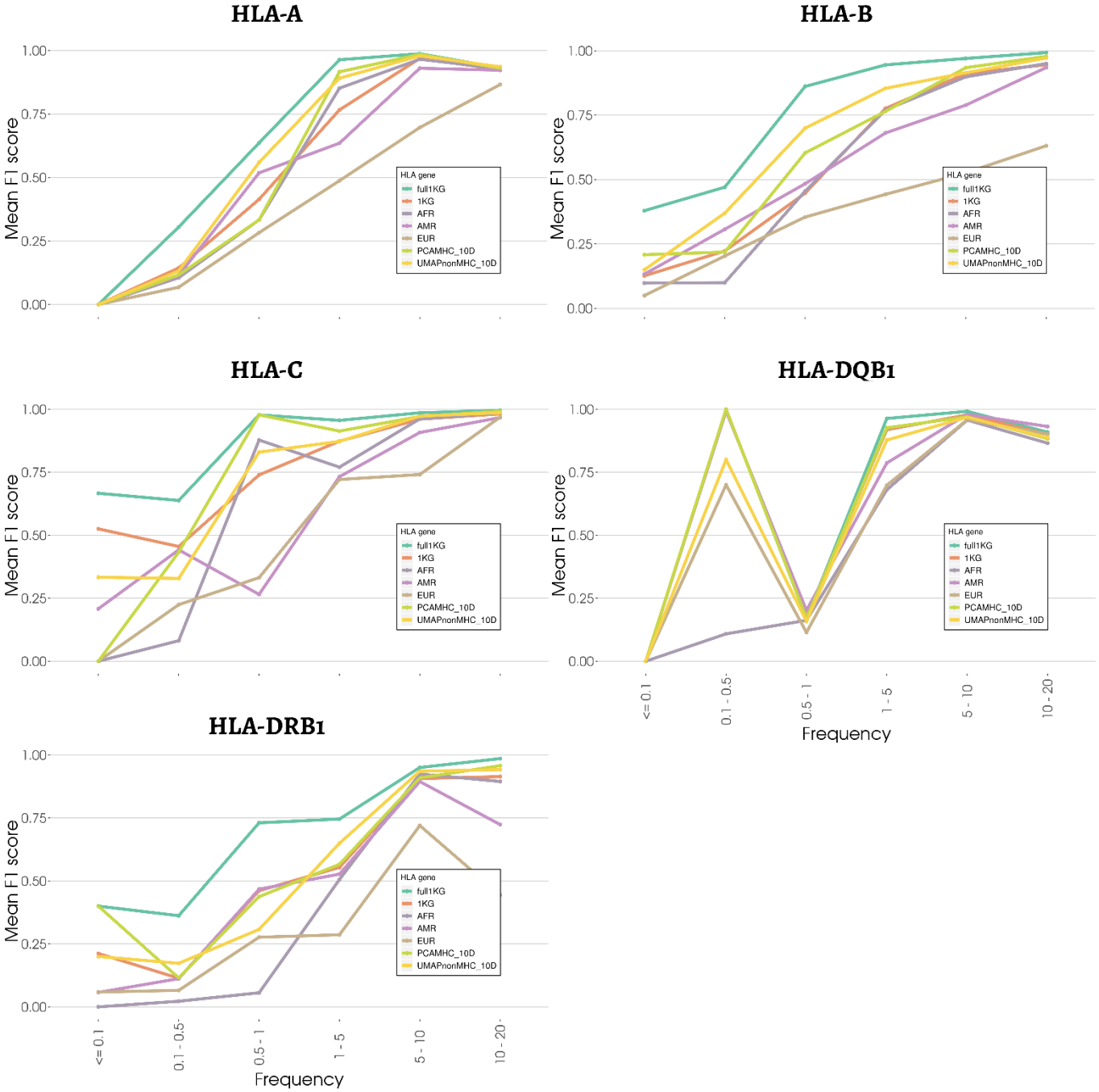


Figure 5 Mean F1-score of HLA alleles imputation, stratified by groups of frequency, for the full 1KG model, super-populations models, and a selection of custom models The custom models PCAMHC_10D and UMAPnotMHC_10D are displayed as they are the most accurate. HLA-C does not have the 10-20% frequency category, therefore, unlike other genes the last two categories correspond to 5-10% and >20%.

superpopulation performs worse for *HLA-DQB1*. It can be explained by the allele HLA-DQB1*06:01 which is represented only once, therefore it has a 0.1 F1-score.

The results show that creating custom reference panels based on a genotypic distance between individuals can improve the outcome compared to multi-ethnic or declared ancestry panels, but larger multi-ethnic reference panels are always more robust. We went further and looked directly at the imputation of *HLA* alleles individually.

When we analyze results allele by allele, taking *HLA-A* (Figure 6) as example, we observe that in most cases custom models performed just as well, or a few points under the full dataset models (e.g. *HLA-A*01:01*, *HLA-A*23:01*). Multiple *HLA* alleles were better predicted with the custom models compared to the multi-ethnic (1KG) and super-population models (*HLA-A**01:02, *HLA-A**80:01), highlighting the interest in the creation of specific reference panels. We found cases where the full1KG model (N=2,504) or super-population models (N=200) were the only ones to predict the allele (e.g. *HLA-A**02:06, *HLA-A**03:02), however, we also found cases where custom models were the only ones to impute correctly the allele (e.g. *HLA-A*02:04*). Zheng et al. (34) showed that 10 copies of an allele were needed in a model in order to be able to impute them. Nine *HLA-A* alleles were present in both the training and target data but were not imputed at all, by any of the models (e.g. *HLA-A*02:11*, *HLA-A*24:03*, *HLA-A*26:08*). Often, the allele was present in only a few individuals in the target data, causing the miscalled allele to weigh a lot in the score. We focused on *HLA-A* for visualization purposes, but the results apply to *HLA-B*, *HLA-C*, and *HLA-DRB1* (Figure S5). Interestingly, in *HLA-DQB1*, we did not find an allele where the best training dataset was not the full dataset, or 200 individuals from 1KG; for *HLA-DQB1*03:01* however, the AMR and EUR super-populations yielded better results. These last examples show how a custom reference panel could help in the imputation of certain HLA alleles, however, since bigger models produce better imputation results overall we would need to be able to know when to select the results from the custom reference panel.

HLA imputation with HIBAG yields post-probabilities for each genotype. We tried to harness the few cases where custom models performed better (in terms of post-probabilities) to obtain hybrid imputation between the full models and the custom model. We chose UMAPnotMHC_10D as it performed the best on multiple *HLA* genes. Unfortunately, the small number of samples in the custom models leads to lower post-probabilities compared to the full model. In the few cases where UMAPnotMHC_10D yielded better post-probabilities, the imputed genotype was not always correct, whereas the less likely genotype imputed by the other model was correct. In a real situation where the HLA alleles of the target data would not be known, there would be no way to choose between the imputed genotype of the two models (Figure S6).

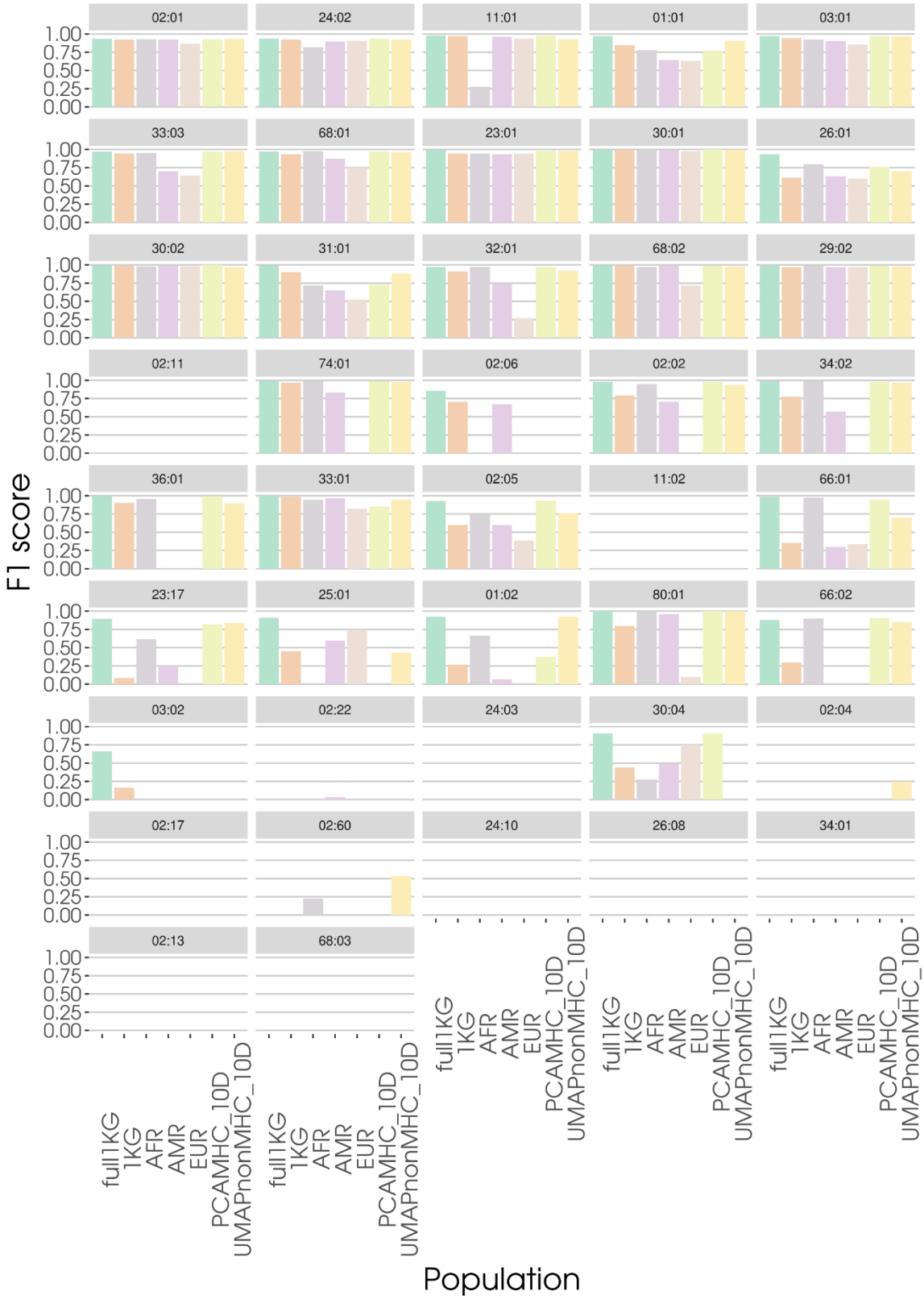


Figure 6 Mean F1-score of each HLA-A alleles (N=42) for the full 1KG training dataset, the African, American, European super-populations datasets, and the most accurate custom reference panels PCAMHC_10D and UMAPnonMHC_10D. Alleles are ordered by decreasing frequency in the 1KG dataset. Those absent from the training dataset have been removed to compute the means, keeping them uniformly lower the results.

Replication with admixed Brazilian individuals from SABE

We replicated our methodology on the longitudinal Health, Well-Being, and Aging cohort (SABE - *Saúde, Bem-estar e Envelhecimento*) to validate the impact of the composition of the models on *HLA* imputation (Figure 7). SABE is an independent dataset of 1,322 individuals from Brazil mostly with African and European admixed ancestry (50). To validate our conclusions, we used the same models as with the CAAPA data, therefore, between 11.6% and 45.1% of the model SNPs were missing in the target data.

Though it probably reduced the imputation score overall, the missing SNP were homogeneous across conditions for each gene, with averages of: 30,0% for *HLA-A*, 14,3% for *HLA-B*, 13,9% for HLA-C, 39,4% for HLA-DQB1, and 39,6% for HLA-DRB1. We also limited our study to the PCAMHC_10D and UMAPnotMHC_10D custom models, as these two models predicted *HLA-A*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1* better, out of all the custom models in the CAAPA dataset.

As with CAAPA, the custom models had nominally higher F1-score than the 1KG model, but only for the *HLA-B* (0.44, 0.50 for PCA and UMAP vs. 0.42 for 1KG) and HLA-DRB1 (0.56 for UMAP vs. 0.48 for 1KG). Overall, the validation with the SABE population showed the same patterns as the CAAPA

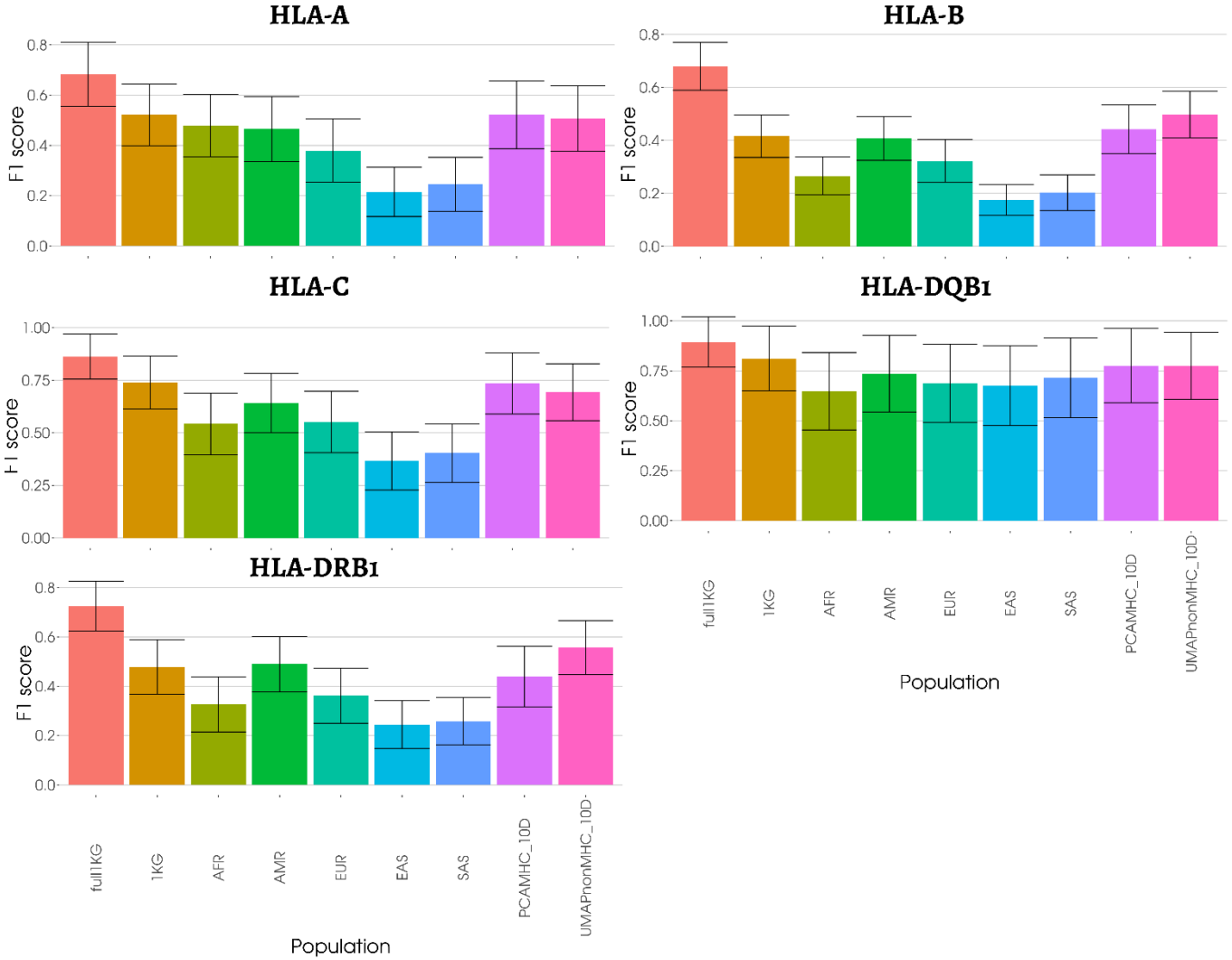


Figure 7 Mean F1-score of SABE's imputed HLA-A, HLA-B, HLA-C, HLA-DQB1 and HLA-DRB1 genotypes, using the full 1KG model, compared to super-populations from 1KG or individuals selected by dimension reduction. Alleles absent from the training dataset have been removed to compute the means, keeping them uniformly lower the results.

population, with a global preference of the full model and multiple cases where the custom reference panels were to be preferred, but presented low post-probabilities genotypes.

Discussion

Our results are a part of a broader dynamic in the HLA community and with the SHLARC whose goals are to provide immunogeneticists reliable tools and reference panels for HLA imputation, thus increasing the power of *HLA* association studies to that of existing GWASs. Our work focuses on improving existing methods of *HLA* imputation by finely accounting for ancestry in the choice of the training model. Our underlying hypothesis was that oversampling individuals with close genetically ancestry to our target individuals, in the reference panel, would increase accuracy for rare *HLA* alleles. In this line of research, we chose to evaluate the imputation of CAAPA, an admixed African-American cohort, using reference panels composed of different combinations of 1,000 Genomes Project individuals: randomly selected, from a super-population or selected for their estimated ancestry by dimension reduction. We showed that ultimately, the number of individuals was the crucial point of HLA imputation. Indeed, the reference panel composed of 2,504 individuals from 1KG systematically had a higher F1-score than other smaller models. Using less individuals for training however, selecting individuals close to the ancestry of the target population was a good strategy and resulted in slightly better HLA imputation F1-scores, compared to multi-ethnic reference panels. The improvement does not concern the rarest or most common alleles which are respectively badly and well imputed by all those models. At the allele level, we expected the full model to impute *HLA* alleles other models would not; we also saw the opposite with custom reference panels capturing a part of the information that was left out in the full model. Unfortunately, we could not conclude on its applicability since the custom reference panels had fewer individuals resulting in lower post-probabilities that rendered a hybrid imputation impossible. Research on SNP to SNP imputation also encounters the problem of lack of diversity for the imputation of rarer alleles, and are working with specific reference panels to enhance imputation accuracy (51,52).

Interestingly, we were also able to make use of UMAP for genomic ancestry representation, as can be also seen in recent research (53–55). It presented a good separation of ancestry groups in two dimensions when using the MHC only, concordant with the frequency difference of *HLA* alleles between populations (56), whereas PCA would fail to clearly separate them in the same conditions. PCA uses SNPs which explain the most variance, unlike UMAP which tries to preserve the topography of the higher dimensions in its reduction, taking into account every SNP available for distance. Besides, we observe distance between individuals that are sometimes higher inside a labeled 1KG population than between population as described in Maróstica et al. (56,57). This representation of this genomic diversity inside the MHC directly impacts how we should construct reference panels in the future and highlights the importance of gathering more data from different ancestry background.

Our work shows the potential interest of population-specific reference panels, as multiple studies showed (40–42,58–62), but strayed further from the geographic definition of the population and rather tried to find a local definition of ancestry to select training datasets. While doing so we also omitted potential sides to the problem and creating limits to our method.

One major setback to *HLA* imputation compared to typing is the impossibility to predict *de novo* alleles, and the difficulty of imputing rare alleles. It is intrinsic to all training machine learning methods and it is especially a problem in HLA, where each gene can have thousands of alleles. In HIBAG for instance, an allele should be present at least 10 times in the training dataset in order to be predicted (34). We saw in this study that this limit can be overcome to a certain extent but still hinders HLA imputation. Additionally, the choice to limit the number of randomly selected individuals was directly linked to the maximum of samples in the smallest super-population (n_{AMR}=347). However, it has led to low

imputation scores. Even though we performed replications, the difference between super-population models and the full dataset, or the custom models, may greatly vary if we were to increase this limit with another multi-ethnic dataset. It is one potential improvement to this work, which may validate or not our findings.

We chose to represent the HLA imputation with the F1-score accuracy as seen in Cook et al. (38). This choice is convenient for the analysis of HLA which we encounter at low frequencies, unbalanced frequencies between the different alleles. We chose to set the F1-score at 0 when a specific allele was not imputed at all (whereas F1-score should be null) in order to represent all alleles in common between the two datasets and weigh negatively this absence of imputation. It has increased the confidence interval of each averaged F1-score and limited the possibility to find statistical differences between them. It is important to note that the F1-score gives a harsher view on HLA imputation because rare alleles have low scores, however, HLA imputation performs very well for common alleles (Figure S3) (32).

Beside methodology, HLA imputation gains a lot of accuracy from the number of samples and the diversity in the reference panels, this is why initiatives looking into expanding the HLA data and creating bigger reference panels, such as Degenhardt et al., are essential to the field (43,44,49). With the SHLARC (45), we advocate for a coordination of such efforts in order to provide multi-ethnic panels of sufficient size, and help researchers do HLA imputation to investigate HLA risk and protection alleles, focusing on the coverage of the globe for data gathering. The evolution of imputation tools will also consequently improve HLA imputation. HLA-IMP*03 (37) and CookHLA (38) showed improved results over the algorithms they are created upon and DeepHLA (39) also showed high accuracy, with a specific focus on rare HLA alleles. Eventually, these efforts will reach a limit and we think the main focus of research should be gathering data around the world.

Conclusion

The SNP-HLA Reference Consortium (SHLARC) wants to contribute to the HLA association analysis community by providing a platform for HLA imputation with exhaustive and diverse reference panels. We hope this will help for association studies to rapidly increase their statistical power and become a natural extension of genome-wide association studies pointing towards HLA association.

Methods

Data processing

Genotyping data of 1KG, CAAPA, and SABE genotypes obtained from whole genome sequencing (49,50,63) were handled with PLINK v1.90b6.21 (64) and went through the same quality control step: removal of A/T and G/C ambiguous SNPs, >2% missing genotypes, and <1% minor allele frequency SNPs were removed. HLA data comprises of two-field alleles for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1* and *HLA-DRB1*, stored in a CSV file.

HLA imputation models were computed on R 3.5.3 with HIBAG v1.19.3 (34) and its complementary package HIBAG.gpu v0.9.1. Training data were subsetted with PLINK to contain only the SNPs also present in the target data for CAAPA. HLA imputation of the SABE data was done with the previously computed models, except for custom models.

HLA imputation metrics

The F1-score is a harmonic mean of sensitivity (for a specific allele, # of correctly predicted allele/# of said alleles in the target dataset) and the positive predictive value (for a specific allele, # of correctly predicted allele/# of predictions of said allele). This score has the property to give an important weight to the coverage of a specific allele prediction. For instance, if a rare allele is present once in a dataset of 100 alleles and not predicted by the model, you would have a 99% accuracy but a F1-score of 0.

HLA imputation models are limited by the pool of HLA alleles in the training dataset, contrary to software based on read alignment which rely on the complete database of known HLA alleles. Therefore, we chose to average the results of all alleles that were both present in the training and target datasets. Additionally, if one of these alleles is not predicted by the model, the PPV, by definition, cannot be computed; in this case, the F1-score is also null. Since we wanted to focus our analysis on rare alleles, we decided to set the F1 scores of such alleles to 0, to visualize the impact of HLA alleles that are in the training dataset but do not manage to impute the ones in the target data.

Dimension reduction

Principal Component Analysis (PCA) is routinely used in population genomics and association studies to study population ancestry. It relies on SNPs to which are attributed different contributions, maximizing the variance in their genotypes. It allows to separate populations along multiple orthogonal axes with different contributions for each SNP. Uniform Manifold Approximation Projection (UMAP), along with t-SNE, is central in single-cell transcriptomics analyses (65,66). Recently, It has also appeared in population genomics publications (53,54). UMAP is based on simplicial topology, so that it identifies sets of neighbors for each individual and tries to preserve them while transforming coordinates into new ones with less dimensions.

We performed dimension reduction after merging 1KG and CAAPA data. We ran PCA with PLINK, and UMAP on the BiRD cluster from Nantes University, using the umap R package. This package does not handle missing data; therefore, we applied the PLINK geno filter with a 0 threshold beforehand to remove any SNP with missing data. We followed the same process with SABE but merged the dataset with both 1KG and CAAPA.

We applied a silhouette score on the coordinates of the CAAPA individuals to identify the preferred number of clusters. We then performed k-means with the number of clusters which had the highest silhouette score. If the maximum score was inferior to 0.4, we chose not to perform clustering because simulations showed different groups would be greatly overlapping.

Acknowledgements

We would like to thank the Centre de Calcul Intensif des Pays de la Loire, for the computing available that was available to us. This work was supported by the ANR PIA-Investment (NExT, SHLARC Project, Nantes Université), the ATIP-Avenir Inserm program, the Région Pays de Loire ConnectTalent. Venceslas Douillard has received funding from the Inserm and Région Pays de la Loire. Nicolas Vince has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 846520.

References

1. Fellay J, Shianna K V, Ge D, Colombo S, Ledergerber B, Weale M, et al. Study of Major Determinants for Host Control of HIV-1. Science (80-) [Internet]. 2007;317(August):944–7. Available from: www.sciencemag.org/cgi/content/full/1143767/DC1

2. Limou S, Zagury J-F. Immunogenetics: Genome-Wide Association of Non-Progressive HIV and Viral Load Control: HLA Genes and Beyond. Front Immunol [Internet]. 2013;4(MAY):1–13. Available from: http://journal.frontiersin.org/article/10.3389/fimmu.2013.00118/abstract

3. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science [Internet]. 2019;365(6460). Available from: http://www.ncbi.nlm.nih.gov/pubmed/31604244

4. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in Covid-19. Nature [Internet]. 2020 Dec 11; Available from: http://www.nature.com/articles/s41586-020-03065-y

5. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature [Internet]. 2021 Jul 8; Available from: http://www.nature.com/articles/s41586-021-03767-x

6. Douillard V, Castelli EC, Mack SJ, Hollenbach JA, Gourraud PA, Vince N, et al. Current HLA Investigations on SARS-CoV-2 and Perspectives [Internet]. Vol. 12, Frontiers in Genetics. 2021. p. 10–6. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2021.774922/full

7. Klein RJ, Zeiss C, Chew EY, Tsai J, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science (80-) [Internet]. 2005 Apr 15;308(5720):385–9. Available from: https://www.science.org/doi/10.1126/science.1109557

8. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science (80-) [Internet]. 2006 Dec;314(5804):1461–3. Available from: https://www.science.org/doi/10.1126/science.1135245

9. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature [Internet]. 2018;562(7726):203–9. Available from: http://www.nature.com/articles/s41586-018-0579-z

10. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017;27(3):S2–8.

11. Hirata M, Kamatani Y, Nagai A, Kiyohara Y, Ninomiya T, Tamakoshi A, et al. Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. J Epidemiol. 2017;27(3):S9–21.

12. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831

diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290–9.

13. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet [Internet]. 2017 Jul;101(1):5–22. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409

14. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet [Internet]. 2019;20(8):467–84. Available from: http://dx.doi.org/10.1038/s41576-019-0127-1

15. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature [Internet]. 2020;577(7789):179–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31915397

16. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet [Internet]. 2018;103(3):338–48. Available from: https://doi.org/10.1016/j.ajhg.2018.07.015

17. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet [Internet]. 2016 Oct 22;48(10):1279–83. Available from: http://www.nature.com/articles/ng.3643

18. Dausset J. Iso-leuco-anticorps. Acta Haematol [Internet]. 1958;20(1–4):156–66. Available from: https://www.karger.com/Article/FullText/205478

19. Jean Dausset. The Major Histocompatibility Complex in Man: past, present and futur concepts. Science (80-) [Internet]. 1981;213(September):55–97. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780124169746000065

20. Concannon P, Chen WM, Julier C, Morahan G, Akolkar B, Erlich HA, et al. Genome-wide scan for linkage to type 1 diabetes in 2,496 multiplex families from the type 1 diabetes genetics consortium. Diabetes. 2009;58(4):1018–22.

21. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18(12):1091–102.

22. Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, et al. Genomewide Association Study of an AIDS-Nonprogression Cohort Emphasizes the Role Played by HLA Genes (ANRS Genomewide Association Study 02). J Infect Dis [Internet]. 2009 Feb;199(3):419–26. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1086/596067

23. Hu Z, Liu Y, Zhai X, Dai J, Jin G, Wang L, et al. New loci associated with chronic hepatitis B virus infection in Han Chinese. Nat Genet [Internet]. 2013;45(12):1499–503. Available from: http://dx.doi.org/10.1038/ng.2809

24. Jiang D, Ma X, Yu H, Cao G, Ding D, Chen H, et al. Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B. Hepatology [Internet]. 2015 Jul 28;62(1):118–28. Available from: https://onlinelibrary.wiley.com/doi/10.1002/hep.27794

25. Vergara C, Thio CL, Johnson E, Kral AH, O'Brien TR, Goedert JJ, et al. Multi-Ancestry Genome-Wide Association Study of Spontaneous Clearance of Hepatitis C Virus. Gastroenterology [Internet]. 2019 Apr;156(5):1496-1507.e7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0016508518354210

26. Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet. 2015;47(10):1107–13.

27. Vince N, Limou S, Daya M, Morii W, Rafaels N, Geffard E, et al. Association of HLA-DRB1∗09:01 with tIgE levels among African-ancestry individuals with asthma. J Allergy Clin Immunol [Internet]. 2020 Jul;146(1):147–55. Available from: https://doi.org/10.1016/j.jaci.2020.01.011

28. Valencia A, Vergara C, Thio CL, Vince N, Douillard V, Grifoni A, et al. Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1. Am J Hum Genet [Internet]. 2022;109(2):299–310. Available from: https://doi.org/10.1016/j.ajhg.2022.01.001

29. Domenighetti C, Douillard V, Sugier P, Sreelatha AAK, Schulte C, Grover S, et al. The Interaction between HLA-DRB1 and Smoking in Parkinson's Disease Revisited. Mov Disord [Internet]. 2022 Jul 10; Available from: https://onlinelibrary.wiley.com/doi/10.1002/mds.29133

30. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Res [Internet]. 2019 Oct 31;48(D1):D948–55. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz950/5610347

31. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol. 2007;68(9):779–88.

32. Meyer D, Nunes K. HLA imputation, what is it good for? Hum Immunol [Internet]. 2017 Mar;78(3):239–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885917300277

33. Douillard V, Castelli EC, Mack SJ, Hollenbach JA, Gourraud P-A, Vince N, et al. Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research. Front Genet [Internet]. 2021 Dec 2;12. Available from: In review

34. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG - HLA genotype imputation with attribute bagging. Pharmacogenomics J [Internet]. 2014;14(2):192–200. Available from: http://dx.doi.org/10.1038/tpj.2013.18

35. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. Tang J, editor. PLoS One [Internet]. 2013 Jun 6;8(6):e64683. Available from: https://dx.plos.org/10.1371/journal.pone.0064683

36. Pappas DJ, Tomich A, Garnier F, Marry E, Gourraud P-A. Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: Consequences of sampling fluctuation and haplotype frequency distribution tail truncation. Hum Immunol [Internet]. 2015 May;76(5):374–80. Available from: http://dx.doi.org/10.1016/j.humimm.2015.01.029

37. Motyer A, Vukcevic D, Dilthey A, Donnelly P, McVean G, Leslie S. Practical Use of Methods for Imputation of HLA Alleles from SNP Genotype Data. bioRxiv. 2016;091009.

38. Cook S, Choi W, Lim H, Luo Y, Kim K, Jia X. Accurate imputation of human leukocyte antigens. Nat Commun [Internet]. (2021):1–11. Available from: http://dx.doi.org/10.1038/s41467-021-21541-5

39. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat Commun [Internet]. 2021;12(1):1–14. Available from: http://dx.doi.org/10.1038/s41467-021-21975-x

40. Okada Y, Momozawa Y, Ashikawa K, Kanai M, Matsuda K, Kamatani Y, et al. Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. Nat Genet [Internet]. 2015;47(7):798–802. Available from: http://dx.doi.org/10.1038/ng.3310

41. Ritari J, Hyvärinen K, Clancy J, Partanen J, Koskela S. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. NAR Genomics Bioinforma [Internet]. 2020 Jun 1;2(2):1–9. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa030/5831010

42. Nordin J, Ameur A, Lindblad-Toh K, Gyllensten U, Meadows JRS. SweHLA: the high confidence HLA typing bio-resource drawn from 1000 Swedish genomes. Eur J Hum Genet [Internet]. 2020 May 16;28(5):627–35. Available from: http://www.nature.com/articles/s41431-019-0559-2

43. Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, et al. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum Mol Genet [Internet]. 2019 Jun 15;28(12):20782092. Available from: https://academic.oup.com/hmg/article/28/12/2078/5261434

44. Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. Nat Genet [Internet]. 2021 Nov 10;53(10):1504–16. Available from: http://dx.doi.org/10.1016/j.oraloncology.2011.05.004%0Ahttp://dx.doi.org/10.1016/S0084-3873(10)79678-4

45. Vince N, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. Genet Epidemiol [Internet]. 2020 Oct 18;44(7):733–40. Available from: https://onlinelibrary.wiley.com/doi/10.1002/gepi.22334

46. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Vol. 526, Nature. 2015. p. 68–74.

47. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell [Internet]. 2022 Sep;185(18):3426-3440.e19. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867422009916

48. Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 2017;45(D1):D854–9.

49. Abi-Rached L, Gouret P, Yeh JH, Cristofaro J Di, Pontarotti P, Picard C, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. PLoS One. 2018;13(10):1–11.

50. Naslavsky MS, Scliar MO, Yamamoto GL, Wang JYT, Zverinova S, Karp T, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from Brazil. Nat Commun [Internet]. 2022 Dec 4;13(1):1004. Available from: https://www.nature.com/articles/s41467-022-28648-3

51. Kals M, Nikopensius T, Läll K, Pärn K, Sikka TT, Suvisaari J, et al. Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. bioRxiv [Internet]. 2019;579201. Available from: https://www.biorxiv.org/content/10.1101/579201v1?rss=1

52. Herzig AF, Velo-Suárez L, Consortium F, Consortium F, Dina C, Redon R, et al. Can imputation in a European country be improved by local reference panels? The example of France. bioRxiv [Internet]. 2022;2022.02.17.480829. Available from: https://www.biorxiv.org/content/10.1101/2022.02.17.480829v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.02.17.480829v1.abstract

53. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet [Internet]. 2020;85–91. Available from: http://dx.doi.org/10.1038/s10038-020-00851-4

54. Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, et al. Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun [Internet]. 2020;11(1):1–11. Available from: http://dx.doi.org/10.1038/s41467-020-15194-z

55. Dai CL, Vazifeh MM, Yeang CH, Tachet R, Wells RS, Vilar MG, et al. Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. Am J Hum Genet [Internet]. 2020;106(3):371–88. Available from: https://doi.org/10.1016/j.ajhg.2020.02.002

56. Maróstica AS, Nunes K, Castelli EC, Silva NSB, Bruce S, Goudet J, et al. How HLA diversity is apportioned : influence of selection and relevance to transplantation. 2022;

57. Lewontin RC. The Apportionment of Human Diversity. In: Evolutionary Biology [Internet]. New York, NY: Springer US; 1972. p. 381–98. Available from: http://link.springer.com/10.1007/978-1-4684-9063-3_14

58. Mimori T, Yasuda J, Kuroki Y, Shibata TF, Katsuoka F, Saito S, et al. Construction of full-length Japanese reference panel of class I HLA genes with single-molecule, real-time sequencing. Pharmacogenomics J [Internet]. 2019;19(2):136–46. Available from: http://dx.doi.org/10.1038/s41397-017-0010-4

59. Luo Y, Kanai M, Choi W, Li X, Yamamoto K, Ogawa K, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. medRxiv [Internet]. 2020;14:2020.07.16.20155606. Available from: https://doi.org/10.1101/2020.07.16.20155606

60. Zhou F, Cao H, Zuo X, Zhang T, Zhang X, Liu X, et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. Nat Genet. 2016;48(7):740–6.

61. Huang Y-H, Khor S-S, Zheng X, Chen H-Y, Chang Y-H, Chu H-W, et al. A high-resolution HLA imputation system for the Taiwanese population: a study of the Taiwan Biobank. Pharmacogenomics J [Internet]. 2020 Oct 11;20(5):695–704. Available from: http://dx.doi.org/10.1038/s41397-020-0156-3

62. Nunes K, Zheng X, Torres M, Moraes ME, Piovezan BZ, Pontes GN, et al. HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. Hum Immunol [Internet]. 2016 Mar;77(3):307–12. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885915005571

63. Daya M, Cox C, Acevedo N, Boorgula MP, Campbell M, Chavan S, et al. Multiethnic genome-wide and HLA association study of total serum IgE level. J Allergy Clin Immunol [Internet]. 2021;148(6):1589–95. Available from: https://doi.org/10.1016/j.jaci.2021.09.011

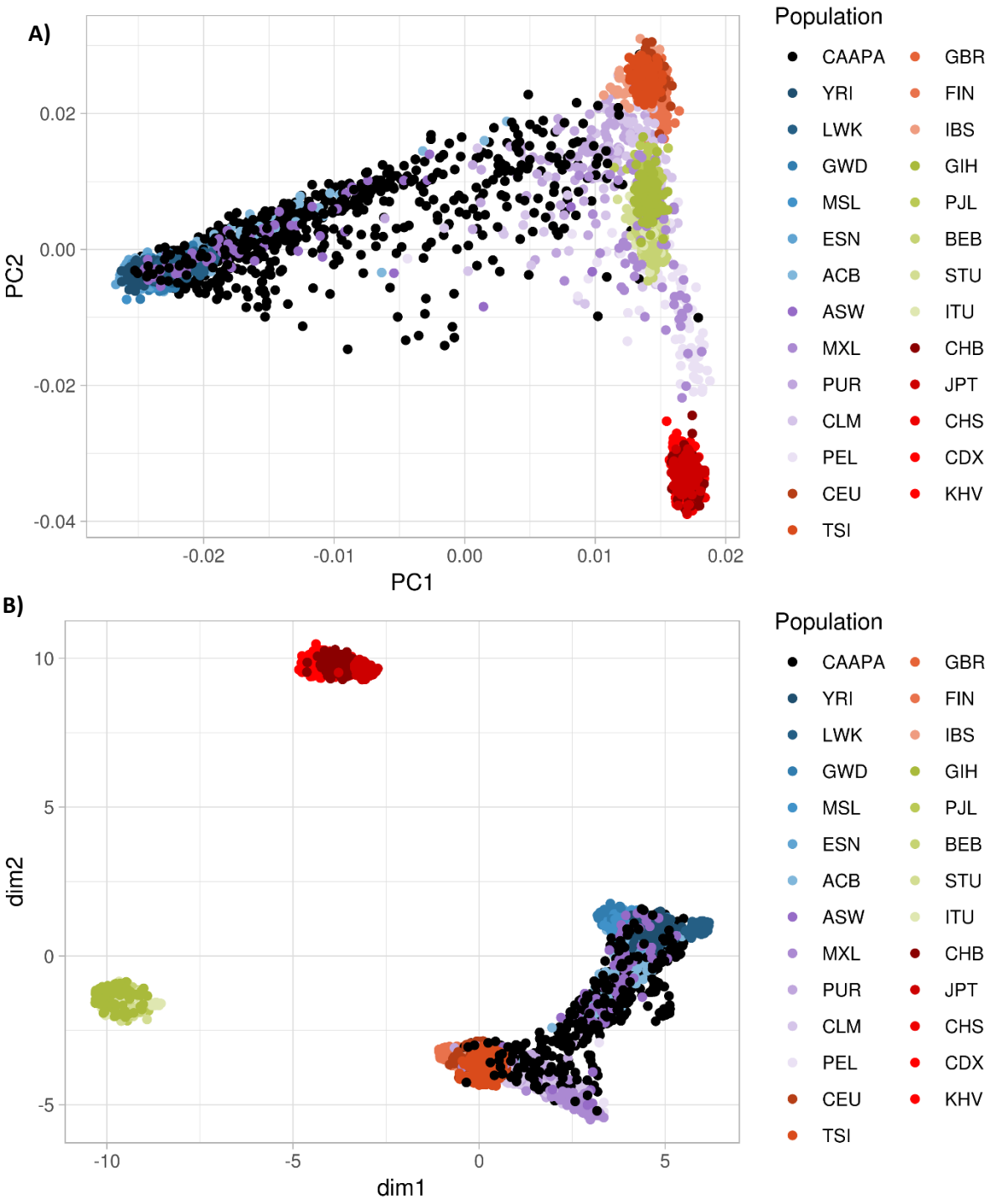
64. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):1–16.

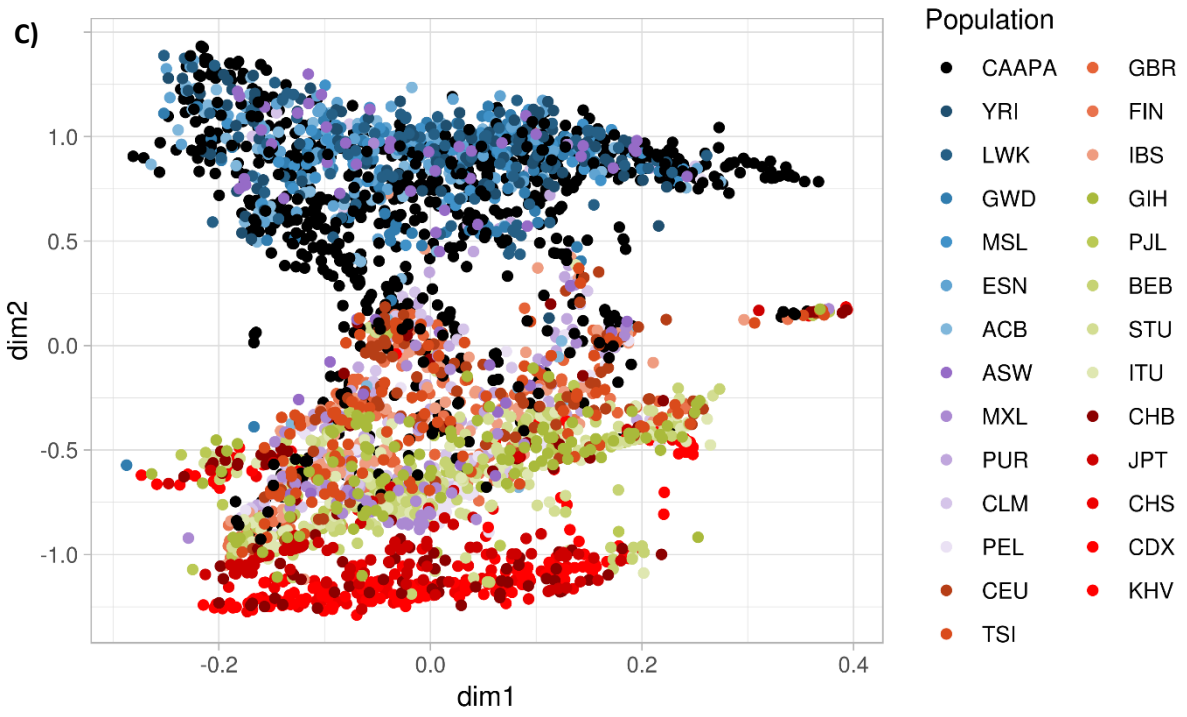
65. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018; Available from: http://arxiv.org/abs/1802.03426

66. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–47.

Supplementary Data

Figure S1: PCA and UMAP representations of a dataset containing both CAAPA and 1KG populations, according to the SNPs used (the chromosome 6 without MHC, or inside the MHC) and the number of dimensions selected (2 or 10). (A) Two dimensions of a PCA based on chromosome 6 SNPs; (B) Two-dimensional UMAP based on chromosome 6 SNPs; (C) Two first dimensions of a ten-dimensional UMAP based on MHC SNPs; (D) Two first dimensions of a ten-dimensional UMAP based on chromosome 6 SNPs.




C)
dim2
dim1
1.0
0.5
0.0
-0.5
-1.0
-0.2
0.0
0.2
0.4
Population
CAAPA
YRI
LWK
GWD
MSL
ESN
ACB
ASW
MXL
PUR
CLM
PEL
CEU
TSI
GBR
FIN
IBS
GIH
PJL
BEB
STU
ITU
CHB
JPT
CHS
CDX
KHV

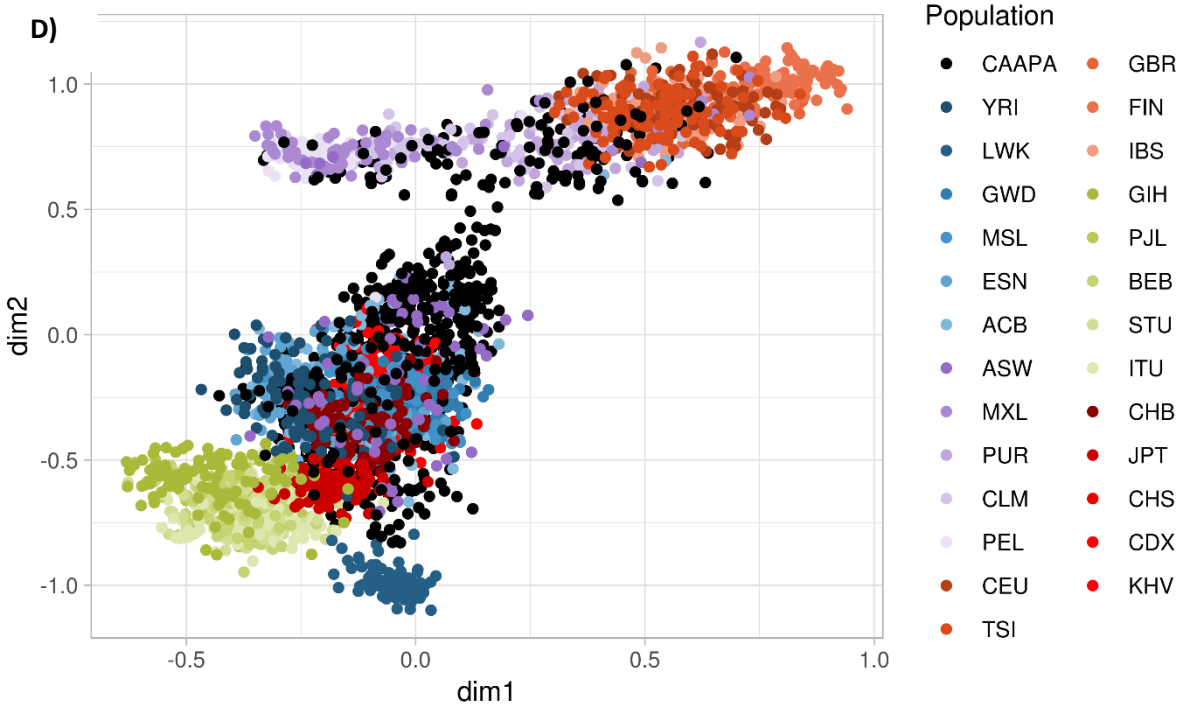

D)
dim2
dim1
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
Population
CAAPA
YRI
LWK
GWD
MSL
ESN
ACB
ASW
MXL
PUR
CLM
PEL
CEU
TSI
GBR
FIN
IBS
GIH
PJL
BEB
STU
ITU
CHB
JPT
CHS
CDX
KHV

Figure S2: Comparison of custom dataset composition, based on the sum of different individuals in the training dataset(s). Some datasets have more than 200 individuals because they represent multiple models of 200 individuals which were close from different subsets of CAAPA. Summing them up does not round to 400 or 600 because some individuals overlap. “Whole” datasets correspond to the closest 200 1KG individuals selected from averaging the positions of all CAAPA individuals in the PCA/UMAP representation, not from subsets.

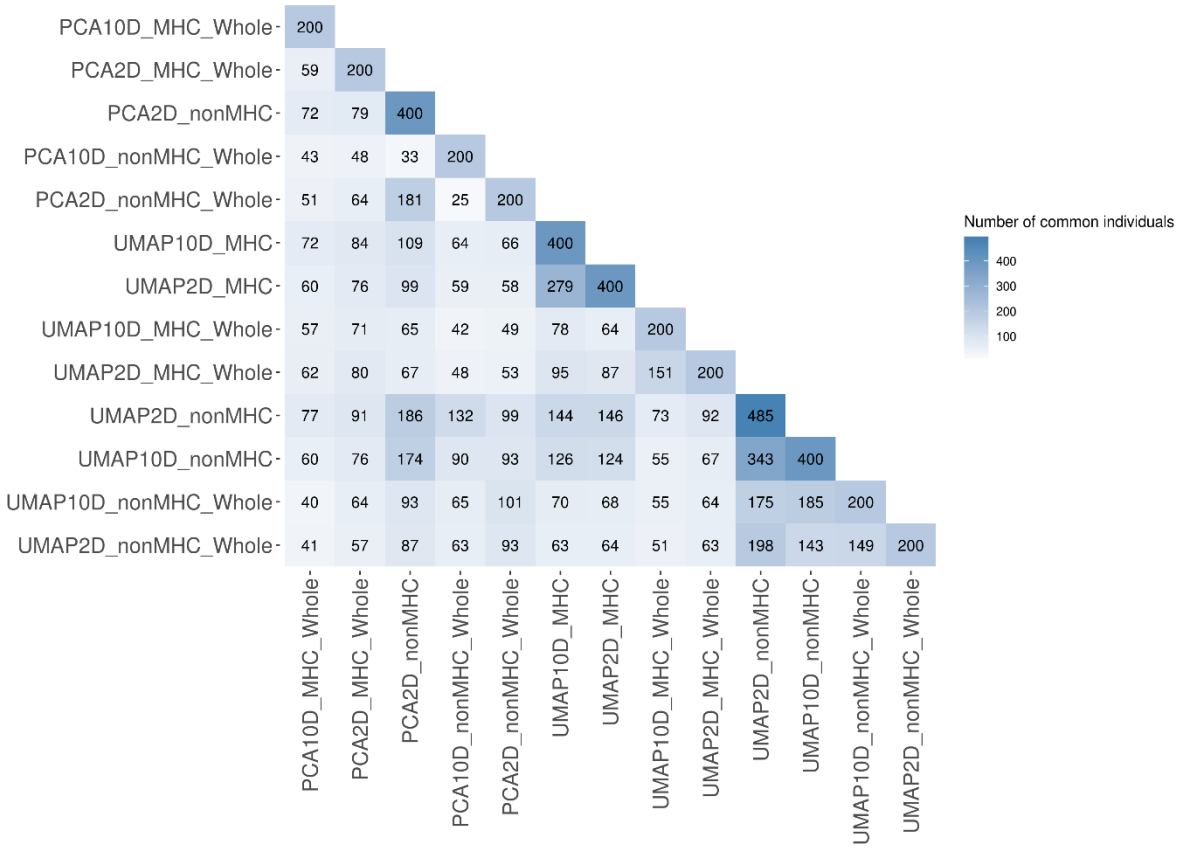


Figure S3: Accuracy measured by mean allelic error rate (1-Accuracy) for CAAPA HLA genotype imputation with the full 1KG, super-populations of 1KG, or the two custom models with the highest accuracy as training models. It represents the same information as accuracy but is easier to visualize, a higher value means a less good HLA imputation. It is obtained by averaging each allele error rate.

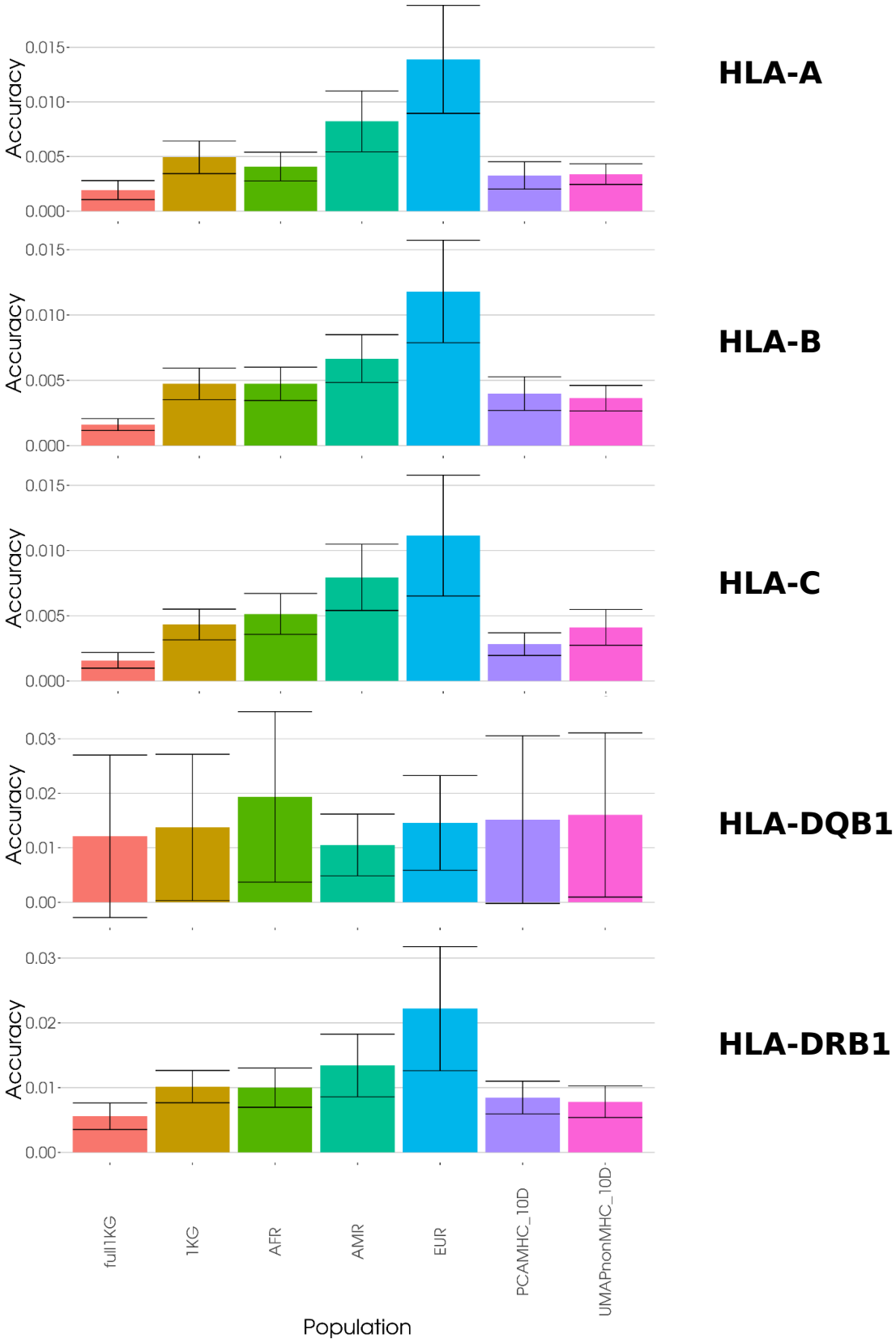


Figure S4: (A) Individual (counted as 0 or 1) and (B) haplotype (individuals can be counted as 0, 0.5, or 1) overall accuracies of CAAPA HLA genotypes predictions using 1KG as a training dataset.

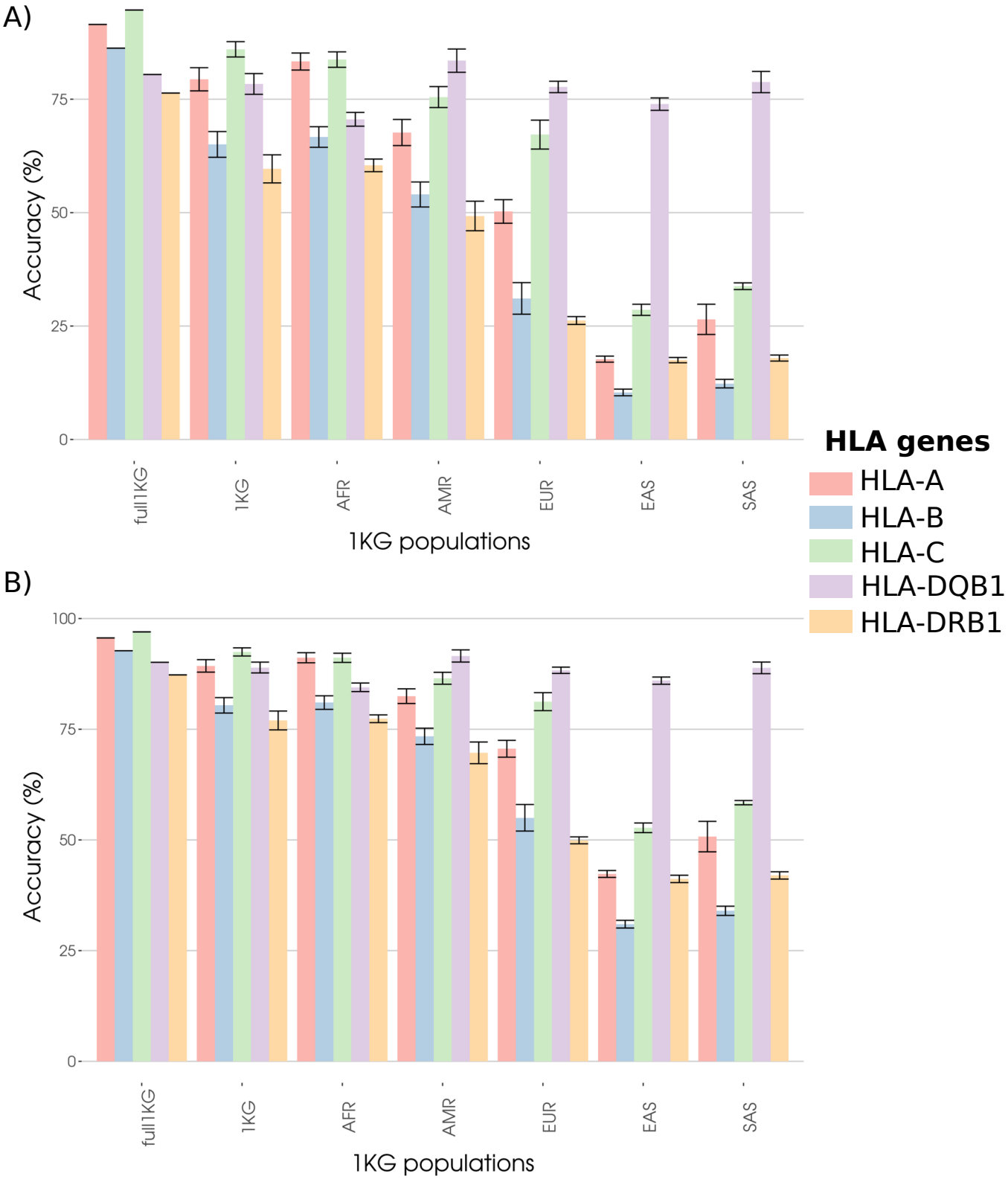
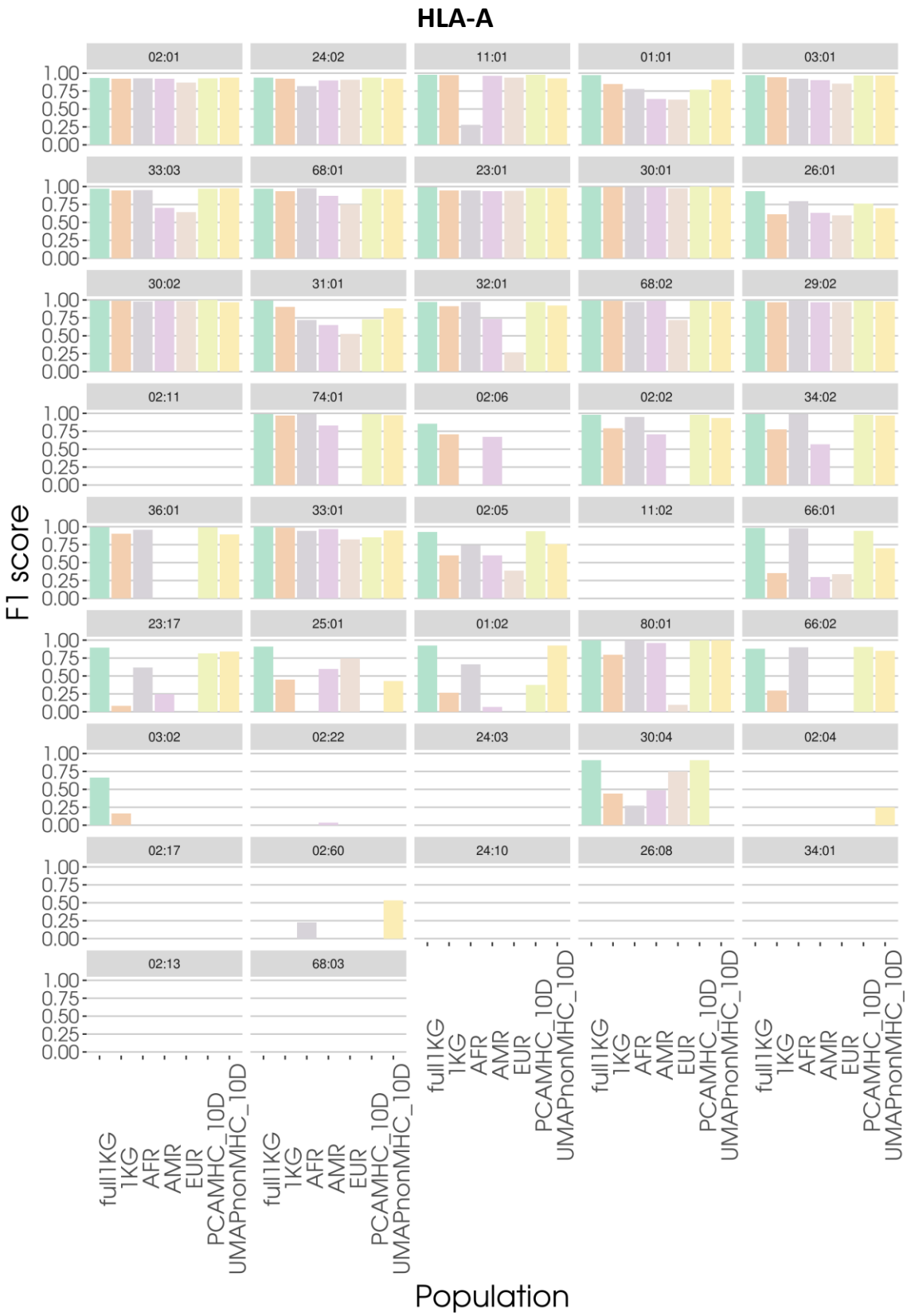
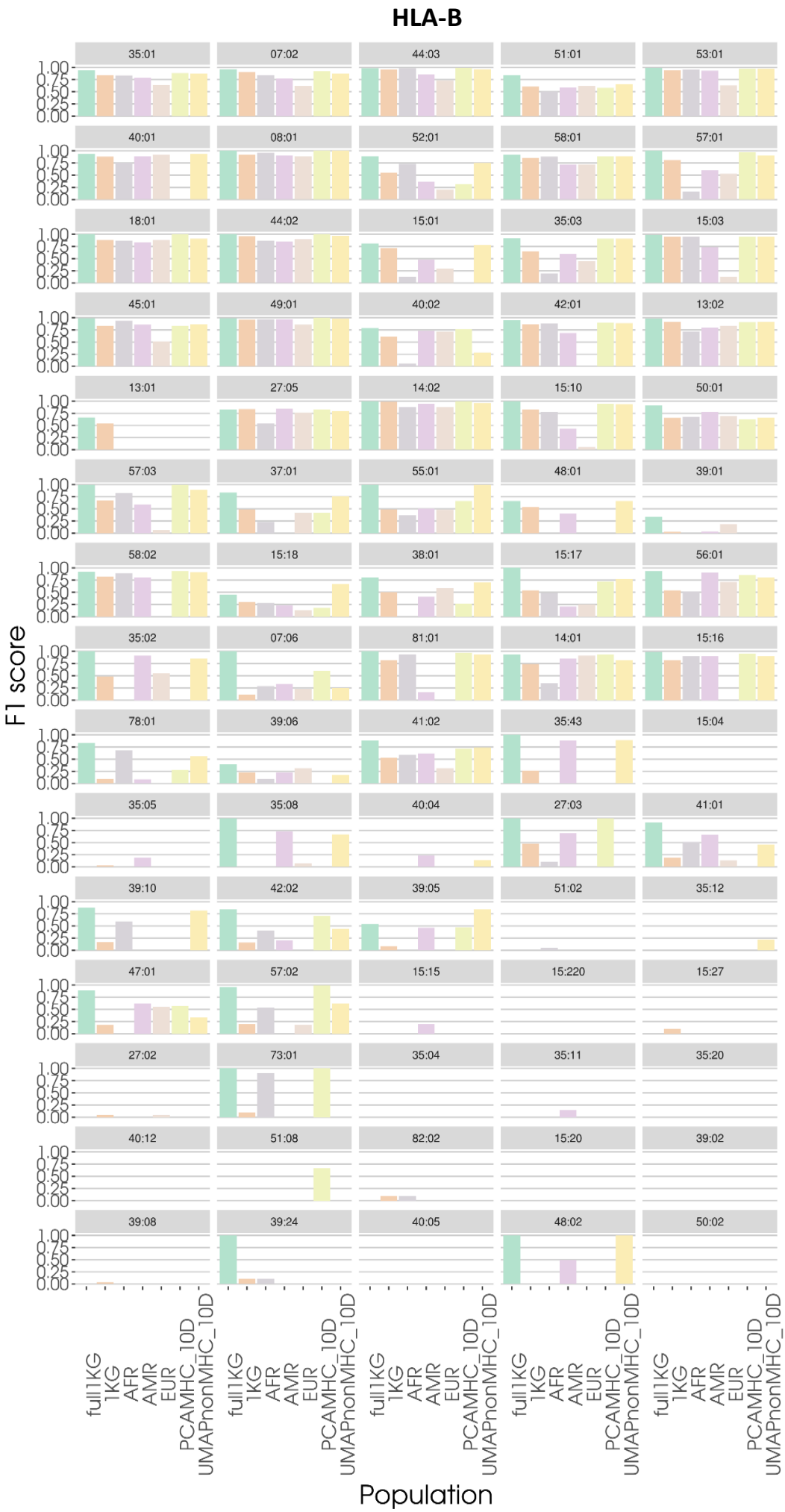


Figure S5: Mean F1-score by allele for each HLA gene *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, and *HLA-DRB1*. We compared the full 1KG dataset, to super-populations, to the most accurate custom models. HLA alleles that were absent from the 1KG dataset have been removed, therefore the alleles with no call at all mostly come from a singleton allele in the test dataset that has been mis-imputed once.




HLA-B
35:01
07:02
44:03
51:01
53:01
40:01
08:01
52:01
58:01
57:01
18:01
44:02
15:01
35:03
15:03
45:01
49:01
40:02
42:01
13:02
13:01
27:05
14:02
15:10
50:01
57:03
37:01
55:01
48:01
39:01
58:02
15:18
38:01
15:17
56:01
35:02
07:06
81:01
14:01
15:16
78:01
39:06
41:02
35:43
15:04
35:05
35:08
40:04
27:03
41:01
39:10
42:02
39:05
51:02
35:12
47:01
57:02
15:15
15:220
15:27
27:02
73:01
35:04
35:11
35:20
40:12
51:08
82:02
15:20
39:02
39:08
39:24
40:05
48:02
50:02
1.00
0.75
0.50
0.25
0.00
F1 score
full1KG
1KG
AFR
AMR
EUR
PCAMHC_10D
UMAPnonMHC_10D
Population


HLA-C
04:01
07:02
06:02
01:02
03:04
07:01
16:01
03:03
15:02
12:02
05:01
12:03
08:01
03:02
14:02
17:01
08:02
02:10
07:18
02:02
07:06
15:05
07:04
16:02
04:03
08:04
18:02
17:03
14:03
18:01
03:05
12:04
12:05
F1 score
1.00
0.75
0.50
0.25
0.00
full1KG
1KG
AFR
AMR
EUR
PCAMHC_10D
UMAPnonMHC_10D
Population


HLA-DQB1
03:01
05:01
02:02
06:02
03:02
03:03
02:01
06:01
04:02
05:02
05:03
03:19
06:03
06:09
06:04
03:04
F1 score
1.00
0.75
0.50
0.25
0.00
full1KG
1KG
AFR
AMR
EUR
PCAMHC_10D
UMAPnonMHC_10D
Population


HLA-DRB1
07:01
15:01
03:01
11:01
09:01
13:01
13:02
01:01
15:02
15:03
10:01
12:02
03:02
04:04
04:05
04:01
14:54
04:03
16:02
12:01
13:03
01:02
08:04
14:04
11:02
08:02
08:01
11:04
04:07
13:04
14:02
04:11
01:03
04:02
16:01
08:06
04:92
13:12
14:01
04:08
11:03
14:03
13:05
13:27
11:10
F1 score
1.00
0.75
0.50
0.25
0.00
full1KG
1KG
AFR
AMR
EUR
PCAMHC_10D
UMAPnonMHC_10D
Population

Figure S6: Global imputation accuracy of *HLA-B* and post-probability comparison between the full 1KG dataset and the one created with three models selected by ten-dimensional UMAP with SNPs outside of the MHC SNPs. The x-axis shows the four cases where no model, one of them, or both, imputed individuals of CAAPA correctly. Percentages refer to the whole CAAPA dataset. Colors shows for each case the number of genotypes which have a better post-probability either in the full or the custom 1KG model.

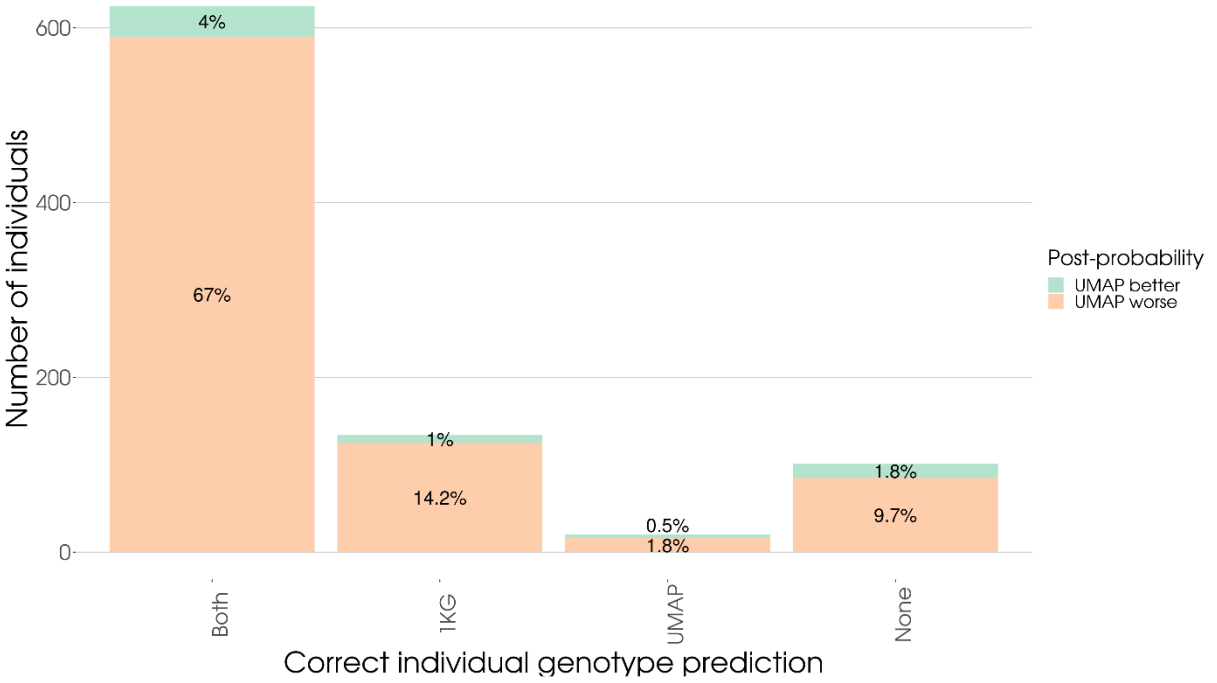


Table S1: Nomenclature of the custom models. Two main dimension reduction methods have been applied to the SNP genotypes datasets, PCA and UMAP. They used the same set of SNP only taking the whole chromosome 6 without the MHC region, or the MHC region alone (29-34Mb) in the models. Then, we ran a k-means clustering method on two or ten dimensions of the reduced dataset to obtain subsamples of each model. For UMAPnotMHC_10D, we identified three clusters with a silhouette score on the ten dimensions, we ran the k-means on the dataset and created a sub-model for each of the groups (UMAPnotMHC_10D_1, UMAPnotMHC_10D_2 & UMAPnotMHC_10D_3). We then pooled the results to compute the accuracies of the whole model with other, non-custom, models.

Name	Method	SNPs used	# of dimensions used for clustering
UMAPnotMHC_2D	UMAP	Chromosome 6 without the MHC	2
UMAPnotMHC_10D			10
UMAPMHC_2D		Only the MHC region	2
UMAPMHC_10D			10
PCAnotMHC_2D	PCA	Chromosome 6 without the MHC	2
PCAnotMHC_10D			10
PCAMHC_2D		Only the MHC region	2
PCAMHC_10D			10

VI.3 - La mise à profit des données génétiques et des outils bioinformatique pour explorer extensivement le rôle du HLA

VI.3.1 - L'association HLA à la maladie de Parkinson change selon le statut tabagique

En addition de l'imputation HLA qui permet d'explorer en profondeur les associations HLA, l'objectif à long terme du SHLARC est de démocratiser l'analyse HLA globalement : de génotypes, d'haplotypes ou encore la prédiction de l'affinité des allèles HLA pour des peptides ; en partageant des outils et des connaissances.

Les associations avec la maladie de Parkinson sont étudiées depuis plusieurs années : certains allèles HLA (376,377) et le statut tabagique auraient un effet protecteur sur la maladie. Le tabagisme modifie les protéines par homocitrullination (*i.e.* la transformation d'un acide aminé lysine en homocitrulline), et notamment l'α-synucléine, une protéine qui s'accumule dans le cerveau des patients atteints de Parkinson. Cependant, il existe également une interaction, une modification de ces associations par un mécanisme inconnu, lorsque les deux facteurs sont présents en même temps (376,378). La nature de cette interaction entre le HLA et le statut tabagique est controversée car Chuang *et al.* montrent une association plus faible avec les deux facteurs alors que Hollenbach *et al.* indiquent que cette association protectrice n'existe que lorsque les deux facteurs se rencontrent.

Les travaux suivants tirent profit de la randomisation mendélienne, une méthode statistique pour identifier les signaux d'association potentiellement causaux en démêlant les associations existantes entre SNP et initiation tabagique, de celles entre les SNP et la maladie de Parkinson. Nos résultats ont été obtenus à partir de génotypes et haplotypes HLA imputés par HIBAG et EasyHLA, respectivement. Ils ont confirmé l'implication de HLA-DRB1 dans la maladie de Parkinson, et plus précisément celle de l'acide aminé Valine en 11ème position. Nous répliquons ainsi les résultats de Chuang *et al*. Nous avons également confirmé ces signaux par la prédiction de l'affinité des allèles HLA pour les peptides de l'α-synucléine qui diminue avec le polymorphisme Valine 11 ou l'homocitrullination. Cependant, la présence de ces deux facteurs empêche cette diminution de la présentation et donc élimine l'effet protecteur face à Parkinson.

VI.3.2 - Article - L'interaction revisitée entre HLA-DRB1 et le tabagisme dans la maladie de Parkinson

BRIEF REPORT

The Interaction between *HLA-DRB1* and Smoking in Parkinson's Disease Revisited

Cloé Domenighetti, PhD,[1] Venceslas Douillard, PhD,[2] Pierre-Emmanuel Sugier, PhD,[1] Ashwin Ashok Kumar Sreelatha, PhD candidate,[3] Claudia Schulte, MSc,[4,5] Sandeep Grover, PhD,[3] Patrick May, PhD,[6] Dheeraj R. Bobbili, PhD,[6] Milena Radivojkov-Blagojevic, MSc,[7] Peter Lichtner, PhD,[7] Andrew B. Singleton, PhD,[8,9] Dena G. Hernandez, PhD,[8] Connor Edsall, PhD candidate,[8] Pierre-Antoine Gourraud, PhD,[2] George D. Mellick, PhD,[10] Alexander Zimprich, MD,[11] Walter Pirker, MD,[12] Ekaterina Rogaeva, PhD,[13] Anthony E. Lang, MD,[13,14,15,16] Sulev Koks, MD, PhD,[17,18] Pille Taba, MD, PhD,[19,20] Suzanne Lesage, PhD,[21] Alexis Brice, MD,[21] Jean-Christophe Corvol, MD, PhD,[21,22] Marie-Christine Chartier-Harlin, PhD,[23] Eugénie Mutez, MD,[23] Kathrin Brockmann, MD,[4,5] Angela B. Deutschländer, MD,[24,25,26] Georges M. Hadjigeorgiou, MD,[27,28] Efthimos Dardiotis, MD,[27] Leonidas Stefanis, MD, PhD,[29,30] Athina Maria Simitsi, MD, PhD,[29] Enza Maria Valente, MD, PhD,[31,32] Simona Petrucci, MD, PhD,[33,34] Stefano Duga, PhD,[35,36] Letizia Straniero, PhD,[35] Anna Zecchinelli, MD,[37] Gianni Pezzoli, MD,[38] Laura Brighina, MD, PhD,[39,40] Carlo Ferrarese, MD, PhD,[39,40] Grazia Annesi, PhD,[41] Andrea Quattrone, MD,[42] Monica Gagliardi, PhD,[43] Hirotaka Matsuo, MD, PhD,[44] Akiyoshi Nakayama, PhD,[44] Nobutaka Hattori, MD, PhD,[45] Kenya Nishioka, MD, PhD,[45] Sun Ju Chung, MD, PhD,[46] Yun Joong Kim, MD, PhD,[47] Pierre Kolber, MD,[48] Bart P.C. van de Warrenburg, MD, PhD,[49] Bastiaan R. Bloem, MD, PhD,[49] Jan Aasly, MD,[50] Mathias Toft, MD, PhD,[51] Lasse Pihlstrøm, MD, PhD,[51] Leonor Correia Guedes, MD, PhD,[52,53] Joaquim J. Ferreira, MD, PhD,[52,54] Soraya Bardien, PhD,[55] Jonathan Carr, PhD,[56] Eduardo Tolosa, MD, PhD,[57,58] Mario Ezquerra, PhD,[59] Pau Pastor, MD, PhD,[60,61] Monica Diez-Fairen, MSc,[60,61] Karin Wirdefeldt, MD, PhD,[62,63] Nancy L. Pedersen, PhD,[63] Caroline Ran, PhD,[64] Andrea C. Belin, PhD,[64] Andreas Puschmann, MD, PhD,[65] Emil Ygland Rödström, MD,[65] Carl E. Clarke, MD,[66] Karen E. Morrison, MD,[67] Manuela Tan, PhD,[68] Dimitri KraincMD, PhD,[69] Lena F. Burbulla, PhD,[69,70,71,72] Matt J. Farrer, PhD,[73] Rejko Krüger, MD,[6,48,74,75] Thomas Gasser, MD,[4,5] Manu Sharma, PhD,[3] Nicolas Vince, PhD,[2] Alexis Elbaz, MD, PhD,[1*] on behalf of the Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) Consortium

[1]*Université Paris-Saclay, UVSQ, Univ. Paris-Sud, Inserm, Team "Exposome, Heredity, Cancer, and Health", CESP, Villejuif, France* [2]*Nantes Université, INSERM, Center for Research in Transplantation and Translational Immunology, Nantes, France* [3]*Centre for Genetic Epidemiology, Institute for Clinical Epidemiology, and Applied Biometry, University of Tubingen, Tübingen, Germany* [4]*Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tubingen, Tübingen, Germany* [5]*German Center for Neurodegenerative Diseases (DZNE), Tubingen, Germany* [6]*Translational Neuroscience, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg* [7]*Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany* [8]*Molecular Genetics Section, Laboratory of Neurogenetics, NIA, NIH, Bethesda, Maryland, USA* [9]*Center For Alzheimer's and Related Dementias, NIA, NIH, Bethesda, Maryland, USA* [10]*Griffith Institute for Drug Discovery, Griffith University, Nathan, Queensland, Australia* [11]*Department of Neurology, Medical University of Vienna, Wien, Austria* [12]*Department of Neurology, Klinik Ottakring, Vienna, Austria* [13]*Tanz Centre for Research in Neurodegenerative Diseases, University of Toronto, Toronto, Ontario, Canada* [14]*Edmond J. Safra Program in Parkinson's Disease, Morton and Gloria Shulman Movement Disorders Clinic, Toronto Western Hospital, UHN, Toronto, Ontario, Canada* [15]*Division of Neurology, University of Toronto, Toronto, Ontario, Canada* [16]*Krembil Brain Institute, Toronto, Ontario, Canada* [17]*Centre for Molecular Medicine and Innovative Therapeutics,*

***Correspondence to:** Dr. Alexis Elbaz, INSERM U1018 CESP, Hôpital Paul Brousse, Bâtiment 15/16, 16 avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France; E-mail: alexis.elbaz@inserm.fr

Relevant conflicts of interest/financial disclosures: A.B.S. reports grants from Department of Defense, during the conduct of the study and grants from The Michael J. Fox Foundation, outside the submitted work. W.P. reports personal fees from Grünenthal, personal fees from AbbVie, personal fees from AOP Orphan, personal fees from Zambon, personal fees and other from Boehringer Ingelheim, personal fees from Stada, and personal fees from UCB Pharma, outside the submitted work. A.E.L. reports personal fees from AbbVie, personal fees from AFFiRis, personal fees from Janssen, personal fees from Biogen, personal fees from Merck, personal fees from Sun Pharma, personal fees from Corticobasal Solutions, personal fees from Sunovion, personal fees from Paladin, personal fees from Lilly, personal fees from Medtronic, personal fees from Theravance, personal fees from Lundbeck, personal fees from Retrophin, personal fees from Roche, and personal fees from PhotoPharmics, outside the submitted work. A. B. reports grants from France Parkinson + FRC, grants from Agence Nationale de Recherche (ANR)-EPIG, grants from ANR-Joint Programming for Neurodegenerative Diseases (JPND), grants from Roger de Spoelberch Foundation (RDS), grants from France Alzheimer, grants from Institut de France, grants from ANR-EPIG, and grants from FMR (maladies rares), outside the submitted work. J.C.C. reports grants from The Michael J. Fox Foundation, Sanofi, and served in advisory boards for Air Liquide, Biogen, Denali, Ever Pharma, Idorsia, Prevail Therapeutic, Theranexus, and UCB, outside the submitted work. M.C.C.H. reports grants from France Parkinson, grants from ANR (MetDePaDi, Synapark), grant from ANR–JPND (TransNeuro), grants from Fondation de France, grants from The Michael J. Fox Foundation, outside the submitted work.

Murdoch University, Murdoch, Australia [18]*Perron Institute for Neurological and Translational Science, Nedlands, Western Australia, Australia* [19]*Department of Neurology and Neurosurgery, University of Tartu, Tartu, Estonia* [20]*Neurology Clinic, Tartu University Hospital, Tartu, Estonia* [21]*Department of Neurologie, Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, INSERM, CNRS, Assistance Publique Hôpitaux de Paris, Paris, France* [22]*Assistance Publique Hôpitaux de Paris, Department of Neurology, CIC Neurosciences, Paris, France* [23]*Univ. Lille, Inserm, CHU Lille, UMR-S 1172 - LilNCog- Centre de Recherche Lille Neurosciences and Cognition, Lille, France* [24]*Department of Neurology, Ludwig Maximilians University of Munich, Munich, Germany* [25]*Department of Neurology, Max Planck Institute of Psychiatry, Munich, Germany* [26]*Department of Neurology and Department of Clinical Genomics, Mayo Clinic Florida, Jacksonville, Florida, USA* [27]*Department of Neurology, Laboratory of Neurogenetics, University of Thessaly, University Hospital of Larissa, Larissa, Greece* [28]*Department of Neurology, Medical School, University of Cyprus, Nicosia, Cyprus* [29]*1st Department of Neurology, Eginition Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece* [30]*Center of Clinical Research, Experimental Surgery, and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece* [31]*Department of Molecular Medicine, University of Pavia, Pavia, Italy* [32]*Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Mondino Foundation, Pavia, Italy* [33]*UOC Medical Genetics and Advanced Cell Diagnostics, S. Andrea University Hospital, Rome, Italy* [34]*Department of Clinical and Molecular Medicine, University of Rome, Rome, Italy* [35]*Department of Biomedical Sciences, Humanitas University, Milan, Italy* [36]*Humanitas Clinical and Research Center, IRCCS, Milan, Italy* [37]*Parkinson Institute, Azienda Socio Sanitaria Territoriale (ASST) Gaetano Pini/CTO, Milan, Italy* [38]*Parkinson Institute, Fontazione Grigioni, Milan, Italy* [39]*Department of Neurology, San Gerardo Hospital, Monza, Italy* [40]*Department of Medicine and Surgery and Milan Center for Neuroscience, University of Milano Bicocca, Milan, Italy* [41]*Institute for Biomedical Research and Innovation, National Research Council, Cosenza, Italy* [42]*Institute of Neurology, Magna Graecia University, Catanzaro, Italy* [43]*Institute of Molecular Bioimaging and Physiology National Research Council, Catanzaro, Italy* [44]*Department of Integrative Physiology and Bio-Nano Medicine, National Defense Medical College, Saitama, Japan* [45]*Department of Neurology, Juntendo University School of Medicine, Tokyo, Japan* [46]*Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea* [47]*Department of Neurology, Yonsei University College of Medicine,*

K.B. reports grants from The Michael J. Fox Foundation, grants from BMBF; personal fees from Zambon, UCB, and Abbvie; and grants from University of Tuebingen, outside the submitted work. L.S. has received the following grants over the past year: PPMI2 (supported by The Michael J. Fox Foundation), IMPRIND-IMI2 Number 116060 (EU, H2020), "Transferring autonomous and non-autonomous cell degeneration 3D models between EU and USA for development of effective therapies for neurodegenerative diseases (ND)-CROSS NEUROD" (H2020-EU 1.3.3., 778003), "Chaperone-Mediated Autophagy in Neurodegeneration" (Hellenic Foundation for Research and Innovation grant HFRI-FM17-3013), and "CMA as a Means to Counteract α-Synuclein Pathology in Non-Human Primates" grant by The Michael J. Fox Foundation (Collaborator). He is co-head and PI at the NKUA of the General Secretariat of Research and Technology (GSRT)-funded grant "National Network of Precision Medicine for Neurodegenerative Diseases." He has served on an Advisory Board for AbbVie, ITF Hellas, and Biogen and has received honoraria from Abbvie and Sanofi. There are no specific disclosures related to the current work. E.M.V. serves as Associate Editor of Journal of Medical Genetics, Section Editor of Pediatric Research, Member of the Editorial Board of Movement Disorders Clinical Practice; grants from the Italian Ministry of Health, CARIPLO Foundation, Pierfranco and Luisa Mariani Foundation, and Telethon Foundation Italy, outside the submitted work. N.H. reports grants from Japan Agency for Medical Research and Development (AMED), Japan Society for the Promotion of Science (JSPS), and the Ministry of Education Culture, Sports, Science, and Technology Japan; grant-in-aid for Scientific Research on Innovative Areas; personal fees and other from Dai-Nippon Sumitomo Pharma, Takeda Pharmaceutical, Kyowa Kirin, GSK K.K, Nippon Boehringer Ingelheim, FP Pharmaceutical Corporation, Eisai, Kissei Pharmaceutical Company, Nihon Medi-physics, Novartis Pharma K.K, Biogen Idec Japan, and AbbVie, from Medtronic, other from Boston Scientific Japan, personal fees and other from Astellas Pharma, grants and other from Ono Pharmaceutical, other from Nihon Pharmaceutical, other from Asahi Kasei Medical, other from Mitsubishi Tanabe Pharma Corporation, personal fees and other from Daiichi Sankyo, other from OHARA Pharmaceutical, other from Meiji Seika Pharma, personal fees from Sanofi K.K., personal fees from Pfizer Japan, personal fees from Alexion Pharmaceuticals, personal fees from Mylan N.V, personal fees from MSD K.K, personal fees from Lund Beck Japan, and other from Hisamitsu Pharmaceutical, outside the submitted work. K.N. reports grants from Japan Society for the Promotion of Science (JSPS), outside the submitted work. P.K. reports other from Centre Hospitalier de Luxembourg; University of Luxembourg, grants from Fonds National de Recherche (FNR), and from null, outside the submitted work. B.P.C.W. reports grants from ZonMW, grants from Hersenstichting, grants from uniQure, other from uniQure, grants from Gossweiler Fund, and grants from Radboud university medical centre, outside the submitted work. B.R.B. reports grants from Netherlands Organization for Health Research and Development, grants from The Michael J. Fox Foundation, grants from Parkinson Vereniging, grants from Parkinson Foundation, grants from Gatsby Foundation, grants from Verily Life Sciences, grants from Horizon 2020, grants from Topsector Life sciences and Health, grants from Stichting Parkinson Fonds, grants from UCB, grants from AbbVie, during the conduct of the study; personal fees from Biogen, personal fees from AbbVie, personal fees from Walk with Path, personal fees from UCB, personal fees from AbbVie, personal fees from Zambon, personal fees from Bial, personal fees from Roche, outside the submitted work; and serves as editor-in-chief of the Journal of Parkinson's Disease and serves on the editorial board of Practical Neurology and Digital Biomarkers. M. Toft reports grants from Research Council of Norway, during the conduct of the study; grants from South-Eastern Norway Regional Health Authority, and grants from The Michael J. Fox, outside the submitted work. L.P. reports grants from Norwegian Health Association, and grants from South-Eastern Norway Regional Health Authority, outside the submitted work. J.J.F. reports grants from GlaxoSmithKline, Grunenthal, Fundação MSD (Portugal), TEVA, MSD, Allergan, Novartis, Medtronic, GlaxoSmithKline, Novartis, Lundbeck, Solvay, BIAL, Merck-Serono, Merz, Ipsen, Biogen, Acadia, Abbvie, and Sunovion Pharmaceuticals, personal fees from Faculdade de Medicina de Lisboa, Campus Neurológico Sénior (CNS), BIAL, and Novartis outside the submitted work. E.T. received honoraria for consultancy from TEVA, Bial, Prevail Therapeutics, Boehringer Ingelheim, Roche, and BIOGEN and has received funding for research from Spanish Network for Research on Neurodegenerative Disorders (CIBERNED), Instituto Carlos III (ISCIII), and The Michael J. Fox Foundation for Parkinson's Research. K.W. reports grants from Swedish Research Council during the conduct of the study. N.L.P. reports grants from Swedish Research Council during the conduct of the study. A.P. reports grants from Parkinsonfonden (The Swedish Parkinson Foundation), grants from ALF (Swedish Government), grants from Region Skåne, Sweden, Skåne University Hospital, Hans-Gabriel och Trolle Wachtmeister Stiftelse för Medicinsk Forskning, Sweden, and Multipark—a strategic research environment at Lund University, during the conduct of the study; and personal fees from Elsevier, outside the submitted work. E.Y.R. reports grants from ALF (Swedish Government), Hans-Gabriel och Trolle Wachtmeister Stiftelse för Medicinsk Forskning, Sweden, and Demensfonden (all in Sweden). M. Tan reports grants from Parkinson's United Kingdom (UK), other from The Michael J. Fox Foundation and University College London, outside the submitted work. R.K. reports grants from FNR and the German Research Council (DFG), non-financial support from AbbVie, Zambon, during the conduct of the study; personal fees from University of Luxembourg; Luxembourg

Seoul, South Korea [48]*Neurology, Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg* [49]*Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Centre, Nijmegen, The Netherlands* [50]*Department of Neurology, St Olav's Hospital and Norwegian University of Science and Technology, Trondheim, Norway* [51]*Department of Neurology, Oslo University Hospital, Oslo, Norway* [52]*Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal* [53]*Department of Neurosciences and Mental Health, Neurology, Hospital de Santa Maria, Centro Hospitalar Universitario Lisboa Norte (CHULN), Lisbon, Portugal* [54]*Laboratory of Clinical Pharmacology and Therapeutics, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal* [55]*Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Stellenbosch, South Africa* [56]*Division of Neurology, Department of Medicine, Faculty of Medicine and Health Sciences, Stellenbosch University, Stellenbosch, South Africa* [57]*Parkinson's Disease and Movement Disorders Unit, Neurology Service, Hospital Clínic de Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain* [58]*Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED: CB06/05/0018-ISCIII) Barcelona, Barcelona, Spain* [59]*Lab of Parkinson Disease and Other Neurodegenerative Movement Disorders, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Institut de Neurociències, Universitat de Barcelona, Barcelona, Catalonia, Spain* [60]*Fundació per la Recerca Biomèdica i Social Mútua Terrassa, Barcelona, Spain* [61]*Movement Disorders Unit, Department of Neurology, Hospital Universitari Mutua de Terrassa, Barcelona, Spain* [62]*Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden* [63]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden* [64]*Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden* [65]*Department of Clinical Sciences Lund, Neurology, Lund University, Skåne University Hospital, Lund, Sweden* [66]*University of Birmingham and Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, United Kingdom* [67]*Faculty of Medicine, Health and Life Sciences, Queens University, Belfast, United Kingdom* [68]*Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom* [69]*Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA* [70]*Metabolic Biochemistry, Biomedical Center (BMC), Faculty of Medicine, Ludwig-Maximilians-Universität München, Munich, Germany* [71]*Munich Cluster for Systems Neurology (SyNergy), Munich, Germany* [72]*German Center for Neurodegenerative Diseases (DZNE), Munich, Germany* [73]*Department of Neurology, McKnight Brain Institute, University of Florida, Gainesville, Florida, USA* [74]*Parkinson's Research Clinic, Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg* [75]*Transversal Translational Medicine, Luxembourg Institute of Health (LIH), Strassen, Luxembourg*

Institute of Health; Centre Hospitalier de Luxembourg, grants from Fonds National de Recherche, Luxembourg (FNR), grants from FNR, grants from FNR, Luxembourg/DFG, grants from FNR, Luxembourg (FNR), personal fees from Desitin/Zambon, personal fees from AbbVie, and personal fees from Medtronic, outside the submitted work. T.G. reports personal fees from UCB Pharma, Novartis, Teva, and MedUpdate, grants from The Michael J. Fox Foundation for Parkinson's Research, Bundesministerium für Bildung und Forschung (BMBF), and DFG, other from JPND program, funded by the European Commission, outside the submitted work; in addition, T.G. has a patent patent number: EP1802749 (A2) *KASPP* (*LRRK2*) gene, its production and use for the detection and treatment of neurodegenerative disorders issued. A. E. reports grants from ANR, The Michael J. Fox foundation, Plan Ecophyto (French Ministry of Agriculture), and France Parkinson, outside the submitted work.

Funding agencies: This study used data from the Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) consortium, conducted under a partnership agreement between 35 studies. The Courage-PD consortium is supported by the EU JPND research (https://www.neurodegenerationresearch.eu/initiatives/annual-calls-for-proposals/closed-calls/risk-factors-2012/risk-factor-call-results/courage-pd/). C.D. is the recipient of a doctoral grant from Université Paris-Saclay, France. P.M. was funded by the FNR, Luxembourg as part of the National Centre of Excellence in Research on Parkinson's disease (NCER-PD, FNR11264123), and the DFG Research Units FOR2715 (INTER/DFG/17/11583046) and FOR2488 (INTER/DFG/19/14429377). A.B.S., D.G.H., and C.E. are funded by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services, project ZO1 AG000949. E.R. is funded by the Canadian Consortium on Neurodegeneration in Aging. S.K. is funded by MSWA. P.T. is the recipient of an Estonian Research Council Grant PRG957. E.M.V. is funded by the Italian Ministry of Health (Ricerca Corrente 2021). S.B. and J.C. are supported by grants from the National Research Foundation of South Africa (106052); the South African Medical Research Council (Self-Initiated Research Grant); and Stellenbosch University, South Africa; they also acknowledge the support of the NRF-DST Centre of Excellence for Biomedical Tuberculosis Research; South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town. P.P. and M.D.F. have received funding from the Spanish Ministry of Science and Innovation (SAF2013-47939-R). K.W. and N.L.P. are funded by the Swedish Research Council (K2002-27X-14,056-02B, 521–2010-2479, 521–2013-2488, and 2017–02175). N.L.P. is funded by the National Institutes of Health (ES10758 and AG 08724). C.R. is funded by the Märta Lundkvist Foundation, Swedish Brain Foundation, Karolinska Institutet Research Fund. A.C.B. from the Swedish Brain Foundation, Swedish Research Council, and Karolinska Institutet Research Funds. M.T is funded by the Parkinson's UK. M.S. was supported by the grants from the German Research Council (DFG/SH 599/6–1), MSA Coalition, and The Michael J. Fox Foundation (USA Genetic Diversity in PD Program: GAP-India Grant ID: 17473). P.G. GEN sample collection was funded by the MRC and UK Medical Research Council (CEC, KEM). The sponsors had no role in the study design, data collection, data analysis, data interpretation, the writing of the report, or the decision to submit the paper for publication.

ABSTRACT: Background: Two studies that examined the interaction between *HLA-DRB1* and smoking in Parkinson's disease (PD) yielded findings in opposite directions.
Objective: To perform a large-scale independent replication of the *HLA-DRB1* × smoking interaction.
Methods: We genotyped 182 single nucleotide polymorphism (SNPs) associated with smoking initiation in 12 424 cases and 9480 controls to perform a Mendelian randomization (MR) analysis in strata defined by *HLA-DRB1*.
Results: At the amino acid level, a valine at position 11 (V11) in *HLA-DRB1* displayed the strongest association with PD. MR showed an inverse association between genetically predicted smoking initiation and PD only in absence of V11 (odds ratio, 0.74, 95% confidence interval, 0.59–0.93, $P_{Interaction} = 0.028$). *In silico* predictions of the influence of V11 and smoking-induced modifications of α-synuclein on binding affinity showed findings consistent with this interaction pattern.
Conclusions: Despite being one of the most robust findings in PD research, the mechanisms underlying the inverse association between smoking and PD remain unknown. Our findings may help better understand this association. © 2022 The Authors. *Movement Disorders* published by Wiley Periodicals LLC on behalf of International Parkinson and Movement Disorder Society

Key Words: Parkinson's disease; smoking; gene-environment interaction; HLA

Genome-wide association studies (GWAS) in Parkinson's disease (PD) identified an association with the human leukocyte antigen (*HLA*) region, in particular with *HLA-DRB1*. Hollenbach et al[1] reported an inverse association of PD with the shared epitope (SE), a combination of amino acids (AA) coded by *HLA-DRB1*, only in the presence of a valine at position 11 (V11). The strongest association in a cross-ethnic GWAS meta-analysis was an inverse association with a histidine at position 13 (H13) in *HLA-DRB1*, strongly correlated with V11.[2] The latest study, with some overlap with the previous two, highlighted three AA (V11, H13, and H33) encoded by *HLA-DRB1* inversely associated with PD.[3]

Following studies showing interactions between smoking and *HLA-DRB1* in other conditions,[4-6] Chuang et al[7] genotyped one single nucleotide polymorphism (SNP) in the *HLA-DRB1* region whose minor G allele is inversely associated with PD (2056 cases, 2723 controls) and reported a significant positive interaction between self-reported smoking and rs660895-G: the inverse association between smoking and PD was stronger in carriers of the AA genotype compared to G-allele carriers.[7] Based on a smaller selected sample (837 cases, 918 controls), the study that identified an inverse association of the SE and V11 combination (SE+V11+) with PD also showed an interaction with smoking, but in the opposite direction: the inverse association between smoking and PD was restricted to SE+V11+ carriers.[1] The authors hypothesized that post-translational modifications of α-synuclein induced by smoking (citrullination/homocitrullination) explained this interaction.

We performed a large-scale independent replication of the *HLA-DRB1* × smoking interaction by performing a Mendelian randomization (MR) analysis using smoking predisposing genes as instrumental variables in strata defined by *HLA-DRB1*.

Subjects and Methods

Courage-PD

The Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) consortium pooled individual-level data from 35 studies and used the Neurochip array to genotype participants (Supplementary Appendix S1). Analyses are based on 26 studies with at least 50 cases or controls of European descent (12 424 cases, 9480 controls); participants' chracteristics are shown in Supplementary Table S1. Additional methods on genotyping and imputation of *HLA* alleles/haplotypes/AA are available as Supplementary Appendix S1. All studies were approved by local ethical committees following procedures of each country.

Smoking Initiation: Two-Sample Mendelian Randomization

Because self-reported smoking was not available in most studies, we used SNPs associated with smoking initiation to perform two-sample MR.[8] Summary statistics for the association between SNPs and smoking initiation (182 SNPs independently associated at $P < 5 \times 10^{-8}$) came from the GWAS and Sequencing Consortium of Alcohol and Nicotine use (n = 1 232 091, European descent) (Supplementary Appendix S1),[9] and those for associations with PD came from Courage-PD (Supplementary Table S2).

In Silico Prediction of Binding Affinity of *HLA-DRB1* Alleles to α-Synuclein

We assessed the binding affinity (nM) of *HLA-DRB1* alleles to α-synuclein derived peptides using NetMHCIIpan 4.0 and predicted whether peptides are strong, weak, or non-binders.[10] After targeting 607 four-digit *HLA-DRB1* alleles, we restricted our analyses to 34 alleles observed in Courage-PD. Of 126 α-synuclein derived peptides,[1] we retained 96 peptides with lysine residues that can be

homocitrullinated to examine the role of smoking-related post-translational modifications. We also performed analyses restricted to a single peptide (Tyrosine 39, Y39) that induces T cell responses in PD patients[11] and was previously used for binding affinity predictions.[2]

Statistical Analyses

We used SAS9.4 (SAS Institute Inc, Cary, NC, USA), STATA16 (StataCorp LP, College Station, TX, USA), and R packages HIBAG[12] and TwoSampleMR[13] (R Foundation for Statistical Computing, Vienna, Austria).

Interaction between Genetically Predicted Smoking Initiation and HLA-DRB1

To perform an independent replication of the *HLA-DRB1* × smoking interaction, we excluded the French study that contributed to identify the interaction between smoking and rs660895 in PD.[7]

We used the random-effects inverse-variance weighted (IVW)[8] approach to perform MR analyses for genetically predicted smoking initiation in two strata defined by the presence of V11 encoded by *HLA-DRB1* alleles (Supplementary Appendix S1). We compared the two MR estimates using the statistic $(\beta_2-\beta_1)/\sqrt{\left(\mathrm{SE}(\beta_2)^2+\mathrm{SE}(\beta_1)^2\right.}$, where β_1 and β_2 are MR estimates in the two strata with variances $\mathrm{SE}(\beta_1)^2$ and $\mathrm{SE}(\beta_2)^2$; this difference represents the interaction between smoking and *HLA-DRB1* and follows a normal distribution. In sensitivity analyses, we used other MR approaches that are less powerful, but more robust to pleiotropy (weighted median-method and mode-based, MR-PRESSO, MR-Lasso)[8]; we also performed analyses after excluding 31 pleiotropic SNPs associated with alcohol drinking and/or body mass index (Supplementary Appendix S1).

As secondary analyses, we ran MR analyses stratified by rs660895[7] and *HLA-DRB1**04,[3] which are both inversely associated with PD and in linkage disequilibrium with V11. Analyses stratified by rs660895 have the advantage that they did not involve *HLA* imputation and are, therefore, based on a larger number of cases and controls than analyses that required *HLA* imputation.

In Silico Prediction of Binding Affinity

To examine the influence of V11 encoded by *HLA-DRB1* alleles and homocitrullination (HC) of α-synuclein derived peptides on binding affinity, we described binding affinity for the four groups defined by the combination of V11 and HC. All 2 × 2 differences were tested using the Wilcoxon non-parametric test corrected for multiple comparisons.[14] We compared the percentage of binding peptides in the four groups using multinomial logistic regression.

Data Availability

Results can be reproduced using the Supplementary Appendix S1.

Results

Supplementary Table S3 shows 19 SNPs from the *HLA* region associated with PD after accounting for multiple comparisons, of which 17 were located near *HLA-DRB1* (including rs660895); none of them was associated with smoking initiation ($P > 0.05$). Among 64 alleles of *HLA* class 2 genes (*HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1*), five were significantly and inversely associated with PD (*HLA-DQA1**03:01, *HLA-DQA1**03:03; *HLA-DQB1**03:02; *HLA-DRB1**04:01, and *HLA-DRB1**04:04) (Supplementary Table S4). The odds ratio (OR) for the association of all *HLA-DRB1**04 alleles combined with PD was of 0.84 (95% confidence interval [CI], 0.78–0.91; $P = 3.9 \times 10^{-6}$). Associations between *DRB1* ~ *HLA-DQB1* haplotypes and PD are shown in Supplementary Table S4.

Among 131 AA encoded by *HLA-DRB1* and 116 by *HLA-DQB1*, 11 AA were associated (9 inversely, 2 positively) with PD and were all encoded by *HLA-DRB1* (Supplementary Table S5). Two AA, V11, and S37, remained significantly associated with PD after a backward stepwise selection procedure, with a stronger association for V11 (OR, 0.85; 95% CI, 0.79–0.92; $P = 2.2 \times 10^{-5}$) than S37 (OR, 1.07; 95% CI, 1.00–1.14; $P = 0.040$). The association of H13 and H33 with PD was explained by V11 (Supplementary Table S6). We found no significant interaction between SE and V11 ($P = 0.29$); only V11 remained associated with PD (OR, 0.81; 95% CI, 0.74–0.89; $P < 10–3$) when both were included in the model (Supplementary Table S7).

The overall association between genetically predicted smoking initiation and PD was of 0.86 (95% CI, 0.73–1.05; $P = 0.10$) without evidence of heterogeneity between SNPs ($P = 0.40$). Compared with 26% (n = 2212) of the controls, 22% (n = 2531) of the cases carried at least one V11 residue. Genetically predicted smoking initiation was inversely associated with PD in the absence of V11 (OR_{IVW}, 0.74; 95%, 0.59–0.93; $P = 0.0092$), but not in its presence (OR_{IVW}, 1.25; 95% CI, 0.83–1.87; $P = 0.29$), with a positive and significant interaction ($P = 0.03$) (Table 1, Fig. 1). There was no significant heterogeneity across SNPs and MR-Presso did not detect pleiotropy (all $P > 0.10$). Results of pleiotropy-robust approaches were consistent with the IVW method, although CIs were generally larger. Similar conclusions were reached after

TABLE 1 *Mendelian randomization analysis of the relation between genetically predicted smoking initiation (182 SNPs) and PD stratified by* HLA-DRB1

	0 allele or AA residue			1/2 alleles or AA residues				
HLA-DRB1	**OR per 1-SD increase in the prevalence of ever smoking (95% CI)**	***P***	***P*-het.**	**OR per 1-SD increase in the prevalence of ever smoking (95% CI)**	***P***	***P*-het.**	**Interaction OR (95% CI)[a]**	***P***
Valine 11[b]	6383 controls, 8812 cases			2212 controls, 2531 cases				
Inverse variance weighted	0.74 (0.59–0.93)	9.2×10^{-3}	0.73	1.25 (0.83–1.87)	0.29	0.40	1.68 (1.06–2.68)	0.0
Weighted median	0.75 (0.53–1.07)	0.11		1.14 (0.61–2.15)	0.68		1.52 (0.75–3.11)	0.26
Weighted mode	0.63 (0.30–1.31)	0.22		1.72 (0.38–7.82)	0.48		2.74 (0.51–14.77)	0.24
MR-Lasso	No invalid SNP ($\lambda = 0.20$)			1.30 (0.87–1.96)	0.20[c]		1.76 (1.10–2.81)	0.020
MR-PRESSO			0.59			0.47		
rs660895-G[d]	6498 controls, 8903 cases			2982 controls, 3521 cases				
Inverse variance weighted	0.73 (0.59–0.91)	4.8×10^{-3}	0.84	1.33 (0.95–1.87)	0.10	0.41	1.83 (1.22–2.74)	3.5×10^{-3}
Weighted median	0.72 (0.52–1.00)	0.05		1.04 (0.62–1.73)	0.89		1.45 (0.78–2.66)	0.24
Weighted mode	0.68 (0.31–1.48)	0.34		0.99 (0.23–4.26)	0.99		1.46 (0.30–7.08)	0.66
MR-Lasso	No invalid SNP ($\lambda = 0.19$)			1.25 (0.89–1.75)	0.20[e]		1.71 (1.14–2.56)	9.1×10^{-3}
MR-PRESSO			0.83			0.40		
*HLA-DRB1**04[b]	6563 controls, 9014 cases			2032 controls, 2329 cases				
Inverse variance weighted	0.73 (0.59–0.92)	6.8×10^{-3}	0.77	1.29 (0.83–2.00)	0.26	0.47	1.75 (1.07–2.87)	0.03
Weighted median	0.70 (0.50–0.97)	0.03		1.16 (0.59–2.29)	0.66		1.67 (0.81–3.46)	0.18
Weighted mode	0.67 (0.30–1.48)	0.32		1.51 (0.34–6.66)	0.59		2.26 (0.38–13.39)	0.34
MR-Lasso	No invalid SNP ($\lambda = 0.20$)			1.18 (0.76–1.83)	0.46[f]		1.61 (0.98–2.64)	0.06
MR-PRESSO			0.67			0.57		

Valine 11 amino acids and *HLA-DRB1**04 alleles were determined using imputation of *HLA* alleles and amino acids based on SNPs from the *HLA* region.
[a]The interaction OR represents the OR in carriers of 1/2 alleles or AA residues divided by the OR in carriers of 0 allele or AA residue.
[b]Total number: 8595 controls, 11,343 cases.
[c]Number of invalid SNPs = 4; $\lambda = 0.17$.
[d]Total number: 9480 controls, 12,424 cases.
[e]Number of invalid SNPs = 4; $\lambda = 0.19$.
[f]Number of invalid SNPs = 11; $\lambda = 0.19$.
SNPs, single nucleotide polymorphism; PD, Parkinson's disease; OR, odds ratio; CI, confidence interval; AA, amino acid; *P*-het., *P* for heterogeneity; λ, tuning parameter for MR-Lasso.

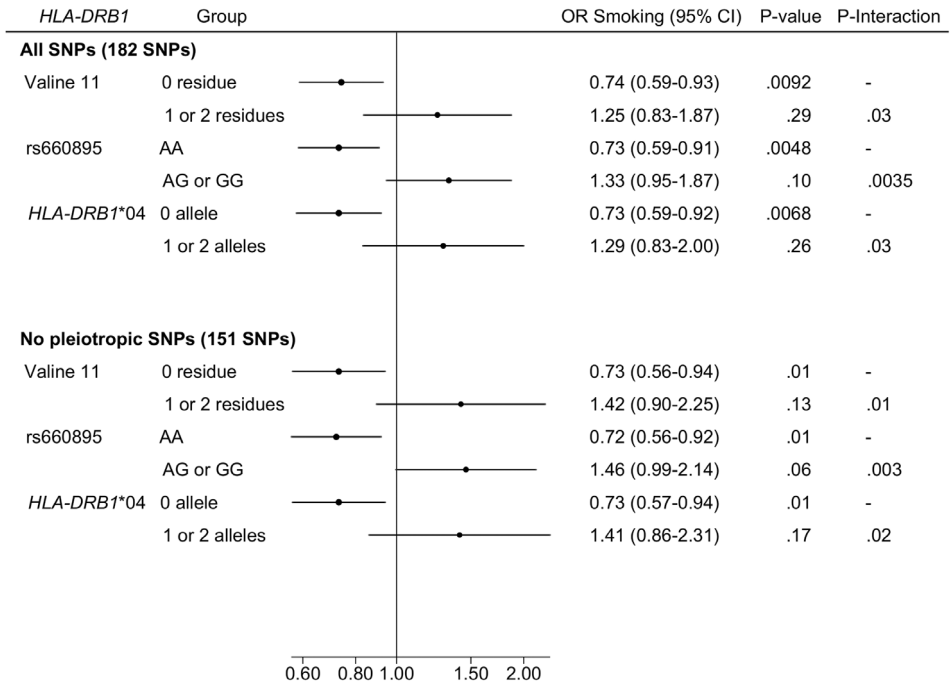


FIG. 1. Forest plot of the association between genetically predicted smoking initiation (inverse variance weighted estimate) and Parkinson's disease stratified by *HLA-DRB1*.

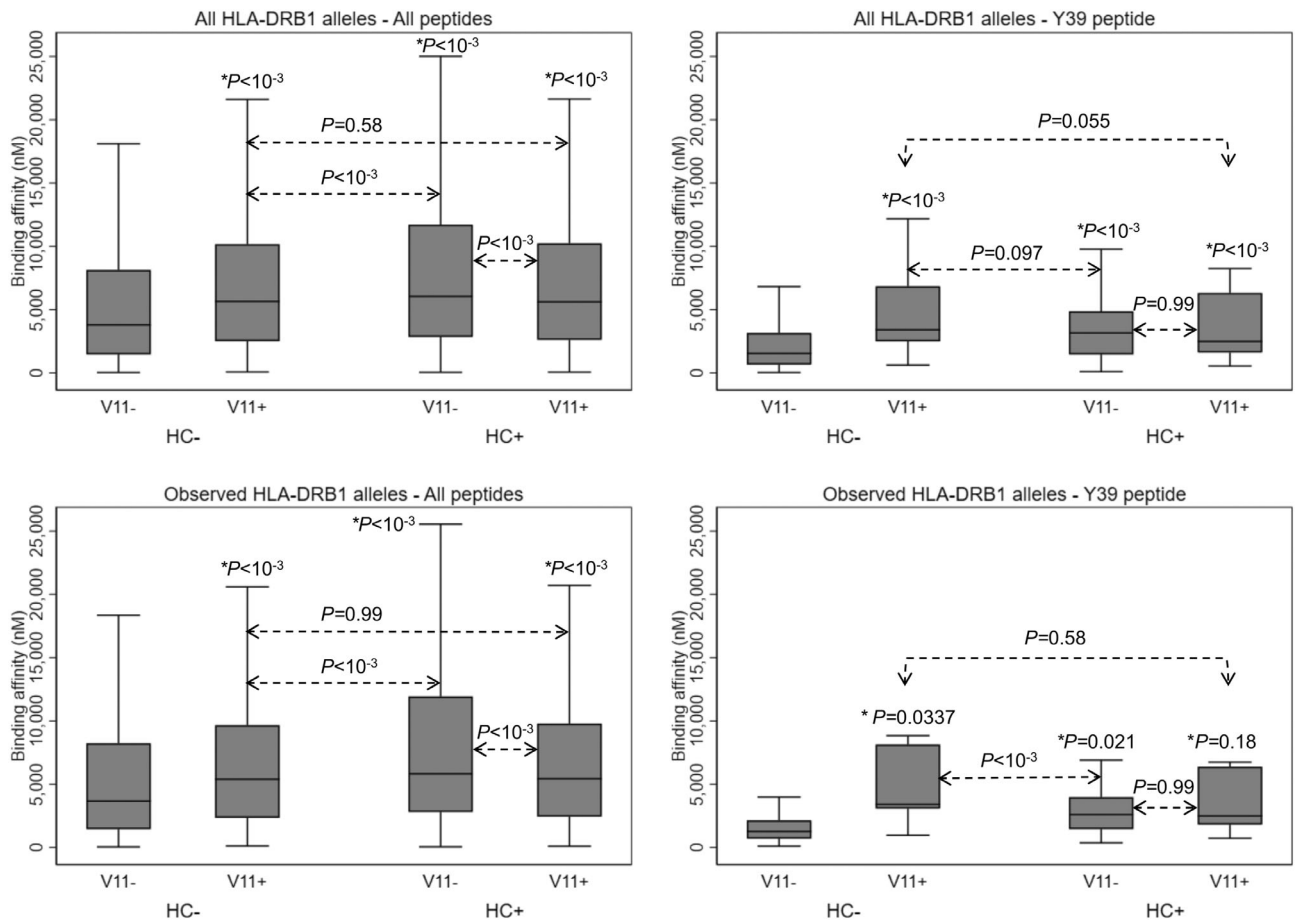


FIG. 2. Prediction of binding affinity (nM) according to the presence of a valine at position 11 (V11) coded by *HLA-DRB1* alleles and homocitrullination (HC) of α-synuclein derived peptides. **P* values for the comparison versus the reference group (V11−HC−).

excluding 31 pleiotropic SNPs (Fig. 1, Supplementary - Table S8). Results were similar in analyses stratified by rs660895 or *HLA-DRB1**04.

Compared to V11−HC−, V11+HC− and V11−HC+ were both associated with decreased binding affinity, with a stronger effect of HC+ than V11+ (Fig. 2, Supplementary - Table S9). Alternatively, in the presence of HC+, V11+ increased binding affinity (all peptides) or had no effect (Y39); HC+ had no effect on binding affinity in the presence of V11+. Analyses of binding and non-binding peptides paralleled these results (Supplementary Table S10).

Discussion

We replicate an interaction between *HLA-DRB1* and smoking,[7] according to which the inverse association between smoking and PD is only present in participants without protective *HLA-DRB1* AA/alleles. *In silico* predictions of binding affinity are consistent with an interaction between V11 and post-translational smoking-induced modifications of α-synuclein derived peptides.

Recent MR studies showed an inverse association between genetically predicted smoking and PD.[15-18] These findings are in favor of a causal role of smoking in PD, but the underlying mechanisms remain unknown and gene–environment interactions analyses may contribute to their understanding. The interaction pattern we found is similar to the interaction between self-reported smoking and rs660895 reported by Chuang et al.[7] Our study represents a fully independent replication using a different approach to define smoking (MR) and SNP-based imputation of *HLA* amino acids that allowed us to examine this interaction at the AA level. Therefore, our findings contradict those from Hollenbach et al[1] who reported an interaction in the opposite direction based on a selected sample of smaller size.

Lower binding affinity for α-synuclein derived peptides is associated with a weaker immune response that may explain decreased PD risk.[19] Our binding affinity analyses are consistent with the interaction pattern we identified. Although V11 and HC both decreased binding affinity for α-synuclein derived peptides in the absence of each other, consistent with the inverse association of V11 and smoking with PD, there was a positive interaction between V11 and HC, whereby both V11 and HC had a weaker or no effect in the presence of each other; this pattern is consistent with the lack of association between smoking and PD in V11 carriers that we found.

We used MR to define genetically predicted smoking initiation, rather than self-reported smoking; MR has the advantage that, provided that a set of assumptions are met, smoking-PD association estimates are less likely to be biased by confounding or reverse causation than association estimates based on self-reported smoking.[8] Another strength of our study compared to Chuang et al[7] is that rather than using a single SNP, we used genome-wide data to impute AA encoded by *HLA-DRB1*. Finally, using an independent dataset, we report similar associations with *HLA* alleles and AA as previous studies.[2,3] One limitation of our *HLA-DRB1* × smoking interaction analyses is that the approach we used allowed us to estimate the association between smoking initiation and PD stratified by *HLA-DRB1*, but did not allow us to estimate the association between *HLA-DRB1* and PD stratified by smoking.

Despite being one of the most robust findings in PD, the mechanisms underlying its inverse association with smoking remain unknown. This work represents the first example of large-scale replication of a gene–environment interaction in PD, and allows proposing a biological mechanism to explain the inverse smoking-PD association, in the context of a larger body of work on the relationship between the immune system and PD.[19]

Acknowledgments: We thank the GWAS and Sequencing Consortium of Alcohol and Nicotine consortium (GSCAN use) for providing summary statistics for this analysis. Additional Courage-PD investigators are: Sophia N. Pchelina (Saint Petersburg, Russia), Thomas Brücke (Wien, Austria), Marie-Anne Loriot (Paris, France), Claire Mulot (Paris, France), Yves Koudou (Villejuif, France), Alain Destée (Lille, France), Georgia Xiromerisiou (Larissa, Greece), Christos Koros (Athens, Greece), Matina Maniati (Athens, Greece), Maria Bozi (Athens, Greece), Micol Avenali (Pavia, Italy), Margherita Canesi (Milan, Italy), Giorgio Sacilotto (Milan, Italy), Michela Zini (Milan, Italy), Roberto Cilia (Milan, Italy), Francesca Del Sorbo (Milan, Italy), Nicoletta Meucci (Milan, Italy), Rosanna Asselta (Milan, Italy), Radha Procopio (Catanzaro, Italy), Clara Hellberg (Lund, Sweden), Manabu Funayama (Tokyo, Japan), Aya Ikeda (Tokyo, Japan), Takashi Matsushima (Tokyo, Japan), Yuanzhe Li (Tokyo, Japan), Hiroyo Yoshino (Tokyo, Japan), Zied Landoulsi (Luxembourg, Luxembourg), Rubén Fernández-Santiago (Barcelona, Spain), Nicholas Wood (London, UK), Huw R. Morris (London, United Kingdom).

Data Availability Statement

Results can be reproduced using the Supplementary material

References

1. Hollenbach JA, Norman PJ, Creary LE, et al. A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. Proc Natl Acad Sci U S A 2019;116:7419–7424.
2. Naito T, Satake W, Ogawa K, et al. Trans-ethnic fine-mapping of the major histocompatibility complex region linked to Parkinson's disease. Mov Disord 2021;36:1805–1814.
3. Yu E, Ambati A, Andersen MS, et al. Fine mapping of the HLA locus in Parkinson's disease in Europeans. NPJ Parkinsons Dis 2021;7:84
4. Hedström AK, Hillert J, Brenner N, et al. DRB1-environment interactions in multiple sclerosis etiology: results from two Swedish case-control studies. J Neurol Neurosurg Psychiatry 2021;92:717–722.
5. Karlson EW, Chang SC, Cui J, et al. Gene-environment interaction between HLA-DRB1 shared epitope and heavy cigarette smoking in predicting incident rheumatoid arthritis. Ann Rheum Dis 2010;69:54–60.
6. Baecklund F, Foo JN, Askling J, et al. Possible interaction between cigarette smoking and HLA-DRB1 variation in the risk of follicular lymphoma. Am J Epidemiol 2017;185:681–687.

7. Chuang YH, Lee PC, Vlaar T, et al. Pooled analysis of the HLA-DRB1 by smoking interaction in Parkinson disease. Ann Neurol 2017;82:655–664.
8. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing Mendelian randomization investigations. Wellcome Open Res 2019;4:186
9. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 2019;51:237–244.
10. Reynisson B, Barra C, Kaabinejadian S, et al. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J Proteome Res 2020;19:2304–2315.
11. Sulzer D, Alcalay RN, Garretti F, et al. T cells from patients with Parkinson's disease recognize α-synuclein peptides. Nature 2017; 546:656–661.
12. Zheng X, Shen J, Cox C, et al. HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J 2014;14:192–200.
13. Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. Elife 2018;7:e34408
14. Douglas CE, Michael FA. On distribution-free multiple comparisons in the one-way analysis of variance. Commun Stat - Theory Methods 1991;20:127–139.
15. Grover S, Lill CM, Kasten M, et al. Risky behaviors and Parkinson disease: a mendelian randomization study. Neurology 2019;93: e1412–e1424.
16. Heilbron K, Jensen MP, Bandres-Ciga S, et al. Unhealthy behaviours and risk of Parkinson's disease: a Mendelian randomisation study. J Parkinsons Dis 2021;11:1981–1993.
17. Dominguez-Baleon C, Ong JS, Scherzer CR, et al. Understanding the effect of smoking and drinking behavior on Parkinson's disease risk: a Mendelian randomization study. Sci Rep 2021;11:13980
18. Domenighetti C, Sugier PE, Sreelatha AAK, et al. Mendelian randomisation study of smoking, alcohol, and coffee drinking in relation to Parkinson's disease. J Parkinsons Dis 2022;12:267–282.
19. Tan EK, Chao YX, West A, et al. Parkinson disease and the immune system - associations, mechanisms and therapeutics. Nat Rev Neurol 2020;16:303–318.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

SGML and CITI Use Only
DO NOT PRINT

Author Roles

(1) Research project: A. Conception, B. Organization, C. Execution; (2) Statistical Analysis: A. Design, B. Execution, C. Review and Critique; (3) Manuscript: A. Writing of the FIRST DRAFT, B. Review and Critique.

C.D.: 1A, 1B, 1C, 2A, 2B, 2C, 3A, 3B
V.D.: 1A, 1B, 1C, 2A, 2C, 3B
P.E.S.: 1A, 1B, 1C, 2A, 2C, 3B
A.A.K.S.: 1B, 1C, 2C, 3B
C.S.: 1B, 1C, 2C, 3B
S.G.: 1B, 1C, 2C, 3B
P.M.: 1B, 1C, 2C, 3B
D.R.B.: 1B, 1C, 2C, 3B
M.R.B.: 1A, 1B, 1C, 2C, 3B
P.L.: 1A, 1B, 1C, 2C, 3B
A.B.S.: 1B, 1C, 2C, 3B
D.G.H.: 1B, 1C, 2C, 3B
C.E.: 1B, 1C, 2C, 3B
P.A.G.: 1B, 1C, 2C, 3B
G.D.M.: 1B, 1C, 2C, 3B
A.Z.: 1B, 1C, 2C, 3B
W.P.: 1B, 1C, 2C, 3B
E.R.: 1B, 1C, 2C, 3B
A.E.L.: 1B, 1C, 2C, 3B
S.K.: 1B, 1C, 2C, 3B
P.T.: 1B, 1C, 2C, 3B
S.L.: 1B, 1C, 2C, 3B
A.B.: 1B, 1C, 2C, 3B
J.C.C.: 1B, 1C, 2C, 3B
M.C.C.H.: 1B, 1C, 2C, 3B
E.M.: 1B, 1C, 2C, 3B
K.B.: 1B, 1C, 2C, 3B
A.B.D.: 1B, 1C, 2C, 3B
G.M.H.: 1B, 1C, 2C, 3B
E.D.: 1B, 1C, 2C, 3B
L.S.: 1B, 1C, 2C, 3B
A.M.S.: 1B, 1C, 2C, 3B
E.M.V.: 1B, 1C, 2C, 3B
S.P.: 1B, 1C, 2C, 3B
S.D.: 1B, 1C, 2C, 3B
L.S.: 1B, 1C, 2C, 3B
A.Z.: 1B, 1C, 2C, 3B
G.P.: 1B, 1C, 2C, 3B
L.B.: 1B, 1C, 2C, 3B
C.F.: 1B, 1C, 2C, 3B
G.A.: 1B, 1C, 2C, 3B
A.Q.: 1B, 1C, 2C, 3B
M.G.: 1B, 1C, 2C, 3B
H.M.: 1B, 1C, 2C, 3B
Y.K.: 1B, 1C, 2C, 3B
N.H.: 1B, 1C, 2C, 3B

K.N.: 1B, 1C, 2C, 3B
S.J.C.: 1B, 1C, 2C, 3B
Y.J.K.: 1B, 1C, 2C, 3B
P.K.: 1B, 1C, 2C, 3B
B.P.C.v.d.W.: 1B, 1C, 2C, 3B
B.R.B.: 1B, 1C, 2C, 3B
J.A.: 1B, 1C, 2C, 3B
M.T.: 1B, 1C, 2C, 3B
L.P.: 1B, 1C, 2C, 3B
L.C.G.: 1B, 1C, 2C, 3B
J.J.F.: 1B, 1C, 2C, 3B
S.B.: 1B, 1C, 2C, 3B
J.C.: 1B, 1C, 2C, 3B
E.T.: 1B, 1C, 2C, 3B
M.E.: 1B, 1C, 2C, 3B
P.P.: 1B, 1C, 2C, 3B
M.D.F.: 1B, 1C, 2C, 3B
K.W.: 1B, 1C, 2C, 3B
N.L.P.: 1B, 1C, 2C, 3B
C.R.: 1B, 1C, 2C, 3B
A.C.B.: 1B, 1C, 2C, 3B
A.P.: 1B, 1C, 2C, 3B
E.Y.R.: 1B, 1C, 2C, 3B
C.E.C.: 1B, 1C, 2C, 3B
K.E.M.: 1B, 1C, 2C, 3B
M.T.: 1B, 1C, 2C, 3B
D.K.: 1B, 1C, 2C, 3B
L.F.B.: 1B, 1C, 2C, 3B
M.J.F.: 1B, 1C, 2C, 3B
R.K.: 1A, 1B, 1C, 2C, 3B
T.G.: 1A, 1B, 1C, 2C, 3B
M.S.: 1A, 1B, 1C, 2C, 3B
N.V.: 1A, 1B, 1C, 2A, 2C, 3B
A.E.: 1A, 1B, 1C, 2A, 2B, 2C, 3A, 3B

Supplementary methods

Courage-PD international consortium

The Courage-PD (COmprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease) international consortium pooled individual-level data from 35 studies on Parkinson's disease from different populations worldwide and used the same array to genotype participants.

The Geo-PD (Genetic Epidemiology Of Parkinson's disease; https://geopd.biomedinfo.org/) consortium represents one of the components of Courage-PD. This consortium aims at conducting collaborative studies on genetic susceptibility in Parkinson's disease; one of its main features is that participating sites are distributed in the five continents, therefore representing a highly diverse population. In addition, several other studies from Europe contributed to Courage-PD. Parkinson's disease was diagnosed using standard criteria (United Kingdom Parkinson's Disease Society Brain Bank - UKPDSBB, Gelb, Bower).[1-3]

All studies were approved by local ethical committees following the procedures of each country, and material transfer agreements were set up between participating sites and the University of Tübingen (Germany).

According to the study's consortium agreement, participating sites contributed DNA and demographic/environmental data. DNAs (25µl of DNA at a concentration of 50 to 100 ng/µl) were shipped for quality control to University of Tübingen (Germany) and genotyped in a central laboratory in Munich (Institute of Human Genetics, Helmholtz Zentrum, Germany). The samples from two sites (Gasser, Morris/Wood) were genotyped at the Laboratory of Neurogenetics (National Institute on Aging, National Institutes of Health, Bethesda, USA). Demographic data were harmonized and collected using a standardized form and cleaned at Inserm U1018 (Villejuif, France).

In a previous study on the relation between genetically-predicted smoking and PD (7369 cases, 7018 controls), we excluded samples overlapping with iPDGC (to allow for an independent replication) and participants with a positive family history of PD or with Mendelian PD mutations (to avoid dilution of effects of environmental exposures).[4] For the present analyses whose main objective was to examine the interaction between genetically-predicted smoking and *HLA-DRB1*, we retained samples overlapping with iPDGC and those with familial PD in order to increase the sample size and the statistical power to detect an interaction, as the number of participants required to detect interactions is larger than for main effects. Analyses are therefore based on 26 studies with at least 50 cases or controls of European descent (12424 cases, 9480 controls; Supplementary Table 1).

Genotyping

The Neurochip chip was used to genotype all the samples.[5] Briefly, this chip is a custom-designed array containing a tagging variant backbone with good genome-wide resolution of about 306,670, complemented with a manually curated custom content comprised of 179,467 variants implicated in diverse neurological diseases, including Parkinson's disease.

Genotyping was performed with an automated protocol according to the manufacturer's instructions (Illumina, San Diego, CA, USA). All arrays were scanned with an Illumina iScan and raw data were analyzed with the Illumina Beeline and GenomeStudio software packages using the manifest file Neuro_Consortium_20013217_A1.bpm. Clustering was performed in GenomeStudio with the GenTrain cluster 2.0 algorithm. Genotypes were post-processed with zCall (DOI: 10.1093/bioinformatics/bts479) to improve detection of rare alleles.

Quality control

Genotyped data exported from Genome Studio to PLINK format were used for quality control and downstream analysis. The pooled dataset from the 35 sites consists of 27,538 subjects. Phenotypic data were missing for 245 subjects and three sites were removed as they only included cases, leaving 26,535 subjects (14,859 cases, 11,431 controls) for quality control using "COMRARE" an automated pipeline under development at the University of Tübingen. The pipeline uses PLINK (https://www.cog-genomics.org/plink) and R scripts (R Foundation for Statistical Computing, Vienna, Austria) and the following steps were implemented separately for each site:

1- Per individual quality control:

- Identification of individuals with elevated missing genotyping rates or outlying heterozygosity rate: individuals with a genotype failure rate ≥4% or heterozygosity rate ± 4 standard deviations from the mean were excluded.

- Identification of Individuals with discordant sex information: the homozygosity rate was calculated for X-linked SNPs for each individual and compared to the expected rate. Participants for whom phenotypic and genotypic sex were discordant were removed.
- Identification of duplicated or related Individuals: pairs of individuals with an identity by descent (IBD) greater than 0.185 were removed.
- Eigensoft software was used to compute principal components in order to correct for population stratification by merging our dataset with HapMap.[6, 7] A scatter plot of the first two principal components was used to identify outliers in each study.

2- *Per marker quality control:*

- Identification of markers with high missing data rate: a call-rate threshold of 96% was used and SNPs with a lower rate were removed.
- SNPs with a significant ($P<10^{-5}$) difference in rates of missing values between cases and controls were removed.
- We excluded variants with a minor allele frequency (MAF) $<5\times10^{-8}$ and those in Hardy-Weinberg disequilibrium ($P<5\times10^{-8}$).

Imputation

After QC, we used the HRC/1000G imputation preparation and checking tool (https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.3.0.zip) to check for Ref/Alt allele assignments, incorrect strands, deviation from allele frequency, and palindromic SNPs. Imputation of autosomal variants was performed separately for each dataset based on based on 271,398 to 373,664 SNPs in each study, through the Michigan Imputation server using the HRC reference panel and the GRCh37/hg19 assembly with a R^2 filter of 0.3. The mean of the number of SNPs available in each study after imputation was 13,710,549 (SD=2,986,478).

Imputation of HLA alleles

Sequence-based typing of *HLA* alleles is considered the gold standard for *HLA* typing, but is time-consuming and cost-prohibitive for research purposes. An alternative approach that has considerably developed over the past 10 years is SNP-based imputation of *HLA* alleles. These methods take advantage of the extended haplotype structure within the *HLA* region to predict *HLA* alleles using dense SNP genotypes such as those available in GWAS datasets and require access to large and ethnically diverse training datasets including both SNP and sequenced *HLA* alleles.[8]

Four-digit alleles of *HLA* class I (*HLA*-A, *HLA*-B, *HLA*-C) and II (*HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1*) genes were imputed based on GWAS data using the HIBAG R package (R Foundation for Statistical Computing, Vienna, Austria) separately in each of the 26 Courage-PD studies including European individuals (12,424 cases, 9,480 controls; Supplementary Table 2) using the European reference panel provided by HIBAG.[8, 9] Given the results of previous studies that highlighted *HLA* class II genes in PD, the remainder of our analyses focused on these genes. Two-digit alleles (*e.g.*, *HLA-DRB1**04) were derived by combining four-digit alleles (*e.g.*, *HLA-DRB1**04:01, *HLA-DRB1**04:02, etc.).

Allelic imputation probabilities were high in all studies and comparable between cases and controls. As recommended in the original publication, we retained alleles with an imputation probability greater than 50% as this threshold represents the best compromise between accuracy and call rate.[8] At this threshold, HIBAG prediction accuracies ranged from 94.8% to 99.2% for four-digit alleles for individuals of European descent, with call rates between 90.1% and 98.8%.[8] A study that imputed *HLA* alleles in Europeans using HIBAG with a 50% threshold compared the results of *HLA-DRB1* imputation to high throughput *HLA* sequencing in 380 PD samples from Norway.[10] There was good agreement between imputed and sequenced alleles: for *HLA-DRB1**04:01 alleles, sensitivity was 96% and specificity was 91%; for *HLA-DRB1**04:04 alleles, sensitivity was 97% and specificity was 92%; for *HLA-DRB1**04 alleles, agreement was excellent (sensitivity, 98%; specificity, 96%).[10] In our study, on average, 3% of participants were excluded from the analyses due to an allelic imputation probability ≤50%; depending on the gene, the number of cases included in the analyses ranged from 11,343 to 12,341 and the number of controls from 8,595 to 9,417. In sensitivity analyses based on a more stringent probability threshold of 80%, we verified that associations between *HLA* alleles and PD remained stable although the number of cases and controls decreased by ~29%.

Haplotype and amino acid imputation of the HLA region

Imputation of *A~B~C~DRB1~DQB1* haplotypes and amino-acids (AA) was performed separately in each of the 26 Courage-PD studies including European individuals using Easy-HLA (https://hla.univ-nantes.fr/).[11] Easy-HLA implements a computational and statistical method of HLA haplotypes inference based on a published reference population containing over 600,000 haplotypes to upgrade missing or partial HLA information; it also provides additional functional annotations including AA of proteins coded by *HLA* alleles.

Subjects with allelic or haplotypic imputation probabilities ≤50% were excluded from subsequent analyses (~15%) which are therefore based on 10,613 cases and 8,056 controls. As for *HLA* alleles, we restricted our analyses to *HLA* class II genes and defined *HLA-DRB1~HLA-DQB1* haplotypes (n=89).

We determined *HLA-DRB1* alleles that carried the shared epitope (SE), a five AA sequence motif in positions 70-74 of HLA-DRβ chains encoded by HLA-DRB1 alleles, using the IPD-IMGT/HLA database (Release 3.44.0, https://www.ebi.ac.uk/ipd/imgt/hla/).

Statistical analyses

Association between HLA and Parkinson's disease

We identified SNPs located in the *HLA* region between positions chr6:28,477,895 and chr6:33,448,264. Of 38,946 SNPs available in the HLA region, we selected 29,515 SNPs available in ≥20 studies. In addition to SNPs with a minor allele frequency (MAF) <1% and in Hardy-Weinberg disequilibrium ($P<5\times10^{-8}$), we also excluded those with low imputation quality ($r^2<0.8$). We examined their association with Parkinson's disease in each study under an additive model using logistic regression adjusted for sex and the first four principal components (PC) using PLINK software (v1.9).[12] We used the GWAMA software[13] to pool summary statistics according to a fixed ($I^2\leq25\%$) or random ($I^2>25\%$) effects model and to calculate ORs and their 95% CI.

Allele frequencies of *HLA* class II genes, haplotypes, and AA were estimated in cases and controls per study and then meta-analysed to estimate pooled frequencies.

The association of Parkinson's disease with alleles of *HLA* class II genes, *DRB1~DQB1* haplotypes, and AA encoded by *HLA-DRB1* and *HLA-DQB1* was estimated under a dominant model using logistic regression adjusted for study, sex, and the first four principal components (PC). Alleles and haplotypes with a pooled frequency ≤0.005 in controls were excluded from the analyses. A total of 64 alleles of *HLA* class II genes (39 were available in 26 studies and 25 in 18-25 studies) and 24 *DRB1~DQB1* haplotypes (12 were available in 26 studies and 12 in 18-25 studies) were considered. We selected 247 AA available in at least 20 studies whose frequency in controls was ≥0.005 and ≤0.995 (226 available in 26 studies); after independent analyses for each AA, we used a backward stepwise selection procedure starting from a model including all AA and their two-way interactions. We also estimated Spearman correlations between AA significantly associated with Parkinson's disease in controls.

For all of the analyses described above, we accounted for multiple testing by calculating adjusted P-values (q-values) based on the false discovery rate (FDR) and considered those ≤0.05 as statistically significant.[14]

In order to assess the independent and joint effects of SE and V11, we constructed a four-class categorical variable corresponding to the combination of these two variables (dominant model); the interaction between the two variables was tested by including a multiplicative interaction term in the model. Given the absence of a significant interaction, we then ran a multivariable model including SE and V11 to estimate their independent effects.

Interaction between HLA-DRB1 and genetically-predicted smoking initiation: stratified two-sample Mendelian randomisation

For our Mendelian randomization analysis, we excluded the Elbaz site as it was included in the initial study that showed an interaction between rs660895 and cigarette smoking in Parkinson's disease.[15]

Mendelian randomization uses genetic variants, mostly SNPs, associated with an exposure to estimate its causal effect on an outcome. In two-sample Mendelian randomization, the SNP-exposure and SNP-outcome association estimated (beta and SE) come from two independent samples. To be valid instrumental variable (IV), the SNPs must verify three assumptions: (i) they must be associated the exposure, (ii) they must not be directly associated with the outcome except through the exposure and (iii) they must not be associated with unmeasured confounders of the exposure-outcome association.[16]

After clumping to retain independent SNPs ($r^2>0.001$, genomic region=10,000 kb, based on European ancestry reference data from the 1000 Genomes Project), we selected a total of 203 SNPs associated with smoking initiation

in the GWAS and Sequencing of Alcohol and Nicotine use consortium (GSCAN),[17] of which 3 were not available in Courage-PD and 18 were excluded (palindromic SNPs with a minor allelic frequency≥0.42; MAF<0.01; SNPs not available in at least 17 studies), leaving 182 SNPs for our Mendelian randomization analyses (F-statistic=205.75; **Supplementary table 2**); these SNPs explain 3% (R^2) of the variance of the exposure (smoking initiation).[4, 18] Of these SNPs, 7 (3.8%) were genotyped and 175 (96.2%) were imputed; when genotyped SNPs were available, we selected them in priority.

We estimated associations between each of the 182 SNPs and PD in strata defined by *HLA-DRB1* (V11, rs660895, *HLA-DRB1*04*). For each site, logistic regression adjusted for sex and the first four principal components was performed for each SNP under an additive genetic model (number of alleles for genotyped SNPs, dosage for imputed SNPs) using PLINK software (version 1.9).[12] Effects size and standard errors from the 26 studies were meta-analysed using the GWAMA software.[13] Those with low heterogeneity across studies (I^2≤25%) were combined using a fixed-effect model, while we used a random-effects model for SNPs with high heterogeneity (I^2>25%).

We used the random effects inverse-variance weighted (IVW)[19] approach in two strata defined by *HLA-DRB1* (V11, rs660895, *HLA-DRB1*04*), based on summary statistics corresponding to the association between SNPs and Parkinson's disease in Courage-PD, and the association between SNPs and smoking initiation in GSCAN. Heterogeneity between IVs was tested using the Cochran's Q-statistic. Compared to analyses stratified by V11 or *HLA-DRB1*04* that relied on imputation of *HLA* alleles, analyses stratified by rs660895 did not require *HLA* imputation and are based on a larger number of cases and controls.

In sensitivity analyses, we used other MR approaches less sensitive to pleiotropic effects.[18, 20] The weighted median-method provides consistent estimates if at least 50% of the SNPs are valid instruments.[21] The weighted mode-based method uses the causal estimate from each SNP to calculate the modal estimate. The largest group of variants with the same causal estimate in the asymptotic limit are considered valid instruments.[22] MR-PRESSO allows to detect outliers (global test), to compute an estimate corrected for horizontal pleiotropy after removing them (if p-global test<0.05), and to test the difference between the original and updated estimates (distortion test).[23] The MR-Egger method can detect directional pleiotropy and provide corrected effect estimates, but it requires the strong and non-verifiable InSIDE (INstrument Strength Independent of Direct Effect) assumption and in some situations can lead to biased estimates and inflated Type I error.[24] We used the I^2_{GX} statistic to quantify the strength of regression dilution bias in SNP-exposure association estimates for MR-Egger, with values <90% indicating violation of the NOME (No Measurement Error) assumption.[25] There was no evidence of directional pleiotropy in any strata (P>0.30) but I^2_{GX} was of 0.66 in our study. For these reasons, we chose to use MR-Lasso instead, another robust method in which an individual intercept is included for each SNP, thus allowing to detect individual pleiotropic SNPs rather than testing for global directional pleiotropy (as MR-Egger does).[26] The penalty term in Lasso regression shrinks the intercepts and forces some of them to be 0; genetic variants whose intercept is 0 are considered as valid instruments and used in the analysis, while the others are considered as pleiotropic variants and are excluded. The degree of shrinkage is determined by the value of the tuning parameter λ. MR-Lasso is less sensitive than MR-Egger to violations of the InSIDE assumption.

We also searched for pleiotropic SNPs associated with PD or with exposures associated with PD (body mass index, alcohol drinking, coffee drinking) at a genome-wide significant level using the Phenoscanner database and phenoscanner R package to examine whether some SNPs were associated with PD or with environmental exposures associated with PD.[27, 28] We identified 7 SNPs associated with alcohol drinking and 26 SNPs associated with BMI (2 in common), and repeated the analyses after excluding 31 SNPs.

References

1. Gelb DJ, Oliver E, Gilman S. Diagnostic criteria for Parkinson disease. Arch Neurol 1999;56:33-39.
2. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry 1988;51:745-752.
3. Bower JH, Maraganore DM, McDonnell SK, Rocca WA. Incidence and distribution of parkinsonism in Olmsted County, Minnesota, 1976-1990. Neurology 1999;52:1214-1220.
4. Domenighetti C, Sugier PE, Sreelatha AAK, et al. Mendelian Randomisation Study of Smoking, Alcohol, and Coffee Drinking in Relation to Parkinson's Disease. J Parkinsons Dis 2022;12:267-282.
5. Blauwendraat C, Faghri F, Pihlstrom L, et al. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol Aging 2017;57:247.e249-247.e213.
6. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-909.
7. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics 2006;2:e190.
8. Zheng X, Shen J, Cox C, et al. HIBAG—HLA genotype imputation with attribute bagging. The Pharmacogenomics Journal 2014;14:192-200.
9. Zheng X. Imputation-Based HLA Typing with SNPs in GWAS Studies. Methods Mol Biol 2018;1802:163-176.
10. Yu E, Ambati A, Andersen MS, et al. Fine mapping of the HLA locus in Parkinson's disease in Europeans. NPJ Parkinsons Dis 2021;7:84.
11. Geffard E, Limou S, Walencik A, et al. Easy-HLA: a validated web application suite to reveal the full details of HLA typing. Bioinformatics 2020;36:2157-2164.
12. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.
13. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 2010;11:288.
14. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440-9445.
15. Chuang YH, Lee PC, Vlaar T, et al. Pooled analysis of the HLA-DRB1 by smoking interaction in Parkinson disease. Ann Neurol 2017;82:655-664.
16. Lawlor DA, Harbord RM, Sterne JA, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med 2008;27:1133-1163.
17. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 2019;51:237-244.
18. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing Mendelian randomization investigations. Wellcome Open Res 2019;4:186.
19. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 2013;37:658-665.
20. Slob EAW, Burgess S. A comparison of robust Mendelian randomization methods using summary data. Genet Epidemiol 2020;44:313-329.
21. Bowden J, Davey SG, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol 2016;40:304-314.
22. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol 2017;46:1985-1998.
23. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet 2018;50:693-698.
24. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol 2017;32:377-389.
25. Bowden J, Del Greco MF, Minelli C, et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. Int J Epidemiol 2016;45:1961-1974.
26. Rees JMB, Wood AM, Dudbridge F, Burgess S. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. PLoS One 2019;14:e0222362.
27. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019;35:4851-4853.
28. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics 2016;32:3207-3209.

Discussion

« Every decision is made in darkness. Only by making a choice can we learn whether it was right or not. »

The Prisoner, Outer Wilds

VII - Discussion

À l'instar de l'évolution rapide des techniques génomiques qui a amené les chercheurs à constamment repenser l'analyse de données pangénomiques, la redécouverte constante du HLA a transformé sa compréhension et son analyse durant ces dix dernières années. À l'intersection des domaines de la bioinformatique et de l'immunogénétique, des algorithmes toujours plus rapides et efficaces se sont succédés pour faciliter le typage HLA à partir de données de séquençage ou le prédire à partir de panels de référence. Ainsi, l'imputation HLA a également commencé à combler le manque de données global en faisant levier des GWAS existantes pour augmenter le pouvoir statistiques des études. Pourtant, elle fait face à de nombreux obstacles liés à l'immense diversité de la région du CMH dans laquelle se trouvent les gènes du HLA. Les travaux de cette thèse et du projet SHLARC s'inscrivent dans un effort continu de la communauté d'immunogénétique pour améliorer l'imputation HLA, ainsi que dans un dessein plus global de démocratisation de l'analyse HLA.

Avec l'imputation HLA d'individus d'ancestralité composite, nous avons lancé un effort méthodologique parallèle à celui de l'augmentation de la taille des panels de référence, dont on peut comparer les efficacités.

VII.1 - La trajectoire de l'imputation HLA

VII.1.1 - La recherche de pistes d'amélioration par la génération de nouvelles données

L'imputation HLA prend sa source dans les modèles d'imputation SNP de 2008 par Leslie *et al.* (317). Ils disposaient déjà d'une grande précision (plus de 95%) dans des populations diverses en se basant sur le projet HapMap, avec environ 500 individus européens, africains et asiatiques (25). La comparaison avec les travaux actuels n'est pourtant plus possible car les données ne sont plus les mêmes et la diversité HLA non plus. En effet, ces modèles nécessitaient des données déjà phasées, mais surtout seulement 3 000 allèles HLA étaient connus, alors que ce nombre a été multiplié par dix en moins de vingt ans. De plus, l'imputation HLA tend maintenant à se concentrer sur des populations sous-représentées, sur un plus grand nombre de gènes HLA et, de fait, sur des allèles plus rares, récupérés grâce à un séquençage plus exhaustif.

Le sujet de cette thèse se concentre sur les populations sous-représentées et d'ancestralité composite. Ces travaux sont construits à partir des observations de la communauté scientifique sur l'impact de la diversité dans l'imputation HLA.

En préambule, il est intéressant de suivre les résultats des études menées sur l'impact de la diversité sur l'imputation des polymorphismes SNP. Herzig *et al.* ont ainsi pu montrer qu'il est possible de complémenter les panels de référence multi-ethniques des plateformes d'imputation par des données

d'une population absente. L'ajout de données d'individus français, alors que le panel de référence contient des européens majoritairement du Royaume-Uni, augmente la précision d'imputation. Notamment, l'imputation hybride, en choisissant le génotype imputé le plus probable des deux modèles, semble améliorer l'imputation des allèles rares (379). Ces résultats sont également répliqués avec une cohorte de 37 000 individus estoniens, où l'association avec les variants rares et codants est largement améliorée (380).

Dans les méthodes actuelles d'imputation HLA, deux stratégies principales ont été mises en place pour améliorer la précision : la création des panels de référence spécifiques ou multi-ethniques. Huang *et al.* ont ainsi construit leur propre panel de référence d'imputation HLA composé d'individus de la biobanque Taiwanaise, sans évaluer néanmoins son efficacité face à d'autres panels (381). D'autres études ont démontré qu'un panel de référence spécifique (ex. japonais) a de meilleures performances d'imputation que des panels européens ou multi-ethniques (382). Ritari *et al.* sont arrivés aux mêmes conclusions avec un panel de référence d'individus finlandais, en montrant aussi que le choix du génotype HLA de plus haute probabilité (entre européens et finlandais) pouvait améliorer les résultats pour HLA-C et HLA-DQA1 notamment (336).

À l'inverse, certains chercheurs ont créé des panels de référence avec le plus de diversité possible. Degenhardt *et al.* ont ainsi réuni 2 276 individus d'Asie, d'Europe et du Moyen-Orient et démontré l'amélioration de l'imputation HLA par rapport à un modèle construit avec les données 1KG seulement (346). Le plus important panel de référence multi-ethnique à ce jour est celui de Luo *et al.* qui regroupe plusieurs biobanques pour 21 546 individus d'ancestralités diverses. Ils ont pu mettre en évidence une meilleure précision d'imputation dans plusieurs populations, en comparaison à un panel de référence européen, même en réduisant la taille du panel artificiellement (383).

Les articles sur les panels de référence spécifiques sont les seuls à comparer leur performance avec celles des panels multi-ethniques, il semblerait ainsi qu'un panel créé pour imputer une population particulière offre une meilleure précision, néanmoins comme cela n'est pas toujours possible, les modèles multi-ethniques restent globalement les plus utilisés. Cette manière d'envisager l'imputation HLA nécessite un effort de typage.

En conséquence, les travaux de cette thèse ont plutôt cherché à changer sa méthodologie, pour s'affranchir de la non-exhaustivité du typage HLA et améliorer l'imputation HLA des population sous-représentées. Nous avons décidé d'évaluer ceci avec une population d'ancestralité composite, comme ont pu le faire Nunes *et al.* (345) et Karnes *et al.* (334) en montrant que l'ajout de la population ciblée dans le jeu de données d'entraînement améliore l'imputation HLA. Par définition, ces populations ne correspondent pas à un seul groupe ethnique, ainsi notre hypothèse principale a été que l'on peut

remplacer ces groupes par une mesure de proximité génétique plus précise, et ainsi inclure les populations mixtes ou non-représentées.

VII.1.2 - La création de panels de références personnalisés et ses limites

Nous avons introduit le concept de panels de références personnalisés de façon à prendre compte plus précisément la diversité HLA des populations à imputer, notamment dans les allèles les plus rares. Les résultats ont montré une supériorité globale des panels de références plus importants en taille, cependant, les panels personnalisés prouvent sur certains allèles qu'ils peuvent amener de l'information. Ainsi, des allèles HLA pourtant présents dans le plus grand panel n'étaient pas imputés, contrairement à notre modèle personnalisé.

Malheureusement, la limite principale de notre étude est le nombre d'individus sélectionnés dans ces panels. En effet, dans un souci d'homogénéité avec la taille des super-populations que nous comparons, nous avons restreint ces modèles à 200 individus. De fait, même lorsque les génotypes HLA imputés étaient corrects dans notre modèle, il était impossible de le prédire en comparant avec les post-probabilités d'un modèle multi-ethnique composé de plus d'individus, comme dans les études d'Herzig *et al.* et Ritari *et al.* (336,379). Par ailleurs, certaines imputations ont été effectuées avec plusieurs modèles différents, ce qui a virtuellement augmenté le nombre d'individus nécessaires pour la phase d'entraînement. Pour contrer ces limitations, les données de SABE ainsi que d'autres jeux de données en cours d'acquisition par le projet pourront permettre d'augmenter la taille de ces modèles personnalisés. Enfin, dans le cas des modèles combinés, la comparaison de l'imputation HLA de chaque partie du jeu de données de validation, plutôt que globalement, permettrait de donner une idée par sous-groupe génétique identifié de la précision obtenue.

Nous avons choisi la réduction de dimension, avec l'ACP et l'UMAP, qui est utilisée en association génétique pour enlever l'ancestralité comme facteur confondant, grâce aux coordonnées réduites des individus. Sachant que les fréquences des allèles HLA sont différentes dans les populations, l'hypothèse est qu'imputer une population à partir d'un panel de référence où les individus lui sont proche génétiquement permettra de maximiser la précision. Nous avons pu montrer que l'UMAP était intéressante pour représenter les populations lorsqu'on se base sur la région du CMH seulement. Son utilisation en dehors de l'analyse en *single cell*, pour la génétique des populations, connait un certain essor (367). Cependant, nous n'avons pas réussi à voir une différence nette de performance par rapport à d'autres panels de références où les individus étaient sélectionnés par ACP : ni en changeant le nombre de dimensions prises en compte, ni en changeant les SNP intégrés. Le choix de la proximité SNP basée sur la région du CMH pourrait être réduite à nouveau pour s'intéresser seulement au locus HLA que l'on veut imputer, afin de capter les haplotypes SNP spécifiques du gène voulu. Aussi, il serait

intéressant d'explorer le changement de représentation de CAAPA lorsque l'UMAP est calculée avec ou sans la population de 1KG. Par ailleurs, nous avons répliqué nos résultats avec la population brésilienne de SABE en tant que jeu de données de validation (plutôt que CAAPA). Cependant, ces résultats se basent sur la réduction de dimension calculée avec la population de CAAPA, ce qui peut altérer les résultats. Sakaue *et al.* ont montré que la combinaison de l'ACP et de l'UMAP affine la représentation des populations génétiques. Les populations éloignées génétiquement sont représentées comme telles et les structures sous-jacentes à chacune de ces populations est également conservée. Cette méthode pourrait ainsi aider à la sélection d'un panel de référence précis (370).

Une des spécificités des panels de référence de HIBAG est de retenir les liens entre les haplotypes SNP et des allèles HLA. Or, deux haplotypes SNP peuvent être associés à un même allèle HLA, et ce indépendamment de la rareté de l'allèle. Degenhardt *et al.* montrent que des erreurs systématiques peuvent apparaître dans l'imputation HLA selon les ancestralités à cause de cette différence de fond haplotypique (346). L'étude préconise des calculs de sensibilité et spécificité selon l'ancestralité et de vérifier les résultats post-imputation pour identifier les haplotypes problématiques (384).

VII.1.3 - Le nouveau monde de l'imputation HLA

Pendant ma thèse, de nouvelles versions d'anciens algorithmes et de nouveaux logiciels d'imputation HLA ont été développés. Bien que HIBAG reste un des outils d'imputation HLA les plus utilisés, pratiques et performants, ces nouveaux logiciels remettent en question sa prépondérance. En effet, CookHLA et DeepHLA sont sortis en 2021 en introduisant des propriétés différentes dans leurs calculs (332,333).

CookHLA est une évolution de l'algorithme de SNP2HLA (321) qui impute les acides aminés à un locus HLA et déduit les allèles. Dans cette version, l'imputation est effectuée par exon (pour les plus polymorphes) pour éviter la réduction du déséquilibre de liaison avec la distance génétique. De plus, à l'instar des travaux de cette thèse, CookHLA se concentre notamment sur l'impact de l'ancestralité sur l'imputation HLA. Cependant, CookHLA utilise une carte génétique qui décrit les relations statistiques entre les différents polymorphismes en prenant en compte la similarité entre les individus, ce qui améliore la précision d'imputation.

DeepHLA quant à lui est une implémentation d'un algorithme de réseaux neuronaux profonds convolutifs et multi-tâches (*multitask convolutional deep learning*) (333). La principale originalité de ce type de *deep learning* est sa capacité à prédire plusieurs variables, ici plusieurs loci HLA, en parallèle, utilisant ainsi les relations statistiques entre les différents loci pour améliorer le résultat. Naito *et al.* ont eux mis leurs efforts sur l'imputation des allèles rares et ont réussi à obtenir de meilleurs résultats par rapport à SNP2HLA et HIBAG.

Bien qu'il manque une comparaison indépendante de ces logiciels entre eux, et avec les autres outils, nous pouvons imaginer qu'ils changeront également le choix d'imputation HLA (385). De plus, l'application des panels de référence personnalisés pourrait toujours être envisagé et tester dans chaque condition. Ces nouveaux algorithmes apportent de nouveaux concepts dans l'imputation HLA mais le problème de la rareté, ou de l'absence d'un allèle, dans un jeu de référence reste insoluble malgré tout. Ceci explique l'importance, en parallèle de ces changements méthodologiques, de la récolte de nouvelles données de génotypes HLA. Pour illustrer cette rareté, la base de données AFND dispose des données de 10 millions d'individus, soit 20 millions d'allèles à un locus donné (163). Si un allèle est retrouvé chez seulement un seul individu dans la base de données, une estimation rapide à partir de sa fréquence donnerait seulement 800 occurrences dans la population mondiale. Les approximations du nombre d'allèles différents dans le monde est de 3 millions par locus, ce qui indique que de nombreux allèles restent à découvrir (171).

En outre, d'autres outils versés dans l'aide à l'imputation HLA et aux analyses qui en découlent viennent s'ajouter à la palette de l'immunogénéticien. Degenhardt a développé un outil pour l'imputation HLA qui évalue également plusieurs métriques pour aider à l'analyse et identifier les allèles potentiellement problématiques (384). HLA-TAPAS, par exemple, intervient à chaque étape de l'analyse, pour la vérification du format des allèles HLA, à la création de panels de référence jusqu'à l'imputation et l'analyse d'association (383). La création de telles procédures autour de l'imputation HLA montre une envie des chercheurs de faciliter l'analyse HLA et de la rendre disponible au plus grand nombre. Tous ces efforts, dont le SHLARC se revendique, ont pour rôle de faire avancer la communauté génétique et immunogénétique sur les questions de l'association du HLA, pour avancer vers une réelle compréhension des mécanismes sous-jacents.

VII.2 - L'évolution de l'analyse HLA et son impact sur le monde de la génomique et de la clinique

VII.2.1 - De l'association génétique des SNP du CMH à une analyse complète du HLA

En 2013, le constat de Trowsdale & Knight sur l'association HLA tempérait les nombreux résultats obtenus dans les GWAS : « It is sobering to reflect that, after several decades, progress has really taken place only on the [resolution of HLA amino acid association], and even then the progress has tended to be refinement more than novel insight » (254), notamment en citant le diabète de type 1 dont l'association est connue depuis 1987 et redécouverte régulièrement depuis (256). Kennedy *et al.* indique aussi que la plupart des associations SNP peinent à indiquer un gène candidat pour les pathologies dans la région du CMH (23).

Grâce à l'imputation HLA j'ai pu participer à des études d'association plus précises révélant des liens entre des haplotypes HLA étendus conservés chez des individus afro-américains dans la sclérose en plaques (386). Par ailleurs, nous avons montré avec Valencia *et al.* l'association de deux acides aminés, Leucine 26 et Proline 55 de HLA-DQB1, avec la clairance spontanée du virus de l'hépatite C (*hepatitis C virus, HCV*), en parallèle d'une étude de la présentation des peptides de l'HCV. Enfin avec Domenighetti *et al.*, nous avons pu combiner des données de génotypes HLA, de prédiction de présentation de peptides, d'haplotypes HLA et des données cliniques pour affiner la compréhension de l'association de HLA-DRB1 avec la maladie de Parkinson, et son interaction avec le statut tabagique (300).

Cependant, la pandémie de COVID-19 a montré la complexité actuelle d'une analyse HLA rigoureuse (212). Si les GWAS et les associations HLA ont déjà fait leurs preuves, elles requièrent un grand nombre d'individus pour prendre en compte la diversité des allèles. Or, dans un contexte de production rapide de résultats, la qualité des associations s'en retrouve diminuée. Le HLA est un candidat évident pour toutes les associations avec des phénotypes immunitaires, cependant, l'absence de lien statistique dans de grandes cohortes semble mettre de côté cette hypothèse. Au contraire, les plus petites études entre le HLA et SARS-CoV-2 montrent des associations, comme le MERS et SARS par le passé, qui sont susceptibles d'être aléatoires (387).

Concernant la réduction de dimension, bien qu'elle semble adaptée et maintenant utilisée en routine dans les études d'associations, une étude récente par Elhaik semble remettre en cause sa fiabilité (388). L'une des faiblesses majeures serait que l'ACP présente des distances très différentes en génétique des populations selon le nombre d'individus dans chaque groupe ancestral, ce qui pourrait fausser entièrement la correction dans les études d'association. Ces résultats extrêmement controversés sont encore récents et nécessitent du temps avant d'être confirmés. Les résultats de distance entre les populations génétiques, obtenus avec l'UMAP dans nos travaux récents, ou ceux de Sakaue *et al.* (370) ouvrent la voie pour un changement de la représentation de l'ancestralité.

Ainsi, pendant mes travaux de thèse, l'imputation HLA a permis d'étendre la place des études d'association HLA, mais il ne faut pas négliger l'effet combiné des analyses HLA plus larges, comme les haplotypes ou la prédiction de l'affinité des allèles pour ajouter des arguments dans la causalité de l'association HLA-pathologie. Ce futur de l'association HLA est également lié à celui de l'imputation HLA, car avec les méthodes de séquençage toujours plus efficaces, il faut se demander comment va évoluer la génération de ces données.

VII.2.2 - Quel futur pour l'imputation HLA ?

La création du SHLARC a permis de rassembler de nombreux acteurs de l'immunogénétique, d'évaluer l'imputation HLA, de trouver des solutions pour prendre mieux en compte l'ancestralité et d'appliquer ceci dans des études d'associations HLA. Le consortium rassemble pour l'instant 3 384 individus, dont près de la moitié est d'ancestralité composite africaine-européenne. De nouvelles données de Finlande, multi-ethniques des États-Unis et de populations diverses s'ajouteront également dans les prochaines années. Le pouvoir d'imputation HLA de ces panels de référence seront disponibles dans un site Internet qui facilitera l'imputation de toutes les populations pour les généticiens.

Des sites d'imputation HLA existent déjà, comme celui du HLA Imputation Portal (HIP), HKImpNet (383) ou HLA-IMPUTER (389,390) mais leur capacité d'imputation est limitée aux panels de référence de base de HIBAG, dont le plus conséquent est le panel européen. D'autres jeux de données, telles que les biobanques de Taiwan (228), du Japon (391), de la Finlande (336) ou celle du Royaume-Uni (392) n'ont pas encore des données HLA complètes et pourront servir de référence dans le futur. Récemment, le consortium TopMed s'est associé avec le serveur d'imputation SNP du Michigan pour proposer son panel de référence pour les SNP, mais aussi pour le HLA (316). Le panel est composé de 24 456 individus, à près de 50% européens mais avec près de 25% d'individus afro-américains. La limitation majeure de ce panel est le format des génotypes HLA qui sont donnés en groupe G, c'est-à-dire qu'ils ne décrivent que les exons les plus polymorphes et ne prédisent pas la séquence entière de l'allèle. Dans le futur, il sera intéressant de faire évoluer la diversité des panels de recherche et fournir la résolution la plus haute possible pour les allèles HLA, avec au minimum deux champs pour décrire entièrement la séquence des allèles.

Il est important de se questionner sur le futur de l'imputation HLA et de sa place dans la communauté d'immunogénomique dans les années à venir. Les motivations qui poussent à utiliser l'imputation HLA sont sa facilité à exécuter, même sur des jeux de données déjà existants, ainsi que son coût moindre. Ces dernières années ont connues une baisse importante du coût de séquençage de génome qui est passée en-dessous des 1 000$ par échantillon (393). De plus, la région du CMH dispose maintenant d'algorithmes d'alignement optimisés qui permettent d'obtenir un typage à partir de génomes ou exomes entiers (342). Malgré cela, l'imputation HLA est toujours d'une importance capitale. En effet, le prix d'un génotypage SNP, qui permet d'obtenir un typage HLA également, reste 5 à 10 fois plus faible qu'un séquençage. De plus, les données SNP d'association déjà générées peuvent toujours être étendues grâce à cette imputation HLA, sans coût supplémentaire. Même si de nombreux allèles rares sont absents des panels de référence actuellement, l'ajout de nombreux individus d'ancestralités diverses au fil des années va permettre de combler ce manque : les allèles rares vont avoir plus d'occurrences et de nouveaux allèles deviendront les allèles rares.

Une interrogation récurrente autour de l'imputation HLA est sa capacité à être transposée en clinique, notamment pour la transplantation. L'incertitude inhérente à l'imputation avec les probabilités liées au génotypes et l'absence de certains allèles dans les panels de référence empêche pour l'instant toute utilisation en transplantation. Afin de réduire cette incertitude, il serait intéressant de la quantifier par rapport à un séquençage classique qui peut également présenter des erreurs. En gardant tous les génotypes les plus probables rendus par HIBAG pour un individu, il serait aussi possible d'évaluer les épitopes de toutes ces prédictions et donner une vue d'ensemble. Les erreurs de génotypes HLA dans l'imputation sont souvent liées à des allèles proches, ainsi, si tous les génotypes imputés ne présentent pas un certain épitope, cela pourrait être considéré comme sûr pour la transplantation.

L'analyse HLA est un domaine à l'intersection de la génétique, de l'immunologie et de la bioinformatique. Toutes les évolutions décrites dans la récolte des données méthodologiques, l'imputation, et la prédiction d'autres paramètres du HLA modifient le paysage immunogénétique, et impactent également d'autres disciplines.

VII.2.3 - Les conséquences globales des avancées technologiques en immunogénomique

Les efforts concentrés sur l'imputation HLA et son analyse ont aussi améliorer la prise en charge de la région entière du CMH et changent la manière d'appréhender la génétique. En effet, d'autres gènes que ceux du HLA sont polymorphes et les motifs de déséquilibre de liaison dans le CMH peuvent servir pour leur imputation. Ainsi, des algorithmes comme MHC*IMP ou DeepHLA ont été testés avec les gènes HLA-E/-F/-G, ou MICA, MICB et TAP (331,333). MICA et MICB (*MHC class I polypeptide related sequence A/B*) sont des protéines homologues aux gènes HLA de classe I mais elles ne se lient pas avec la β2-microglobuline et ne présentent pas de peptides. Elles ont néanmoins deux cents allèles chacune et interagissent avec la protéine NKG2D des cellules NK (394).

En extension des logiciels de prédiction de liaison des peptides aux allèles HLA, le logiciel ERGO-II intègre également la variabilité des TCR (395). Les TCR ont un système de polymorphisme à part, avec des réarrangements génomiques de différentes régions, V, D et J, qui incorporent des variations aux jonctions. Pour l'activation des lymphocytes T, l'ensemble TCR-peptide-HLA est essentiel, et ERGO-II permet de prédire l'affinité de chaque membre du complexe grâce à des réseaux de neurones. Ces interactions sont essentielles pour comprendre l'activation des lymphocytes mais également la sélection des lymphocytes T dans le thymus, car Sharon *et al.* ont montré que le polymorphisme des allèles HLA était associé à la diversité des TCR chez un individu (396).

En dehors de la région du CMH, les algorithmes d'imputation ont trouvé une autre utilité sur le chromosome 19 où sont situés les gènes KIR. Ce sont non seulement des gènes très polymorphes avec

des centaines d'allèles mais il existe une diversité de gènes différents qui sont présents ou absents selon les haplotypes portés par l'individu. Ainsi des outils tels que KIBAG (397), KIR*IMP (330), ou par Ritari *et al.* (398), reprennent des algorithmes ou des méthodes existantes avec le HLA pour créer des panels de référence SNP-KIR. Cependant, la majorité des algorithmes d'imputation KIR se concentrent sur le contenu en gènes KIR mais ne vont pas au niveau de l'allèle.

En prenant du recul sur la biologie, comme l'exemple de 23andMe et des données génétiques évoqué au début de cette thèse, la croissance exponentielle de la récolte de données immunogénomiques pose des questions sur leur confidentialité. Notamment, la base de données AFND centralise les données HLA qui sont des données de santé personnelles et identifiantes : ceci nécessite des autorisations. Avec Sayadi *et al.* nous avons décrit comment distribuer les calculs de fréquence d'allèles HLA sur le site en gardant les données dans leurs centres respectifs (399). Pour finir, ces efforts de récolte de données HLA à travers le globe, menées par le SHLARC ou d'autres projets ne doivent pas s'effectuer au détriment des populations dans lesquelles les données sont récupérées. Lors des conférences de l'European Federation for Immunogenetics et l'American Society of Human Genetics, de nombreuses interventions sur la diversité génétique en population ont rappelé que ces récoltes sont des collaborations internationales qui doivent être faite avec l'accord et dans l'intérêt des populations, non comme une fin en soi. Il est impératif de penser ces collaborations et de réfléchir en amont à leur utilisation et à leurs conséquences.

Conclusion

“I learned a lot, by the end of everything. The past is past, now, but that’s… you know, that’s okay! It’s never really gone completely. The future is always built on the past, even if we won’t get to see it. Still, it’s um, time for something new, now.”

Riebeck, Outer Wilds

VIII - Conclusion

Lors de ces trois années de doctorat en Bioinformatique, j'ai pu appréhender la thématique de l'immunogénétique dans toute sa complexité. Je me suis tout particulièrement intéressé à l'impact de la diversité génétique de la région du CMH sur les modèles statistiques d'imputation des molécules de HLA. Mon travail a consisté à évaluer des méthodes mathématiques de réduction de dimensions afin de sélectionner des individus pour créer des modèles spécifiques de certaines populations génétiques, peu représentées ou mélangées génétiquement. Pour compléter cette expertise, j'ai également pu me plonger dans l'analyse HLA pour accompagner lors de l'étude d'autres paramètres immunogénomiques HLA, jusqu'à leur association avec des pathologies.

La Bioinformatique est par essence un domaine transversal de la science entre la biologie et l'informatique. Au-delà de la composante génomique importante de mes travaux, ils s'inscrivent aussi dans une réflexion autour de la génétique des populations et de la prise en compte de l'ancestralité génétique dans les algorithmes de prédiction et les études qui en découlent. En effet, la majorité des études d'association en génome entier se concentrent sur des populations européennes, ce qui a créé avec le temps un biais dans les données disponibles pour les algorithmes de prédiction. L'objectif de ces travaux est de réduire ce biais pour l'imputation HLA.

Or, la création de panels de référence pour l'imputation HLA par apprentissage statistique est une tâche qui nécessite de mobiliser des infrastructures informatiques dédiés. La mise en place de tels outils dans un environnement de calculs adapté est de plus en plus courante pour générer des données ou modéliser des relations entre génétique et pathologies en biologie. Nous avons donc collaboré avec les superordinateurs de BiRD (Institut de Recherche en Santé de Nantes), du Liger (Ecole Centrale de Nantes) et du CCIPL (Nantes Université) pour mener à bien ces travaux.

Concernant les données utiles pour l'imputation HLA, nous avons tiré parti de données récentes de séquençage par le projet 1,000 Genomes et de celles du consortium CAAPA, pour évaluer l'importance de l'ancestralité des données dans l'imputation HLA. Ces premiers résultats ont montré le statut crucial du volume des données SNP et HLA dans l'imputation HLA. De plus, pour imputer une partie du jeu de données de CAAPA, nous avons montré que l'utilisation de groupes ethniques auto-définis pour l'imputation HLA (comme afro-américains) ne rivalise pas avec une réelle proximité génétique. Pour une population d'ancestralité composite comme CAAPA, cela peut s'expliquer par des différences de proportions de chaque ancestralité dans le génome des individus et des fréquences de polymorphismes différentes. De grands panels de référence multi-ethniques d'imputation HLA existent déjà, mais pour l'instant leur diversité reste limitée, notamment comparée à l'immense

diversité HLA. Dans le futur, le projet SHLARC se dotera de nouveaux jeux de données dans des populations peu représentées encore dans le paysage génomique.

Pour contourner les limites de l'imputation HLA dans des populations peu représentées, sans recourir au séquençage de nouveaux individus, cette thèse s'est donc penchée sur la création de panels de références spécifiques à l'aide de stratégies de réduction de dimension. La méthode de l'UMAP s'est tout d'abord montrée d'un intérêt particulier pour représenter la région génomique du CMH, ouvrant la voie pour une utilisation plus large en génomique des populations et dans les études d'association. Pour la création de panels de références spécifiques, la sélection d'individus par réduction de dimension a montré des limites de performances par rapport à des panels plus importants et multi-ethniques. Cependant, leurs meilleures performances sur des allèles très spécifiques pourrait être exploitée. L'imputation HLA gagnerait probablement en précision à utiliser ces deux méthodes combinées, après augmentation de la taille des modèles spécifiques, pour augmenter la précision des résultats. L'intégration de nouvelles données et de nouveaux algorithmes pour l'imputation HLA de populations génomiques non-européennes doit continuer à être développée. Il est néanmoins nécessaire pour les nouvelles données obtenues de travailler en collaboration avec les chercheurs issus de ces populations étudiées, et de continuer à s'interroger sur l'utilisation de ces données sensibles.

Enfin, dans le cadre de la dissémination de l'analyse HLA dans la communauté scientifique, nous avons pu participer à plusieurs études d'associations. Notre capacité à fournir une imputation HLA de qualité avec nos données disponibles ainsi que l'ajout d'une expertise sur d'autres outils HLA comme les haplotypes ou la prédiction des peptides présentés par les allèles a permis de faciliter la compréhension d'associations HLA avec le virus de l'hépatite C ou la maladie de Parkinson. D'un point de vue général, l'analyse HLA ne peut donc pas se limiter à l'association statistique : le peptidome HLA, l'association SNP et HLA, les haplotypes, les structures moléculaires, les autres paramètres immunogénomiques comme les TCR ou les KIR, et des autres du gène du CMH doivent intégrer la boîte à outil de l'analyse HLA (143).

Ce projet SHLARC va perdurer dans le temps pour accueillir de nouveaux jeux de données, suivre l'évolution des algorithmes d'imputation HLA et propose de nouvelles méthodologies pour améliorer les résultats de l'imputation. Dans le futur, le SHLARC va également mettre à disposition des panels de référence pour l'imputation HLA sur un site Web accessible à tous. Ce site en construction compte centraliser les données récoltées pour permettre aux chercheurs d'imputer des génotypes HLA en utilisant les populations les plus proches génétiquement.

Les prémisses de cette thèse ont commencé il y a cinq ans pendant un stage de Master 2 et j'ai eu la chance de pouvoir les poursuivre depuis trois ans pour produire les résultats présentés ici. D'un point

de vue personnel, il a parfois été compliqué de motiver la recherche en comparaison avec les nouveautés dans la littérature, mais le temps pris pour remettre en question le projet lui a permis de mûrir. De plus, le partage de l'expertise construite au fil des années avec d'autres chercheurs a pu solidifier la confiance dans les résultats obtenus et l'intérêt de ces travaux dans le cadre de l'analyse HLA. Enfin, parallèlement au travail de recherche, le travail de communication au grand public a été une source de motivation. Nous avons eu l'opportunité de présenter l'intérêt de la bioinformatique en génétique par le biais d'une idée de jeu, et par de nombreuses heures de réflexion, de tests, de colle, de découpage, de code, d'un *escape game* en collaboration avec l'Inserm. Celui-ci a permis d'explorer un autre pan de la communication scientifique, également essentiel.

Pour conclure, les développements technologiques de ces dernières années en génomique humaine continuent de révolutionner la biologie, grâce à leur interaction avec l'informatique et la statistique dans les algorithmes d'apprentissage statistique. Ce mélange des domaines nécessite plus que jamais une communication claire entre les différents experts pour que de nouvelles idées émergent. Cette collaboration permettra d'élucider les relations complexes entre la diversité des génomes humains et les différents phénotypes, pathogènes ou non.

IX - Références biblilographiques

1. Inserm. Divers et variants – C'est quoi un mutant ? [Internet]. 2021. Available from: https://www.inserm.fr/c-est-quoi/divers-et-variants-c-est-quoi-mutant-🐭/
2. Labrunie F. Test génétiques : quel cadre législatif pour la France ? [Internet]. Numerama. 2018. Available from: https://www.numerama.com/sciences/366523-tests-genetiques-quel-cadre-legislatif-pour-la-france.html
3. Inserm. Tests génétiques "récréatifs" : Juste un jeu ? [Internet]. Inserm, le magazine n°42. 2019. Available from: https://www.inserm.fr/actualite/tests-genetiques-recreatifs-juste-jeu/
4. Servick K. Can 23andMe have it all? Science (80-). 2015;349(6255):1472–7.
5. Smith LP. The Spectrum of Genetic Testing. Semin Oncol Nurs. 2019;35(1):11–21.
6. Phillips BL, Callaghan C. The immunology of organ transplantation. Surg (United Kingdom) [Internet]. 2020;38(7):353–60. Available from: https://doi.org/10.1016/j.mpsur.2020.04.008
7. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. Malina D, editor. N Engl J Med [Internet]. 2020 Aug 27;383(9):874–82. Available from: http://www.nejm.org/doi/10.1056/NEJMms2004740
8. Borrell LN, Elhawary JR, Fuentes-Afflick E, Witonsky J, Bhakta N, Wu AHB, et al. Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. Malina D, editor. N Engl J Med [Internet]. 2021 Feb 4;384(5):474–80. Available from: http://www.nejm.org/doi/10.1056/NEJMms2029562
9. Darwin C. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. 1859;544. Available from: http://93.174.95.29/_ads/E09AFB7ABD01C7B508462C7DADA04A39
10. Mendel G. Experiments in Plant Hybridization. J R Hortic Soc. 1865;(1865):3–47.
11. Nasmyth K. The magic and meaning of Mendel's miracle. Nat Rev Genet. 2022;23(7):447–52.
12. Koltzoff N. The Structure of the Chromosomes in the Salivary Glands of Drosophila. Science (80-) [Internet]. 1934 Oct 5;80(2075):312–3. Available from: https://www.science.org/doi/10.1126/science.80.2075.312
13. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev [Internet]. 1994;58(3):563–602. Available from: https://mmbr.asm.org/content/58/3/563
14. WILKINS MHF, STOKES AR, WILSON HR. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. Nature [Internet]. 1953 Apr 25;171(4356):738–40. Available from: https://www.nature.com/articles/171738a0
15. WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature [Internet]. 1953 Apr 25;171(4356):737–8. Available from: https://www.nature.com/articles/171737a0
16. FRANKLIN RE, GOSLING RG. Molecular Configuration in Sodium Thymonucleate. Nature [Internet]. 1953 Apr 25;171(4356):740–1. Available from: https://www.nature.com/articles/171740a0
17. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.
18. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature [Internet]. 2020;577(7789):179–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31915397
19. Slatkin M. Linkage disequilibrium - Understanding the evolutionary past and mapping the

medical future. Nat Rev Genet. 2008;9(6):477–85.

20. Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. Nat Genet. 1992;2(3):204–11.

21. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science (80-) [Internet]. 2001 Feb 16;291(5507):1304–51. Available from: sftp://cerca@192.168.2.5/home/cerca/Desktop/data/laptop_files/info/biologia/homo_sapiens/human_genome/Celera_genoma.pdf%5Cnpapers2://publication/uuid/21C9A6AC-3A9B-4931-BB7D-CE922633B16B

22. Manfredi E, Tusell L, Vitezica ZG. Prediction of complex traits: Conciliating genetics and statistics. J Anim Breed Genet. 2017;134(3):178–83.

23. Kennedy AE, Ozbek U, Dorak MT. What has GWAS done for HLA and disease associations? Int J Immunogenet [Internet]. 2017 Oct;44(5):195–211. Available from: https://onlinelibrary.wiley.com/doi/10.1111/iji.12332

24. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al. The complete sequence of a human genome. Science (80-) [Internet]. 2022 Apr;376(6588):44–53. Available from: https://www.science.org/doi/10.1126/science.abj6987

25. Tanaka T. The International HapMap Project. Nature [Internet]. 2003 Dec;426(6968):789–96. Available from: http://www.nature.com/articles/nature02168

26. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Vol. 526, Nature. 2015. p. 68–74.

27. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature [Internet]. 2015 Oct 1;526(7571):75–81. Available from: http://www.nature.com/articles/nature15394

28. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020;48(D1):D941–7.

29. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell [Internet]. 2022 Sep;185(18):3426-3440.e19. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867422009916

30. Lewontin RC. The Apportionment of Human Diversity. In: Evolutionary Biology [Internet]. New York, NY: Springer US; 1972. p. 381–98. Available from: http://link.springer.com/10.1007/978-1-4684-9063-3_14

31. Carlson J, Harris K. The apportionment of citations: A scientometric analysis of Lewontin 1972. Philos Trans R Soc B Biol Sci. 2022;377(1852).

32. Maróstica AS, Nunes K, Castelli EC, Silva NSB, Bruce S, Goudet J, et al. How HLA diversity is apportioned : influence of selection and relevance to transplantation. 2022;

33. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol [Internet]. 2019;2(1). Available from: http://dx.doi.org/10.1038/s42003-018-0261-x

34. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. Nat Genet [Internet]. 2020;52(3):242–3. Available from: http://dx.doi.org/10.1038/s41588-020-0580-y

35. Ju D, Hui D, Hammond DA, Wonkam A, Tishkoff SA. Importance of Including Non-European Populations in Large Human Genetic Studies to Enhance Precision Medicine. Annu Rev Biomed Data Sci [Internet]. 2022 Aug 10;5(1):321–39. Available from: https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-122220-112550

36. ASHG. ASHG Denounces Attempts to Link Genetics and Racial Supremacy. Am J Hum Genet. 2018;103(5):636.

37. Joshi RS, Rigau M, García-Prieto CA, Castro de Moura M, Piñeyro D, Moran S, et al. Look-alike humans identified by facial recognition algorithms show genetic similarities. Cell Rep. 2022;40(8).

38. Gorer PA. The Detection of Antigenic Differences in Mouse Erythrocytes by the Employment of Immune Sera. Br J Exp Pathol. 1936;17(1):42.

39. Gorer PA. The genetic and antigenic basis of tumour transplantation. J Pathol Bacteriol [Internet]. 1937 May;44(3):691–7. Available from: https://onlinelibrary.wiley.com/doi/10.1002/path.1700440313

40. Dausset J. Iso-leuco-anticorps. Acta Haematol [Internet]. 1958;20(1–4):156–66. Available from: https://www.karger.com/Article/FullText/205478

41. Payne R, Rolfs MR. Fetomaternal Leukocyte Incompatibility. J Clin Invest [Internet]. 1958 Dec 1;37(12):1756–63. Available from: http://www.jci.org/articles/view/103768

42. VAN ROOD JJ, EERNISSE JG, VAN LEEUWEN A. Leucocyte Antibodies in Sera from Pregnant Women. Nature [Internet]. 1958 Jun;181(4625):1735–6. Available from: https://www.nature.com/articles/1811735a0

43. Ceppellini R, Mattiuz PL, Scudeller G, Visetti M. Experimental allotransplantation in man. I. The role of the HL-A system in different genetic combinations. Transplant Proc [Internet]. 1969 Mar;1(1):385–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4944245

44. Amos DB, Seigler HF, Southworth JG, Ward FE. Skin graft rejection between subjects genotyped for HL-A. Transplant Proc [Internet]. 1969 Mar;1(1):342–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4944238

45. Thorsby E. A short history of HLA. Tissue Antigens. 2009;74(2):101–16.

46. Terasaki PI, Marchioro TL, Starz TE. Sero-typing of human lymphocyte antigens: Preliminary Trials on Long-Term Kidney Homograft Survivors. Histocompat Test. 1965;83–96.

47. Holscher CM, Jackson KR, Segev DL. Transplanting the Untransplantable. Am J Kidney Dis [Internet]. 2020;75(1):114–23. Available from: https://doi.org/10.1053/j.ajkd.2019.04.025

48. Daikeler T, Hügle T, Farge D, Andolina M, Gualandi F, Baldomero H, et al. Allogeneic hematopoietic SCT for patients with autoimmune diseases. Bone Marrow Transplant. 2009;44(1):27–33.

49. Burt RK, Loh Y, Pearce W, Beohar N, Barr WG, Craig R, et al. Clinical applications of blood-derived and marrow-derived stem cells for nonmalignant diseases. JAMA - J Am Med Assoc. 2008;299(8):925–36.

50. Geffard E, Limou S, Walencik A, Daya M, Watson H, Torgerson D, et al. Easy-HLA: a validated web application suite to reveal the full details of HLA typing. Valencia A, editor. Bioinformatics [Internet]. 2020 Apr 1;36(7):2157–64. Available from: https://academic.oup.com/bioinformatics/article/36/7/2157/5637224

51. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, et al. IPD-MHC 2.0: An improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res. 2017;45(D1):D860–4.

52. Trowsdale J, Campbell RD. Mouse MHC Genes and Products . Curr Protoc Immunol. 1998;27(1):1–7.

53. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nat Rev Genet [Internet]. 2004 Dec;5(12):889–99. Available from: http://www.nature.com/articles/nrg1489

54. Herberg JA, Sgouros J, Jones T, Copeman J, Humphray SJ, Sheer D, et al. Genomic analysis of the Tapasin gene, located close to the TAP loci in the MHC. Eur J Immunol. 1998;28(2):459–67.

55. Ortmann B, Copeman J, Lehner PJ, Sadasivan B, Herberg JA, Grandea AG, et al. A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes. Science (80-). 1997;277(5330):1306–9.

56. Teng MS, Stephens R, Pasquier L Du, Freeman T, Lindquist JA, Trowsdale J. A humanTAPBP (TAPASIN)-related gene,TAPBP-R. Eur J Immunol [Internet]. 2002 Apr;32(4):1059–68. Available from: https://onlinelibrary.wiley.com/doi/10.1002/1521-4141(200204)32:4%3C1059::AID-IMMU1059%3E3.0.CO;2-G

57. Noris M, Remuzzi G. Overview of complement activation and regulation. Semin Nephrol [Internet]. 2013;33(6):479–92. Available from: http://dx.doi.org/10.1016/j.semnephrol.2013.08.001

58. Kalliolias GD, Ivashkiv LB. TNF biology, pathogenic mechanisms and emerging therapeutic strategies. Nat Rev Rheumatol [Internet]. 2016 Jan 10;12(1):49–62. Available from: http://www.nature.com/articles/nrrheum.2015.169

59. Janke C, Magiera MM. The tubulin code and its role in controlling microtubule properties and functions. Nat Rev Mol Cell Biol [Internet]. 2020;21(6):307–26. Available from: http://dx.doi.org/10.1038/s41580-020-0214-3

60. Marshall JS, Warrington R, Watson W, Kim HL. An introduction to immunology and immunopathology. Allergy, Asthma Clin Immunol [Internet]. 2018;14(s2):1–10. Available from: https://doi.org/10.1186/s13223-018-0278-1

61. Netea MG, Balkwill F, Chonchol M, Cominelli F, Donath MY, Giamarellos-Bourboulis EJ, et al. A guiding map for inflammation. Nat Immunol [Internet]. 2017;18(8):826–31. Available from: http://dx.doi.org/10.1038/ni.3790

62. Farber DL, Netea MG, Radbruch A, Rajewsky K, Zinkernagel RM. Immunological memory: Lessons from the past and a look to the future. Nat Rev Immunol [Internet]. 2016;16(2):124–8. Available from: http://dx.doi.org/10.1038/nri.2016.13

63. Netea MG, Schlitzer A, Placek K, Joosten LAB, Schultze JL. Innate and Adaptive Immune Memory: an Evolutionary Continuum in the Host's Response to Pathogens. Cell Host Microbe [Internet]. 2019;25(1):13–26. Available from: https://doi.org/10.1016/j.chom.2018.12.006

64. Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol. 2006;24:419–66.

65. Fabre JW. Regulation of MHC expression. Immunol Lett. 1991;29(1–2):3–8.

66. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, et al. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. J Immunol. 2015;194(1):5–11.

67. Araki Y, Fann M, Wersto R, Weng N. Histone Acetylation Facilitates Rapid and Robust Memory CD8 T Cell Response through Differential Expression of Effector Molecules (Eomesodermin and Its Targets: Perforin and Granzyme B). J Immunol. 2008;180(12):8102–8.

68. Elliott T, Cerundolo V, Elvin J, Townsend A. Peptide-induced conformational change of the class I heavy chain. Nature. 1991;351(6325):402–6.

69. Michalek MT, Grant EP, Gramm C, Goldberg AL, Rock KL. A role for the ubiquitin-dependent proteolytic pathway in MHC class l-restricted antigen presentation. Nature [Internet]. 1993 Jun 10;363(6429):552–4. Available from: https://www.nature.com/articles/363552a0

70. Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, et al. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. Cell. 1994;78(5):761–71.

71. Reits E, Griekspoor A, Neijssen J, Groothuis T, Jalink K, Van Veelen P, et al. Peptide Diffusion, Protection, and Degradation in Nuclear and Cytoplasmic Compartments before Antigen Presentation by MHC Class I. Immunity. 2003;18(1):97–108.

72. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules [Internet]. Vol. 37, Trends in Immunology. Elsevier Ltd; 2016. p. 724–37. Available from: http://dx.doi.org/10.1016/j.it.2016.08.010

73. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class i and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12):823–36.

74. Wearsch PA, Cresswell P. Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer. Nat Immunol. 2007;8(8):873–81.

75. Cresswell P, Bangia N, Dick T, Diedrich G. The nature of the MHC class I peptide loading complex. Immunol Rev. 1999;172:21–8.

76. Nguyen TT, Chang SC, Evnouchidou I, York IA, Zikos C, Rock KL, et al. Structural basis for antigenic peptide precursor processing by the endoplasmic reticulum aminopeptidase ERAP1. Nat Struct Mol Biol. 2011;18(5):604–13.

77. Garstka MA, Fish A, Celie PHN, Joosten RP, Janssen GMC, Berlin I, et al. The first step of peptide selection in antigen presentation by MHC class I molecules. Proc Natl Acad Sci U S A. 2015;112(5):1505–10.

78. Romieu-Mourez R, François M, Boivin M-N, Stagg J, Galipeau J. Regulation of MHC Class II Expression and Antigen Processing in Murine and Human Mesenchymal Stromal Cells by IFN-γ, TGF-β, and Cell Density. J Immunol. 2007;179(3):1549–58.

79. Mosmann TR, Coffman RL. TH1 and TH2 cells: Different patterns of lymphokine secretion lead to different functional properties. Annu Rev Immunol. 1989;7:145–73.

80. Korn T, Bettelli E, Oukka M, Kuchroo VK. IL-17 and Th17 cells. Annu Rev Immunol. 2009;27:485–517.

81. Bednar KJ, Lee JH, Ort T. Tregs in Autoimmunity : Insights Into Intrinsic Brake Mechanism Driving Pathogenesis and Immune Homeostasis. 2022;13(June):1–8.

82. Cresswell P, Roche PA. Invariant chain-MHC class II complexes: Always odd and never invariant. Immunol Cell Biol. 2014;92(6):471–2.

83. Neefjes J. CIIV, MIIC and other compartments for MHC class II loading. Eur J Immunol. 1999;29(5):1421–5.

84. Denzin LK, Cresswell P. HLA-DM induces clip dissociation from MHC class II αβ dimers and facilitates peptide loading. Cell. 1995;82(1):155–65.

85. Denzin LK, Fallas JL, Prendes M, Yi W. Right place, right time, right peptide: DO keeps DM focused. Immunol Rev. 2005;207:279–92.

86. Sercarz EE, Maverakis E. MHC-guided processing: Binding of large antigen fragments. Nat Rev Immunol. 2003;3(8):621–9.

87. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet [Internet]. 2009 Jan 9;54(1):15–39. Available from: http://www.nature.com/articles/jhg20085

88. Beck S, Geraghty D, Inoko H, Rowen L, Aguado B, Bahram S, et al. Complete sequence and gene map of a human major histocompatibility complex. Nature [Internet]. 1999 Oct;401(6756):921–3. Available from: http://www.nature.com/articles/44853

89. Petrie EJ, Clements CS, Lin J, Sullivan LC, Johnson D, Huyton T, et al. CD94-NKG2A recognition of human leukocyte antigen (HLA)-E bound to an HLA class I leader sequence. J Exp Med.

2008;205(3):725–35.

90. Trowsdale J. Genetic and Functional Relationships between MHC and NK Receptor Genes HLA class I and NK receptors are encoded within. Immunity. 2001;15:363–74.

91. Pos W, Sethi DK, Call MJ, Schulze MSED, Anders AK, Pyrdol J, et al. Crystal structure of the HLA-DM-HLA-DR1 complex defines mechanisms for rapid peptide selection. Cell [Internet]. 2012;151(7):1557–68. Available from: http://dx.doi.org/10.1016/j.cell.2012.11.025

92. Denzin LK, Sant'Angelo DB, Hammond C, Surman MJ, Cresswell P. Negative Regulation by HLA-DO of MHC Class II-Restricted Antigen Processing. Science (80-) [Internet]. 1997 Oct 3;278(5335):106–9. Available from: https://www.science.org/doi/10.1126/science.278.5335.106

93. Singh RK, Singh D, Yadava A, Srivastava AK. Molecular fossils "pseudogenes" as functional signature in biological system. Genes and Genomics [Internet]. 2020;42(6):619–30. Available from: https://doi.org/10.1007/s13258-020-00935-7

94. Robinson J, Soormally AR, Hayhurst JD, Marsh SGE. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. Hum Immunol [Internet]. 2016;77(3):233–7. Available from: http://dx.doi.org/10.1016/j.humimm.2016.01.020

95. Jordier F, Gras D, De Grandis M, D'Journo XB, Thomas PA, Chanez P, et al. HLA-H: Transcriptional Activity and HLA-E Mobilization. Front Immunol. 2020;10(January):1–7.

96. Würfel FM, Wirtz RM, Winterhalter C, Taffurelli M, Santini D, Mandrioli A, et al. HLA-J, a Non-Pseudogene as a New Prognostic Marker for Therapy Response and Survival in Breast Cancer. Geburtshilfe Frauenheilkd. 2020;80(11):1123–33.

97. Lyu L, Yao J, Wang M, Zheng Y, Xu P, Wang S, et al. Overexpressed Pseudogene HLA-DPB2 Promotes Tumor Immune Infiltrates by Regulating HLA-DPB1 and Indicates a Better Prognosis in Breast Cancer. Front Oncol. 2020;10(August).

98. Paganini J, Abi-Rached L, Gouret P, Pontarotti P, Chiaroni J, Di Cristofaro J. HLAIb worldwide genetic diversity: New HLA-H alleles and haplotype structure description. Mol Immunol [Internet]. 2019;112(February):40–50. Available from: https://doi.org/10.1016/j.molimm.2019.04.017

99. Carlini F, Ferreira V, Buhler S, Tous A, Eliaou JF, René C, et al. Association of HLA-A and non-classical HLA class i alleles. PLoS One. 2016;11(10):1–17.

100. Geraghty DE, Pei J, Lipsky B, Hansen JA, Taillon-Miller P, Bronson SK, et al. Cloning and physical mapping of the HLA class I region spanning the HLA-E-to-HLA-F interval by using yeast artificial chromosomes. Proc Natl Acad Sci U S A. 1992;89(7):2669–73.

101. Shukla H, Gillespie GA, Srivastava R, Collins F, Chorney MJ. A class I jumping clone places the HLA-G gene approximately 100 kilobases from HLA-H within the HLA-A subregion of the human MHC. Genomics. 1991;10(4):905–14.

102. Gleimer M, Wahl AR, Hickman HD, Abi-Rached L, Norman PJ, Guethlein LA, et al. Although Divergent in Residues of the Peptide Binding Site, Conserved Chimpanzee Patr-AL and Polymorphic Human HLA-A*02 Have Overlapping Peptide-Binding Repertoires. J Immunol [Internet]. 2011 Feb 1;186(3):1575–88. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

103. Williams F, Curran MD, Middleton D. Characterisation of a novel HLA-A pseudogene, HLA-BEL, with significant sequence identity with a gorilla MHC class I gene. Tissue Antigens. 1999;54(4):360–9.

104. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, et al. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. Immunogenetics. 2008;60(1):1–18.

105. Mitchell S, Vargas J, Hoffmann A. Signaling via the NFκB system. WIREs Syst Biol Med [Internet]. 2016 May 16;8(3):227–41. Available from: https://onlinelibrary.wiley.com/doi/10.1002/wsbm.1331

106. Pouw RB, Ricklin D. Tipping the balance: intricate roles of the complement system in disease and therapy. Semin Immunopathol [Internet]. 2021;43(6):757–71. Available from: https://doi.org/10.1007/s00281-021-00892-7

107. IMGT-HLA. IMGT-HLA statistics [Internet]. 2022 [cited 2022 Sep 28]. Available from: https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics/

108. Dausset J. Milestones in Blood Transfusion and Immunohaematology - The Birth of MAC. Vox Sang [Internet]. 1984 Jul;45(1):89–90. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1423-0410.1983.tb04129.x

109. Parham P. Molecular definition of the transplantation antigens. FEBS J. 2018;285(15):2728–45.

110. Park I, Terasaki P. Origins of the first HLA specificities. Hum Immunol [Internet]. 2000 Mar;61(3):185–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885999001548

111. Moalic V, Ferec C. Typage HLA, méthodes d'analyses et applications cliniques. Presse Med. 2005;34(15):1101–8.

112. Hurley CK. Naming HLA diversity: A review of HLA nomenclature. Hum Immunol [Internet]. 2021;82(7):457–65. Available from: https://doi.org/10.1016/j.humimm.2020.03.005

113. Meral B. Bone Marrow and Stem Cell Transplantation [Internet]. Beksaç M, editor. Methods in Molecular Biology. New Jersey: Humana Press; 2007. X, 313. (Methods in Molecular Biology; vol. 134). Available from: http://link.springer.com/10.1007/978-1-4614-9437-9

114. Cao K, Chopek M, Fernández-Viña MA. High and intermediate resolution DNA typing systems for class I HLA-A, B, C genes by hybridization with sequence-specific oligonucleotide probes (SSOP). Rev Immunogenet. 1999;1(2):177–208.

115. Erlich H. HLA DNA typing: Past, present, and future. Tissue Antigens. 2012;80(1):1–11.

116. Cesbron-Gautier A, Simon P, Achard L, Cury S, Follea G, Bignon J. Technologie Luminex : application aux typages HLA par biologie moléculaire (PCR-SSO) et à l'identification des anticorps anti-HLA. 2004;62:93–8.

117. Olerup O, Zetterquist H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: An alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens [Internet]. 1992 May;39(5):225–35. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1399-0039.1992.tb01940.x

118. Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, et al. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res. 1989;17(7):2503–16.

119. Elsermans V, Guidicelli LG, Walencik A. Les méthodes de typage HLA. Le Courr la Transplant. 2018;XVIII:66–71.

120. Santamaria P, Lindstrom AL, Boyce-Jacino MT, Myster SH, Barbosa JJ, Faras AJ, et al. HLA class I sequence-based typing. Hum Immunol. 1993;37(1):39–50.

121. Santamaria P, Boyce-Jacino MT, Lindstrom AL, Barbosa JJ, Faras AJ, Rich SS. HLA class II "typing": Direct sequencing of DRB, DQB, and DQA genes. Hum Immunol. 1992;33(2):69–81.

122. Voorter CEM, Palusci F, Tilanus MGJ. Sequence-Based Typing of HLA: An Improved Group-Specific Full-Length Gene Sequencing Approach. In: Methods in Molecular Biology [Internet]. 2014. p. 101–14. Available from: http://link.springer.com/10.1007/978-1-4614-9437-9_7

123. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16 th IHIW : Review of HLA typing by NGS. Int J Immunogenet [Internet]. 2013 Feb;40(1):72–6. Available from: http://doi.wiley.com/10.1111/iji.12024

124. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.

125. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008;18(7):1051–63.

126. Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–26.

127. Gabriel C, Fürst D, Faé I, Wenda S, Zollikofer C, Mytilineos J, et al. HLA typing by next-generation sequencing - getting closer to reality. Tissue Antigens. 2014;83(2):65–75.

128. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. J Hum Genet [Internet]. 2015;60(11):665–73. Available from: http://dx.doi.org/10.1038/jhg.2015.102

129. Vayntrub TA, Mack SJ, Fernandez-Viña MA. Preface: 17th International HLA and Immunogenetics Workshop. Hum Immunol [Internet]. 2020 Feb;81(2–3):52–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885920300471

130. Jekarl DW, Lee GD, Yoo J Bin, Kim JR, Yu H, Yoo J, et al. HLA-A , -B, -C, -DRB1 allele and haplotype frequencies of the Korean population and performance characteristics of HLA typing by next-generation sequencing. HLA [Internet]. 2021 Mar 2;97(3):188–97. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14167

131. Baier DM, Hofmann JA, Fischer H, Rall G, Stolze J, Ruhner K, et al. Very low error rates of NGS-based HLA typing at stem cell donor recruitment question the need for a standard confirmatory typing step before donor work-up. Bone Marrow Transplant [Internet]. 2019;54(6):928–30. Available from: http://dx.doi.org/10.1038/s41409-018-0411-2

132. Nilsson LL, Funck T, Kjersgaard ND, Hviid TVF. Next-generation sequencing of HLA-G based on long-range polymerase chain reaction. HLA [Internet]. 2018 Sep;92(3):144–53. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13342

133. Ralazamahaleo M, Andreani M, Giustiniani P, Guidicelli G, Visentin J. Characterization of the novel HLA-DQA1*01:01:05 allele by sequencing-based typing. HLA [Internet]. 2019 Aug 22;94(2):172–3. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13569

134. Loginova M, Smirnova D, Kutyavina S, Paramonov I, Zarubin M. The novel HLA-A allele, HLA-A*01:354 , identified in a Buryat individual. HLA [Internet]. 2021 May 22;97(5):435–6. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14170

135. Ananeva A, Leksina Y, Andryushkina A, Shagimardanova E. The novel HLA-A*02:941 allele was identified during high-resolution HLA typing. HLA [Internet]. 2021 Feb 13;97(2):136–8. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14088

136. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, et al. HLA Typing for the Next Generation. PLoS One [Internet]. 2015 May 27;10(5):e0127153. Available from: https://dx.plos.org/10.1371/journal.pone.0127153

137. De Santis D, Truong L, Martinez P, D'Orsogna L. Rapid high-resolution <scp>HLA</scp> genotyping by <scp>MinION</scp> Oxford nanopore sequencing for deceased donor organ allocation. HLA [Internet]. 2020 Aug 26;96(2):141–62. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13901

138. Mosbruger TL, Dinou A, Duke JL, Ferriola D, Mehler H, Pagkrati I, et al. Utilizing nanopore sequencing technology for the rapid and comprehensive characterization of eleven HLA loci;

addressing the need for deceased donor expedited HLA typing. Hum Immunol [Internet]. 2020;81(8):413–22. Available from: https://doi.org/10.1016/j.humimm.2020.06.004

139. Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. Nat Genet [Internet]. 2015;47(6):682–8. Available from: http://dx.doi.org/10.1038/ng.3257

140. Allen FH, Amos DB, Batchelor JR, Bodmer WF, Ceppellini R, Dausset J, et al. Nomenclature for factors of the HL-a system. Bull World Health Organ [Internet]. 1968;39(3):483–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/5303912

141. Albert E. Nomenclature for factors of the HLA system, 1987. Tissue Antigens [Internet]. 1988 Oct;32(4):177–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3217934

142. Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, et al. Nomenclature for factors of the HLA system, 2010. Tissue Antigens [Internet]. 2010;75(4):291–455. Available from: http://doi.wiley.com/10.1111/j.1399-0039.2010.01466.x

143. Douillard V, Castelli EC, Mack SJ, Hollenbach JA, Gourraud P-A, Vince N, et al. Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research. Front Genet [Internet]. 2021 Dec 2;12. Available from: In review

144. Sidney J, del Guercio MF, Southwood S, Engelhard VH, Appella E, Rammensee HG, et al. Several HLA alleles share overlapping peptide specificities. J Immunol [Internet]. 1995 Jan 1;154(1):247–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7527812

145. del Guercio MF, Sidney J, Hermanson G, Perez C, Grey HM, Kubo RT, et al. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J Immunol [Internet]. 1995;154(2):685–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7529283

146. Wang M, Claesson MH. Classification of Human Leukocyte Antigen (HLA) Supertypes. In: Methods in molecular biology (Clifton, NJ) [Internet]. 2014. p. 309–17. Available from: http://link.springer.com/10.1007/978-1-4939-1115-8_17

147. Petersdorf EW, Anasetti C, Martin PJ, Gooley T, Radich J, Malkki M, et al. Limits of HLA mismatching in unrelated hematopoietic cell transplantation. Blood [Internet]. 2004;104(9):2976–80. Available from: http://dx.doi.org/10.1182/blood-2004-04-1674

148. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood [Internet]. 2007;110(13):4576–83. Available from: http://dx.doi.org/10.1182/blood-2007-06-097386

149. Morishima Y, Kashiwase K, Matsuo K, Azuma F, Morishima S, Onizuka M, et al. Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. Blood. 2015;125(7):1189–97.

150. Spellman S, Setterholm M, Maiers M, Noreen H, Oudshoorn M, Fernandez-Viña M, et al. Advances in the Selection of HLA-Compatible Donors: Refinements in HLA Typing and Matching over the First 20 Years of the National Marrow Donor Program Registry. Biol Blood Marrow Transplant. 2008;14(9 SUPPL.):37–44.

151. Geffard E, Boussamet L, Walencik A, Delbos F, Limou S, Gourraud P, et al. HLA-EPI : A new EPIsode in exploring donor/recipient epitopic compatibilities. HLA [Internet]. 2022 Feb 16;99(2):79–92. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14505

152. Tambur AR, Claas FHJ. HLA epitopes as viewed by antibodies: What Is it all about? Am J Transplant. 2015;15(5):1148–54.

153. Laune D, Molina F, Ferrieres G, Mani JC, Cohen P, Simon D, et al. Systematic exploration of the antigen binding activity of synthetic peptides isolated from the variable regions of immunoglobulins. J Biol Chem [Internet]. 1997;272(49):30937–44. Available from:

http://dx.doi.org/10.1074/jbc.272.49.30937

154. Duquesnoy RJ, Marrari M, Tambur AR, Mulder A, da Mata Sousa LCD, da Silva AS, et al. First report on the antibody verification of HLA-DR, HLA-DQ and HLA-DP epitopes recorded in the HLA Epitope Registry. Hum Immunol [Internet]. 2014;75(11):1097–103. Available from: http://dx.doi.org/10.1016/j.humimm.2014.09.012

155. Rimando J, Slade M, DiPersio JF, Westervelt P, Gao F, Liu C, et al. HLA epitope mismatch in haploidentical transplantation is associated with decreased relapse and delayed engraftment. Blood Adv. 2018;2(24):3590–601.

156. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007;2(8).

157. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class i binding prediction beyond humans. Immunogenetics. 2009;61(1):1–13.

158. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE. IMGT/HLA Database - A sequence database for the human major histocompatibility complex. Nucleic Acids Res. 2001;29(1):210–3.

159. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Res [Internet]. 2019 Oct 31;48(D1):D948–55. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz950/5610347

160. Robinson J, Waller MJ, Stoehr P, Marsh SGE. IPD - The Immuno Polymorphism Database. Nucleic Acids Res. 2005;33(DATABASE ISS.):523–6.

161. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res [Internet]. 2015 Jan 28;43(D1):D423–31. Available from: http://academic.oup.com/nar/article/43/D1/D423/2438496/The-IPD-and-IMGTHLA-database-allele-variant

162. Middleton D, Menchaca L, Rood H, Komerofsky R. New allele frequency database: http://www.allelefrequencies.net. Tissue Antigens [Internet]. 2003 May;61(5):403–7. Available from: http://doi.wiley.com/10.1034/j.1399-0039.2003.00062.x

163. Gonzalez-Galarza FF, McCabe A, Santos EJM Dos, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. 2020;48(D1):D783–8.

164. LIENERT K, PARHAM P. Evolution of MHC class I genes in higher primates. Immunol Cell Biol [Internet]. 1996 Aug;74(4):349–56. Available from: http://doi.wiley.com/10.1038/icb.1996.62

165. Hughes AL, Yeager M, Carrington M. Peptide binding function and the paradox of HLA disease associations. Immunol Cell Biol. 1996;74(5):444–8.

166. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW. Evolutionarily conserved amino acids that control TCR-MHC interaction. Annu Rev Immunol. 2008;26(V):171–203.

167. Peruzzi M, Wagtmann N, Long EO. A p70 killer cell inhibitory receptor specific for several HLA-B allotypes discriminates among peptides bound to HLA-B*2705. J Exp Med. 1996;184(4):1585–90.

168. Vivian JP, Duncan RC, Berry R, O'Connor GM, Reid HH, Beddoe T, et al. Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. Nature [Internet]. 2011;479(7373):401–5. Available from: http://dx.doi.org/10.1038/nature10517

169. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Guethlein LA, Hilton HG, Pando MJ, et al. Co-evolution of Human Leukocyte Antigen (HLA) Class I Ligands with Killer-Cell Immunoglobulin-Like Receptors (KIR) in a Genetically Diverse Population of Sub-Saharan Africans. PLoS Genet.

2013;9(10).

170. Parham P, Adams EJ, Arnett KL. The Origins of HLA-A,B,C Polymorphism. Immunol Rev. 1995;143(1):141–80.

171. Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, et al. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. Keating BJ, editor. PLOS Genet [Internet]. 2017 Jun 26;13(6):e1006862. Available from: https://dx.plos.org/10.1371/journal.pgen.1006862

172. Parham P, Lawlor DA. Evolution of class I major histocompatibility complex genes and molecules in humans and apes. Hum Immunol. 1991;30(2):119–28.

173. Davis DM, Mandelboim O, Luque I, Baba E, Boyson J, Strominger JL. The transmembrane sequence of human histocompatibility leukocyte antigen (HLA)-C as a determinant in inhibition of a subset of natural killer cells. J Exp Med. 1999;189(8):1265–74.

174. Drake LA, Drake JR. A triad of molecular regions contribute to the formation of two distinct MHC class II conformers. Mol Immunol [Internet]. 2016;74:59–70. Available from: http://dx.doi.org/10.1016/j.molimm.2016.04.010

175. Castelli EC, Mendes-Junior CT, Deghaide NHS, De Albuquerque RS, Muniz YCN, Simes RT, et al. The genetic structure of 3'untranslated region of the HLA-G gene: Polymorphisms and haplotypes. Genes Immun. 2010;11(2):134–41.

176. Garcia A, Milet J, Courtin D, Sabbagh A, Massaro JD, Castelli EC, et al. Association of HLA-G 3'UTR polymorphisms with response to malaria infection: A first insight. Infect Genet Evol. 2013;16:263–9.

177. Iturrieta-Zuazo I, Rita CG, García-Soidán A, de Malet Pintos-Fonseca A, Alonso-Alarcón N, Pariente-Rodríguez R, et al. Possible role of HLA class-I genotype in SARS-CoV-2 infection and progression: A pilot study in a cohort of Covid-19 Spanish patients. Clin Immunol [Internet]. 2020 Oct;219(January):108572. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1521661620307324

178. Rohn H, Schwich E, Tomoya Michita R, Schramm S, Dolff S, Gäckler A, et al. HLA-G 3' untranslated region gene variants are promising prognostic factors for BK polyomavirus replication and acute rejection after living-donor kidney transplant. Hum Immunol. 2020;81(4):141–6.

179. Ramsuran V, Hernández-Sanchez PG, O'hUigin C, Sharma G, Spence N, Augusto DG, et al. Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes. J Immunol [Internet]. 2017 Mar 15;198(6):2320–9. Available from: http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.1601679

180. Lima THA, Souza AS, Porto IOP, Paz MA, Veiga-Castelli LC, Oliveira MLG, et al. HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. HLA [Internet]. 2019 Feb 22;93(2–3):65–79. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13474

181. René C, Lozano C, Villalba M, Eliaou J-F. 5' and 3' untranslated regions contribute to the differential expression of specific HLA-A alleles. Eur J Immunol [Internet]. 2015 Dec;45(12):3454–63. Available from: https://onlinelibrary.wiley.com/doi/10.1002/eji.201545927

182. Kulpa DA, Collins KL. The emerging role of HLA-C in HIV-1 infection. Immunology. 2011;134(2):116–22.

183. Vince N, Li H, Ramsuran V, Naranbhai V, Duh FM, Fairfax BP, et al. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. Am J Hum Genet [Internet]. 2016;99(6):1353–8. Available from: http://dx.doi.org/10.1016/j.ajhg.2016.09.023

184. Klitz W, Hedrick P, Louis EJ. New reservoirs of HLA alleles: Pools of rare variants enhance immune defense. Trends Genet [Internet]. 2012;28(10):480–6. Available from: http://dx.doi.org/10.1016/j.tig.2012.06.007

185. Gonzalez-Galarza FF, Mack SJ, Hollenbach J, Fernandez-Vina M, Setterholm M, Kempenich J, et al. 16th IHIW: Extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles. Int J Immunogenet. 2013;40(1):60–5.

186. Middleton D, Gonzalez F, Fernandez-Vina M, Tiercy JM, Marsh SGE, Aubrey M, et al. A bioinformatics approach to ascertaining the rarity of HLA alleles. Tissue Antigens. 2009;74(6):480–5.

187. Hurley CK, Kempenich J, Wadsworth K, Sauter J, Hofmann JA, Schefzyk D, et al. Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. HLA [Internet]. 2020 Jun;95(6):516–31. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13811

188. Gonzalez-Galarza FF, McCabe A, Melo dos Santos EJ, Jones AR, Middleton D. A snapshot of human leukocyte antigen (HLA) diversity using data from the Allele Frequency Net Database. Hum Immunol [Internet]. 2021;82(7):496–504. Available from: https://doi.org/10.1016/j.humimm.2020.10.004

189. Ciurea SO, Zhang MJ, Bacigalupo AA, Bashey A, Appelbaum FR, Aljitawi OS, et al. Haploidentical transplant with posttransplant cyclophosphamide vs matched unrelated donor transplant for acute myeloid leukemia. Blood. 2015;126(8):1033–40.

190. Bashey A, Zhang X, Jackson K, Brown S, Ridgeway M, Solh M, et al. Comparison of Outcomes of Hematopoietic Cell Transplants from T-Replete Haploidentical Donors Using Post-Transplantation Cyclophosphamide with 10 of 10 HLA-A, -B, -C, -DRB1, and -DQB1 Allele-Matched Unrelated Donors and HLA-Identical Sibling Donors: A Mul. Biol Blood Marrow Transplant [Internet]. 2016;22(1):125–33. Available from: http://dx.doi.org/10.1016/j.bbmt.2015.09.002

191. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol. 2007;68(9):779–88.

192. Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA Diversity in the 1000 Genomes Dataset. Colombo GI, editor. PLoS One [Internet]. 2014 Jul 2;9(7):e97282. Available from: https://dx.plos.org/10.1371/journal.pone.0097282

193. Flajnik MF, Kasahara M. Comparative genomics of the MHC: Glimpses into the evolution of the adaptive immune system. Immunity. 2001;15(3):351–62.

194. Bernstein RM, Schluter SF, Bernstein H, Marchalonis JJ. Primordial emergence of the recombination activating gene 1 (RAG1): Sequence of the complete shark gene indicates homology to microbial integrases. Proc Natl Acad Sci U S A. 1996;93(18):9454–9.

195. Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H. Comparative genomic analysis of the MHC: The evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol Rev. 2002;190(1):95–122.

196. Kaufman J, Milne S, Göbel TWF, Walker BA, Jacob JP, Auffray C, et al. The chicken B locus is a minimal essential major histocompatibility complex. Nature. 1999;401(6756):923–5.

197. Kaufman J. Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. Trends Immunol [Internet]. 2018;39(5):367–79. Available from: http://dx.doi.org/10.1016/j.it.2018.01.001

198. Brandt DYC, César J, Goudet J, Meyer D. The Effect of Balancing Selection on Population Differentiation: A Study with HLA Genes. G3 Genes|Genomes|Genetics [Internet]. 2018 Aug 1;8(8):2805–15. Available from:

https://academic.oup.com/g3journal/article/8/8/2805/6026912

199. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc B Biol Sci [Internet]. 2010 Apr 7;277(1684):979–88. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2009.2084

200. Doherty PC, Zinkernagel RM. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature. 1975;256(5512):50–2.

201. Takahata N, Nei M. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics. 1990;124(4):967–78.

202. Hill AVS. HLA Associations with Malaria in Africa: Some Implications for MHC Evolution. Mol Evol Major Histocompat Complex. 1991;403–20.

203. Cutting GR. Cystic fibrosis genetics: From molecular understanding to clinical application. Nat Rev Genet. 2015;16(1):45–56.

204. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet [Internet]. 2017 Jul;101(1):5–22. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409

205. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet [Internet]. 2019;20(8):467–84. Available from: http://dx.doi.org/10.1038/s41576-019-0127-1

206. Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet [Internet]. 2018 Sep 13;50(9):1219–24. Available from: http://www.nature.com/articles/s41588-018-0183-z

207. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362–7.

208. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–12.

209. Price P, Witt C, Allock R, Sayer D, Garlepp M, Kok CC, et al. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol Rev [Internet]. 1999 Feb;167(1):257–74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1600-065X.1999.tb01398.x

210. Profaizer T, Eckels D. HLA alleles and drug hypersensitivity reactions. Int J Immunogenet. 2012;39(2):99–105.

211. Gomes ER, Demoly P. Epidemiology of hypersensitivity drug reactions. Curr Opin Intern Med. 2005;4(5):487–94.

212. Douillard V, Castelli EC, Mack SJ, Hollenbach JA, Gourraud PA, Vince N, et al. Current HLA Investigations on SARS-CoV-2 and Perspectives [Internet]. Vol. 12, Frontiers in Genetics. 2021. p. 10–6. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2021.774922/full

213. Sanchez-Mazas A. A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. Swiss Med Wkly [Internet]. 2020 Apr 16;150(April):w20214. Available from: https://doi.emh.ch/smw.2020.20214

214. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol [Internet]. 2018 May 2;18(5):325–39. Available from: http://dx.doi.org/10.1038/nri.2017.143

215. Fellay J, Shianna K V, Ge D, Colombo S, Ledergerber B, Weale M, et al. Study of Major Determinants for Host Control of HIV-1. Science (80-) [Internet]. 2007;317(August):944–7. Available from: www.sciencemag.org/cgi/content/full/1143767/DC1

216. Pereyra F, Jia X, McLaren PJ, Telenti A, De Bakker PIW, Walker BD. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation - SUPP. Science (80-). 2010;330(6010):1551–7.

217. Pelak K, Goldstein DB, Walley NM, Fellay J, Ge D, Shianna K V., et al. Host Determinants of HIV-1 Control in African Americans. J Infect Dis [Internet]. 2010 Apr 15;201(8):1141–9. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1086/651382

218. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, et al. Influence of HLA-C Expression Level on HIV Control. Science (80-) [Internet]. 2013 Apr 5;340(6128):87–91. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1232685

219. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PIW, Walker BD, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science (80-) [Internet]. 2010 Dec 10;330(6010):1551–7. Available from: https://www.science.org/doi/10.1126/science.1195271

220. Bardeskar NS, Mania-Pramanik J. HIV and host immunogenetics: unraveling the role of HLA-C. HLA [Internet]. 2016 Nov;88(5):221–31. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.12882

221. Lavanchy D. The global burden of hepatitis C. Liver Int. 2009;29(SUPPL. 1):74–81.

222. Duggal P, Thio CL, Wojcik GL, Goedert JJ, Mangia A, Latanich R, et al. Genome-wide association study of spontaneous resolution of hepatitis C virus infection: Data from multiple cohorts. Ann Intern Med. 2013;158(4):235–45.

223. Vergara C, Thio CL, Johnson E, Kral AH, O'Brien TR, Goedert JJ, et al. Multi-Ancestry Genome-Wide Association Study of Spontaneous Clearance of Hepatitis C Virus. Gastroenterology [Internet]. 2019 Apr;156(5):1496-1507.e7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0016508518354210

224. Valencia A, Vergara C, Thio CL, Vince N, Douillard V, Grifoni A, et al. Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1. Am J Hum Genet [Internet]. 2022;109(2):299–310. Available from: https://doi.org/10.1016/j.ajhg.2022.01.001

225. Adebamowo SN, Adeyemo AA. Classical HLA alleles are associated with prevalent and persistent cervical high-risk HPV infection in African women. Hum Immunol [Internet]. 2019;80(9):723–30. Available from: https://doi.org/10.1016/j.humimm.2019.04.011

226. Chen Y, Liao Y, Yuan K, Wu A, Liu L. HLA-A, -B, -DRB1 Alleles as Genetic Predictive Factors for Dengue Disease: A Systematic Review and Meta-Analysis. Viral Immunol [Internet]. 2019 Apr;32(3):121–30. Available from: https://www.liebertpub.com/doi/10.1089/vim.2018.0151

227. Sawai H, Nishida N, Khor S-S, Honda M, Sugiyama M, Baba N, et al. Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepatocellular carcinoma. Sci Rep [Internet]. 2018 Dec 21;8(1):7958. Available from: http://www.nature.com/articles/s41598-018-26217-7

228. Huang Y-H, Liao S-F, Khor S-S, Lin Y-J, Chen H-Y, Chang Y-H, et al. Large-scale genome-wide association study identifies HLA class II variants associated with chronic HBV infection: a study from Taiwan Biobank. Aliment Pharmacol Ther [Internet]. 2020 Aug;52(4):682–91. Available from: http://doi.wiley.com/10.1111/apt.15887

229. Spínola H. HLA Loci and Respiratory Infectious Diseases. J Respir Res [Internet]. 2016;2(3):56–66. Available from: http://www.ghrnet.org/index.php/jrr/article/view/1639

230. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med [Internet]. 2020;46(5):846–8. Available from: https://doi.org/10.1007/s00134-020-05991-x

231. Zhou Y, Fu B, Zheng X, Wang D, Zhao C, Qi Y, et al. Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients. Natl Sci Rev [Internet]. 2020 Jun 1;7(6):998–1002. Available from: https://academic.oup.com/nsr/article/7/6/998/5804736

232. Worldometer. COVID-19 CORONAVIRUS PANDEMIC [Internet]. [cited 2022 Sep 22]. Available from: https://www.worldometers.info/coronavirus/

233. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human leukocyte antigen susceptibility map for SARS-CoV-2. J Virol [Internet]. 2020;94(13):1–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32303592

234. Romero-López JP, Carnalla-Cortés M, Pacheco-Olvera DL, Ocampo-Godínez JM, Oliva-Ramírez J, Moreno-Manjón J, et al. A bioinformatic prediction of antigen presentation from SARS-CoV-2 spike protein revealed a theoretical correlation of HLA-DRB1*01 with COVID-19 fatality in Mexican population: An ecological approach. J Med Virol [Internet]. 2020;0–2. Available from: http://dx.doi.org/10.1002/jmv.26561

235. Correale P, Mutti L, Pentimalli F, Baglio G, Saladino RE, Sileri P, et al. HLA-B*44 and C*01 Prevalence Correlates with Covid19 Spreading across Italy. Int J Mol Sci [Internet]. 2020 Jul 23;21(15):5205. Available from: https://www.mdpi.com/1422-0067/21/15/5205

236. Pisanti S, Deelen J, Gallina AM, Caputo M, Citro M, Abate M, et al. Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. J Transl Med [Internet]. 2020;18(1):1–16. Available from: https://doi.org/10.1186/s12967-020-02515-5

237. Ishii T. Human Leukocyte Antigen (HLA) Class I Susceptible Alleles Against COVID-19 Increase Both Infection and Severity Rate. Cureus [Internet]. 2020 Dec 23;12(12). Available from: https://www.cureus.com/articles/35178-human-leukocyte-antigen-hla-class-i-susceptible-alleles-against-covid-19-increase-both-infection-and-severity-rate

238. Sakuraba A, Haider H, Sato T. Population Difference in Allele Frequency of HLA-C*05 and Its Correlation with COVID-19 Mortality. Viruses [Internet]. 2020 Nov 20;12(11):1333. Available from: https://www.mdpi.com/1999-4915/12/11/1333

239. Khor S, Omae Y, Nishida N, Sugiyama M, Kinoshita N, Suzuki T, et al. HLA-A*11:01:01:01, HLA-C*12:02:02:01-HLA-B*52:01:02:02, Age and Sex Are Associated With Severity of Japanese COVID-19 With Respiratory Failure. Front Immunol [Internet]. 2021 Apr 22;12(April). Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.658570/full

240. Novelli A, Andreani M, Biancolella M, Liberatoscioli L, Passarelli C, Colona VL, et al. HLA allele frequencies and susceptibility to COVID-19 in a group of 99 Italian patients. HLA [Internet]. 2020 Nov 3;96(5):610–4. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14047

241. Deb P, Zannat K, Talukder S, Bhuiyan AH, Jilani MSA, Saif-Ur-Rahman KM. Association of HLA gene polymorphism with susceptibility, severity, and mortality of COVID-19: A systematic review. HLA [Internet]. 2022 Apr 15;99(4):281–312. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14560

242. Dobrijević Z, Gligorijević N, Šunderić M, Penezić A, Miljuš G, Tomić S, et al. The association of human leucocyte antigen (HLA) alleles with COVID - 19 severity: A systematic review and meta - analysis. 2022;(April).

243. The Severe Covid-19 GWAS Group. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med [Internet]. 2020 Oct 15;383(16):1522–34. Available from: http://www.nejm.org/doi/10.1056/NEJMoa2020283

244. Degenhardt F, Ellinghaus D, Juzenas S, Lerga-Jaso J, Wendorff M, Maya-Miles D, et al. Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Hum Mol Genet [Internet]. 2022 Jul 15;(March):4–7. Available from:

http://dx.doi.org/10.1016/j.mee.2009.03.089

245. Astbury S, Reynolds CJ, Butler DK, Muñoz-Sandoval DC, Lin KM, Pieper FP, et al. HLA-DR polymorphism in SARS-CoV-2 infection and susceptibility to symptomatic COVID-19. Immunology. 2022;(January):68–77.

246. Wang F, Huang S, Gao R, Zhou Y, Lai C, Li Z, et al. Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. Cell Discov [Internet]. 2020; Available from: http://dx.doi.org/10.1038/s41421-020-00231-4

247. Weiner J, Suwalski P, Holtgrewe M, Rakitko A, Thibeault C, Müller M, et al. Increased risk of severe clinical course of COVID-19 in carriers of HLA-C*04:01. eClinicalMedicine. 2021;40.

248. Gutiérrez-Bautista JF, Rodriguez-Nicolas A, Rosales-Castillo A, López-Ruz MÁ, Martín-Casares AM, Fernández-Rubiales A, et al. Study of HLA-A, -B, -C, -DRB1 and -DQB1 polymorphisms in COVID-19 patients. J Microbiol Immunol Infect [Internet]. 2022 Jun;55(3):421–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1684118221001833

249. De Marco R, Faria TC, Mine KL, Cristelli M, Medina-Pestana JO, Tedesco-Silva H, et al. HLA-A homozygosis is associated with susceptibility to COVID-19. HLA [Internet]. 2021 Aug 29;98(2):122–31. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14349

250. Augusto DG, Yusufali T, Peyser ND, Butcher X, Marcus GM, Olgin JE, et al. HLA-B*15:01 is associated with asymptomatic SARS-CoV-2 infection. medRxiv Prepr Serv Heal Sci [Internet]. 2021;1(510). Available from: http://www.ncbi.nlm.nih.gov/pubmed/34031661%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8142661

251. Nguyen A, Yusufali T, Hollenbach JA, Nellore A, Thompson RF. Minimal observed impact of HLA genotype on hospitalization and severity of SARS-CoV-2 infection. HLA [Internet]. 2022 Jun 24;99(6):607–13. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.14574

252. Shachar S Ben, Barda N, Manor S, Israeli S, Dagan N, Carmi S, et al. MHC Haplotyping of SARS-CoV-2 Patients: HLA Subtypes Are Not Associated with the Presence and Severity of COVID-19 in the Israeli Population. J Clin Immunol [Internet]. 2021 May 29; Available from: https://doi.org/10.1007/s10875-021-01071-x

253. Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. 2017;18(1).

254. Trowsdale J, Knight JC. Major Histocompatibility Complex Genomics and Human Disease. Annu Rev Genomics Hum Genet [Internet]. 2013 Aug 31;14(1):301–23. Available from: http://www.annualreviews.org/doi/10.1146/annurev-genom-091212-153455

255. Hu X, Deutsch AJ, Lenz TL, Onengut-Gumuscu S, Han B, Chen W-M, et al. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. Nat Genet [Internet]. 2015 Aug 13;47(8):898–905. Available from: http://www.nature.com/articles/ng.3353

256. Todd JA, Bell JI, McDevitt HO. HLA-DQbeta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature. 1987;329:599–604.

257. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet. 2012;44(12):1341–8.

258. Okada Y, Han B, Tsoi LC, Stuart PE, Ellinghaus E, Tejasvi T, et al. Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. Am J Hum Genet. 2014;95(2):162–72.

259. Mignot E, Kimura A, Latlermann A, Lin X, Yasunaga S, Mueller-Eckhardt G, et al. Extensive HLA class II studies in 58 non-DRB 1*15 (DR2) narcoleptic patients with cataplexy. Tissue Antigens.

1997;49(4):329–41.

260. Chabas D, Taheri S, Renier C, Mignot E. The Genetics of Narcolepsy. Annu Rev Genomics Hum Genet. 2003;4:459–83.

261. Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet. 2015;47(10):1107–13.

262. Patsopoulos NA, Barcellos LF, Hintzen RQ, Schaefer C, van Duijn CM, Noble JA, et al. Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. PLoS Genet. 2013;9(11).

263. Chao MJ, Barnardo MCNM, Lincoln MR, Ramagopalan S V., Herrera BM, Dyment DA, et al. HLA class I alleles tag HLA-DRB1*1501 haplotypes for differential risk in multiple sclerosis susceptibility. Proc Natl Acad Sci U S A. 2008;105(35):13069–74.

264. Belbasis L, Bellou V, Evangelou E, Ioannidis JPA, Tzoulaki I. Environmental risk factors and multiple sclerosis: An umbrella review of systematic reviews and meta-analyses. Lancet Neurol [Internet]. 2015;14(3):263–73. Available from: http://dx.doi.org/10.1016/S1474-4422(14)70267-4

265. Harkiolaki M, Holmes SL, Svendsen P, Gregersen JW, Jensen LT, McMahon R, et al. T Cell-Mediated Autoimmune Disease Due to Low-Affinity Crossreactivity to Common Microbial Peptides. Immunity [Internet]. 2009;30(3):348–57. Available from: http://dx.doi.org/10.1016/j.immuni.2009.01.009

266. Lang HLE, Jacobsen H, Ikemizu S, Andersson C, Harlos K, Madsen L, et al. A functional and structural basis for TCR cross-reactivity in multiple sclerosis. Nat Immunol. 2002;3(10):940–3.

267. Madsen LS, Andersson EC, Jansson L, Krogsgaard M, Andersen CB, Engberg J, et al. A humanized model for multiple sclerosis using HLA-DR2 and a human T- cell receptor. Nat Genet. 1999;23(3):343–7.

268. Hahn M, Nicholson MJ, Pyrdol J, Wucherpfennig KW. Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor. Nat Immunol. 2005;6(5):490–6.

269. Sulzer D, Alcalay RN, Garretti F, Cote L, Kanter E, Agin-Liebes J, et al. T cells from patients with Parkinson's disease recognize α-synuclein peptides. Nature. 2017;546(7660):656–61.

270. Kodavali C V., Watson AM, Prasad KM, Celik C, Mansour H, Yolken RH, et al. HLA associations in schizophrenia: Are we re-discovering the wheel? Am J Med Genet Part B Neuropsychiatr Genet [Internet]. 2014 Jan;165(1):19–27. Available from: https://onlinelibrary.wiley.com/doi/10.1002/ajmg.b.32195

271. Trowsdale J. The MHC, disease and selection. Immunol Lett [Internet]. 2011;137(1–2):1–8. Available from: http://dx.doi.org/10.1016/j.imlet.2011.01.002

272. Sekar A, Bialas AR, De Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4 Schizophrenia Working Group of the Psychiatric Genomics Consortium HHS Public Access. Nature Febr [Internet]. 2016;11(5307589):177–83. Available from: http://www.nature.com/authors/editorial_policies/license.html#terms

273. Chaix R, Cao C, Donnelly P. Is mate choice in humans MHC-dependent? PLoS Genet. 2008;4(9):1–5.

274. Winternitz J, Abbate JL, Huchard E, Havlíček J, Garamszegi LZ. Patterns of MHC-dependent mate selection in humans and nonhuman primates: a meta-analysis. Mol Ecol. 2017;26(2):668–88.

275. Sacchi N, Castagnetta M, Miotti V, Garbarino L, Gallina A. High-resolution analysis of the HLA-A, -B, -C and -DRB1 alleles and national and regional haplotype frequencies based on 120 926 volunteers from the Italian Bone Marrow Donor Registry. HLA [Internet]. 2019 Sep 3;94(3):285–95. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tan.13613

276. Schmidt AH, Sauter J, Baier DM, Daiss J, Keller A, Klussmeier A, et al. Immunogenetics in stem cell donor registry work: The DKMS example (Part 1). Int J Immunogenet [Internet]. 2020 Feb 6;47(1):13–23. Available from: https://onlinelibrary.wiley.com/doi/10.1111/iji.12471

277. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372(26):2509–20.

278. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science (80-) [Internet]. 2015 Apr 3;348(6230):124–8. Available from: https://www.science.org/doi/10.1126/science.aaa1348

279. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science (80-). 2015;350(6257):207–11.

280. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. N Engl J Med. 2014;371(23):2189–99.

281. Peters B, Bui HH, Frankild S, Nielsen M, Lundegaard C, Kostem E, et al. A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput Biol. 2006;2(6):0574–84.

282. Sidney J, Southwood S, Oseroff C, Guercio M, Sette A, Grey HM. Measurement of MHC/Peptide Interactions by Gel Filtration. Curr Protoc Immunol. 1999;31(1):1–19.

283. Ramarathinam SH, Croft NP, Illing PT, Faridi P, Purcell AW. Employing proteomics in the study of antigen presentation: an update. Expert Rev Proteomics [Internet]. 2018;15(8):637–45. Available from: https://doi.org/10.1080/14789450.2018.1509000

284. Becker FG, Cleary M, Team RM, Holtermann H, The D, Agenda N, et al. The HLA FactsBook [Internet]. Vol. 7, Syria Studies. Elsevier; 2000. 37–72 p. Available from: https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

285. Nielsen M, Lund O, Buus S, Lundegaard C. MHC Class II epitope predictive algorithms. Immunology. 2010;130(3):319–28.

286. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. Brief Bioinform [Internet]. 2020 Jul 15;21(4):1119–35. Available from: https://academic.oup.com/bib/article/21/4/1119/5511798

287. Niu L, Cheng H, Zhang S, Tan S, Zhang Y, Qi J, et al. Structural basis for the differential classification of HLA-A*6802 and HLA-A*6801 into the A2 and A3 supertypes. Mol Immunol [Internet]. 2013;55(3–4):381–92. Available from: http://dx.doi.org/10.1016/j.molimm.2013.03.015

288. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak HS, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol. 2017;13(8):e1005725.

289. Gfeller D, Guillaume P, Michaux J, Pak H-S, Daniel RT, Racle J, et al. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. J Immunol. 2018;201(12):3705–16.

290. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res [Internet]. 2020 Jul 2;48(W1):W449–54. Available from: https://academic.oup.com/nar/article/48/W1/W449/5837056

291. Klein RJ, Zeiss C, Chew EY, Tsai J, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science (80-) [Internet]. 2005 Apr 15;308(5720):385–9. Available from: https://www.science.org/doi/10.1126/science.1109557

292. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science (80-) [Internet]. 2006 Dec;314(5804):1461–3. Available from: https://www.science.org/doi/10.1126/science.1135245

293. Bednar M. DNA microarray technology and application. Med Sci Monit. 2000;6(4):796–800.

294. Deelen J, Beekman M, Uh HW, Broer L, Ayers KL, Tan Q, et al. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. Hum Mol Genet. 2014;23(16):4420–32.

295. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet [Internet]. 2019 Apr 29;51(4):592–9. Available from: http://www.nature.com/articles/s41588-019-0385-z

296. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet [Internet]. 2006 Oct 24;38(10):1166–72. Available from: http://www.nature.com/articles/ng1885

297. Colombo S, Rauch A, Rotger M, Fellay J, Martinez R, Fux C, et al. The HCP5 Single-Nucleotide Polymorphism: A Simple Screening Tool for Prediction of Hypersensitivity Reaction to Abacavir. J Infect Dis [Internet]. 2008 Sep 15;198(6):864–7. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1086/591184

298. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature [Internet]. 1988 Sep;335(6186):167–70. Available from: http://www.nature.com/articles/335167a0

299. Vince N, Limou S, Daya M, Morii W, Rafaels N, Geffard E, et al. Association of HLA-DRB1∗09:01 with tIgE levels among African-ancestry individuals with asthma. J Allergy Clin Immunol [Internet]. 2020 Jul;146(1):147–55. Available from: https://doi.org/10.1016/j.jaci.2020.01.011

300. Domenighetti C, Douillard V, Sugier P, Sreelatha AAK, Schulte C, Grover S, et al. The Interaction between HLA-DRB1 and Smoking in Parkinson's Disease Revisited. Mov Disord [Internet]. 2022 Jul 10; Available from: https://onlinelibrary.wiley.com/doi/10.1002/mds.29133

301. Ferreira LMR, Meissner TB, Tilburgs T, Strominger JL. HLA-G: At the Interface of Maternal–Fetal Tolerance. Trends Immunol [Internet]. 2017;38(4):272–86. Available from: http://dx.doi.org/10.1016/j.it.2017.01.009

302. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.

303. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. Biometrical J. 2021;63(5):915–47.

304. Ba R, Geffard E, Douillard V, Simon F, Mesnard L, Vince N, et al. Surfing the Big Data Wave: Omics Data Challenges in Transplantation. Transplantation. 2022;106(2):E114–25.

305. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;339(7713):157–60.

306. Ewens WJ. Mathematics, genetics and evolution. Quant Biol. 2013;1(1):9–31.

307. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Publ Gr [Internet]. 2010;11(7):499–511. Available from: http://dx.doi.org/10.1038/nrg2796

308. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6).

309. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906–13.

310. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.

311. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.

312. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet [Internet]. 2007;81(5):1084–97. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929707638828

313. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet [Internet]. 2018;103(3):338–48. Available from: https://doi.org/10.1016/j.ajhg.2018.07.015

314. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.

315. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet [Internet]. 2016 Oct 22;48(10):1279–83. Available from: http://www.nature.com/articles/ng.3643

316. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290–9.

317. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. J Hum Genet [Internet]. 2008;82(January):48–56. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929707000079

318. Meyer D, Nunes K. HLA imputation, what is it good for? Hum Immunol [Internet]. 2017 Mar;78(3):239–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885917300277

319. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J [Internet]. 2018 May 25;18(3):367–76. Available from: http://dx.doi.org/10.1038/tpj.2017.7

320. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG - HLA genotype imputation with attribute bagging. Pharmacogenomics J [Internet]. 2014;14(2):192–200. Available from: http://dx.doi.org/10.1038/tpj.2013.18

321. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. Tang J, editor. PLoS One [Internet]. 2013 Jun 6;8(6):e64683. Available from: https://dx.plos.org/10.1371/journal.pone.0064683

322. Dilthey A, Leslie S, Moutsianas L, Shen J, Cox C, Nelson MR, et al. Multi-Population Classical HLA Type Imputation. PLoS Comput Biol. 2013;9(2).

323. Xie M, Li J, Jiang T. Accurate HLA type inference using a weighted similarity graph. BMC Bioinformatics [Internet]. 2010;11(SUPPL. 11):S10. Available from: http://www.biomedcentral.com/1471-2105-11-S11

324. Setty MN, Gusev A, Pe'Er I. HLA type inference via haplotypes identical by descent. J Comput Biol. 2011;18(3):483–93.

325. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics [Internet]. 2011 Apr 1;27(7):968–72. Available from: https://academic.oup.com/bioinformatics/article-

lookup/doi/10.1093/bioinformatics/btr061

326. Li SS, Wang H, Smith A, Zhang B, Zhang XC, Schoch G, et al. Predicting multiallelic genes using unphased and flanking single nucleotide polymorphisms. Genet Epidemiol [Internet]. 2011 Feb;35(2):85–92. Available from: https://onlinelibrary.wiley.com/doi/10.1002/gepi.20549

327. Paunić V, Steinbach M, Kumar V, Maiers M. Prediction of HLA genes from SNP data and HLA haplotype frequencies. Proc - 12th IEEE Int Conf Data Min Work ICDMW 2012. 2012;964–71.

328. Paunić V, Steinbach M, Madbouly A, Kumar V. Amb-EM: A SNP-based prediction of HLA alleles using ambiguous HLA Data. ACM BCB 2014 - 5th ACM Conf Bioinformatics, Comput Biol Heal Informatics. 2014;104–13.

329. Motyer A, Vukcevic D, Dilthey A, Donnelly P, McVean G, Leslie S. Practical Use of Methods for Imputation of HLA Alleles from SNP Genotype Data. bioRxiv. 2016;091009.

330. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, et al. Imputation of KIR Types from SNP Variation Data. Am J Hum Genet [Internet]. 2015;97(4):593–607. Available from: http://dx.doi.org/10.1016/j.ajhg.2015.09.005

331. Squire DM, Motyer A, Ahn R, Nititham J, Huang Z-M, Oksenberg JR, et al. MHC*IMP – Imputation of Alleles for Genes in the Major Histocompatibility Complex. bioRxiv [Internet]. 2020; Available from: http://dx.doi.org/10.1101/2020.01.24.919191

332. Cook S, Choi W, Lim H, Luo Y, Kim K, Jia X. Accurate imputation of human leukocyte antigens. Nat Commun [Internet]. (2021):1–11. Available from: http://dx.doi.org/10.1038/s41467-021-21541-5

333. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat Commun [Internet]. 2021;12(1):1–14. Available from: http://dx.doi.org/10.1038/s41467-021-21975-x

334. Karnes JH, Shaffer CM, Bastarache L, Gaudieri S, Glazer AM, Steiner HE, et al. Comparison of HLA allelic imputation programs. Tang J, editor. PLoS One [Internet]. 2017 Feb 16;12(2):e0172444. Available from: https://dx.plos.org/10.1371/journal.pone.0172444

335. Kuniholm MH, Xie X, Anastos K, Xue X, Reimers L, French AL, et al. Human leucocyte antigen class I and II imputation in a multiracial population. Int J Immunogenet [Internet]. 2016 Dec;43(6):369–75. Available from: https://onlinelibrary.wiley.com/doi/10.1111/iji.12292

336. Ritari J, Hyvärinen K, Clancy J, Partanen J, Koskela S. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. NAR Genomics Bioinforma [Internet]. 2020 Jun 1;2(2):1–9. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa030/5831010

337. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 Genes|Genomes|Genetics [Internet]. 2015 May 1;5(5):931–41. Available from: https://academic.oup.com/g3journal/article/5/5/931/6025555

338. Bauer DC, Zadoorian A, Wilson LOW, Thorne NP. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Brief Bioinform [Internet]. 2016 Nov 1;19(2):bbw097. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw097

339. Castelli EC, Paz MA, Souza AS, Ramalho J, Mendes-Junior CT. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. Hum Immunol [Internet]. 2018 Sep;79(9):678–84. Available from: https://doi.org/10.1016/j.humimm.2018.06.010

340. Klasberg S, Surendranath V, Lange V, Schöfl G. Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping. Transfus Med

Hemotherapy [Internet]. 2019;46(5):312–25. Available from: https://www.karger.com/Article/FullText/502487

341. Dilthey AT, Mentzer AJ, Carapito R, Cutland C, Cereb N, Madhi SA, et al. HLA*LA—HLA typing from linearly projected graph alignments. Berger B, editor. Bioinformatics [Internet]. 2019 Nov 1;35(21):4394–6. Available from: https://academic.oup.com/bioinformatics/article/35/21/4394/5426702

342. Chen J, Madireddi S, Nagarkar D, Migdal M, Vander Heiden J, Chang D, et al. In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales. Brief Bioinform [Internet]. 2021 May 20;22(3):1–11. Available from: https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa223/5906908

343. Hsieh AR, Chang SW, Chen PL, Chu CC, Hsiao CL, Yang WS, et al. Predicting HLA genotypes using unphased and flanking single-nucleotide polymorphisms in Han Chinese population. BMC Genomics [Internet]. 2014;15(1):1–13. Available from: BMC Genomics

344. Vince N, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. Genet Epidemiol [Internet]. 2020 Oct 18;44(7):733–40. Available from: https://onlinelibrary.wiley.com/doi/10.1002/gepi.22334

345. Nunes K, Zheng X, Torres M, Moraes ME, Piovezan BZ, Pontes GN, et al. HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. Hum Immunol [Internet]. 2016 Mar;77(3):307–12. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0198885915005571

346. Degenhardt F, Wendorff M, Wittig M, Ellinghaus E, Datta LW, Schembri J, et al. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum Mol Genet [Internet]. 2019 Jun 15;28(12):20782092. Available from: https://academic.oup.com/hmg/article/28/12/2078/5261434

347. Khor SS, Yang W, Kawashima M, Kamitsuji S, Zheng X, Nishida N, et al. High-Accuracy imputation for HLA class i and II genes based on high-resolution SNP data of population-specific references. Pharmacogenomics J. 2015;15(6):530–7.

348. Blanton RE. Population Genetics and Molecular Epidemiology of Eukaryotes. Sadowsky M, Riley LW, editors. Microbiol Spectr [Internet]. 2018 Nov 2;6(6):139–48. Available from: https://journals.asm.org/doi/10.1128/microbiolspec.AME-0002-2018

349. Okazaki A, Yamazaki S, Inoue I, Ott J. Population genetics: past, present, and future. Hum Genet [Internet]. 2021;140(2):231–40. Available from: https://doi.org/10.1007/s00439-020-02208-5

350. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature [Internet]. 2017 Jan 19;541(7637):302–10. Available from: https://www.nature.com/articles/nature21347

351. Browning SR, Browning BL. Identity by descent between distant relatives: Detection and applications. Annu Rev Genet. 2012;46:617–33.

352. Gklambauer. Identity by descent [Internet]. 2014. Available from: https://en.wikipedia.org/wiki/Identity_by_descent

353. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132(2):583–9.

354. WRIGHT S. THE GENETICAL STRUCTURE OF POPULATIONS. Ann Eugen [Internet]. 1949 Jan;15(1):323–54. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1949.tb02451.x

355. Holsinger KE, Weir BS. Genetics in geographically structured populations: Defining, estimating and interpreting FST. Nat Rev Genet. 2009;10(9):639–50.

356. Ginsburgh V, Weber S. The Palgrave Handbook of Economics and Language [Internet]. Ginsburgh V, Weber S, editors. London: Palgrave Macmillan UK; 2016. 9–25 p. Available from: http://link.springer.com/10.1007/978-1-137-32505-1

357. Weir BS, Goudet J. A Unified Characterization of Population Structure and Relatedness. Genetics [Internet]. 2017 Aug 1;206(4):2085–103. Available from: https://academic.oup.com/genetics/article/206/4/2085/6072590

358. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.

359. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science (80-). 2002;298(5602):2381–5.

360. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.

361. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet [Internet]. 2013;93(2):278–88. Available from: http://dx.doi.org/10.1016/j.ajhg.2013.06.020

362. Naslavsky MS, Scliar MO, Yamamoto GL, Wang JYT, Zverinova S, Karp T, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. Nat Commun. 2022;13(1):1–11.

363. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet [Internet]. 2010 Jul 15;11(7):459–63. Available from: http://www.nature.com/articles/nrg2813

364. CAAPA. Genome-wide association study of asthma in individuals of African ancestry reveals novel asthma susceptibility loci. 2017;21.

365. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93.

366. Abegaz F, Chaichoompu K, Génin E, Fardo DW, König IR, John JMM, et al. Principals about principal components in statistical genetics. Brief Bioinform. 2019;20(6):2200–16.

367. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet [Internet]. 2020;85–91. Available from: http://dx.doi.org/10.1038/s10038-020-00851-4

368. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–47.

369. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat Biotechnol [Internet]. 2021;39(February). Available from: http://dx.doi.org/10.1038/s41587-020-00809-z

370. Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, et al. Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun [Internet]. 2020;11(1):1–11. Available from: http://dx.doi.org/10.1038/s41467-020-15194-z

371. Dai Y, Pei G, Zhao Z, Jia P. A convergent study of genetic variants associated with Crohn's disease: Evidence from GWAS, gene expression, methylation, eQTL and TWAS. Front Genet. 2019;10(APR):1–13.

372. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018; Available from: http://arxiv.org/abs/1802.03426

373. CAAPA. The CAAPA Project [Internet]. 2012. Available from: https://www.caapa-project.org/

374. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of

admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun. 2016;7.

375. Breiman L. Out-of-bag estimation. Tech Report Univ Calif. 1996;1–13.

376. Hollenbach JA, Norman PJ, Creary LE, Damotte V, Montero-Martin G, Caillier S, et al. A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. Proc Natl Acad Sci [Internet]. 2019 Apr 9;116(15):7419–24. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1821778116

377. Naito T, Satake W, Ogawa K, Suzuki K, Hirata J, Foo JN, et al. Trans-Ethnic Fine-Mapping of the Major Histocompatibility Complex Region Linked to Parkinson's Disease. Mov Disord. 2021;36(8):1805–14.

378. Chuang Y-H, Lee P-C, Vlaar T, Mulot C, Loriot M-A, Hansen J, et al. Pooled analysis of the HLA-DRB1 by smoking interaction in Parkinson disease. Ann Neurol [Internet]. 2017 Nov;82(5):655–64. Available from: file:///C:/Users/Carla Carolina/Desktop/Artigos para acrescentar na qualificação/The impact of birth weight on cardiovascular disease risk in the.pdf

379. Herzig AF, Velo-Suárez L, Consortium F, Consortium F, Dina C, Redon R, et al. Can imputation in a European country be improved by local reference panels? The example of France. bioRxiv [Internet]. 2022;2022.02.17.480829. Available from: https://www.biorxiv.org/content/10.1101/2022.02.17.480829v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.02.17.480829v1.abstract

380. Kals M, Nikopensius T, Läll K, Pärn K, Sikka TT, Suvisaari J, et al. Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. bioRxiv [Internet]. 2019;579201. Available from: https://www.biorxiv.org/content/10.1101/579201v1?rss=1

381. Huang Y-H, Khor S-S, Zheng X, Chen H-Y, Chang Y-H, Chu H-W, et al. A high-resolution HLA imputation system for the Taiwanese population: a study of the Taiwan Biobank. Pharmacogenomics J [Internet]. 2020 Oct 11;20(5):695–704. Available from: http://dx.doi.org/10.1038/s41397-020-0156-3

382. Okada Y, Momozawa Y, Ashikawa K, Kanai M, Matsuda K, Kamatani Y, et al. Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. Nat Genet [Internet]. 2015;47(7):798–802. Available from: http://dx.doi.org/10.1038/ng.3310

383. Luo Y, Kanai M, Choi W, Li X, Yamamoto K, Ogawa K, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. medRxiv [Internet]. 2020;14:2020.07.16.20155606. Available from: https://doi.org/10.1101/2020.07.16.20155606

384. Degenhardt F. HLA Pipeline [Internet]. Available from: https://github.com/ikmb/HLApipePublic

385. Naito T, Okada Y. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. Semin Immunopathol [Internet]. 2022;44(1):15–28. Available from: https://doi.org/10.1007/s00281-021-00901-9

386. Goodin DS, Oksenberg JR, Douillard V, Gourraud P-A, Vince N. Genetic susceptibility to multiple sclerosis in African Americans. Montgomery CG, editor. PLoS One [Internet]. 2021 Aug 9;16(8):e0254945. Available from: http://dx.doi.org/10.1371/journal.pone.0254945

387. Sanchez-Mazas A. HLA studies in the context of coronavirus outbreaks. Swiss Med Wkly [Internet]. 2020 Apr 16;150(April):w20248. Available from: https://doi.emh.ch/smw.2020.20248

388. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated [Internet]. Vol. 12, Scientific Reports. Nature Publishing

Group UK; 2022. 1–35 p. Available from: https://doi.org/10.1038/s41598-022-14395-4

389. Shen J. HLA-IMPUTER: an easy to use web application for HLA imputation and association analysis using population-specific reference panels. Bioinformatics.

390. Tran L, Mack SJ, The COVID-19 HLA & Immunogenetics Consortium. HLA Imputation Portal [Internet]. Available from: https://database-hlacovid19.org/shiny/HLA-Imputation-Portal/

391. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017;27(3):S2–8.

392. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature [Internet]. 2018;562(7726):203–9. Available from: http://www.nature.com/articles/s41586-018-0579-z

393. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [Internet]. 2021. Available from: www.genome.gov/sequencingcostsdata

394. Neuchel C, Fürst D, Tsamadou C, Schrezenmeier H, Mytilineos J. Extended loci histocompatibility matching in HSCT—Going beyond classical HLA. Int J Immunogenet. 2021;48(4):299–316.

395. Springer I, Tickotsky N, Louzoun Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front Immunol. 2021;12(April):1–11.

396. Sharon E, Sibener L V, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. Nat Genet [Internet]. 2016 Sep 1;48(9):995–1002. Available from: http://www.nature.com/articles/ng.3625

397. Khor S-S, Zheng X, Ishitani A, Azuma F, Pyo C-W, Omae Y, et al. P052 A novel algorithm for KIR copy number imputation by KIBAG and consolidation of population specific KIR references in HKimpNet (HLA & KIR imputation network). Hum Immunol [Internet]. 2019;80(2019):93. Available from: https://doi.org/10.1016/j.humimm.2019.07.104

398. Ritari J, Hyvärinen K, Partanen J, Koskela S. KIR gene content imputation from single-nucleotide polymorphisms in the Finnish population. PeerJ. 2022;10.

399. Sayadi S, Douillard V, Vince N, Südholt M, Gourraud P-A. Distributing human leukocyte antigen (HLA) database in histocompatibility: a shift in HLA data governance. Explor Immunol [Internet]. 2022 Nov 1;2(6):749–59. Available from: https://www.explorationpub.com/Journals/ei/Article/100380

X - Liste des communications scientifiques

X.1 - Publications

- Montassier E, Al-Ghalith GA, Mathé C, Le Bastard Q, **Douillard V,** Garnier A, Guimon R, Raimondeau B, Touchefeu Y, Duchalais E, Vince N, Limou S, Gourraud p-A, Laplaud DA, Nicot AB, Soulilou J-P & Berthelot L. ***Distribution of Bacterial α1,3-Galactosyltransferase Genes in the Human Gut Microbiome.*** *Front Immunol [Internet].* 2020 Jan 13;10. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2019.03000/full
- Vince N, Limou S, Daya M, Morii W, Rafaels N, Geffard E, **Douillard V,** Walencik A, Boorgula MP, Chavan S, Vergara C, Ortega VE, Wilson JG, Lange LA, Watson H, Nicolae DL, Meyers DA, Hansel NN, Ford JG, Faruque MU, Bleecker ER, Campbell M, Beaty TH, Ruczinski I, Mathias RA, Taub MA, Ober C, Noguchi E, Barnes KC, Torgerson D, Gourraud P-A. ***Association of HLA-DRB1*09:01 with tIgE levels among African-ancestry individuals with asthma.*** J Allergy Clin Immunol [Internet]. 2020 Jul;146(1):147–55. Available from: https://doi.org/10.1016/j.jaci.2020.01.011
- Vince N*, **Douillard V*,** Geffard E, Meyer D, Castelli EC, Mack SJ, Limou S & Gourraud P-A. ***SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics***. *Genet Epidemiol [Internet]*. 2020 Oct 18;44(7):733–40. Available from: https://onlinelibrary.wiley.com/doi/10.1002/gepi.22334
- Goodin DS, Oksenberg JR, Douillard V, Gourraud P-A, Vince N. ***Genetic susceptibility to multiple sclerosis in African Americans.*** Montgomery CG, editor. PLoS One [Internet]. 2021 Aug 9;16(8):e0254945. Available from: http://dx.doi.org/10.1371/journal.pone.0254945
- Ba R., Geffard E., **Douillard V.,** Simon F., Mesnard L., Vince N., Gourraud P. A., & Limou S. (2021). ***Surfing the Big Data Wave: Omics Data Challenges in Transplantation.*** Transplantation, 10.1097/TP.0000000000003992. Advance online publication. https://doi.org/10.1097/TP.0000000000003992
- **Douillard V,** Castelli EC, Mack SJ, Hollenbach JA, Gourraud P, Vince N, et al. ***Current HLA Investigations on SARS-CoV-2 and Perspectives.*** *Front Genet*

[Internet]. 2021 Nov 29;12(November):10–6. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2021.774922/full

- **Douillard V,** Castelli EC, Mack SJ, Hollenbach JA, Gourraud P-A, Vince N, et al. ***Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research.*** Front Genet [Internet]. 2021 Dec 2;12. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2021.774916/full
- Valencia A, Vergara C, Thio C. L, Vince N, **Douillard V,** Grifoni A, Cox A L, Johnson E, Kral A H, O'Brien T R, Goedert J J, Mangia A, Piazzolla V, Mehta S H, Kirk G D, Kim A Y, Lauer G M, Chung R T, Peters M G, Khakoo S I, Alric L, Cramp M E, Donfield S M, Edlin B R, Busch M P, Alexander G, Rosen H R, Murphy E L, Wojcik G L, Carrington M, Gourraud P-A, Sette A, Thomas D L, and Duggal P. (2022). ***Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1.*** American journal of human genetics, S0002-9297(22)00001-5. Advance online publication. https://doi.org/10.1016/j.ajhg.2022.01.001
- Domenighetti C, **Douillard V,** Sugier PE, Sreelatha AAK, Schulte C, Grover S, May P, Bobbili DR, Radivojkov-Blagojevic M, Lichtner P, Singleton AB, Hernandez DG, Edsall C, Gourraud PA, Mellick GD, Zimprich A, Pirker W, Rogaeva E, Lang AE, Koks S, Taba P, Lesage S, Brice A, Corvol JC, Chartier-Harlin MC, Mutez E, Brockmann K, Deutschländer AB, Hadjigeorgiou GM, Dardiotis E, Stefanis L, Simitsi AM, Valente EM, Petrucci S, Duga S, Straniero L, Zecchinelli A, Pezzoli G, Brighina L, Ferrarese C, Annesi G, Quattrone A, Gagliardi M, Matsuo H, Nakayama A, Hattori N, Nishioka K, Chung SJ, Kim YJ, Kolber P, van de Warrenburg BPC, Bloem BR, Aasly J, Toft M, Pihlstrøm L, Correia Guedes L, Ferreira JJ, Bardien S, Carr J, Tolosa E, Ezquerra M, Pastor P, Diez-Fairen M, Wirdefeldt K, Pedersen NL, Ran C, Belin AC, Puschmann A, Ygland Rödström E, Clarke CE, Morrison KE, Tan M, KraincMD D, Burbulla LF, Farrer MJ, Krüger R, Gasser T, Sharma M, Vince N, Elbaz A; Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) Consortium. ***The Interaction between HLA-DRB1 and Smoking in Parkinson's Disease Revisited.*** Movement Disorders. 2022 Jul 10. doi: 10.1002/mds.29133

- Cloé Domenighetti, **Venceslas Douillard**, Pierre-Emmanuel Sugier, Nicolas Vince, Alexis Elbaz. ***Une valine 11 codée par le gène HLA-DRB1 est associée inversement au risque de maladie de Parkinson et interagit avec l'initiation du tabagisme.*** Revue Neurologique. Volume 178, Supplement, 2022, Page S15, ISSN 0035-3787, https://doi.org/10.1016/j.neurol.2022.02.151.
- Sayadi S, **Douillard V**, Vince N, Südholt M, Gourraud P-A. ***Distributing HLA database in histocompatibility : a shift in HLA data governance***. International Journal of Immunogenetics. Accepté.
- Durand A, Winkler CA, Vince N, **Douillard V**, Geffard E, Binns-Roemer E, Ng DK; Gourraud P-A, Reidy K, Warady B, Furth S, Kopp JB, Kaskel FJ, Limou S. ***Identification of Novel Genetic Risk Factors for Focal Segmental Glomerulosclerosis in Children Highlights the Immune System Role.*** American Journal of Kidney Disease. Accepté.
- **Douillard V**, dos Santos Brito Silva N, Limou S, Gourraud P-A, Launay E, Castelli E. C, Vince N, The SNP-HLA Reference Consortium (SHLARC). ***Exploring HLA imputation of admixed population with dimension reduction.*** En préparation.

X.2 - Communication orales

- Nantes Actualités Transplantation (NAT) 2018, Nantes, France, 31 Mai-1 Juin - *Missing data & imputation : how to deal with incomplete or simply unavailable information ?*
- EFI 2019, Lisbon, Portugal, 8-11 Mai – *Navigating the treacherous waters of HLA imputation with the SHLARC*, Vince N, **Douillard V,** Limou S, Gourraud P-A (nomination meilleur abstract)
- EFI 2020/2021, conférence en ligne, 21-23 Avril - *Sailing towards the new horizon of immunogenomics along with the SNP-HLA Reference Consortium*, **Douillard V,** Limou S, Gourraud P-A, Vince N
- International HLA & Immunogenetics Workshop 2021, conférence en ligne, 26 Mars 2021 – *Building reference panels for admixed population: the impact of diversity on HLA imputation*
- IHIW 2022, Noordwijkerhout, the Netherlands, 11-15 Mai – *Improving HLA imputation in admixed population with UMAP dimension reduction*

- EFI 2022, Amsterdam, The Netherlands, 17-20 Mai – *Improving HLA imputation in admixed population with UMAP dimension reduction*

X.3 - Posters

- Laboratoire d'Excellence Immunotherapy-Graft-Oncology (LabEx-IGO) 2018, Nantes, France, 17-18 Avril - *Harnessing the power of functional immunogenomics parameters to discover new associations with diseases*
- *European Federation for Immunogenetics* (EFI) 2018, Venice, Italy, 9-12 Mai - *SNP-HLA Reference Consortium: HLA and SNP data sharing for promoting HLA-centric analyses in genomics*, Vince N, **Douillard V,** Geffard E, Limou S, Gourraud P-A
- *Immune Polymorphism and Population Dynamics Workshop 2019*, New Orleans, LA, USA, 27-30 Octobre – *Navigating the treacherous waters of HLA imputation with the SHLARC*, Vince N, **Douillard V,** Limou S, Gourraud P-A
- Journées Ouvertes Biologie, Informatique, Mathématiques (JOBIM) 2019, Nantes, France, 2-5 Juillet - *Navigating the treacherous waters of HLA imputation with the SHLARC*, Vince N, **Douillard V,** Geffard E, Limou S, Gourraud P-A
- *American Society of Human Genetics 2020*, conférence en ligne, 27-30 Octobre - *Sailing towards the new horizon of immunogenomics along with the SNP-HLA Reference Consortium*, Nicolas Vince, **Douillard V,** Sophie Limou, Pierre Antoine Gourraud
- NAT-LabExIGO 2020/2021, conférence en ligne, 31 Mai- 1 Juin - *Sailing towards the new horizon of immunogenomics along with the SNP-HLA Reference Consortium*

X.4 - Autres communications

- Limou S, **Douillard V**. Écriture d'un article de blog pour le COVID-19 Host Genetics Initiative, 15 Décembre 2021 – *Exploring the puzzle of HLA immunogenetics in COVID-19 patients*
- Ba R, **Douillard V,** Durand A, Garnier A, Geffard E. Cocréateur & maître du jeu pour Transplant'Action, un escape game basé sur la bioinformatique et l'immunogénétique présenté au grand public à la Fête des Sciences 2020, la Nuit Blanche des Chercheurs 2020, les Utopiales 2020 et inScience 2022. Partenariat avec l'Inserm pour améliorer et partager le jeu à plusieurs établissements avec Da Silva P & Sicot C.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

Nantes Université

Titre : Imputation HLA dans des populations d'ancestralité composite grâce à des méthodes de réduction de dimension

Mots clés : HLA ; immunogénomique ; imputation ; réduction de dimension ; UMAP ; association

Résumé : La génomique humaine a rapidement évolué cette dernière décennie grâce aux avancées technologiques de génotypage et de séquençage, permettant l'essor des études d'associations en génome entier. Ces études ont mis en avant la région du Complexe Majeur d'Histocompatibilité (CMH) comme impliquée dans de nombreuses pathologies infectieuses et autoimmunes, et notamment le système HLA, molécules centrales de l'immunité.
La diversité génétique de la région du CMH complexifie son étude détaillée. Ainsi des méthodes d'inférence statistique du HLA à partir de polymorphismes simples de l'ADN se sont développées. Ce travail s'articule autour du SNP-HLA Reference Consortium (SHLARC) qui vise à récolter des données génétiques diverses afin d'améliorer les méthodes d'imputation HLA qui sont actuellement optimisées pour des populations européennes.
L'accès à une infrastructure de calcul dédiée est nécessaire pour créer rapidement des modèles d'imputation HLA, et leur performance dépend du nombre de polymorphismes et d'individus disponibles. De plus, en exploitant des algorithmes de réduction de dimension, comme l'UMAP, pour synthétiser les distances génétiques entre les individus, il est possible de créer des modèles d'imputation HLA spécifiques d'une population génétique qui améliorent la prédiction dans les populations d'ancestralité composite ou peu représentées.
Le SHLARC ouvre ainsi la porte à l'imputation HLA pour toutes les populations génétiques. En conséquence, il facilite la conduite d'études d'association HLA. Avec d'autres pans de l'analyse HLA, il permet d'identifier les mécanismes biologiques exacts à l'origine du du lien entre le HLA et des pathologies.

Title: HLA imputation in admixed populations using dimension reduction

Keywords: HLA; immunogenomics; imputation; dimension reduction; UMAP; association

Abstract: Human genomics quickly evolved in the last decade thanks to technological advances in genotyping and sequencing, allowing for the growth of genome-wide association studies. Those studies repetitively brought light on the Major Histocompatibility Complex (MHC), and especially the HLA molecule which is key in immunity, for its association in infectious and autoimmune pathologies.
The genetic diversity of the MHC technically hinders its investigation, therefore, statistical inference methods of the HLA were developed. This work revolves around the SNP-HLA Reference Consortium (SHLARC) which aims to gather genetic data from diverse populations in order to improve HLA imputation methods, currently optimised for European populations.
Dedicated computation infrastructures are mandatory to rapidly create HLA imputation models, and their performance is linked to the number of polymorphisms and individuals available. Moreover, exploiting dimension reduction algorithms, such as UMAP, to synthesize genetic distances between individuals, it is possible to create population-specific HLA imputation models which improve prediction in admixed and underrepresented populations.
The SHLARC opens the door to HLA imputation for every genetic population. Consequently, it allows to conduct HLA association studies. Along with additional HLA studies, it helps identifying the biologicial mechanisms linking HLA and pathologies.